Understanding the genetic architecture of fatty liver disease

Oveis Jamialahmadi

Department of Molecular and Clinical Medicine
Institute of Medicine
Sahlgrenska Academy, University of Gothenburg

UNIVERSITY OF GOTHENBURG

Gothenburg 2023

Cover illustration: Oveis Jamialahmadi

Understanding the genetic architecture of fatty liver disease

oveis.jamialahmadi@wlab.gu.se

ISBN 978-91-8069-035-5 (PRINT)
ISBN 978-91-8069-036-2 (PDF)
http://hdl.handle.net/2077/73763

Printed in Borås, Sweden 2023
Printed by Stema Specialtryck AB

To my family

"I then understood that all forms of knowledge which do not unite these conditions (imperviousness to doubt) do not deserve any confidence, because they are not beyond the reach of doubt, and what is not impregnable to doubt cannot constitute certitude."

Ghazali (1058-1111 CE), The Deliverance from Error.

Understanding the genetic architecture of fatty liver disease

Oveis Jamialahmadi

Department of Molecular and Clinical Medicine, Institute of Medicine
Sahlgrenska Academy, University of Gothenburg
Gothenburg, Sweden

ABSTRACT

Non-alcoholic fatty liver disease (NAFLD) is currently the most common chronic liver disease, ranging from simple steatosis to more severe conditions, namely non-alcoholic steatohepatitis (NASH), liver fibrosis, cirrhosis, and hepatocellular carcinoma (HCC). NAFLD has a strong genetic component, and its heritability depends on environmental factors and ethnicity. So far, genome-wide association studies were able to explain only a small fraction of its heritability, indicating the presence of missing heritability. Moreover, despite more than 20% of the general population and more than 70% of individuals with obesity have fatty liver, only a minority of individuals will progress to end stage liver disease. In the first study, we used polygenic risk scores (PRS) based on 5 known common genetic determinants of NAFLD to stratify the risk of HCC in individuals with dysmetabolism. We showed the ability of our PRS to predict the full spectrum of NAFLD and HCC both in high-risk individuals and in the general population. Additionally, we demonstrated a causal association between genetic predisposition to hepatic steatosis and HCC using a Mendelian Randomization approach. In the second study, we performed an exome-wide association study of alanine aminotransferase (ALT), a biomarker of liver fat and damage, to identify other genetic determinants of fatty liver disease. We found two missense variants on *GPAM* and *APOE* genes, robustly associated with liver fat content and chronic liver disease. Finally, in the third study, we performed a gene-environment-wide interaction study (GEWIS) of ALT to evaluate the role of gene-environment interactions in fatty liver disease susceptibility and to identify new genetic determinant of NAFLD. We found a new *locus* interacting with body mass index (BMI), the strongest environmental risk factor for NAFLD, associated with liver fat and chronic liver disease, but not with ALT.

In conclusion, these findings strongly support a causal relationship between liver fat accumulation and severe liver disease. Moreover, new genetic determinants of fatty liver identified by our analyses may be used for risk

stratification of advanced liver disease and HCC, and exploited as potential drug targets.

Keywords: Non-alcoholic fatty liver disease, Polygenic risk score, Genome-wide association studies, Mendelian Randomization.

ISBN 978-91-8069-035-5 (PRINT)
ISBN 978-91-8069-036-2 (PDF)

SAMMANFATTNING PÅ SVENSKA

Icke-alkoholrelaterad fettleversjukdom (NAFLD) är för närvarande den vanligaste kroniska leversjukdomen, vilken inkluderar allt från enkel steatos till svårare tillstånd såsom icke-alkoholrelaterad steatohepatit (NASH), leverfibros, cirros och hepatocellulärt karcinom (HCC). NAFLD har en stark genetisk komponent och dess ärftlighet beror på miljöfaktorer och etnicitet. Hittills har genomomfattande associationsstudier (GWAS) bara kunnat förklara en liten del av dess ärftlighet vilket innebär att en stor del av ärftligheten är okänd. Trots att mer än 20 % av befolkningen och mer än 70 % av individerna med fetma har fettlever, kommer endast en minoritet av dessa att utveckla slutstadiet av leversjukdom. I den första studien använde vi polygena riskpoäng (PRS) baserade på 5 etablerade genetiska determinatorer för NAFLD för att stratifiera risken för HCC hos individer med metabolt syndrom. Vi visade förmågan hos vår PRS att förutsäga hela spektrumet av NAFLD och HCC både hos högriskindivider och befolkningen som helhet. Dessutom visade vi ett orsakssamband mellan genetisk predisposition för leversteatos och HCC med hjälp av en mendelsk randomiseringsmetod. I den andra studien utförde vi en exomomfattande associationsstudie av alaninaminotransferas (ALT), en biomarkör för leverfett och leverskador för att identifiera andra genetiska determinatorer för fettleversjukdom. Vi hittade två missense-varianter på *GPAM*- och *APOE*-gener, starkt associerade med mängden leverfett och kronisk leversjukdom. I den tredje studien utförde vi en gen-miljöomfattande interaktionsstudie (GEWIS) av ALT för att utvärdera rollen av gen-miljöinteraktioner och dess koppling till fettleversjukdomar och för att identifiera ny genetisk determinant för NAFLD. Vi hittade ett nytt lokus som interagerar med body mass index (BMI), den starkaste miljöriskfaktorn för NAFLD som är associerad med leverfett och kronisk leversjukdom, men inte med ALT.

Sammanfattningsvis stöder dessa upptäckter starkt ett orsakssamband mellan ackumulering av leverfett och allvarlig leversjukdom. Dessutom kan de nya genetiska determinanterna för fettlever som identifierats av våra analyser användas för riskstratifiering av avancerad leversjukdom och HCC, samt utnyttjas som potentiella läkemedelsmål.

LIST OF PAPERS

This thesis is based on the following studies, referred to in the text by their Roman numerals.

I. Bianco C*, Jamialahmadi O*, Pelusi S*, Baselli G, Dongiovanni P, Zanoni I, Santoro L, Maier S, Liguori A, Meroni M, Borroni V, D'Ambrosio R, Spagnuolo R, Alisi A, Federico A, Bugianesi E, Petta S, Miele L, Vespasiani-Gentilucci U, Anstee QM, Stickel F, Hampe J, Fischer J, Berg T, Fracanzani AL, Soardo G, Reeves H, Prati D, Romeo S, Valenti L. Non-invasive stratification of hepatocellular carcinoma risk in non-alcoholic fatty liver using polygenic risk scores.

Journal of Hepatology, 2021. doi: 10.1016/j.jhep.2020.11.024.

II. Jamialahmadi O, Mancina RM, Ciociola E, Tavaglione F, Luukkonen PK, Baselli G, Malvestiti F, Thuillier D, Raverdy V, Männistö V, Pipitone RM, Pennisi G, Prati D, Spagnuolo R, Petta S, Pihlajamäki J, Pattou F, Yki-Järvinen H, Valenti L, Romeo S. Exome-Wide Association Study on Alanine Aminotransferase Identifies Sequence Variants in the GPAM and APOE Associated with Fatty Liver Disease.

Gastroenterology, 2021. doi: 10.1053/j.gastro.2020.12.023.

III. Jamialahmadi O, Mancina RM, Ciociola E, Valenti L, Romeo S. Gene-BMI-wide Interaction Study of Alanine Aminotransferase Identifies UXBN2A/CYP7A1 as a novel locus for Fatty Liver Disease.

Manuscript.

CONTENT

ABBREVIATIONS

ALT	Alanine aminotransferase
APOE	Apolipoprotein E
AST	aspartate aminotransferase
BMI	Body mass index
CAP	Controlled attenuation parameter
CS	Credible set
eQTL	Expression quantitative trait loci
FLD	Fatty liver disease
GEWIS	Gene-environment-wide interaction study
GRM	Genetic relationship matrix
GWAS	Genome-wide association studies
λ_{GC}	Genomic control
GCKR	Glucokinase regulator
GPAM	Glycerol-3-phosphate acyltransferase 1, mitochondrial
HCC	Hepatocellular carcinoma
HDL	High-density lipoproteins
HSD17B13	Hydroxysteroid 17β- dehydrogenase
IBD	Identity by descent
IV	Instrumental variable
LDSC	LD-score regression

LMM	Linear mixed model
LD	Linkage disequilibrium
LDL	Low-density lipoproteins
MRI	Magnetic resonance imaging
MRS	Magnetic resonance spectroscopy
MBOAT7	Membrane bound O-acyltransferase domain-containing 7
MR	Mendelian Randomization
MAF	Minor allele frequency
MARC1	Mitochondrial Amidoxime Reducing Component 1
NOME	NO Measurement Error
NAFLD	Non-alcoholic fatty liver disease
NASH	Non-alcoholic steatohepatitis
PNPLA3	Patatin-like phospholipase domain-containing 3
PI	Phosphatidylinositol
PRS	Polygenic risk scores
PIP	Posterior inclusion probability
pLoF	Predicted loss-of-function
PCA	Principal component analysis
PDFF	Proton density fat fraction
RCT	Randomized controlled trials
SNP	Single nucleotide polymorphism

TM6SF2	Transmembrane 6 superfamily member 2
T2D	Type 2 diabetes
VLDL	Very-low-density lipoprotein

1 INTRODUCTION

Aim of this thesis is to provide a better understanding of the genetic structure of fatty liver disease (FLD), and further examines the applicability of genetic modulators of FLD to stratify the risk of progression to hepatocellular carcinoma (HCC). The thesis consists of three studies in which well-known genetic modulators of FLD were used to evaluate a causal relationship between liver fat content and HCC[1]. Furthermore, novel genetic modulators of FLD were discovered using both exome-wide association and genome-BMI-wide interaction approaches on alanine aminotransferase (ALT)[2].

1.1 NON-ALCOHOLIC FATTY LIVER DISEASE

Non-alcoholic fatty liver disease (NAFLD) is the most common chronic liver disease in developed countries with a bidirectional association with various features of metabolic syndrome[3]. It has been estimated that NAFLD affects around 21-33% of the global population, and the burden is expected to deteriorate in parallel with the global prevalence of type 2 diabetes, obesity and metabolic syndrome, which in turn may progress to end-stage liver disease and HCC[4]. While a minority of individuals with NAFLD (~10%) may further develop liver-related complications, the risk stratification of these individuals among NAFLD population remains an important challenge[3].

NAFLD is described by the accumulation of triglycerides in the form of lipid droplets in more than 5% of hepatocytes (steatosis) and in the absence of other secondary causes, including excessive alcohol intake, viral hepatitis, steatogenic medications, or hereditary liver diseases[5-7]. NAFLD covers a broad spectrum of disorders, including simple steatosis (with or without mild inflammation), non-alcoholic steatohepatitis (NASH), characterized by lobular inflammation and hepatocellular injury due to ballooning, liver fibrosis (collagen deposition), cirrhosis (advanced fibrosis) and HCC (Figure 1)[3, 8, 9]. NAFLD is a heterogenous condition, and therefore, there exists a high interindividual variability in the disease progression and response to treatment. Nonetheless, recent advances in genomics and metabolomics have shed some light on different aspects of the disease[3].

While the natural course of NAFLD is bidirectional and depends on the disease activity and fibrosis stage, advanced fibrosis (stages 3 and 4) is the main prognostic feature of all-cause and liver-related mortality and events[3, 10].

Nonetheless, cardiovascular disease and extrahepatic malignancies are still the leading causes of death in NAFLD patients, most probably due to coexisting metabolic comorbidities between NAFLD and cardiovascular disease[11].

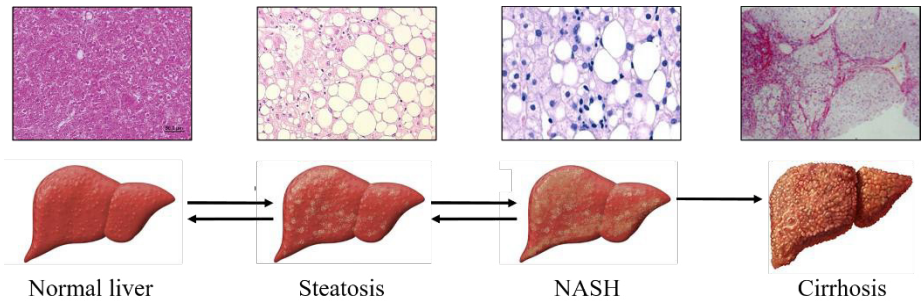

Figure 1. The disease spectrum of fatty liver disease. Images are courtesy of Professor Luca Valenti, University of Milan, Italy. Abbreviations: NASH, non-alcoholic steatohepatitis.

1.1.1 EPIDEMIOLOGY

The global prevalence of NAFLD in 2016 was 25.2%, and varied from 13.5% to 31.8% in Africa and Middle East, respectively. Obesity, type 2 diabetes, hyperlipidemia, metabolic syndrome and hypertension are among the main metabolic comorbidities associated with the disease[12]. Notably, around 43-64% of individuals with type 2 diabetes and 80% with obesity are diagnosed with NAFLD[3]. Of note, approximately 20% of individuals with NAFLD are lean (body mass index, BMI < 25 kg/m^2), and around 40% are non-obese (BMI < 30 kg/m^2)[13].

Moreover, NASH prevalence among NAFLD patients with available biopsy varies from 20% to 59%, of whom ~10-29% develop cirrhosis over 1-4 decades, and between 4 to 27% of NASH-induced cirrhotic patients develop HCC. Hence, regular screening and surveillance for HCC are recommended in individuals with NASH-related cirrhosis[7, 12, 14]. In addition, while there are a large number of non-cirrhotic NAFLD patients who develop HCC, they are not yet included in routine HCC screening guidelines due to lower incidence rate as compared to NASH-related cirrhosis[14].

The expected increase in NAFLD and NASH prevalence due to an increase in type 2 diabetes and obesity on one hand, and an increase in the prevalence of

childhood obesity worldwide on the other hand, further underscore health-care and economic burden of fatty liver disease[3, 5, 14, 15].

1.1.2 NAFLD DIAGNOSIS

NAFLD is diagnosed by detecting liver steatosis after the exclusion of other liver diseases (e.g. Wilson's disease or acquired lipodystrophy) or medications (e.g. steroids and tamoxifen) using non-invasive methods (e.g. clinical risk scores or imaging) or invasive histological assessments[3, 5, 14, 16].

Serum biomarkers of liver function, such as alanine aminotransferase (ALT) and aspartate aminotransferase (AST), can be elevated in NAFLD and NASH. However, they are not sensitive to fibrosis progression and many patients with NAFLD have normal levels of liver enzymes[3, 5].

Liver steatosis is usually diagnosed by abdominal ultrasonography. However, this technology is not sensitive to mild steatosis (< 30%) or presence of advanced fibrosis in NASH patients, and operationally challenging in individuals with central obesity. Moreover, liver fibrosis can be estimated by measuring liver stiffness using vibration-controlled transient elastography (e.g. FibroScan or ARFI) or magnetic resonance elastography (MRE). Liver steatosis can also be estimated simultaneously by measuring controlled attenuation parameter (CAP)[3, 5, 16, 17].

Unlike ultrasonography, magnetic resonance imaging (MRI)-based methods, such as MRI-derived proton density fat fraction (MRI-PDFF), are highly sensitive to small amounts of liver fat contents. It has been shown that MRI-PDFF is more sensitive and accurate than both CAP and histologically determined steatosis grade, and well correlated with measurements from magnetic resonance spectroscopy (MRS)[16, 17].

Conventionally, liver biopsy has been the gold standard technique to detect histological aspects of NAFLD. However, this approach suffers from limitations, such as sampling bias, intra-/inter-observer variability and invasiveness[18]. Nonetheless, liver biopsy is the ultimate necessity when the extent of hepatic fibrosis or presence of cirrhosis cannot be clearly deduced from non-invasive methods[5].

1.1.3 ENVIRONMENTAL RISK FACTORS OF NAFLD

Obesity, insulin resistance, type 2 diabetes, metabolic syndrome, familial disorders (e.g. hypobetalipoproteinemia or lipodystrophy), dietary habits, lack of exercise, socioeconomic factors, smoking and alcohol consumption are among the main risk factors and environmental triggers predisposing to NAFLD[19-22].

Obesity (visceral) is a major risk factor for different features of metabolic syndrome, and is defined by BMI, where BMI greater than or equal 25 and 30 denote overweight and obesity, respectively. However, there are ongoing debates on better accuracy of waist circumference in measuring the visceral obesity[20].

It has been also shown that alcohol consumption (even moderately) and obesity synergistically increase NAFLD risk (i.e. dual etiology fatty liver disease)[19]. Furthermore, the crosstalk between liver and gut may also contribute to metabolic aberrations in NAFLD, as evidenced by studies reporting changes in gut microbiota compositions in NAFLD patients. An increase in the intestinal permeability (leaky gut) due to the impaired intestinal barrier function may aggravate the inflammation through bacteria-produced metabolites such as short-chain fatty acids or lipopolysaccharides[3, 5, 23].

It should also be noted that, NAFLD is a complex disorder with both genetic and environmental components; however, the exact contribution of each component is unknown and may be influenced by ethnicity, geography and the interplay between these components (gene-environment interactions)[19, 22, 24].

1.2 GENOME-WIDE ASSOCIATION STUDIES

Complex traits and disorders, such as diabetes or NAFLD, are influenced by numerous genetic variants and environmental factors, as opposed to Mendelian disorders with single genetic defects[25]. Since sequencing the human reference genome in 2003, genome-wide association studies (GWAS) have been the standard hypothesis-free tools to examine the association between millions of single nucleotide polymorphisms (SNPs) across the entire genome and biological phenotypes, especially complex traits (Figure 2). To date, more than 3,700 genome-wide association studies (GWAS) have detected more than 50,000 genome-wide significant (P < 5E-8) genetic variants[26-30]. These findings have shaped our understanding of genetic architecture of several complex traits and their genetic susceptibility[28].

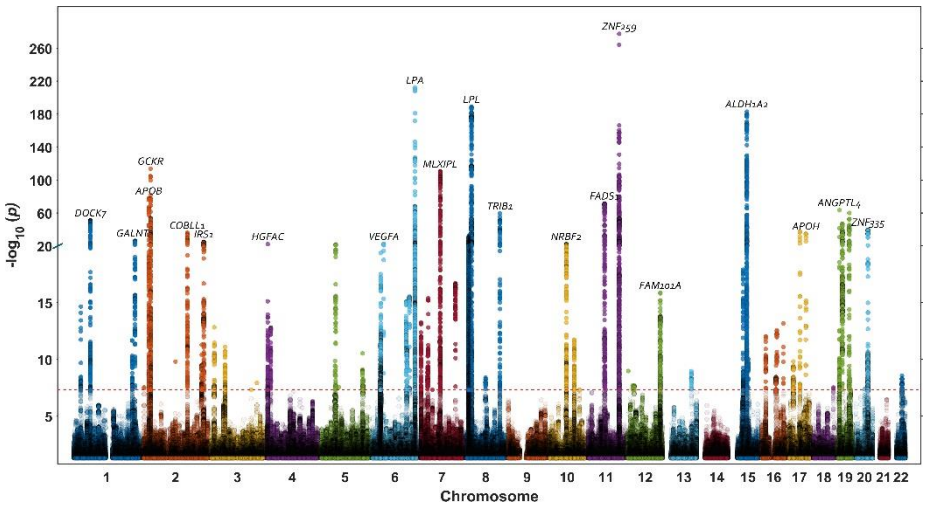

Figure 2. An example of GWAS of a polygenic trait (average diameter of very-low-density lipoprotein (VLDL) particles) in the UK Biobank study. Y-axis shows -log10 of p-values for the association of each common SNP with VLDL diameter, and X-axis represents the physical position in base pair on chromosomes. Dashed line shows the genome-wide significant threshold of 5E-8. Top loci have been marked with their nearest coding gene symbols.

While GWAS have already detected multiple well-replicated *loci*, they still cannot fully explain the "missing heritability" observed in complex traits, that is, identified common SNPs explain only a fraction of heritability of familial clustering and often have small effects on the trait under study[25, 31]. Despite this

fact, several drugs approved by US Food and Drug Administration target genes with common variants of modest effects, which may warrant the continuation of GWAS with larger sample sizes[28]. For instance, genetic *loci* on genes targeted by thiazolidinediones (for type 2 diabetes) or statins (lipid lowering medications) account for less than 1% of the trait variation[32]. In addition to the importance of GWAS in identifying potential drug targets or trait-associated *loci*, findings from GWAS provide invaluable resources for causal inference using Mendelian Randomization and estimation of SNP heritability[28, 33].

Mere SNP-trait associations cannot directly pinpoint the underlying mechanisms or genes related to the phenotypic variations, and hence the direct biological implications of GWAS findings may not be straightforward[28]. More than 90% of disease-associated genetic variants are found in non-coding regions, and only around 1/3 of genes targeted by causal variants are the nearest gene to the top GWAS hits[25, 28]. One striking example is the association of intronic variants at *FTO locus* with obesity, which was originally thought to be the gene responsible for the observed effect on obesity. However, follow-up studies showed that indeed this *locus* interacts with *Irx3* promoter (thousands of kb away) in a mouse model. This was later confirmed by showing the association between the intronic *locus* and *IRX3* expression in human brain samples[25].

Fine-mapping of multiple SNP-trait associations using either frequentist or Bayesian approaches can potentially narrow down the multiple associations to a fewer set of putative causal variants (i.e. credible set)[25, 29]. Moreover, combining GWAS signals with expression quantitative trait *loci* (eQTL) data via colocalization can potentially identify genes (eGene) which their mRNA levels are influenced by genetic variations from GWAS[25, 29, 34]. However, it is worth noting that dissecting the functional impact of variants with small effect sizes is extremely challenging, converging to Fisher's "infinitesimal model" of infinite causal variants with unidentifiable functional effects on the trait. Hence, eGenes of variants with small effects have a small contribution to the disease etiology[35].

Despite the numerous achievements of GWA studies so far, their applicability can be more extended by increasing sample size (e.g. at population level), deep phenotyping, and covering other neglected ethnicities. Moreover, gene-gene and gene-environment interactions may aid in discovering more undetected associations[28, 31].

One important limitation of GWAS is multiple comparison problem, which necessitates stringent thresholds. The GWAS significance threshold

commonly used is based on the Bonferroni correction of 1 million independent tests (i.e. 5E-8 with a false positive rate of 5%)[28]. This stringent threshold in practice limits the ability of most GWAS to explain the heritability of many traits[31]. Increasing sample size (e.g. by employing large consortia), or reducing the number of tests (e.g. by taking a candidate gene approach or selecting only putative loss-of-function variants) can to some extent mitigate this issue[2, 28]. Indeed, the conundrum of missing heritability can be partially explained by the hypothesis that many SNPs of modest effect sizes cannot reach the significance threshold[31]. Moreover, many traits are highly influenced by environmental risk factors; therefore, gene-gene and gene-environment interactions are expected to explain a proportion of the underlying heritability. However, one should not forget that due to shared environment or epistatic effects, many twin-based estimates of heritability may be biased[28, 36].

1.2.1 POLYGENIC RISK SCORES

Genetic risk profiling was originally employed to identify at-risk individuals in familial diseases or Mendelian disorders, such as mutations in *BRCA1* and *BRCA2* genes in breast cancer, or mutations in *CTFR* in cystic fibrosis[37]. In addition, as alluded before, many genetic variants of small effect sizes have been discovered from GWA studies to be associated with a wide range of complex traits, and this number is expected to increase with population-based GWAS and meta-analyses. While the tiny effects of these variants suggest they cannot serve as risk predictive tools, these genetic variants can be combined for risk stratification or prognostic prediction purposes in form of polygenic risk scores (PRS), which reflect the genetic predisposition (probabilistic susceptibility) at an individual level[37-39] (Figure 3).

PRS are calculated by weighted (GWAS marginal effects) sum of risk alleles at several genetic *loci*[40]. So far, multiple algorithms have been developed to calculate PRS, aiming at inclusion of variants in terms of maximizing the explained phenotypic variance or disease discrimination[38, 39]. Individuals with high PRS can then benefit from personalized health management strategies or lifestyle modifications[35, 40]. There are already ongoing investigations on clinical applications of PRS for risk prediction and stratification. Notably, a longitudinal study of coronary artery disease in the UK Biobank has reported a superior performance of PRS in predicting the disease incidence compared to common clinical risk factors (e.g. age, sex, blood pressure, smoking, etc.). When extending these findings to 13 million middle-aged individuals from

UK, inclusion of PRS in the common risk prediction algorithm QRISK shifted around 500,000 individuals from low to high risk for statin prescription[41].

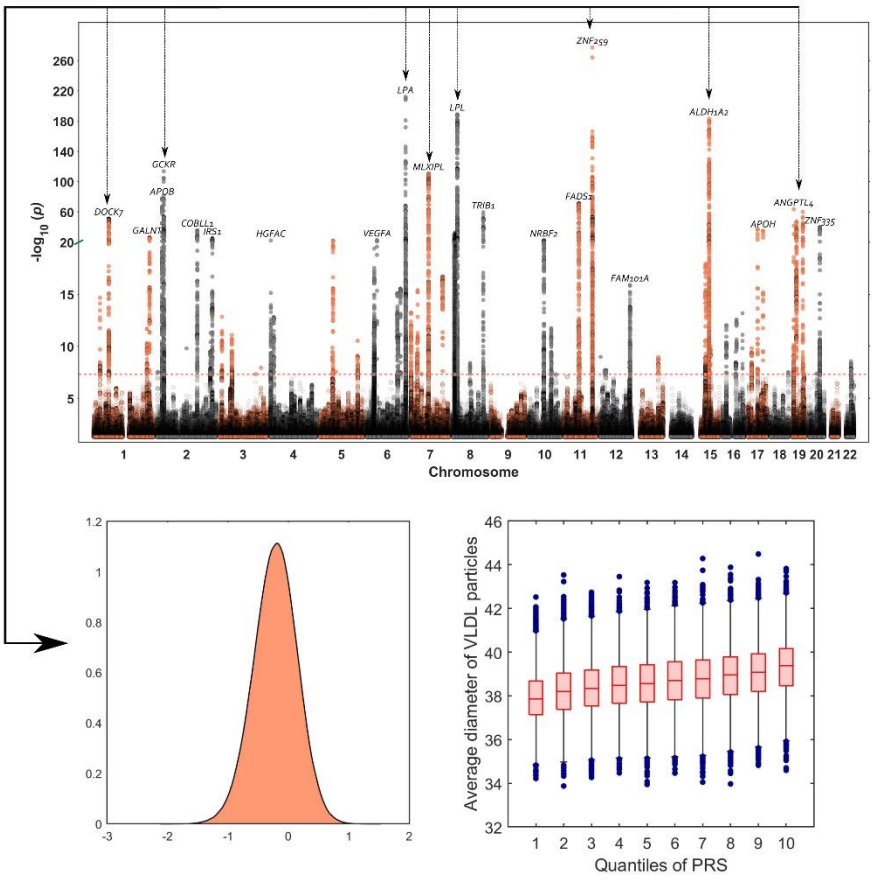

Figure 3. An example of PRS. A polygenic risk score for average diameter of VLDL particles in Europeans from UK Biobank was calculated from the GWAS depicted in Figure 2 after linkage disequilibrium (LD) clumping. Lower panel shows the distribution of the calculated polygenic score along with the quantile plot of 10 quantiles of the score versus the mean of the trait (y-axis).

1.3 MENDELIAN RANDOMIZATION

Observational epidemiological studies rarely can identify a causal relationship between an exposure (risk factor) and outcome due to the presence of several unmeasured confounders (distortion of exposure-outcome association because of common causes), reverse causality and other potential biases[42, 43]. The protective effect of antioxidant and vitamin supplements on cardiovascular disease is one example of such spurious causal inferences via observational studies[44]. Mendelian Randomization (MR) on the other hand employs genetic variants as instrumental variables (IVs) to examine the causal association between an exposure of interest and an outcome[42, 45]. Conceptually, MR relies on the Mendel's second law, also called random assortment law, which states that the inheritance of one trait is independent of the others[46]. This random allocation of genetic variants at conception guards against the confounding biases in observational studies, and mimics randomized controlled trials (RCT)[44]. Furthermore, genetic variants are not influenced by disorders and therefore, are protected against reverse causation[44]. However, while genetic variants are generally free of confounding factors associated with observational studies, extra caution is required against introducing confounding via population stratification[43].

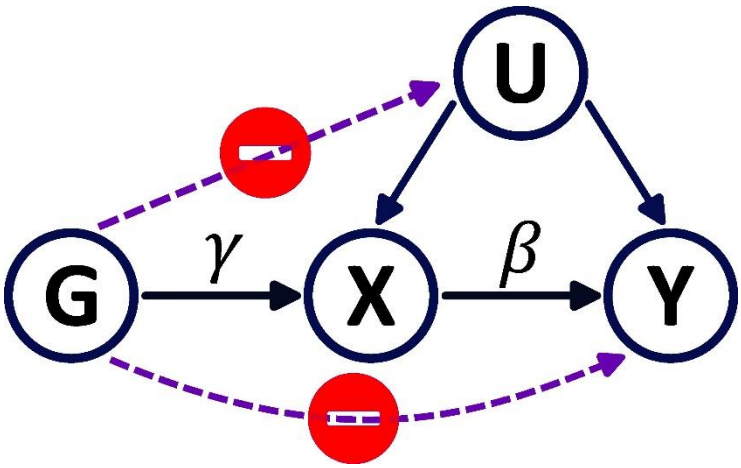

Figure 4. Schematic representation of directed acyclic graph (DAG) of MR analysis. Exposure X causes outcome Y with causal effect size of β, given that genetic variant G is a valid instrument associated with the exposure (γ), affects Y only through X, and does not associate with measured or unmeasured confounders (U). Dashed lines represent forbidden paths violating instrumental variable assumptions.

MR relies on three main assumptions (Figure 4) that genetic variants should 1) be robustly associated with the exposure (relevance), 2) influence the outcome only via the exposure (exclusion restriction), and 3) not share any common causes with the outcome (independence)[42-45].

Thanks to GWAS, several genetic instruments can be selected based on variants which robustly associate with the exposure of interest (e.g. $P < 5E-8$). Moreover, single variants can be combined as PRS to increase the statistical power. Nonetheless, it is always recommended to examine the relevance assumption via F-statistics (typically > 10) or other appropriate metrics such as proportion of variance explained (R^2) to account for the sample size used in the MR study[42, 44]. Lack of genetic instrument-exposure strength in turn results in weak instrument bias, violating the so-called "NO Measurement Error (NOME)" assumption of MR, which assumes the association of genetic instruments-exposure are estimated with no measurement error[47].

While one important assumption of MR is exclusion restriction, many genetic variants may influence the outcome under investigation via pathways independent of exposure-outcome pathway (so-called horizontal pleiotropy). In such cases, where the genetic instruments show pleiotropic effects, the causal estimates can be biased[44]. Several approaches have been developed to tackle this issue including MR-Egger, mode- and median-based estimators[48].

The number of studies employing MR to study the causal relationships between modifiable exposures and disease outcomes has been growing increasingly, from 1 to 800 studies in 2003 and 2020, respectively. MR studies cover a wide range of applications such as clinical, socioeconomic or environmental, and include examples from the effect of IL-6 receptor blockage on COVID-19 to education/intelligence effect on Alzheimer's disease risk[44, 49].

1.4 GENETIC SUSCEPTIBILITY TO NAFLD

As discussed before, NAFLD is a complex trait with both genetic and environmental components. The estimated heritability (i.e. the amount of the phenotypic variance explained by genetic factors) of NAFLD based on twin studies and familial aggregation ranges from 20% to 70%, which in turn depends on the environmental factors, ethnicity or study design[22, 24, 27]. Moreover, in a recent twin study, the heritability (additive) of hepatic fat content measured by MRI-PDFF was estimated to be around 50%[50].

Estimated SNP-heritability (h^2_g) of liver fat content measured by MRI-PDFF in the UK Biobank (for ~50,000 Europeans) is 17% (unpublished data). Although this estimate represents only common variants under an additive genetic model (narrow-sense), it suggests the presence of missing heritability, which as mentioned previously, can be explained by larger GWAS, rare variant (gene-based) analyses or exploring gene-gene or gene-environment interactions[28, 31]. Indeed, Stender et al. have shown the robust interaction of three genetic modulators of NAFLD, namely *PNPLA3* I148M, *TM6SF2* E167K and *GCKR* P446L, and adiposity (BMI) in predicting the entire spectrum of NAFLD, ranging from steatosis to inflammation and cirrhosis[51]. Nevertheless, this study was limited to only few well-known common variants and did not examine the interaction effect at exome- or genome-wide level.

1.4.1 COMMON GENETIC DETERMINANTS OF NAFLD

GWAS have shaped our understanding of genetic architecture of NAFLD within the past decade by examining the association of millions of SNPs across the genome[8, 27]. So far, at least 6 genetic *loci* have been identified to be robustly associated with NAFLD susceptibility and progression, namely variants on patatin-like phospholipase domain-containing 3 (*PNPLA3*), transmembrane 6 superfamily member 2 (*TM6SF2*), glucokinase regulator (*GCKR*), membrane bound O-acyltransferase domain-containing 7 (*MBOAT7*), Mitochondrial Amidoxime Reducing Component 1 (*MARC1*), and hydroxysteroid 17β-dehydrogenase (*HSD17B13*)[8, 22, 24, 27, 52, 53]. While the first 5 genes are involved in hepatic fat metabolism, the splice variant on *HSD17B13* protects against liver fibrosis and HCC development, independently of liver fat accumulation[24].

PNPLA3

In 2008, the first NAFLD GWAS identified the most robust genetic determinant of fatty liver disease, namely *PNPLA3 I148M* (rs738409 C>G)[54]. This missense variant, which encodes for an isoleucine to methionine substitution at position 148, associates with the full spectrum of the disease and within different ethnicities (Europeans, Asians and Hispanics)[27]. PNPLA3 is involved in remodeling of intracellular lipid droplets and has been shown to have acylglycerol O-acyltransferase and triacylglycerol lipase activity[27]. Additionally, it shows retinyl ester activity in hepatic stellate cells[55]. Unlike the wild-type PNPLA3, the mutant protein shows no lipase activity and evades degradation, which results in an accumulation on lipid droplets and interfering with triglycerides mobilization and turnover[56]. This in turn leads to sequestering gene identification-58 (CGI-58), a cofactor for triglycerides hydrolase activity of adipose triglyceride lipase (ATGL)[57]. In line with this, *Pnpla3* silencing in I148M knock-in mice mitigated liver fat accumulation, inflammation and fibrosis in steatogenic-fed mice, suggesting pharmacological silencing of the mutant protein may be a potential therapeutic option[58, 59].

GCKR

A GWAS on computed tomography (CT) measured steatosis in 2011 identified a missense variant on *GCKR* (rs1260326, P446L) to be associated with CT steatosis and histology based NAFLD[52]. The variant increases glucose influx by increasing the activity of glucokinase (GK), resulting in an elevated *de novo* lipogenesis (owing to an increase in synthesis of malonyl-CoA) and higher susceptibility to NAFLD[22, 24, 27, 60]. Nonetheless, the variant is associated with lower insulin resistance, conferring protection against diabetes[22].

TM6SF2

An exome-wide association study in 2014 identified a missense variant on *TM6SF2* (rs58542926 C>T) associated with higher liver fat content stored in intracellular lipid droplets, and resulting in lower expression at mRNA and protein level[61]. While the exact function of the protein is not clear, TM6SF2 is a membrane protein involved in lipoprotein secretion by modulating qualitative enrichment of VLDL triglyceride content[24, 27]. Therefore, the carriers of this mutation are at risk of NAFLD due to less lipidation and impaired VLDL secretion, diverting the lipid flux towards the synthesis of lipid droplets[22, 24, 60]. Conversely, by reducing the circulating lipoproteins, this variant confers protection against cardiovascular disease[22, 24].

MBOAT7

A GWAS in 2015 identified a missense variant (rs641738 C>T) at *TMC4-MBOAT7 locus* conferring higher risk of alcohol-related liver cirrhosis[62]. One year later, the association of this variants was shown with different spectrum of NAFLD, namely increased liver fat content, severity of liver damage and liver fibrosis. It has also been shown that the variant in fact influences *MBOAT7* expression levels, an endomembrane protein of six transmembrane domains, and not *TMC4*[63]. The variant likely modulates the remodeling of phosphatidylinositol (PI) by incorporating arachidonic acid (a polyunsaturated fatty acid, PUFA) into lysophospholipids via Lands' cycle in the liver[24, 27, 63, 64]. Hence, a reduction in PI bound arachidonic acid results in accumulating of saturated PI and ensuing triglycerides synthesis[24]. Very recently, a large meta-analysis of more than 1 million individuals further verified the association of this genetic variant with the presence and severity of NAFLD[65].

HSD17B13

While all the identified common genetic variants increased the susceptibility to the disease, a protective splice donor variant on *HSD17B13* (rs72613567:TA, adenine insertion adjacent to the donor splice site) was discovered in an exome-wide association study of ALT in 2018. The variant was associated with a lower risk of chronic liver disease and protected against the steatosis progression towards NASH[66]. While the exact mechanisms by which this variant confers a protective effect are not clear yet, it has been suggested that the variant influences NAFLD through the retinol dehydrogenase activity of HSD17B13 in lipid droplets. Hence, the loss-of-function variant results in a reduction in hepatic stellate cell activity, inflammation and fibrogenesis independently of the liver fat content[24, 27, 60].

MARC1

Recently, a GWAS using UK Biobank data found a missense variant on *MARC1* (rs2642438, A165T) protecting against all-cause cirrhosis and fatty liver[53]. While its physiological implications in NAFLD are not clear, a recent study has shown that the variant increases hepatic phosphatidylcholines, with a lipid profile similar to that of the *HSD17B13* variant[67].

Common genetic determinant of fatty liver and hepatic lipid metabolism

It is important to note that genetic determinants predisposing to both alcoholic and non-alcoholic fatty liver are largely overlapping, suggesting the shared genetic architecture of these two conditions[8, 22, 24, 64]. Hence, both alcoholic and non-alcoholic fatty liver can be considered as different manifestations of the same condition, that is fatty liver disease (FLD)[22]. Furthermore, almost all the common variants with a robust association with NAFLD, are involved in lipid metabolism, including lipid trafficking, compartmentalization and remodeling (Figure 5)[22, 24, 60].

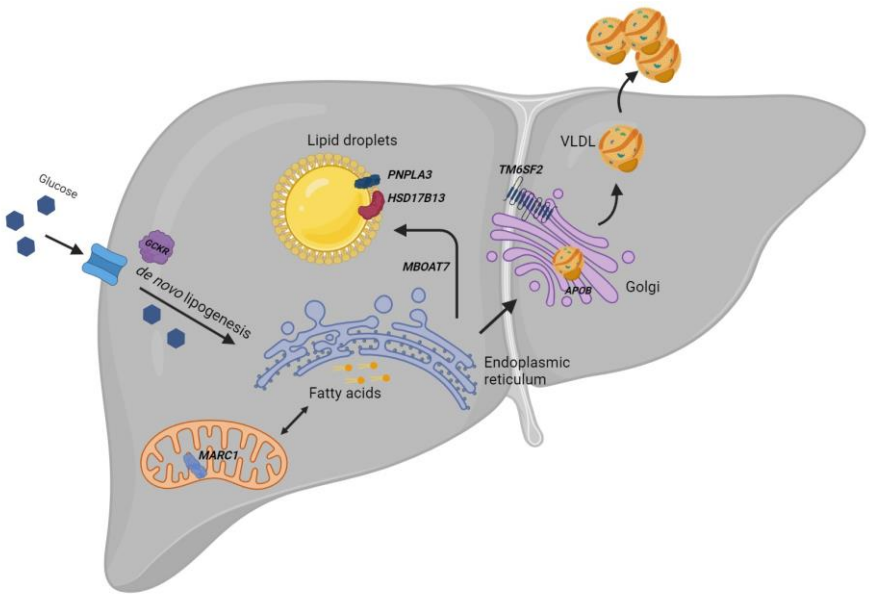

Figure 5. The metabolic pathways of common genetic determinants of NAFLD. The proteins have been shown based on their putative biological functions and their presumed contribution to the disease pathogenesis (created by Biorender).

1.4.2 NAFLD: THE USE OF GENETICS IN CAUSALITY AND RISK STRATIFICATION

While epidemiological studies have already suggested the association of obesity with cardiometabolic disorders and NAFLD, shedding light on causal

relationships between these traits is beyond the reach of observational studies[27]. As alluded before, MR tackles the limitations of observational studies, and is capable of inferring causal relationships between exposure-outcome of interest using robust genetic instruments. Of note, an MR study has shown a causal association between BMI and liver enzyme levels, type 2 diabetes and cardiovascular disease, using the most robust genetic determinant of obesity at *FTO locus*[68].

In another seminal MR study using a polygenic risk score of four genetic determinants of NAFLD (*PNPLA3, TM6SF2, GCKR* and *MBOAT7*), a causal association was found between liver fact content and liver inflammation, ballooning and fibrosis. These findings suggest a causal role of long-term accumulation of hepatic fat on NAFLD-related liver damage[69]. This in turn implies a fallacy in the notion of liver fat content being "benign", since "simple" steatosis is a causal player in the progression towards chronic liver disease[22]. This adverse effect of liver fat accumulation on liver damage can be seen also at a population level (Figure 6, unpublished data), where liver fat content is tightly correlated with different spectra of the fatty liver disease and HCC.

In addition, genetic determinants of NAFLD in form of PRS can potentially be used to delineate the heterogeneity of the disease by risk stratification, and to predict the disease progression[27, 70]. Compared to other biomarkers and predictors, genetic determinants represent a lifetime burden, and since their effects on the outcome are influenced by environmental factors, such as insulin resistance, obesity and lifestyle, they offer a reliable tool for liver disease screening[70]. Several examples of the use of PRS as non-invasive tools in risk stratification and long-term prediction of liver disease complications have been reported, including prediction of NAFLD progression towards NASH/fibrosis, clinical decision making, stratification of advanced fibrosis, and finding potential drug targets[24, 70]. Furthermore, it has been shown that common genetic variants can be integrated with clinical scores to improve the risk prediction for severe liver disease in individuals with an intermediate/high

predicted risk by clinical fibrosis scores, especially in those with metabolic risk factors[71].

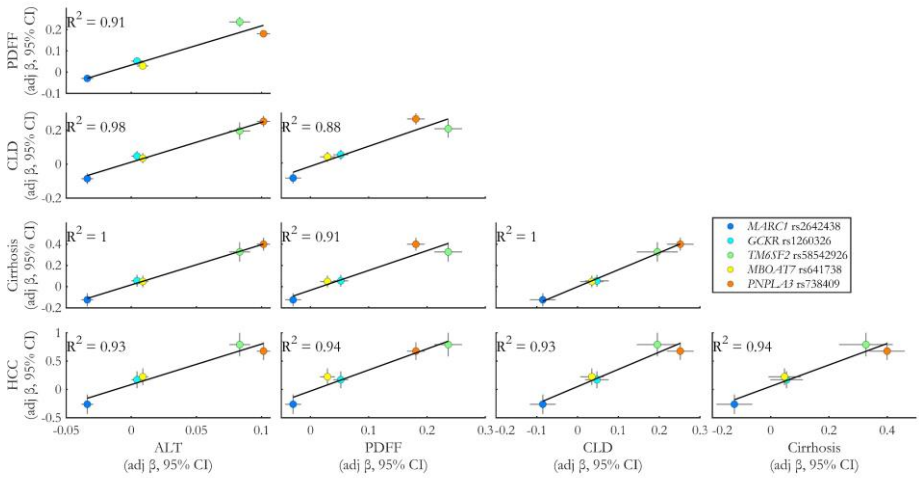

Figure 6. Correlation between effect of common genetic modulators of liver fat content and fatty liver disease at a population level. Effects of genetic variants in 347,000 unrelated Europeans from UK Biobank were estimated using a regression analysis adjusted for age, sex, body mass index and principal components of ancestry. R^2 indicates coefficient of determination between effect sizes. CLD, chronic liver disease; PDFF, proton density fat fraction; ALT, alanine aminotransferase.

2 AIMS

The overall aim of this thesis was to further understand the genetic basis of fatty liver disease using population-based studies and cohorts of at-risk individuals.

The aim of PAPER I was to exploit the main well-known genetic modulators of fatty liver disease to:

- Stratify the risk of individuals to hepatocellular carcinoma (HCC).
- Examine the causal relationship between non-alcoholic fatty liver disease and HCC.

The aim of PAPER II was to identify new common genetic modulators of fatty liver disease to:

- Better understand how heritability contributes to the fatty liver disease predisposition.
- Examine the associations with MRI-derived hepatic fat content and chronic liver disease at a population level.
- Replicate the associations with histologically proven fatty liver disease.

The aim of PAPER III was to partially explain the missing heritability of fatty liver disease due to gene-environment interactions by:

- Identifying common genetic variants with robust interaction with BMI in determining the ALT levels.
- Examining the associations with MRI-derived hepatic fat content and chronic liver disease at a population level.

3 METHODOLOGICAL CONSIDERATIONS

In this section, the general approaches used to perform GWA analyses, including statistical and computational considerations, sample and marker quality controls used for **PAPER I**, **PAPER II**, and **PAPER III** will be briefly discussed. Detailed description of methods and approaches for each paper can be found in the papers attached to this thesis.

3.1 SUBJECTS

The association studies in all papers were performed using one of the largest population-based biomedical databases: the UK Biobank. Other European-based cohorts of at-risk individuals were used for replication of main findings in papers II and III. In paper I, the UK Biobank was utilized to show the generalizability of the findings at a population-level.

3.1.1 UK BIOBANK

The UK Biobank is a large population-based study comprising more than 500,000 adult individuals aged between 40-69 years at recruitment, who visited 22 recruitment centers throughout the United Kingdom between 2006 and 2014 (ukbiobank.ac.uk). The UK Biobank study received ethical approval from the National Research Ethics Service Committee North West Multi-Centre Haydock (reference 16/NW/0274). Extensive phenotypic data includes hospital diagnoses, self-reported data based on questionnaires (e.g. lifestyle, socio-demographic, and health-related information), and physical measurements. Blood, saliva and urine samples for participants were also stored for genetic, metabolomic and proteomic analyses. Moreover, participants gave consent for follow-up using health-related records, such as cancer and death registries and inpatient hospital records[72, 73].

UK Biobank participants were genotyped using two similar arrays: the UK BiLEVE (~50,000 participants) or UK Biobank Axiom array (the remaining ~450,000 participants) with more than 95% overlap. Following sample-/marker-based quality controls, genotyped data were imputed based on the 1000 Genomes Phase 3, UK10K haplotype, and Haplotype Reference Consortium (HRC) reference panels[72].

3.1.2 SAMPLE QUALITY CONTROL

While the UK Biobank is mostly composed of "White" individuals (self-reported ethnic backgrounds), a proportion of participants (~6%) belonged to other non-European ethnicities. Furthermore, initially Bycroft et al. used principal component analysis (PCA) of directly genotyped participants to define a British subset within the broader group of "White" indiviuals[72]. Here, in order to widen the so-called "White British" subset (337,000 participants), we defined the European subset of the UK Biobank by adding individuals who self-reported as being "Irish" or "any other White background" to the "White British" subset. We next used the top 6 principal components (PCs) of ancestry originally calculated by the UK Biobank, and removed outliers as described below:

$$d_i = \sum_{k=1}^{6} \frac{(PC_{ik} - M_k)^2}{E_k}$$

Where E_k and M_k are the eigenvalue and mean of kth PC, calculated in the subset of "White British" individuals, respectively. PC_{ik} is the kth principal component of White individual i. The above formula is equivalent to drawing a six-dimensional ellipse centered among "White British" individuals[74]. Those individuals with a distance more than 7 standard deviation (in "White British" subset) were then excluded ($d_i > 49$), resulting in ~436,000 European individuals. In the original publication of the UK Biobank, along with multiple ensuing reports, a Bayesian outlier detection algorithm (implemented in R package *aberrant*) had been employed to select the largest cluster based on a number of top PCs[72]. Following this approach also resulted in a very similar set of European samples as detected by our approach, which was based on Neale group at the Broad Institute[74].

Further quality controls were applied based on information provided by the UK Biobank, and we further excluded individuals:

1- With excessive relatives (more than 10 putative third-degree relatives)
2- With a mismatch between the self-reported and genetically inferred gender
3- Putative sex chromosome aneuploidy (individuals with sex chromosome configurations that are not either XX or XY)
4- Who were identified by the UK Biobank as outliers based on heterozygosity and missingness (genotyping rate < 0.98).

5- With a withdrawn consent (accruing data updating regularly by the UK Biobank).

Figure 7 shows the 6 top PCs of European subset defined here along with "white British" subset from the UK Biobank.

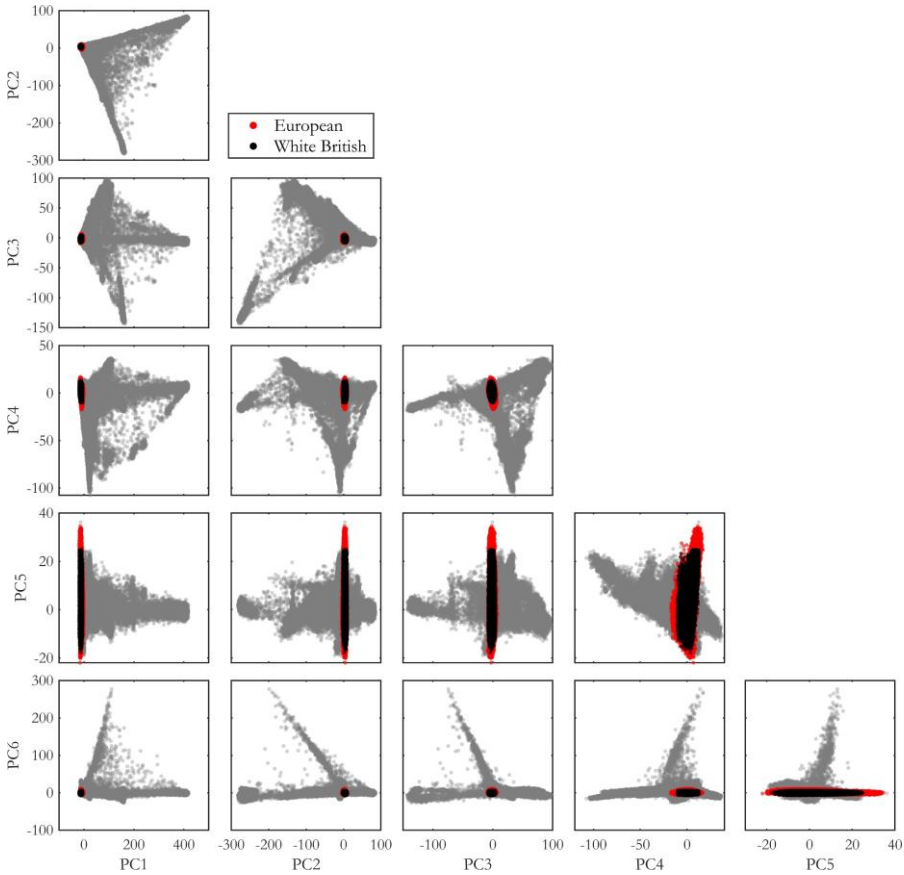

Figure 7. Definition of European subset of UK Biobank using the top 6 genomic PCs. Each dot represents the PC score for each individual in UK Biobank for whom PCA was performed. Red and black denote European subset defined here and White British individuals originally identified by the UK Biobank, respectively.

This set of European samples were used for GWAS when the association study method was adjusted for cryptic relatedness or population stratification by modeling the relatedness (kinship) matrix, either by mixed models

implemented in BOLT-LMM and SAIGE, or whole-genome regression approach implemented in REGENIE[75-78] (see statistical analysis section). For single variant association studies in paper I and some analyses in papers II and III, we further removed related individuals using the pairwise kinship coefficients initially calculated by the UK Biobank using KING software[72]. The kinship coefficient (φ), the conventional measure of relatedness, is the probability of two randomly homologous alleles drawn from two individuals being identical by descent (IBD). The expected value of φ degrades by a factor of 0.5 per each degree of relatedness, with φ being zero for two unrelated individuals. For this purpose, the set of maximal unrelated individuals were identified, so that the minimum set individuals with third degree or closer relatives were excluded from each family network. This is equivalent to exclusion of individuals with a pairwise estimated coefficient $\geq 0.5^{(9/2)}$ [72, 79].We first created graph objects for each family (set of related individuals) resulting in a total of 58,382 graph objects. Next, maximum subset of unrelated individuals per each family graph was determined by identifying the maximal cliques of complement family graphs using the Bron-Kerbosch algorithm[80]. This in turn is equivalent to identifying maximal set of independent vertices of a graph (i.e. unrelated individuals). In case of multiple independent sets per each family graph, the set comprising of individuals with minimum mean of genotype missing rate was picked. Figure 8 depicts an example of few configurations of family graphs within the UK Biobank.

This resulted in further exclusion of approximately 69,000 related individuals resulting in the final set of 365,449 unrelated Europeans. Compared to the simpler approach of excluding of one individual from each pair of individuals with a kinship coefficient of $\geq 0.5^{9/2}$, the maximal unrelated Europeans had 7,833 more individuals resulting in a larger sample size.

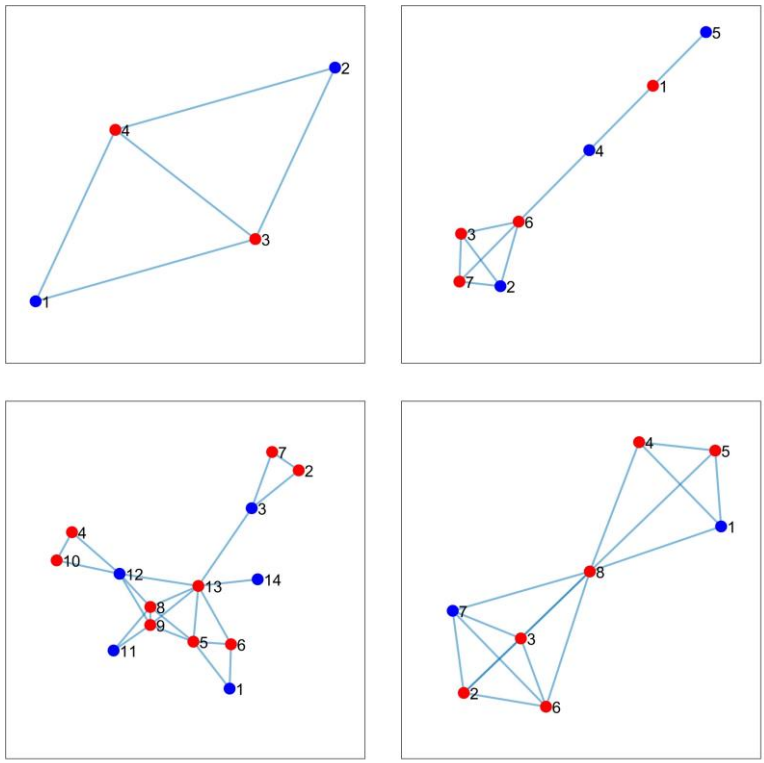

Figure 8. An example of four family graphs in the UK Biobank. Each panel shows a different configuration, where individuals and relatedness were shown as vertices and edges of the graph, respectively. Blue colored individuals represent the maximal set of unrelated individuals per each family group.

3.1.3 MARKER QUALITY CONTROL

From approximately 97 million imputed variants, from HRC, UK10K, and 1KG, we only kept common variants with a minor allele frequency (MAF) > 0.01, imputation INFO score > 0.8, and Hardy–Weinberg equilibrium (HWE) $P > 10^{-10}$. This resulted in a set of approximately 9,400,000 high quality common variants. This set of variants were used for GWA and genome-by-BMI interaction studies in papers II and III.

For linear and logistic mixed models with both SAIGE and BOLT-LMM tools in paper II, a genetic relationship matrix (GRM) was required to adjust for cryptic relatedness and population stratification. To prepare a set of high-quality variants for GRM, we used subset of directly genotyped variants after excluding the variants if falling in any of these categories:

- Positioned on long-range linkage disequilibrium (LD) and major histocompatibility complex (MHC) regions
- Missingness > 0.01
- MAF < 0.01
- Hardy–Weinberg equilibrium (HWE) $P < 10^{-15}$.

Next, LD pruning with a windows size of 500,000 base pairs and pairwise $r^2 < 0.1$ was performed to keep only independent markers resulting in a final set of 146,883 markers[81]. This set of high-quality markers was also used in paper III in the whole-genome regression approach implemented in REGENIE to capture the genetic component of the traits under study.

3.2 STATISTICAL ANALYSES

The main consideration concerning all the papers, in particular papers II and III, is the choice of proper statistical model to perform genetic association analyses. Depending on the trait under study, i.e. binary (e.g. hepatocellular carcinoma) or continuous (glucose levels), a linear or logistic regression model is often employed to test the association between genetic variants and the trait of interest, respectively.

Moreover, we assumed an additive genetic model for both large-scale GWAS and single variant association analysis, meaning that the heterozygote effect is in between of the other two homozygotes (assuming autosomal bi-allelic variants). This was due to a couple of reasons, including simplicity, being the most widely employed genetic model in GWAS, and the fact that additive effects show the largest contribution to the risk of complex traits compared to other models (dominant, recessive or epistatic effects)[82, 83].

To adjust for possible confounders, we considered two components; first, physiologically relevant traits including age, sex and BMI, and second, the effect of population stratification and batch effects, which was accounted for by adjusting for genomic principal components and genotyping array batches, respectively.

The above-mentioned multiple (linear or logistic) regression models can be written as following generalized linear models:

$$f(\mu_i) = \alpha_0 + \alpha_1 X_i + \beta g_i$$

Where $f(\mu)$ is the link function and equals μ for continuous and *logit*(μ) for binary traits, with $\mu = E(Y)$, and Y being the trait under study. α_0 is the intercept, X matrix of covariates and $g \in \{0,1,2\}$ is the allele count for individual *i* under the additive model. This model implies that additional copy of one allele affects (increases or decreases) the mean (continuous) or log odds (of a binary trait) additively. Assuming a bi-allelic variant with *A* being the effect (or risk) allele and *a* being the alternate allele, one can code (*aa, aA, AA*) genotypes as (0, 1, 1) or (0, 0, 1) for dominant and receive models, respectively.

3.2.1 MIXED MODELS

Multiple regression models can suffer from the power issues because the model assumes independence between individuals; hence, population stratification and sample relatedness should be handled (i.e. restricting to unrelated samples of one specific ethnicity) prior to the analyses. This is especially the case for the UK Biobank study, where high relatedness (~30%) is present[72]. Linear mixed models (LMMs) and the recent implementation of generalized linear mixed models[77], offer a greater power to discover new associations, and to better control the genotypic confounding effects, that is population stratification and cryptic relatedness[84]. This is achieved by including an additional random effect to model the correlation among individuals, and modeling (environmental) covariates as fixed-effects. Hence, under the standard infinitesimal model for a quantitative trait, we can write:

$$Y = X\alpha + G_s\beta_s + g + \epsilon$$

$$g \sim N(0, \sigma_A^2 \psi)$$

$$\epsilon \sim N(0, \sigma_e^2 I)$$

Where X and α are the design matrix of covariates, and their corresponding (fixed) effect sizes, G_s is allele count of genetic variant under study with β_s being its (fixed) effect size. g is a random effect explaining the polygenic effect, where σ_A^2 is additive genetic variance and ψ is the genetic relatedness matrix (GRM) estimated from a set of high-quality genotyped markers (see 3.1.3). ϵ is the vector of residuals, with a variance of σ_e^2 (non-genetic variance

or environmental effect[76]), and I is the identity matrix. In case of logistic mixed models, a logit link function replaces Y, the vector of phenotype under test[77]:

$$logit(\mu_i) = X_i\alpha + G_{s,i}\beta_s + g_i$$

The computational bottleneck of above formulations is calculating the empirical kinship matrix (GRM), and then estimating the variance components (σ_A^2 and σ_e^2). In paper II, we used BOLT-LMM and SAIGE for continuous and binary traits, respectively, due to their computational performance. Moreover, SAIGE offers saddle point approximation (SPA) to adjust test scores in case of imbalance case-control ratios, which is particularly useful when defining the disease outcome using electronic health records such as ICD-10 codes[76, 77].

3.2.2 WHOLE-GENOME REGRESSION MODEL

Recently, inspired by mixed-models, a new machine-learning approach, REGENIE, was developed, which showed to be dramatically faster than other mixed-model competitors[78]. REGENIE implements Ridge regression to derive the polygenic basis of a trait, which can be used as another covariate to adjust for population and relatedness confounding effects. Moreover, REGENIE implements Firth's penalized logistic regression to adjust for less common variants and imbalance case-control ratios, shown to have a better Type 1 error rate than SPA adjustment[78]. Since REGENIE allow to test gene-environment (GxE) interactions, we used it in paper III.

3.3 OTHER CONSIDERATIONS

Confounding bias. Typically, the ratio of observed and theoretical median of test statistics under null, called genomic control (λ_{GC}), is used to assess the presence of confounders in GWAS summary statistics. Generally, large λ_{GC} values (> 1) suggest the presence of some confounding resulting in inflating test statistics (many significant findings) across the genome. This inflation in test statistics (in practice, -log10 of p-values are used instead of chi-square test statistics) can be visualized in a quantile-quantile (QQ) plot. Nevertheless, λ_{GC} cannot differentiate between true polygenicity (many associations across the genome) and confounding in large-scale GWAS. This problem however, can be handled via LD-score regression (LDSC) analysis, where the intercept of LDSC analysis indicates the confounding bias[82]

Inference of proton density fat fraction. On the basis of an existing deep learning framework (https://github.com/tarolangner/mri-biometry) we used neck-to-knee body MRI images from a two-point Dixon technique to infer proton density fat fraction (PDFF)[85, 86]. Briefly, individuals were scanned with a Siemens MAGNETOM Aera 1.5-T MRI scanner using a 6-minute dual-echo Dixon Vibe protocol, providing a water and fat separated volumetric dataset covering neck to knees. A single multi-echo slice was further acquired to analyze the liver PDFF[87]. We slightly modified the original pipeline to parallelize the training and inference processes over ~50,000 MRI images.

We used reference dataset of approximately 10,000 MRI-derived PDFF using gradient echo imaging protocol to train the ResNet-50 convolutional neural network (CNN)[88] with a regression layer (Figure 9). Reference PDFF dataset was split into training (70%) and validation (30%) sets, and both coefficient of determination (R^2 = 0.963) and mean absolute error (MAE = 0.632) on the validation set outperformed the previously trained model on a similar but smaller reference dataset. Furthermore, the inferred PDFF values were compared to two independent datasets which were released later by the UK Biobank; first, fat referenced (FR) PDFF available for ~16,000 individuals derived from IDEAL imaging protocol, and second, PDFF dataset on ~15,500 individuals measured using a three-point Dixon technique[89]. The network was trained and validated under an NVIDIA RTX A6000 GPU in PyTorch, with default hyperparameters as in the original framework (e.g. batch size and initial learning rate with Adam optimizer). Although not shown here, we trained and validated the networks with the two above-mentioned independent datasets, since we speculated that this may enhance the predictive performance;

however, the results were comparable to the initial ground truth data we used to train the model.

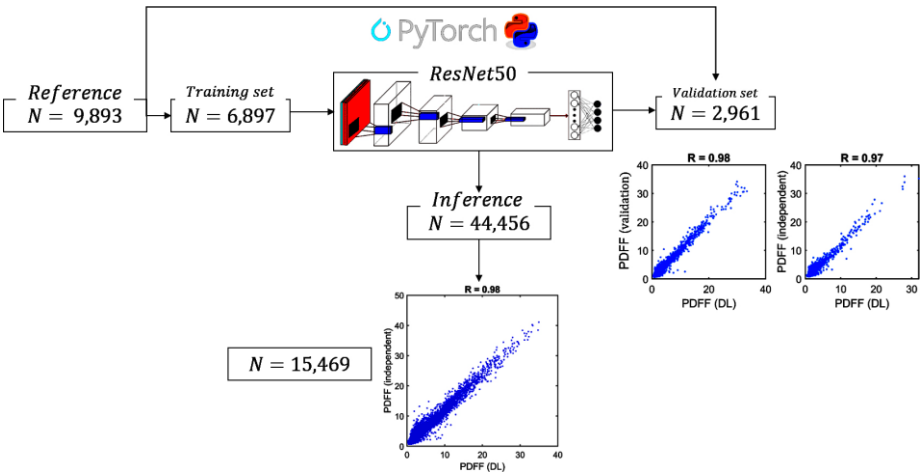

Figure 9. Schematic representation of training, validating and inference of liver fat content measured as PDFF in UK Biobank from neck-to-knee MRI images.

Computational considerations. The overall statistical and computational processes, including sample and genetic quality controls, data handling and analysis were performed in MATLAB (Mathworks), with R, Python, Java and C++ shared libraries and wrappers whenever necessary. This was done by developing a toolbox, MAtlab GEnetics (MAGE) (unpublished) with several functionalities and wrappers, providing reproducible, high-performance (parallelized with hybrid CPU-GPU capability), and flexible data analysis pipelines. A number of these functionalities include:

- Wrappers around PLINK, GCTA[90], BOLT-LMM, SAIGE and REGENIE for quality control, handling of genetic data and GWA analysis.
- Wrappers around annotation tools: VEP and SnpEff [91], fine-mapping tools: PolyFun[92], FINEMAP[93] and SuSiE[94], co-localization: coloc[95] (heavily parallelized with tabix supported), LDSC tool, pathway analysis: MAGMA[96], Mendelian Randomization tools: MendelianRandomization[97] and TwoSampleMR[98], multi-trait meta-analysis: MTAG[99].

- APIs for gnomAD, Ensembl REST, Open Targets[100], GWAS Catalog[101], PhenoScanner[102] and openGWAS[98].

Complex traits with multilevel inclusion/exclusion criteria can be defined easily in a matter of seconds, and association analyses can be performed either internally (with convenient stratification) or via available GWAS tool wrappers. This allowed us to write workflows similar to workflow description language (WDL) where a complete GWAS with downstream analyses can be run in parallel, without no manual interference.

4 RESULTS AND DISCUSSION

4.1 PAPER I

HCC is the major form of primary liver cancer and the second leading cause of cancer-related mortality worldwide, accounting for ~90% of primary liver cancer cases. Several lines of epidemiological evidence suggest the etiological role of NAFLD in the development of HCC[103, 104]. The global burden of NAFLD is expected to deteriorate in parallel with the global prevalence of type 2 diabetes, obesity and metabolic syndrome, which in turn may progress to liver disease and HCC[4, 15]. Existing guidelines focus on HCC surveillance in those with cirrhotic NAFLD, and should be considered in patients with advanced fibrosis[105]. Yet, a reliable non-invasive biomarker to stratify the risk of HCC in individuals with dysmetabolism without severe fibrosis is still an unmet medical need.

In this study, we exploited 5 well-known genetic modulators of fatty liver disease in order to stratify the risk of individuals to HCC, and further to unravel the causal relationship between NAFLD and HCC[1]. Hence, we used two PRS namely, PRS-HFC (PRS of hepatic fat content) calculated from genetic variants in *PNPLA3, TM6SF2, MBOAT7* and *GCKR*, and PRS-5, which was further adjusted for *HSD17B13*.

4.1.1 RESULTS OF PAPER I

Causal relationship between hepatic fat and HCC. Instrumental variable regression analysis adjusted for age, sex, BMI and T2D suggested a causal association between NAFLD (two-sample Mendelian randomization framework) and HCC (PRS-HFC OR = 1.35, 1.18–1.58, P = 1E-5). This association remained significant after adjusting for severe fibrosis ($P < 0.05$); however, the causal effect size reduced by 37-41%. Mediation analysis further showed that almost half of PRS causal effect was mediated through severe liver fibrosis ($P < 1E-16$).

PRS predict the full spectrum of NAFLD. PRS-HFC and PRS-5 were both associated with full spectrum of NAFLD and HCC. In NAFLD cohort, PRS were associated with an OR ~12 for severe liver fibrosis ($P < 1E-27$ for both) and 9 for HCC ($P < 1E-13$). While this association was still significant after adjusting for age, sex, BMI and T2D ($P < 0.01$), the association seemed not to be independent of severe liver fibrosis ($P > 0.1$). Similarly, in UK Biobank,

PRS were associated with liver cirrhosis (P < 1E-32, OR ~ 4) and HCC (P < 1E-16, OR ~ 15). Furthermore, the association with HCC was independent of liver cirrhosis (P < 1E-7, OR ~ 6.5, adjusted for cirrhosis). Overall, the association of PRS-5 with severe liver disease and HCC was stronger than PRS-HFC in all the cohorts under study.

Diagnostic accuracy for HCC. The area under the receiver-operating characteristic curve (AUROC) for HCC in NAFLD cohort was 0.64 and 0.65 for PRS-HFC and PRS-5, respectively. Two similar optimal cut-off points of ≥ 0.532 and ≥ 0.495 (PRS positive) were calculated for PRS-HFC and PRS-5, respectively. At these cut-offs, the sensitivity was 43% with an approximate 80% specificity. When validating this cut-off in a population-based study (UK Biobank), both PRS had the similar AUROC (0.63) to that in NAFLD cohort, but with a lower sensitivity (27%) but higher specificity (90%). These metrics were higher in those with metabolic risk factors such as obesity and T2D (~40% sensitivity and 90% specificity).

4.1.2 DISCUSSION TO PAPER I

The principal aim of this study was to assess the capability of main common genetic modulators of fatty liver disease in form of weighted PRS to predict HCC development in both NAFLD-based and population-based cohorts. More importantly, the causal effect of genetic modulators of hepatic fat accumulation on HCC shown here suggests that therapies targeting liver fat content reduction may be beneficial to protect against HCC development. The PRS calculated here is based on a handful of known genetic markers of NAFLD, which can be coupled with other non-genetic risk predictors to further improve the diagnostic accuracy of HCC development. The positive likelihood ratio at optimal cut-offs also suggested 2 to 4-fold increase in odds of being at risk of developing HCC given a positive PRS.

Another important aspect of this study was the focus not on severe liver fibrosis and cirrhosis, but on those individuals with NAFLD or other metabolic risk factors accounting for a larger proportion of the population. While compared to NASH-related cirrhosis[14], this subgroup accounts for a minority of HCC cases and it is absent from current HCC screening programs since there exists no cost-effective tool to predict HCC risk among these individuals.

We did not detect any horizontal pleiotropy in the causal association between genetic predisposition to NAFLD and HCC; therefore, there was no evidence showing that the genetic instruments influence HCC via another pathway

independent of hepatic fat. Nevertheless, we observed some heterogeneity suggesting either pleiotropy or model misspecification. This heterogeneity was mitigated after excluding *GCKR* variant, for which its protection against T2D was the putative source of heterogeneity, conferring mild protection against HCC while increasing the risk of NAFLD (hence, inconsistent with other genetic instruments and being an outlier). Furthermore, the association between the genetically determined NAFLD and HCC was partially mediated by severe fibrosis, consistent with HCC incidence in those without severe fibrosis or cirrhosis.

The overall predictive performance of PRS in predicting HCC was moderate in our study (AUC = 0.65 in driving cohort) with a slightly better performance in those with T2D in the UK Biobank (AUC = 0.7). Nonetheless, genetic variants remain unchanged throughout the individual's life span as opposed to other environmental risk factors and clinical biomarkers. Thus, such a genetic diagnostic tool can be used as an initial cost-effective step for HCC risk stratification. Unsurprisingly, positive likelihood ratios were larger in obese individuals from the UK Biobank consistently with a previous finding showing the presence of gene-BMI interactions in predicting NAFLD[51].

4.2 PAPER II

Fatty liver disease (FLD) is the most common cause of chronic liver disease, and is expected to become the leading cause of end-stage liver disease in the next decade[15, 22]. Despite the important role of environmental risk factors, hepatic fat content has a strong genetic component[24]. While previous genome- and exome-wide association studies identified several genetic determinants of FLD in *PNPLA3*, *TM6SF2*, *GCKR*, *MBOAT7* and *MARC1*, they account only for a small fraction of the overall heritability of hepatic fat content, suggesting the presence of missing heritability[27, 52-54, 61, 106].

In an attempt to better understand the genetic architecture and predisposition to FLD, we aimed to detect other genetic predictors of this disease. Inspired by a previous discovery of a protein-truncating variant on *HSD17B13* protecting against FLD, we performed an exome-wide association analysis of ALT, which is commonly related to hepatic fat and liver damage[66, 107, 108]. We restricted the study only to predicted loss-of-function (pLoF) and missense variants on the exome. We hypothesized that the putative deleterious effect on the protein function may increase the chance of finding new genetic *loci* associated with the target trait by both enriching for causal variants and reducing the number of examined variants[109, 110]. We began with exome-wide association study of ALT in Europeans from the UK Biobank and further replicated (internal replication) the significant *loci* with measured PDFF, and also in three independent cohorts (external replication).

4.2.1 RESULTS OF PAPER II

Exome-wide association analysis of ALT. Starting from an approximately 9 million imputed common (MAF > 1%) and high-quality imputed (INFO score > 0.8) variants, annotation with VEP and SnpEff resulted in ~40,000 missense and loss-of-function variants (jointly called pLoF). Individuals with a measured PDFF were excluded before performing association analysis on ALT. Following LD clumping and conditional analysis, we identified 190 independent genetic variants significantly associated with ALT levels (Bonferroni threshold of 1.47E-6) among which ~19% (36 variants) have been previously shown to be associated with different stages of fatty liver disease or lipoproteins. Moreover, gene-set enrichment analysis of genes with at least one significant association showed an overrepresentation of genes involved in metabolic liver diseases, lipid homeostasis and triglyceride metabolism. This set of genes was mostly expressed in liver and hepatocytes.

Association of independent variants with PDFF. We next examined the association of the 190 independent genetic variants with PDFF (n = 8930) in the UK Biobank using a linear mixed-model approach, and 8 variants remained significant after the Bonferroni correction. Among this set, 5 were well-known genetic modulators of FLD, namely variants on *PNPLA3, TM6SF2, MARC1* and *MBOAT7/TMC4*. We identified 3 new missense genetic variants in *Apolipoprotein E* (*APOE*, rs429358), *Glycerol-3-phosphate acyltransferase 1, mitochondrial* (*GPAM*, rs2792751) and *olfactory receptor family 12 subfamily D member 2* (*OR12D2,* rs3128853) among which *GPAM* rs2792751 (P = 0.001 and P = 0.051, respectively, encoding p.Val43Ile) and *APOE* rs429358 (P = 0.003 and P = 0.014, respectively, encoding for p.Cys112Arg), but not *OR12D2* rs3128853 were associated with chronic liver disease and cirrhosis. These two variants were also associated with circulating lipoprotein levels.

Replication in external cohorts. We next examined the association of *GPAM* rs2792751 and *APOE* rs429358 with severity of liver steatosis in 3 independent cohorts comprising around 2600 Europeans with existing liver biopsy and at-risk for FLD (from Italy, France and Finland). The association between severity of steatosis and each variant was tested using an ordinal logistic regression model adjusted for age, sex, BMI and allele count of *PNPLA3* rs738409. Meta-analysis of the associations from these cohorts was consistent with an association of both genetic variants with severity of steatosis (fixed-effect OR = 1.21 P = 0.002 for *GPAM* rs2792751, and OR = 0.76 P = 0.002 for *APOE* rs429358). We however were not able to observe any significant pooled association with liver inflammation, ballooning or fibrosis.

Transcriptomic analysis. Gene-set enrichment analysis by examining the liver transcriptome from another independent cohort comprising of 125 obese Italian individuals, showed an upregulation of lipid metabolism and a down-regulation of inflammation for *GPAM* rs2792751, with an opposite regulatory impact for *APOE* rs429358. Nevertheless, no significant difference was observed between carriers and non-carriers of both genetic variants, nor could we find any *in silico* deleterious effect on protein function.

4.2.2 DISCUSSION TO PAPER II

Here we reported a large-scale association study between pLoF variants and ALT levels in Europeans from the UK Biobank, and identified 190 independent genetic *loci*, the largest set of missense and nonsense genetic predictors of ALT. LD score regression analysis further showed the

polygenicity of this trait, with an additive narrow-sense heritability of 13%. Almost 20% of independent genetic variants associated with ALT were reported to be associated with other spectra of FLD, with majority being associated also with lipoprotein levels. Since ALT is clinically regarded as a biomarker of hepatocellular damage[108], this in turn suggests a tight connection between lipoprotein metabolism and liver damage. The enrichment of lipid handling processes with this set of genes, along with the causal link between hepatic fat content and advanced liver disease[69], further support this notion.

After examining the association between this set of 190 independent variants and MRI-PDFF, we identified two missense variants on *APOE* and *GPAM*, respectively protecting and predisposing to steatosis, chronic liver disease and cirrhosis at the population level. We further replicated the association with severity of steatosis in 3 European cohorts of individuals at-risk for fatty liver disease. We however, could not confirm any consistent association with inflammation, ballooning or fibrosis. Due to small effect size, this can be attributed to lower power of these cohorts to detect the association.

The minor allele of rs2792751 *GPAM* (43Ile) was associated with higher hepatic fat content. GPAM (also known as GPAT1 located on outer membrane of mitochondria) catalyzes the committing step of glycerolipids synthesis by esterifying acyl-CoA activated fatty acids to glycerol-3-phosphate. In addition, GPAM is the important link between *de novo* lipogenesis and triglycerides synthesis, owing to a surplus in dietary carbohydrates in the liver[111]. *Gpam* knockout mice showed a reduced triglyceride content in the liver, while the opposite was observed when overexpressing this gene[112, 113]. When comparing the expression levels between carriers and non-carriers of this variant, we observed an upregulation in lipid catabolic pathways in liver, with no significant expression over genotypes. This may be regarded as a compensatory mechanism to the increase in hepatic triglycerides synthesis.

On the other hand, *APOE* rs429358 was associated with a lower hepatic fat content, higher plasma LDL and triglycerides level. When comparing different genotypes of this variant, transcriptomic data showed a reduction in the biological processes related to lipids, without any changes in mRNA levels. ApoE is involved in lipoprotein metabolism and uptake of circulatory lipids into the liver, and liver is the main source of circulating ApoE (in VLDL or high-density lipoproteins, HDL)[114]. Moreover, the variant was robustly associated with increased risk of Alzheimer's disease (main genetic risk variant) and dyslipidemia (consistent with our findings). Therefore, one possible hypothesis could be that the variant results in a loss-of-function

activity of ApoE, where the circulating lipoproteins cannot be taken up into the liver.

4.3 PAPER III

Current GWAS have been able to find multiple associations between common variants and complex traits at the population level. The number of discovered signals is yet to be increased with effect sizes tend to be smaller as a result of greater power of large-scale studies. This situation seems to converge to Fisher's "infinitesimal model" of infinite variants, each with a small effect on the target phenotype[35]. Nonetheless, these large sets of common variants still cannot fully explain the variance of many complex traits[31].

In addition, multiple traits are influenced by environmental exposures, where the interaction between genetic variants themselves or environment may explain a proportion of the underlying missing heritability[28]. In fact, Stender et al. have shown the robust interaction of three genetic modulators of NAFLD, namely *PNPLA3* I148M, *TM6SF2* E167K, and *GCKR* P446L and adiposity (BMI) in predicting the entire spectrum of NAFLD, ranging from steatosis to inflammation and cirrhosis[51]. Nevertheless, this study was limited to only few well-known common variants and did not examine the interaction effect at a genome-wide level. One of the main bottlenecks in estimating the gene-environment interaction effects is that a considerably larger sample size is needed compared to main additive genetic effects in typical GWAS[115]. The availability of a population-based large-scale study such as the UK Biobank can potentially aid in discovering such interactions. Here, by taking a whole-genome regression approach[78], we aimed to detect gene-BMI interaction effects in predicting ALT levels, a marker of liver fat and damage, at a genome-wide level. Similar to our previous study, we also examined the association of significant genetic markers with MRI-PDFF and other live-related traits in European individuals from the UK Biobank.

4.3.1 RESULTS OF PAPER III[1]

Gene-environment-wide interaction study (GEWIS) of ALT. we restricted our analysis to Europeans from the UK Biobank after excluding those with an existing MRI-PDFF measurement, available for approximately 40,000 individuals. We used a whole-genome regression approach implemented in REGENIE software[78] to estimate the interaction effects between approximately 9 million common variants (MAF > 1%) with high imputation

[1] The findings of this study are still in manuscript form. Hence, the locus name of new finding has been replaced with [X] and its nearby gene with [Y]; however, the disclosed locus name can be found at the end of this thesis frame.

score (> 0.8) and BMI as the environmental exposure, in predicting ALT levels. Following LD clumping and conditional analysis, we identified 13 significant genetic *loci* at a genome-wide level (< 5E-8), out of which 11 were new gene-BMI interacting *loci*, except for *PNPLA3* and *TM6SF2*. One of these *loci* was not even associated with ALT (no interaction term in the model), and two only with a nominal significance ($P_{conditional}$ > 5E-8).

Fine-mapping of independent loci. After excluding the *HLA-B locus* on chromosome 6 due to its complex LD structure, we performed a functionally informed fine-mapping to determine the putative set of causal variants at each locus. Per-SNP heritability estimates from PolyFun tool were used as prior causal probabilities in two fine-mapping tools, sum of single effects (SuSiE) and FINEMAP[92-94]. For *TRIB1* and *TOR1B loci*, the first 95% credible set (CS) contained only 1 putative causal variant with a posterior inclusion probability (PIP) or > 0.95 for both SuSiE and FINEMAP. The size of first CSs for other *loci* was varying between 3 to 65 variants, reflecting the uncertainty in determining the causal signals at some *loci*. The number of CSs also varied from 1 to 3, suggesting the presence of multiple causal variants at some *loci*[116].

The association with liver fat content. Eight out of 12 lead genetic variants on *PNPLA3, TM6SF2, APOE, MARC1, TRIB1, COBLL1, GPAM* and *TOR1B* had been shown by others[2, 53, 54, 61, 117-120] and us to be associated with ALT levels and hepatic fat content. Based on this observation, we hypothesized that the 3 new *loci* may be associated with hepatic fat content. After correcting for multiple testing problem, we observed a significant association at locus [X] (see the attached manuscript) with PDFF (Benjamini-Hochberg FDR P < 0.05). The lead variant at this locus was also present in the first 95% CS of fine-mapped variants from GEWIS of ALT (r^2 = 0.95).

***[X] locus, liver disease and other metabolic traits*.** The lead variant at this locus was also associated with higher risk of chronic liver disease and higher fat content, but not associated with ALT levels. It was also associated with an increase in triglycerides, LDL and cholesterol levels. To examine the effect of this *locus* on gene expression patterns of nearby genes, we performed a Bayesian co-localization between summary statistics from GEWI analysis and expression quantitative trait *loci* (eQTL) from 49 tissues from GTEx project processed by eQTL Catalogue[95, 121, 122]. At this *locus*, we observed a consistent evidence of colocalization between interaction effect and gene expression of *[X]* in the liver.

4.3.2 DISCUSSION TO PAPER III

Here, we performed a gene-by-BMI wide association study of ALT levels in the UK Biobank and identified 13 interacting *loci* at a genome-wide level. Eight *loci* have already been reported to associate with liver fat content and ALT levels. Since the majority of interacting *loci* were reported to associate with fatty liver disease, we examined the association of newly identified *loci* with hepatic fat content (measured as MRI-PDFF). We found a new *locus* associated with hepatic fat content and chronic liver disease. In addition, we found another *locus* at *COBLL1*, which was very recently reported to be associated with fatty liver disease in a large-scale multiancestry study[120].

In fine-mapping of interacting *loci* different number of credible sets with varying number of putative causal variants were found. Specifically, while we could not find any association with liver fat content and chronic liver disease at *DPM3* and *GIPR loci,* a missense variant at the latter had the highest PIP, suggesting being the causal variant for the observed interaction effect in the first credible set. Interestingly, the same variant has been shown to be the putative causal variant for glycemic and obesity-related traits[123].

The lead variant at [X] locus, was associated with an increase in triglycerides and LDL levels suggesting that the concomitant increase in hepatic fact content is probably due to an increase in triglycerides synthesis or lipoprotein secretion and not to retention. [Y], a major regulator of cholesterol homeostasis and bile acids biosynthesis, is also a nearby gene. While fine-mapping of interaction effects for this *locus* shows a down-stream variant near [Y] to have the largest PIP in the first CS, we detected a colocalization evidence between liver-specific eQTL and interaction summary statistics only for [X], and not [Y]. Hence, it is possible that [Y] is the gene causing the effect observed on the hepatic fat content. Nonetheless, we observed a colocalization between hepatic fat content GWAS and GEWIS of ALT at this *locus* (not shown), suggesting that the causal signal for gene-BMI interaction of ALT most probably is the variant associating with hepatic fat content.

5 CONCLUSION AND FUTURE PERSPECTIVES

The main objective of this thesis was to better understand the genetic basis of fatty liver disease by both exploring the polygenic diagnostic capability of common genetic modulators, and identifying other genetic components of this complex trait. We started by utilizing the current known genetic players of fatty liver disease in form of weighted PRS, to answer questions regarding the causality and predictive ability of such scores for HCC risk screening. Specifically, in **paper I**, we evaluated the ability of genetic predisposition to liver fat accumulation and NAFLD to predict HCC development in at-risk individuals and in the general population. We further determined the diagnostic accuracy of PRS thresholds to show the increased genetic risk of HCC. Our findings showed a causal link between hepatic fat accumulation and HCC, which may suggest therapies aiming at reducing liver fat may prove beneficial in preventing HCC[22]. Therefore, PRS can be a non-invasive tool to predict the risk of HCC in individuals with NAFLD and dysmetabolism, independently of severe liver fibrosis. Nonetheless, the poor to moderate performance of PRS in risk prediction necessitates further studies to incorporate PRS using whole GWAS summary statistics, or evaluate their combination with other metabolic risk factors and clinical scores.

In **paper II**, we identified two new missense variants on *GPAM* and *APOE* robustly associated with liver damage and steatosis both at population level and in those at risk for liver disease. The majority of ALT associated *loci* (~80%) were novel, and most of those previously reported were associated with lipoproteins and lipid metabolism. In line with these findings and study design, in **paper III**, we attempted to examine the gene-environment interaction effects associated with fatty liver disease susceptibility. For this purpose, we conducted a gene-BMI association study in predicting ALT levels, which identified a new *locus* associated with chronic liver disease and hepatic fat content. Interestingly, this *locus* did not associate with ALT, and therefore, could not be detected via conventional GWA studies.

One important key concept underlying all these studies is the presence of a tight relationship between liver damage, hepatic fat content and lipoprotein metabolism. With the availability of more complex models and wide range of genetic information, such as whole-exome/-genome sequencing data, one can potentially find yet more missing pieces of the heritability puzzle of fatty liver disease.

ACKNOWLEDGEMENT

There are several people who supported me in different ways, and I would be remiss if I did not express my gratitude to:

My supervisor, **Stefano** for offering me this opportunity to work in such a great environment. It was an honor for me to learn from you at different levels. Your critical thinking mindset, open-mindedness towards new ideas, and passion about genetics have always inspired me.

Our collaborators, especially professor **Luca Valenti** for his brilliant thoughts and insights during the past four years. Also, thanks to the other great collaborators: **Umberto Vespasiani-Gentilucci, Federica Tavaglione**, **Antonio De Vincentis**, **Guido Baselli** and **Cristiana Bianco**.

My [past and present] friends at Lab 13. **Rosy** (AKA Rosin, Roseli, Roselin, Rosè, and Perf…) and **Piero** (AKA Master of coffee, Pieriano, vieni vieni..) thanks for your invaluable friendship, what you taught, and for all the memorable moments we had together. **Andrea** (AKA Andra, Cardo Macceo and Coach), you were the first person I met here five years ago, charming and cheerful. Thanks for the amazing memories! **Kavi** (AKA, Kafi, Kaf/aɪ/ta and Baby K), thanks for all the panic, jolliness and [sometimes] kindness. **Ester** (AKA Ster), thanks for being such a sincere friend, and [sometimes endless] joyful chats. **Grazia** (better to skip the nicknames), thanks for all the valuable discussions, I've learnt a lot from you (especially about N/ASH). **Angela**, thanks for your kindness and generosity, especially delicious cakes! And our two newest members **Francesca** and **Tanmoy**, welcome!

My friends and colleagues at WLAB. **Matias** (AKA Candyman and The Glorious), thanks for the wholehearted friendship and all the [>95% nonsensical] discussions. **Matthias**, thanks for being my [introvert] friend (nonsensical chats remain the same here). **Chrissy** (AKA Chrisis), thanks for all the fun, detailed description of [almost] everything, and being my Subway buddy.

Also, thanks to my other (past and present) friends and all WLAB for their warmth, laughs and creating a pleasant environment, especially, **Barbara, Wilhelm, Kassem** (AKA the Chef)**, Malik, Shafaat, Kristina, Gaohua**, **Alba**, **Vagner**, **Ismena**, **Ara** and **Tony**.

A huge thanks also to all unknown scientists, researchers and developers at **Stack Exchange** big family for their scientific benevolence and the tremendous effort!

در نهایت، و مهمتر از همه، سپاس از خانواده و به ویژه **سهیلا**ی عزیزم، که در تمام این سال ها پشتیبان و بالاتر از همه دوست من بود. البته که واژه ها تکلفند و از این رو، زیاده نویسی نمی کنم. ممنون به خاطر «بودن»♥ .

REFERENCES

1. Bianco C, Jamialahmadi O, Pelusi S, et al. Non-invasive stratification of hepatocellular carcinoma risk in non-alcoholic fatty liver using polygenic risk scores. J Hepatol 2021;74:775-782.
2. Jamialahmadi O, Mancina RM, Ciociola E, et al. Exome-Wide Association Study on Alanine Aminotransferase Identifies Sequence Variants in the GPAM and APOE Associated With Fatty Liver Disease. Gastroenterology 2021;160:1634-1646.e7.
3. Powell EE, Wong VW, Rinella M. Non-alcoholic fatty liver disease. Lancet 2021;397:2212-2224.
4. Carlsson B, Lindén D, Brolén G, et al. Review article: the emerging role of genetics in precision medicine for patients with non-alcoholic steatohepatitis. Aliment Pharmacol Ther 2020;51:1305-1320.
5. Huang TD, Behary J, Zekry A. Non-alcoholic fatty liver disease: a review of epidemiology, risk factors, diagnosis and management. Intern Med J 2020;50:1038-1047.
6. Younossi ZM, Rinella ME, Sanyal AJ, et al. From NAFLD to MAFLD: Implications of a Premature Change in Terminology. Hepatology 2021;73:1194-1198.
7. Cohen JC, Horton JD, Hobbs HH. Human fatty liver disease: old questions and new insights. Science 2011;332:1519-23.
8. Eslam M, Valenti L, Romeo S. Genetics and epigenetics of NAFLD and NASH: Clinical impact. J Hepatol 2018;68:268-279.
9. Tilg H, Moschen AR. Evolution of inflammation in nonalcoholic fatty liver disease: the multiple parallel hits hypothesis. Hepatology 2010;52:1836-46.
10. Taylor RS, Taylor RJ, Bayliss S, et al. Association Between Fibrosis Stage and Outcomes of Patients With Nonalcoholic Fatty Liver Disease: A Systematic Review and Meta-Analysis. Gastroenterology 2020;158:1611-1625.e12.
11. Targher G, Byrne CD, Tilg H. NAFLD and increased risk of cardiovascular disease: clinical associations, pathophysiological mechanisms and pharmacological implications. Gut 2020;69:1691-1705.
12. Younossi ZM, Koenig AB, Abdelatif D, et al. Global epidemiology of nonalcoholic fatty liver disease-Meta-analytic assessment of prevalence, incidence, and outcomes. Hepatology 2016;64:73-84.
13. Ye Q, Zou B, Yeo YH, et al. Global prevalence, incidence, and outcomes of non-obese or lean non-alcoholic fatty liver disease: a systematic review and meta-analysis. Lancet Gastroenterol Hepatol 2020;5:739-752.

14. Loomba R, Friedman SL, Shulman GI. Mechanisms and disease consequences of nonalcoholic fatty liver disease. Cell 2021;184:2537-2564.
15. Estes C, Anstee QM, Arias-Loste MT, et al. Modeling NAFLD disease burden in China, France, Germany, Italy, Japan, Spain, United Kingdom, and United States for the period 2016-2030. J Hepatol 2018;69:896-904.
16. Wong VW, Adams LA, de Lédinghen V, et al. Noninvasive biomarkers in NAFLD and NASH - current progress and future promise. Nat Rev Gastroenterol Hepatol 2018;15:461-478.
17. Tamaki N, Ajmera V, Loomba R. Non-invasive methods for imaging hepatic steatosis and their clinical importance in NAFLD. Nat Rev Endocrinol 2022;18:55-66.
18. Davison BA, Harrison SA, Cotter G, et al. Suboptimal reliability of liver biopsy evaluation has implications for randomized clinical trials. J Hepatol 2020;73:1322-1332.
19. Younossi Z, Anstee QM, Marietti M, et al. Global burden of NAFLD and NASH: trends, predictions, risk factors and prevention. Nat Rev Gastroenterol Hepatol 2018;15:11-20.
20. Younossi ZM. Non-alcoholic fatty liver disease - A global public health perspective. J Hepatol 2019;70:531-544.
21. Juanola O, Martínez-López S, Francés R, et al. Non-Alcoholic Fatty Liver Disease: Metabolic, Genetic, Epigenetic and Environmental Risk Factors. Int J Environ Res Public Health 2021;18.
22. Romeo S, Sanyal A, Valenti L. Leveraging Human Genetics to Identify Potential New Treatments for Fatty Liver Disease. Cell Metab 2020;31:35-45.
23. Albillos A, de Gottardi A, Rescigno M. The gut-liver axis in liver disease: Pathophysiological basis for therapy. J Hepatol 2020;72:558-577.
24. Trépo E, Valenti L. Update on NAFLD genetics: From new variants to the clinic. J Hepatol 2020;72:1196-1209.
25. Gallagher MD, Chen-Plotkin AS. The Post-GWAS Era: From Association to Function. Am J Hum Genet 2018;102:717-730.
26. Mills MC, Rahal C. A scientometric review of genome-wide association studies. Commun Biol 2019;2:9.
27. Eslam M, George J. Genetic contributions to NAFLD: leveraging shared genetics to uncover systems biology. Nat Rev Gastroenterol Hepatol 2020;17:40-52.
28. Tam V, Patel N, Turcotte M, et al. Benefits and limitations of genome-wide association studies. Nat Rev Genet 2019;20:467-484.
29. Visscher PM, Wray NR, Zhang Q, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet 2017;101:5-22.

30. Loos RJF. 15 years of genome-wide association studies and no signs of slowing down. Nat Commun 2020;11:5900.
31. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. Nature 2009;461:747-53.
32. Hirschhorn JN. Genomewide association studies--illuminating biologic pathways. N Engl J Med 2009;360:1699-701.
33. Finucane HK, Bulik-Sullivan B, Gusev A, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet 2015;47:1228-35.
34. Reay WR, Cairns MJ. Advancing the use of genome-wide association studies for drug repurposing. Nat Rev Genet 2021;22:658-671.
35. Crouch DJM, Bodmer WF. Polygenic inheritance, GWAS, polygenic risk scores, and the search for functional variants. Proc Natl Acad Sci U S A 2020;117:18924-18933.
36. Zaitlen N, Pasaniuc B, Sankararaman S, et al. Leveraging population admixture to characterize the heritability of complex traits. Nat Genet 2014;46:1356-62.
37. Li R, Chen Y, Ritchie MD, et al. Electronic health records and polygenic risk scores for predicting disease risk. Nat Rev Genet 2020;21:493-502.
38. Choi SW, Mak TS, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. Nat Protoc 2020;15:2759-2772.
39. Wand H, Lambert SA, Tamburro C, et al. Improving reporting standards for polygenic scores in risk prediction studies. Nature 2021;591:211-219.
40. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. Nat Rev Genet 2018;19:581-590.
41. Wray NR, Lin T, Austin J, et al. From Basic Science to Clinical Application of Polygenic Risk Scores: A Primer. JAMA Psychiatry 2021;78:101-109.
42. Bowden J, Holmes MV. Meta-analysis and Mendelian randomization: A review. Res Synth Methods 2019;10:486-496.
43. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Hum Mol Genet 2014;23:R89-98.
44. Richmond RC, Davey Smith G. Mendelian Randomization: Concepts and Scope. Cold Spring Harb Perspect Med 2022;12.
45. Emdin CA, Khera AV, Kathiresan S. Mendelian Randomization. JAMA 2017;318:1925-1926.
46. Smith GD, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? Int J Epidemiol 2003;32:1-22.
47. Zheng J, Baird D, Borges MC, et al. Recent Developments in Mendelian Randomization Studies. Curr Epidemiol Rep 2017;4:330-345.

48. Davies NM, Holmes MV, Davey Smith G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. BMJ 2018;362:k601.
49. Skrivankova VW, Richmond RC, Woolf BAR, et al. Strengthening the Reporting of Observational Studies in Epidemiology Using Mendelian Randomization: The STROBE-MR Statement. JAMA 2021;326:1614-1621.
50. Loomba R, Schork N, Chen CH, et al. Heritability of Hepatic Fibrosis and Steatosis Based on a Prospective Twin Study. Gastroenterology 2015;149:1784-93.
51. Stender S, Kozlitina J, Nordestgaard BG, et al. Adiposity amplifies the genetic risk of fatty liver disease conferred by multiple loci. Nat Genet 2017;49:842-847.
52. Speliotes EK, Yerges-Armstrong LM, Wu J, et al. Genome-wide association analysis identifies variants associated with nonalcoholic fatty liver disease that have distinct effects on metabolic traits. PLoS Genet 2011;7:e1001324.
53. Emdin CA, Haas ME, Khera AV, et al. A missense variant in Mitochondrial Amidoxime Reducing Component 1 gene and protection against liver disease. PLoS Genet 2020;16:e1008629.
54. Romeo S, Kozlitina J, Xing C, et al. Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease. Nat Genet 2008;40:1461-5.
55. Pirazzi C, Valenti L, Motta BM, et al. PNPLA3 has retinyl-palmitate lipase activity in human hepatic stellate cells. Hum Mol Genet 2014;23:4077-85.
56. BasuRay S, Smagris E, Cohen JC, et al. The PNPLA3 variant associated with fatty liver disease (I148M) accumulates on lipid droplets by evading ubiquitylation. Hepatology 2017;66:1111-1124.
57. Wang Y, Kory N, BasuRay S, et al. PNPLA3, CGI-58, and Inhibition of Hepatic Triglyceride Hydrolysis in Mice. Hepatology 2019;69:2427-2441.
58. Lindén D, Ahnmark A, Pingitore P, et al. Pnpla3 silencing with antisense oligonucleotides ameliorates nonalcoholic steatohepatitis and fibrosis in Pnpla3 I148M knock-in mice. Mol Metab 2019;22:49-61.
59. Valenti LVC, Cherubini A. To Be or Not to Be: The Quest for Patatin-Like Phospholipase Domain Containing 3 p.I148M Function. Hepatology 2021;74:2942-2944.
60. Sookoian S, Pirola CJ, Valenti L, et al. Genetic Pathways in Nonalcoholic Fatty Liver Disease: Insights From Systems Biology. Hepatology 2020;72:330-346.
61. Kozlitina J, Smagris E, Stender S, et al. Exome-wide association study identifies a TM6SF2 variant that confers susceptibility to nonalcoholic fatty liver disease. Nat Genet 2014;46:352-6.

62. Buch S, Stickel F, Trépo E, et al. A genome-wide association study confirms PNPLA3 and identifies TM6SF2 and MBOAT7 as risk loci for alcohol-related cirrhosis. Nat Genet 2015;47:1443-8.
63. Caddeo A, Jamialahmadi O, Solinas G, et al. MBOAT7 is anchored to endomembranes by six transmembrane domains. J Struct Biol 2019;206:349-360.
64. Mancina RM, Dongiovanni P, Petta S, et al. The MBOAT7-TMC4 Variant rs641738 Increases Risk of Nonalcoholic Fatty Liver Disease in Individuals of European Descent. Gastroenterology 2016;150:1219-1230.e6.
65. Teo K, Abeysekera KWM, Adams L, et al. rs641738C>T near MBOAT7 is associated with liver fat, ALT and fibrosis in NAFLD: A meta-analysis. J Hepatol 2021;74:20-30.
66. Abul-Husn NS, Cheng X, Li AH, et al. A Protein-Truncating HSD17B13 Variant and Protection from Chronic Liver Disease. N Engl J Med 2018;378:1096-1106.
67. Luukkonen PK, Juuti A, Sammalkorpi H, et al. MARC1 variant rs2642438 increases hepatic phosphatidylcholines and decreases severity of non-alcoholic fatty liver disease in humans. J Hepatol 2020;73:725-726.
68. Fall T, Hägg S, Mägi R, et al. The role of adiposity in cardiometabolic traits: a Mendelian randomization analysis. PLoS Med 2013;10:e1001474.
69. Dongiovanni P, Stender S, Pietrelli A, et al. Causal relationship of hepatic fat with liver damage and insulin resistance in nonalcoholic fatty liver. J Intern Med 2018;283:356-370.
70. Bianco C, Tavaglione F, Romeo S, et al. Genetic risk scores and personalization of care in fatty liver disease. Curr Opin Pharmacol 2021;61:6-11.
71. De Vincentis A, Tavaglione F, Jamialahmadi O, et al. A Polygenic Risk Score to Refine Risk Stratification and Prediction for Severe Liver Disease by Clinical Fibrosis Scores. Clin Gastroenterol Hepatol 2022;20:658-673.
72. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature 2018;562:203-209.
73. Sudlow C, Gallacher J, Allen N, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLOS Medicine 2015;12:e1001779.
74. Palmer DS, Zhou W, Abbott L, et al. Analysis of genetic dominance in the UK Biobank. bioRxiv 2022:2021.08.15.456387.
75. Loh PR, Kichaev G, Gazal S, et al. Mixed-model association for biobank-scale datasets. Nat Genet 2018;50:906-908.
76. Loh PR, Tucker G, Bulik-Sullivan BK, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat Genet 2015;47:284-90.

77. Zhou W, Nielsen JB, Fritsche LG, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat Genet 2018;50:1335-1341.
78. Mbatchou J, Barnard L, Backman J, et al. Computationally efficient whole-genome regression for quantitative and binary traits. Nat Genet 2021;53:1097-1103.
79. Manichaikul A, Mychaleckyj JC, Rich SS, et al. Robust relationship inference in genome-wide association studies. Bioinformatics 2010;26:2867-73.
80. Bron C, Kerbosch J. Algorithm 457: finding all cliques of an undirected graph. Communications of the ACM 1973;16:575-577.
81. Chang CC, Chow CC, Tellier LC, et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 2015;4:7.
82. Bulik-Sullivan BK, Loh PR, Finucane HK, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet 2015;47:291-5.
83. Balding DJ. A tutorial on statistical methods for population association studies. Nat Rev Genet 2006;7:781-91.
84. Uffelmann E, Huang QQ, Munung NS, et al. Genome-wide association studies. Nature Reviews Methods Primers 2021;1:59.
85. Langner T, Strand R, Ahlström H, et al. Large-Scale Inference of Liver Fat with Neural Networks on UK Biobank Body MRI, In Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, Cham, 2020//, Springer International Publishing, 2020.
86. Langner T, Gustafsson FK, Avelin B, et al. Uncertainty-aware body composition analysis with deep regression ensembles on UK Biobank MRI. Comput Med Imaging Graph 2021;93:101994.
87. Linge J, Whitcher B, Borga M, et al. Sub-phenotyping Metabolic Disorders Using Body Composition: An Individualized, Nonparametric Approach Utilizing Large Data Sets. Obesity (Silver Spring) 2019;27:1190-1199.
88. Langner T, Strand R, Ahlström H, et al. Large-scale biometry with interpretable neural network regression on UK Biobank body MRI. Sci Rep 2020;10:17752.
89. Parisinos CA, Wilman HR, Thomas EL, et al. Genome-wide and Mendelian randomisation studies of liver MRI yield insights into the pathogenesis of steatohepatitis. J Hepatol 2020;73:241-251.
90. Yang J, Lee SH, Goddard ME, et al. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 2011;88:76-82.
91. Cingolani P, Platts A, Wang lL, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 2012;6:80-92.

92. Weissbrod O, Hormozdiari F, Benner C, et al. Functionally informed fine-mapping and polygenic localization of complex trait heritability. Nat Genet 2020;52:1355-1363.
93. Benner C, Spencer CC, Havulinna AS, et al. FINEMAP: efficient variable selection using summary data from genome-wide association studies. Bioinformatics 2016;32:1493-501.
94. Wang G, Sarkar A, Carbonetto P, et al. A simple new approach to variable selection in regression, with application to genetic fine mapping. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2020;82:1273-1300.
95. Giambartolomei C, Vukcevic D, Schadt EE, et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet 2014;10:e1004383.
96. de Leeuw CA, Mooij JM, Heskes T, et al. MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput Biol 2015;11:e1004219.
97. Yavorska OO, Burgess S. MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. Int J Epidemiol 2017;46:1734-1739.
98. Hemani G, Zheng J, Elsworth B, et al. The MR-Base platform supports systematic causal inference across the human phenome. Elife 2018;7.
99. Turley P, Walters RK, Maghzian O, et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. Nat Genet 2018;50:229-237.
100. Ghoussaini M, Mountjoy E, Carmona M, et al. Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. Nucleic Acids Res 2021;49:D1311-D1320.
101. Buniello A, MacArthur JAL, Cerezo M, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res 2019;47:D1005-D1012.
102. Kamat MA, Blackshaw JA, Young R, et al. PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. Bioinformatics 2019;35:4851-4853.
103. Anstee QM, Reeves HL, Kotsiliti E, et al. From NASH to HCC: current concepts and future challenges. Nat Rev Gastroenterol Hepatol 2019;16:411-428.
104. Dongiovanni P, Romeo S, Valenti L. Hepatocellular carcinoma in nonalcoholic fatty liver: role of environmental and genetic factors. World J Gastroenterol 2014;20:12945-55.
105. Loomba R, Lim JK, Patton H, et al. AGA Clinical Practice Update on Screening and Surveillance for Hepatocellular Carcinoma in Patients With Nonalcoholic Fatty Liver Disease: Expert Review. Gastroenterology 2020;158:1822-1830.

106. Mancina RM, Dongiovanni P, Petta S, et al. The MBOAT7-TMC4 Variant rs641738 Increases Risk of Nonalcoholic Fatty Liver Disease in Individuals of European Descent. Gastroenterology 2016;150:1219-1230.e6.
107. Sattar N, Forrest E, Preiss D. Non-alcoholic fatty liver disease. BMJ 2014;349:g4596.
108. Kim WR, Flamm SL, Di Bisceglie AM, et al. Serum activity of alanine aminotransferase (ALT) as an indicator of health and disease. Hepatology 2008;47:1363-70.
109. Rivas MA, Pirinen M, Conrad DF, et al. Human genomics. Effect of predicted protein-truncating genetic variants on the human transcriptome. Science 2015;348:666-9.
110. Emdin CA, Khera AV, Chaffin M, et al. Analysis of predicted loss-of-function variants in UK Biobank identifies variants protective for disease. Nat Commun 2018;9:1613.
111. Coleman RA. It takes a village: channeling fatty acid metabolism and triacylglycerol formation via protein interactomes. J Lipid Res 2019;60:490-497.
112. Lindén D, William-Olsson L, Ahnmark A, et al. Liver-directed overexpression of mitochondrial glycerol-3-phosphate acyltransferase results in hepatic steatosis, increased triacylglycerol secretion and reduced fatty acid oxidation. FASEB J 2006;20:434-43.
113. Neschen S, Morino K, Hammond LE, et al. Prevention of hepatic steatosis and hepatic insulin resistance in mitochondrial acyl-CoA:glycerol-sn-3-phosphate acyltransferase 1 knockout mice. Cell Metab 2005;2:55-65.
114. Jiang ZG, Robson SC, Yao Z. Lipoprotein metabolism in nonalcoholic fatty liver disease. J Biomed Res 2013;27:1-13.
115. Kerin M, Marchini J. Inferring Gene-by-Environment Interactions with a Bayesian Whole-Genome Regression Model. Am J Hum Genet 2020;107:698-713.
116. Zou Y, Carbonetto P, Wang G, et al. Fine-mapping from summary data with the "Sum of Single Effects" model. PLoS Genet 2022;18:e1010299.
117. Fairfield CJ, Drake TM, Pius R, et al. Genome-Wide Association Study of NAFLD Using Electronic Health Records. Hepatol Commun 2022;6:297-308.
118. Ghodsian N, Abner E, Emdin CA, et al. Electronic health record-based genome-wide meta-analysis provides insights on the genetic architecture of non-alcoholic fatty liver disease. Cell Rep Med 2021;2:100437.
119. Liu Y, Basty N, Whitcher B, et al. Genetic architecture of 11 organ traits derived from abdominal MRI using deep learning. Elife 2021;10.
120. Vujkovic M, Ramdas S, Lorenz KM, et al. 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 2022;54:761-771.
121. Kerimov N, Hayhurst JD, Peikova K, et al. A compendium of uniformly processed human gene expression and splicing quantitative trait loci. Nat Genet 2021;53:1290-1299.
122. Consortium G. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science 2020;369:1318-1330.
123. Bowker N, Hansford R, Burgess S, et al. Genetically Predicted Glucose-Dependent Insulinotropic Polypeptide (GIP) Levels and Cardiovascular Disease Risk Are Driven by Distinct Causal Variants in the. Diabetes 2021;70:2706-2719.