3 Chapter 3: The Role of Rare Variation in Circulating Metabolic Biomarkers

3.1 Introduction

Metabolic measurements reflect an individual's endogenous biochemical processes and environmental exposures [256, 257]. Many circulating lipids, lipoproteins and metabolites have been previously implicated in the development of cardiovascular disease (CVD) [258-261] or used as biomarkers for disease diagnosis or prognosis [262, 263]. High circulating levels of total cholesterol (TC) and low-density lipoprotein (LDL) cholesterol, for example, have been associated with increased risk of coronary heart disease (CHD)[264]. On the other hand, circulating levels of high-density lipoprotein (HDL) cholesterol have been regarded as protective factors against CHD [265]. Despite the observed association between low HDL levels and increased CHD risk, a causal role for HDL levels was more unclear before genetic studies, due to potential confounding by other CHD risk factors correlated with low HDL, like increased plasma triglycerides (TG) [266].

In the diagnostic setting, metabolites like creatinine and branched chain amino acids (valine, leucine and isoleucine) are helpful biomarkers for diseases such a kidney disease [267] , or hyperinsulinism [268-270]. Understanding the genetic influence on circulating levels of these metabolic biomarkers can help us gain insight into the biological processes regulating these traits, lead to improve aetiological understanding of CVD and identify novel potential therapeutic drug targets. Notable examples of candidate drug targets identified via genetic approaches are *LDLR* [271, 272], *APOB* [273, 274] and *PCSK9* [275, 276]. Mipomersen, a commercially available *APOB* inhibitor, has already shown association with reduction in cardiovascular events in patients with hypercholesterolaemia [277] and two

PCSK9 inhibitors: alirocumab and evolocumab have been shown to reduce risk of myocardial infranction (MI) and stroke in clinical trials [278].

Genome-wide association studies (GWAS) focusing on traditionally measured lipid traits have greatly expanded our knowledge of lipid biology and to date 250 loci have been robustly associated with total cholesterol (TC), high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), and/or triglycerides (TG) [84, 116, 279-285]. Through these studies it has been found that most loci identified in European populations contribute to the genetic architecture of lipid traits across global populations [116], that there are metabolic links between blood lipids and type 2 diabetes, blood pressure, waist-hip-ratio and BMI [280], and more recently that some mechanisms of lowering LDL-C increase type 2 diabetes (T2D) risk [84]. Mendelian randomisation (MR) approaches have also used information gained through GWAS to examine the causal link between low HDL levels and CVD where findings suggest that low HDL levels are not causal for CVD since many studies report no association between CVD and genetically lowered levels of HDL [110-114]. These MR approaches have also been used to identify a potential causal link between increased plasma urate levels and CVD [286], although other studies measuring serum urate levels have not found that link [287].

In addition to this, more detailed metabolic profiling using high resolution nuclear magnetic resonance (NMR) measurements, has proven helpful to find additional lipid and small molecule metabolism-associated loci with smaller sample sizes, and to assess pleiotropic effects of previously established loci [38, 173, 288]. An example of this, is a novel link between the *LPA* locus and very-low-density lipoprotein (VLDL) metabolism (measured by using high resolution NMR), with effect sizes twice as large as those found for traditionally

measured lipid traits like LDL-C and TC, suggesting these measurements are better at capturing the underlying biological processes in lipid metabolism than traditionally measured lipid traits. In this same study, by constructing a genetic risk score using variants associated with Lp(a) levels and using a Mendelian randomisation approach the authors were able to demonstrate a causal link between increased Lp(a) levels and overall lipoprotein metabolism [173].

Despite the increased usage of exome arrays which have been used at scale to capture low-frequency and rare coding variation contributing to lipid and amino acid metabolism [84, 282-284, 288, 289], large-scale sequencing studies have the added value of assessing rare variation at single nucleotide resolution across the whole genome, or exome, including the detection of private variants which could have large effects on protein function. These approaches enabled, for example, the discovery of inactivating variants in key proteins which are models for drug target antagonism such as ANGPTL4, where carriers of a missense E40K variant and other inactivating variants had reduced risk of CHD [290, 291].

Notwithstanding the progress made in recent years in understanding the genetic aetiology of a number of traditional lipid traits, at the time of this analysis, there were no studies coupling NMR measurements with sequencing data to explore the role of rare genetic variants in the metabolism of high resolution lipid, lipoprotein and metabolite traits. In this chapter, I address this gap in knowledge by examining the contribution of rare variation (MAF <1%) to 226 serum metabolic measurements in the INTERVAL cohort which consist of healthy blood donors residing in the UK. This project was done in collaboration with Dr Adam Butterworth's group at the University of Cambridge. My work involved QCing of sequencing and phenotype data as well as all analytical aspects of the study.

3.2 Chapter aims

The overall aim of this chapter is to explore how coupling next generation sequencing (NGS) and high resolution metabolic measurements can help us gain new insights into metabolic biomarker biology through rare variant analyses. To do this, I took advantage of the INTERVAL cohort, which is comprised of healthy blood donors who have been deeply phenotyped and who also have genome-wide array data. In my project I used data from a subset of 7,142 participants with NMR measurements and NGS data to:

I. Identify novel loci, genes and/or gene sets associated with metabolic biomarkers.
II. Identify effector transcripts at established GWAS loci for traditionally measured lipid traits.
III. Assess the contribution of genes known to be involved in lipoprotein metabolism to the tails of the phenotype distribution of lipoprotein and glyceride traits in a healthy population.

3.3 Methods

3.3.1 Participants

The INTERVAL cohort consists of 47,393 predominantly healthy blood donors in the UK [292]. This study was the result of a collaboration between the Universities of Cambridge and Oxford and the NHS Blood and Transplant Unit. The study was set up to determine the optimum intervals between donations for men and women without affecting the overall health of blood donors. Individuals were asked to fill an online general questionnaire every six months containing basic lifestyle and health-related information. At the time of this

study, a different set of biomarker assays were performed on blood samples collected on the first visit and those collected on the 2 year follow-up visit. All individuals have been genotyped using the Affymetrix UK Biobank Axiom Array and imputed using a combined UK10K-1000G Phase III imputation panel [293]. A subset of 4,502 individuals was selected for whole-exome sequencing (WES) [294] and another subset of 3,762 was selected for whole-genome sequencing (WGS). There was an overlap of 54 individuals in both datasets.

3.3.2 Sequencing and genotype calling

WES and WGS were performed at the Wellcome Sanger Institute (WSI) sequencing facility, with read alignment and genotype calling performed by the Human Genetics Informatics (HGI) group at Sanger. For WES sheared DNA was prepared for Illumina paired-end sequencing and enriched for target regions using Agilent's SureSelect Human All Exon V5 capture technology (Agilent Technologies; Santa Clara, California, USA). The exome capture library preparation was sequenced using the Illumina HiSeq 2000 platform as paired-end 75 bp reads. Reads were aligned to the GRCh37 human reference genome using BWA (v0.5.10) [295]. GATK HaplotypeCaller v3.4 [296] was used for variant calling and recalibration. For WGS sheared DNA was prepared for Illumina paired-end sequencing. Sequencing was performed using the Illumina HiSeq X platform as paired-end 75 bp reads. Reads were aligned to the GRCh38 human reference genome using mostly BWA (v.0.7.12) although a subset of samples was aligned with v.0.7.13 or v.0.7.15. GATK HaplotypeCaller v3.5 was used for variant calling and recalibration. I extracted coordinates from the VCF files that mapped to regions targeted in the WES. I then used custom scripts to transform coordinates of variants to GRCh37 human reference.

3.3.3 Sample QC

I performed sample QC for WES using the same filters Tarjinder Singh used on a previous release of the INTERVAL WES dataset [294]. Sample QC for WGS was performed by Kousik Kundu, Klaudia Walter and I. For WES data I filtered out samples based on the following criteria: i) withdrawn consent; ii) estimated contamination >3% according to the software VerifyBamID [297]; iii) sex inferred from genetic data different from sex supplied ; iv) non-European samples after manual inspection of clustering in 1000G principal components analysis (PCA) and choosing cutoffs on the first 2 PCs; v) heterozygosity outliers (samples +/- 3 SD away from the mean number of heterozygous counts); vi) non-reference homozygosity outliers (samples +/- 3 SD away from the mean number of non-reference homozygous counts); vii) outlier Ti/TV rates (transition to transversion ratio +/- 3 SD away from the mean ratio); viii) excess singletons (number of singleton variants >3 SD from the cohort mean). After quality control 4,070 WES samples were kept for downstream analyses. For WGS data we filtered out samples based on the following criteria: i) estimated contamination >2% according to software VerifyBamID; ii) non-reference discordance (NRD) with genotype data on the same samples >4%; iii) European population outliers from PCA (PC1>0 and minimum PC2); iv) heterozygosity outliers (samples +/- 3 SD away from the mean number of heterozygous counts); v) number of third-degree relatives (proportion IBD (PI_HAT) >0.125) > 18, vi) overlap with WES. After quality control 3,670 WGS samples were kept for further analyses.

3.3.4 Variant QC

For variants with MAF>1% I used the following thresholds to exclude variants: i) VQSR: 99.90% tranche for WES and 99% tranche for WGS; ii) missingness >3%; iii) HWE $p<1x10^{-5}$. For variants with MAF≤1% the following thresholds were used: i) VQSR: 99.90% tranche for WES, 99% tranche for WGS SNPs and 90% tranche for WGS indels; ii) GQ: <20 for SNPs and <60 for indels; iii) DP <2; iv) AB>15 & <80 for heterozygous variants. After genotype-level QC (GQ,DP,AB) only variants with <3% missingness were kept. 1,716,946 variants were kept in the final WES release and 1,724,250 in the final WGS release.

3.3.5 Phenotype QC

A total of 230 metabolic biomarkers were produced by the serum NMR metabolomics platform (Nightingale Health Ltd.) [298] on 46,097 blood samples from the INTERVAL cohort collected on the first visit. Phenotyping was performed by Antti J. Kangas (Nightingale Health Ltd.). I performed phenotype QC on the raw phenotypes. Glucose, lactose, pyruvate and acetate were excluded initially due to unreliable measurements according to platform provider. Conjugated linoleic acid and conjugated linoleic acid to total fatty acid ratio were set to missing for 3,585 samples showing signs of peroxidation. Creatinine levels were set to missing for 1,993 samples with isopropyl alcohol signals. Glutamine levels were set to missing for 347 samples that showed signs of glutamine to glutamate degradation. Samples with more than 30% missingness or identified as EDTA plasma were removed. After this step, for each pair of related samples (PI_HAT>0.125) I kept only one, preferentially keeping samples with the lowest missingness in WES or lowest NRD in WGS. Phenotypes were rank-based inverse normalised for all individuals. Clare Oliver-Williams assessed which technical

covariates influenced phenotype levels and determined centre, processing duration and month of donation were possible sources of batch effects. I then separately performed linear regression for WES and WGS adjusting for age, gender, centre, processing duration, month of donation and 10 PCs. Residuals from both WES and WGS linear regressions were used as the outcome variables in all subsequent analyses. After this final step I kept 3,741 samples in the WES dataset and 3,420 samples in the WGS dataset for downstream analyses.

3.3.6 Single point analyses

Power calculations to define MAF threshold for single point analyses were done using Quanto [234]. I used the WES data as a discovery dataset and performed association analyses using SNPTEST v2.5.2 [226] under an additive model. Variants were taken forward for validation if $p < 1x10^{-5}$. I then performed association analyses using SNPTEST on the WGS data which was used as a validation dataset. Results were subsequently meta-analysed using a fixed-effects model in METAL [238]. Genome-wide significance threshold was calculated as: $0.05/ (276,563*19)=9.52x10^{-9}$, where 276,563 is the number of tested variants with MAF>0.1% and 19 is the number of PCs explaining >95% of the variance of 226 metabolic biomarkers , an approach previously used in similar studies using the same NMR platform [38, 173]. A signal was considered to replicate if after meta-analysis it met the following criteria: i) it met the defined genome-wide significance threshold ($9.52x10^{-9}$); and ii) it was nominally significant ($p<0.05$) in the validation dataset (WGS). After this step, to define loci, I performed clumping using PLINK [223] based on the lowest p for each variant on any trait-association using an $r^2=0.2$ and a window size of 1Mb.

3.3.7 Gene-based analyses

I annotated coding variant consequences with VEP [50] using Ensembl gene set version 75 for the hg19/GRCh37 human genome assembly. Loss-of–function (LoF) variants were annotated with a VEP plugin: LOFTEE (https://github.com/konradjk/loftee). This plugin uses distance to end of transcript and other in-frame splice sites, non-canonical splice site information and size of introns to remove LoF that are less likely to have a damaging impact on protein structure. I downloaded M-CAP scores and extracted all missense variants with AC>=1 in the WES or WGS datasets [51]. Two different nested tests were used to group rare variants into testable gene units: predicted to be high confidence LoF by LOFTEE in any transcript of the gene, and the same LoF variants plus rare (MAF <1%) missense variants, mapping to any transcript of the gene, predicted to be likely deleterious by M-CAP (M-CAP score >0.025) (MCAP+LoF). M-CAP uses a machine learning algorithm integrating multiple annotations (e.g base-pair conservation, amino acid conservation, chemical properties of substituted amino acid, etc) to predict the pathogenicity of rare (MAF <1%) missense variants.

I performed rare-variant aggregation tests as implemented in the SKAT-O R package [52, 53]. For the LoF tests, I performed a burden test (rho=1) whereas for the MCAP+LoF tests I used the optimal unified approach (method="optimal.adj"). Genes were taken forward for validation if $p<5x10^{-3}$.

To increase power in my analyses I also implemented a strategy to incorporate information from the multiple phenotypes measured in our dataset, by adjusting for correlated phenotypes, which has been shown to increase power in single point association analyses [30]. To minimise chances of a false positive association I only adjusted for phenotypes as

covariates at the validation stage ensuring evidence of association in discovery stage was present without adjustment for covariates. In order for a metabolic biomarker to be selected as a covariate in the validation stage, the following conditions had to be met: i) no evidence of genetic correlation ($p>0.05$) with outcome using publicly available summary statistics from Kettunen et al. (2016) [25]; ii) phenotypic correlation in our dataset >10%; iii) not belonging to same metabolic biomarker supergroup as outcome (**Table 3.1**). This approach resulted in 99 eligible NMR traits for which other traits could be used as covariates. METASKAT [54] was used to perform meta-analysis using the same parameters as in discovery. A signal was considered to replicate if: i) it met the Bonferroni corrected gene-level significance threshold ($p < 1.32x10^{-7}$); ii) >2 variants were tested; iii) it was nominally significant in the unadjusted test for WGS (i.e without adjusting for correlated traits). The Bonferroni corrected gene-level significance threshold was chosen after adjusting the standard gene-level significance threshold ($2.5x10^{-6}$) for 19 PCs. To test if a single variant was driving an observed association, I performed leave-one-out analysis for all variants contributing to the test. An association was considered to be driven by a single variant if, after removing it, the test resulted in a non-significant association ($p>0.05$).

Traits	Description	Supergroup	Single point tests	Gene tests	Gene-set tests	Enrichment near GWAS signals tests	Tails tests
XXL-VLDL-P	Concentration of chylomicrons and extremely large VLDL particles	Lipid and lipoprotein	X	X	X		X
XXL-VLDL-L	Total lipids in chylomicrons and extremely large VLDL	Lipid and lipoprotein	X	X	X		X
XXL-VLDL-PL	Phospholipids in chylomicrons and extremely large VLDL	Lipid and lipoprotein	X	X	X		X
XXL-VLDL-C	Total cholesterol in chylomicrons and extremely large VLDL	Lipid and lipoprotein	X	X	X	X	X
XXL-VLDL-CE	Cholesterol esters in chylomicrons and extremely large VLDL	Lipid and lipoprotein	X	X	X	X	X
XXL-VLDL-FC	Free cholesterol in chylomicrons and extremely large VLDL	Lipid and lipoprotein	X	X	X	X	X
XXL-VLDL-TG	Triglycerides in chylomicrons and extremely large VLDL	Lipid and lipoprotein	X	X	X	X	X
XL-VLDL-P	Concentration of very large VLDL particles	Lipid and lipoprotein	X	X	X		X
XL-VLDL-L	Total lipids in very large VLDL	Lipid and lipoprotein	X	X	X		X
XL-VLDL-PL	Phospholipids in very large VLDL	Lipid and lipoprotein	X	X	X		X

Traits	Description	Supergroup	Single point tests	Gene tests	Gene-set tests	Enrichment near GWAS signals tests	Tails tests
XL-VLDL-C	Total cholesterol in very large VLDL	Lipid and lipoprotein	X	X	X	X	X
XL-VLDL-CE	Cholesterol esters in very large VLDL	Lipid and lipoprotein	X	X	X	X	X
XL-VLDL-FC	Free cholesterol in very large VLDL	Lipid and lipoprotein	X	X	X	X	X
XL-VLDL-TG	Triglycerides in very large VLDL	Lipid and lipoprotein	X	X	X	X	X
L-VLDL-P	Concentration of large VLDL particles	Lipid and lipoprotein	X	X	X		X
L-VLDL-L	Total lipids in large VLDL	Lipid and lipoprotein	X	X	X		X
L-VLDL-PL	Phospholipids in large VLDL	Lipid and lipoprotein	X	X	X		X
L-VLDL-C	Total cholesterol in large VLDL	Lipid and lipoprotein	X	X	X	X	X
L-VLDL-CE	Cholesterol esters in large VLDL	Lipid and lipoprotein	X	X	X	X	X
L-VLDL-FC	Free cholesterol in large VLDL	Lipid and lipoprotein	X	X	X	X	X
L-VLDL-TG	Triglycerides in large VLDL	Lipid and lipoprotein	X	X	X	X	X
M-VLDL-P	Concentration of medium VLDL particles	Lipid and lipoprotein	X	X	X		X
M-VLDL-L	Total lipids in medium VLDL	Lipid and lipoprotein	X	X	X		X
M-VLDL-PL	Phospholipids in medium VLDL	Lipid and lipoprotein	X	X	X		X
M-VLDL-C	Total cholesterol in medium VLDL	Lipid and lipoprotein	X	X	X	X	X
M-VLDL-CE	Cholesterol esters in medium VLDL	Lipid and lipoprotein	X	X	X	X	X
M-VLDL-FC	Free cholesterol in medium VLDL	Lipid and lipoprotein	X	X	X	X	X
M-VLDL-TG	Triglycerides in medium VLDL	Lipid and lipoprotein	X	X	X	X	X
S-VLDL-P	Concentration of small VLDL particles	Lipid and lipoprotein	X	X	X		X
S-VLDL-L	Total lipids in small VLDL	Lipid and lipoprotein	X	X	X		X
S-VLDL-PL	Phospholipids in small VLDL	Lipid and lipoprotein	X	X	X		X
S-VLDL-C	Total cholesterol in small VLDL	Lipid and lipoprotein	X	X	X	X	X
S-VLDL-CE	Cholesterol esters in small VLDL	Lipid and lipoprotein	X	X	X	X	X
S-VLDL-FC	Free cholesterol in small VLDL	Lipid and lipoprotein	X	X	X	X	X
S-VLDL-TG	Triglycerides in small VLDL	Lipid and lipoprotein	X	X	X	X	X
XS-VLDL-P	Concentration of very small VLDL particles	Lipid and lipoprotein	X	X	X		X
XS-VLDL-L	Total lipids in very small VLDL	Lipid and lipoprotein	X	X	X		X
XS-VLDL-PL	Phospholipids in very small VLDL	Lipid and lipoprotein	X	X	X		X
XS-VLDL-C	Total cholesterol in very small VLDL	Lipid and lipoprotein	X	X	X	X	X
XS-VLDL-CE	Cholesterol esters in very small VLDL	Lipid and lipoprotein	X	X	X	X	X
XS-VLDL-FC	Free cholesterol in very small VLDL	Lipid and lipoprotein	X	X	X	X	X
XS-VLDL-TG	Triglycerides in very small VLDL	Lipid and lipoprotein	X	X	X	X	X
IDL-P	Concentration of IDL particles	Lipid and lipoprotein	X	X	X		
IDL-L	Total lipids in IDL	Lipid and lipoprotein	X	X	X		
IDL-PL	Phospholipids in IDL	Lipid and lipoprotein	X	X	X		
IDL-C	Total cholesterol in IDL	Lipid and lipoprotein	X	X	X	X	
IDL-CE	Cholesterol esters in IDL	Lipid and	X	X	X	X	

Traits	Description	Supergroup	Single point tests	Gene tests	Gene-set tests	Enrichment near GWAS signals tests	Tails tests
		lipoprotein					
IDL-FC	Free cholesterol in IDL	Lipid and lipoprotein	X	X	X	X	
IDL-TG	Triglycerides in IDL	Lipid and lipoprotein	X	X	X	X	
L-LDL-P	Concentration of large LDL particles	Lipid and lipoprotein	X	X	X	X	X
L-LDL-L	Total lipids in large LDL	Lipid and lipoprotein	X	X	X	X	X
L-LDL-PL	Phospholipids in large LDL	Lipid and lipoprotein	X	X	X	X	X
L-LDL-C	Total cholesterol in large LDL	Lipid and lipoprotein	X	X	X	X	X
L-LDL-CE	Cholesterol esters in large LDL	Lipid and lipoprotein	X	X	X	X	X
L-LDL-FC	Free cholesterol in large LDL	Lipid and lipoprotein	X	X	X	X	X
L-LDL-TG	Triglycerides in large LDL	Lipid and lipoprotein	X	X	X	X	X
M-LDL-P	Concentration of medium LDL particles	Lipid and lipoprotein	X	X	X	X	X
M-LDL-L	Total lipids in medium LDL	Lipid and lipoprotein	X	X	X	X	X
M-LDL-PL	Phospholipids in medium LDL	Lipid and lipoprotein	X	X	X	X	X
M-LDL-C	Total cholesterol in medium LDL	Lipid and lipoprotein	X	X	X	X	X
M-LDL-CE	Cholesterol esters in medium LDL	Lipid and lipoprotein	X	X	X	X	X
M-LDL-FC	Free cholesterol in medium LDL	Lipid and lipoprotein	X	X	X	X	X
M-LDL-TG	Triglycerides in medium LDL	Lipid and lipoprotein	X	X	X	X	X
S-LDL-P	Concentration of small LDL particles	Lipid and lipoprotein	X	X	X	X	X
S-LDL-L	Total lipids in small LDL	Lipid and lipoprotein	X	X	X	X	X
S-LDL-PL	Phospholipids in small LDL	Lipid and lipoprotein	X	X	X	X	X
S-LDL-C	Total cholesterol in small LDL	Lipid and lipoprotein	X	X	X	X	X
S-LDL-CE	Cholesterol esters in small LDL	Lipid and lipoprotein	X	X	X	X	X
S-LDL-FC	Free cholesterol in small LDL	Lipid and lipoprotein	X	X	X	X	X
S-LDL-TG	Triglycerides in small LDL	Lipid and lipoprotein	X	X	X	X	X
XL-HDL-P	Concentration of very large HDL particles	Lipid and lipoprotein	X	X	X	X	X
XL-HDL-L	Total lipids in very large HDL	Lipid and lipoprotein	X	X	X	X	X
XL-HDL-PL	Phospholipids in very large HDL	Lipid and lipoprotein	X	X	X	X	X
XL-HDL-C	Total cholesterol in very large HDL	Lipid and lipoprotein	X	X	X	X	X
XL-HDL-CE	Cholesterol esters in very large HDL	Lipid and lipoprotein	X	X	X	X	X
XL-HDL-FC	Free cholesterol in very large HDL	Lipid and lipoprotein	X	X	X	X	X
XL-HDL-TG	Triglycerides in very large HDL	Lipid and lipoprotein	X	X	X	X	X
L-HDL-P	Concentration of large HDL particles	Lipid and lipoprotein	X	X	X	X	X
L-HDL-L	Total lipids in large HDL	Lipid and lipoprotein	X	X	X	X	X
L-HDL-PL	Phospholipids in large HDL	Lipid and lipoprotein	X	X	X	X	X
L-HDL-C	Total cholesterol in large HDL	Lipid and lipoprotein	X	X	X	X	X
L-HDL-CE	Cholesterol esters in large HDL	Lipid and lipoprotein	X	X	X	X	X
L-HDL-FC	Free cholesterol in large HDL	Lipid and lipoprotein	X	X	X	X	X

Traits	Description	Supergroup	Single point tests	Gene tests	Gene-set tests	Enrichment near GWAS signals tests	Tails tests
L-HDL-TG	Triglycerides in large HDL	Lipid and lipoprotein	X	X	X	X	X
M-HDL-P	Concentration of medium HDL particles	Lipid and lipoprotein	X	X	X	X	X
M-HDL-L	Total lipids in medium HDL	Lipid and lipoprotein	X	X	X	X	X
M-HDL-PL	Phospholipids in medium HDL	Lipid and lipoprotein	X	X	X	X	X
M-HDL-C	Total cholesterol in medium HDL	Lipid and lipoprotein	X	X	X	X	X
M-HDL-CE	Cholesterol esters in medium HDL	Lipid and lipoprotein	X	X	X	X	X
M-HDL-FC	Free cholesterol in medium HDL	Lipid and lipoprotein	X	X	X	X	X
M-HDL-TG	Triglycerides in medium HDL	Lipid and lipoprotein	X	X	X	X	X
S-HDL-P	Concentration of small HDL particles	Lipid and lipoprotein	X	X	X	X	X
S-HDL-L	Total lipids in small HDL	Lipid and lipoprotein	X	X	X	X	X
S-HDL-PL	Phospholipids in small HDL	Lipid and lipoprotein	X	X	X	X	X
S-HDL-C	Total cholesterol in small HDL	Lipid and lipoprotein	X	X	X	X	X
S-HDL-CE	Cholesterol esters in small HDL	Lipid and lipoprotein	X	X	X	X	X
S-HDL-FC	Free cholesterol in small HDL	Lipid and lipoprotein	X	X	X	X	X
S-HDL-TG	Triglycerides in small HDL	Lipid and lipoprotein	X	X	X	X	X
XXL-VLDL-PL_%	Phospholipids to total lipds ratio in chylomicrons and extremely large VLDL	Lipid and lipoprotein	X	X	X		
XXL-VLDL-C_%	Total cholesterol to total lipids ratio in chylomicrons and extremely large VLDL	Lipid and lipoprotein	X	X	X	X	
XXL-VLDL-CE_%	Cholesterol esters to total lipids ratio in chylomicrons and extremely large VLDL	Lipid and lipoprotein	X	X	X	X	
XXL-VLDL-FC_%	Free cholesterol to total lipids ratio in chylomicrons and extremely large VLDL	Lipid and lipoprotein	X	X	X	X	
XXL-VLDL-TG_%	Triglycerides to total lipids ratio in chylomicrons and extremely large VLDL	Lipid and lipoprotein	X	X	X	X	
XL-VLDL-PL_%	Phospholipids to total lipds ratio in very large VLDL	Lipid and lipoprotein	X	X	X		
XL-VLDL-C_%	Total cholesterol to total lipids ratio in very large VLDL	Lipid and lipoprotein	X	X	X	X	
XL-VLDL-CE_%	Cholesterol esters to total lipids ratio in very large VLDL	Lipid and lipoprotein	X	X	X	X	
XL-VLDL-FC_%	Free cholesterol to total lipids ratio in very large VLDL	Lipid and lipoprotein	X	X	X	X	
XL-VLDL-TG_%	Triglycerides to total lipids ratio in very large VLDL	Lipid and lipoprotein	X	X	X	X	
L-VLDL-PL_%	Phospholipids to total lipds ratio in large VLDL	Lipid and lipoprotein	X	X	X		
L-VLDL-C_%	Total cholesterol to total lipids ratio in large VLDL	Lipid and lipoprotein	X	X	X	X	
L-VLDL-CE_%	Cholesterol esters to total lipids ratio in large VLDL	Lipid and lipoprotein	X	X	X	X	
L-VLDL-FC_%	Free cholesterol to total lipids ratio in large VLDL	Lipid and lipoprotein	X	X	X	X	
L-VLDL-TG_%	Triglycerides to total lipids ratio in large VLDL	Lipid and lipoprotein	X	X	X	X	
M-VLDL-PL_%	Phospholipids to total lipds ratio in medium VLDL	Lipid and lipoprotein	X	X	X		
M-VLDL-C_%	Total cholesterol to total lipids ratio in medium VLDL	Lipid and lipoprotein	X	X	X	X	
M-VLDL-CE_%	Cholesterol esters to total lipids ratio in medium VLDL	Lipid and lipoprotein	X	X	X	X	
M-VLDL-FC_%	Free cholesterol to total lipids ratio in medium VLDL	Lipid and lipoprotein	X	X	X	X	
M-VLDL-TG_%	Triglycerides to total lipids ratio in medium VLDL	Lipid and lipoprotein	X	X	X	X	
S-VLDL-PL_%	Phospholipids to total lipds ratio in small VLDL	Lipid and lipoprotein	X	X	X		
S-VLDL-C_%	Total cholesterol to total lipids ratio in	Lipid and	X	X	X	X	

Traits	Description	Supergroup	Single point tests	Gene tests	Gene-set tests	Enrichment near GWAS signals tests	Tails tests
	small VLDL	lipoprotein					
S-VLDL-CE_%	Cholesterol esters to total lipids ratio in small VLDL	Lipid and lipoprotein	X	X	X	X	
S-VLDL-FC_%	Free cholesterol to total lipids ratio in small VLDL	Lipid and lipoprotein	X	X	X	X	
S-VLDL-TG_%	Triglycerides to total lipids ratio in small VLDL	Lipid and lipoprotein	X	X	X	X	
XS-VLDL-PL_%	Phospholipids to total lipds ratio in very small VLDL	Lipid and lipoprotein	X	X	X		
XS-VLDL-C_%	Total cholesterol to total lipids ratio in very small VLDL	Lipid and lipoprotein	X	X	X	X	
XS-VLDL-CE_%	Cholesterol esters to total lipids ratio in very small VLDL	Lipid and lipoprotein	X	X	X	X	
XS-VLDL-FC_%	Free cholesterol to total lipids ratio in very small VLDL	Lipid and lipoprotein	X	X	X	X	
XS-VLDL-TG_%	Triglycerides to total lipids ratio very small VLDL	Lipid and lipoprotein	X	X	X	X	
IDL-PL_%	Phospholipids to total lipds ratio in IDL	Lipid and lipoprotein	X	X	X		
IDL-C_%	Total cholesterol to total lipids ratio in IDL	Lipid and lipoprotein	X	X	X	X	
IDL-CE_%	Cholesterol esters to total lipids ratio in IDL	Lipid and lipoprotein	X	X	X	X	
IDL-FC_%	Free cholesterol to total lipids ratio in IDL	Lipid and lipoprotein	X	X	X	X	
IDL-TG_%	Triglycerides to total lipids ratio in IDL	Lipid and lipoprotein	X	X	X	X	
L-LDL-PL_%	Phospholipids to total lipds ratio in large LDL	Lipid and lipoprotein	X	X	X	X	
L-LDL-C_%	Total cholesterol to total lipids ratio in large LDL	Lipid and lipoprotein	X	X	X	X	
L-LDL-CE_%	Cholesterol esters to total lipids ratio in large LDL	Lipid and lipoprotein	X	X	X	X	
L-LDL-FC_%	Free cholesterol to total lipids ratio in large LDL	Lipid and lipoprotein	X	X	X	X	
L-LDL-TG_%	Triglycerides to total lipids ratio in large LDL	Lipid and lipoprotein	X	X	X	X	
M-LDL-PL_%	Phospholipids to total lipds ratio in medium LDL	Lipid and lipoprotein	X	X	X	X	
M-LDL-C_%	Total cholesterol to total lipids ratio in medium LDL	Lipid and lipoprotein	X	X	X	X	
M-LDL-CE_%	Cholesterol esters to total lipids ratio in medium LDL	Lipid and lipoprotein	X	X	X	X	
M-LDL-FC_%	Free cholesterol to total lipids ratio in medium LDL	Lipid and lipoprotein	X	X	X	X	
M-LDL-TG_%	Triglycerides to total lipids ratio in medium LDL	Lipid and lipoprotein	X	X	X	X	
S-LDL-PL_%	Phospholipids to total lipds ratio in small LDL	Lipid and lipoprotein	X	X	X	X	
S-LDL-C_%	Total cholesterol to total lipids ratio in small LDL	Lipid and lipoprotein	X	X	X	X	
S-LDL-CE_%	Cholesterol esters to total lipids ratio in small LDL	Lipid and lipoprotein	X	X	X	X	
S-LDL-FC_%	Free cholesterol to total lipids ratio in small LDL	Lipid and lipoprotein	X	X	X	X	
S-LDL-TG_%	Triglycerides to total lipids ratio in small LDL	Lipid and lipoprotein	X	X	X	X	
XL-HDL-PL_%	Phospholipids to total lipds ratio in very large HDL	Lipid and lipoprotein	X	X	X	X	
XL-HDL-C_%	Total cholesterol to total lipids ratio in very large HDL	Lipid and lipoprotein	X	X	X	X	
XL-HDL-CE_%	Cholesterol esters to total lipids ratio in very large HDL	Lipid and lipoprotein	X	X	X	X	
XL-HDL-FC_%	Free cholesterol to total lipids ratio in very large HDL	Lipid and lipoprotein	X	X	X	X	
XL-HDL-TG_%	Triglycerides to total lipids ratio in very large HDL	Lipid and lipoprotein	X	X	X	X	
L-HDL-PL_%	Phospholipids to total lipds ratio in large HDL	Lipid and lipoprotein	X	X	X	X	
L-HDL-C_%	Total cholesterol to total lipids ratio in large HDL	Lipid and lipoprotein	X	X	X	X	
L-HDL-CE_%	Cholesterol esters to total lipids ratio in large HDL	Lipid and lipoprotein	X	X	X	X	

Traits	Description	Supergroup	Single point tests	Gene tests	Gene-set tests	Enrichment near GWAS signals tests	Tails tests
L-HDL-FC_%	Free cholesterol to total lipids ratio in large HDL	Lipid and lipoprotein	X	X	X	X	
L-HDL-TG_%	Triglycerides to total lipids ratio in large HDL	Lipid and lipoprotein	X	X	X	X	
M-HDL-PL_%	Phospholipids to total lipds ratio in medium HDL	Lipid and lipoprotein	X	X	X	X	
M-HDL-C_%	Total cholesterol to total lipids ratio in medium HDL	Lipid and lipoprotein	X	X	X	X	
M-HDL-CE_%	Cholesterol esters to total lipids ratio in medium HDL	Lipid and lipoprotein	X	X	X	X	
M-HDL-FC_%	Free cholesterol to total lipids ratio in medium HDL	Lipid and lipoprotein	X	X	X	X	
M-HDL-TG_%	Triglycerides to total lipids ratio in medium HDL	Lipid and lipoprotein	X	X	X	X	
S-HDL-PL_%	Phospholipids to total lipds ratio in small HDL	Lipid and lipoprotein	X	X	X	X	
S-HDL-C_%	Total cholesterol to total lipids ratio in small HDL	Lipid and lipoprotein	X	X	X	X	
S-HDL-CE_%	Cholesterol esters to total lipids ratio in small HDL	Lipid and lipoprotein	X	X	X	X	
S-HDL-FC_%	Free cholesterol to total lipids ratio in small HDL	Lipid and lipoprotein	X	X	X	X	
S-HDL-TG_%	Triglycerides to total lipids ratio in small HDL	Lipid and lipoprotein	X	X	X	X	
VLDL-D	Mean diameter for VLDL particles	Lipid and lipoprotein	X	X	X		
LDL-D	Mean diameter for LDL particles	Lipid and lipoprotein	X	X	X	X	X
HDL-D	Mean diameter for HDL particles	Lipid and lipoprotein	X	X	X	X	X
Serum-C	Serum total cholesterol	Lipid and lipoprotein	X	X	X	X	X
VLDL-C	Total cholesterol in VLDL	Lipid and lipoprotein	X	X	X	X	X
Remnant-C	Remnant cholesterol (non-HDL, non-LDL-cholesterol)	Lipid and lipoprotein	X	X	X	X	X
LDL-C	Total cholesterol in LDL	Lipid and lipoprotein	X	X	X	X	X
HDL-C	Total cholesterol in HDL	Lipid and lipoprotein	X	X	X	X	X
HDL2-C	Total cholesterol in HDL2	Lipid and lipoprotein	X	X	X	X	X
HDL3-C	Total cholesterol in HDL3	Lipid and lipoprotein	X	X	X	X	X
EstC	Esterified cholesterol	Lipid and lipoprotein	X	X	X	X	X
FreeC	Free cholesterol	Lipid and lipoprotein	X	X	X	X	X
Serum-TG	Serum total triglycerides	Lipid and lipoprotein	X	X	X	X	X
VLDL-TG	Triglycerides in VLDL	Lipid and lipoprotein	X	X	X	X	X
LDL-TG	Triglycerides in LDL	Lipid and lipoprotein	X	X	X	X	X
HDL-TG	Triglycerides in HDL	Lipid and lipoprotein	X	X	X	X	X
DAG	Diacylglycerol	Lipid and lipoprotein	X	X	X		
DAG/TG	Ratio of diacylglycerol to triglycerides	Lipid and lipoprotein	X	X	X	X	
TotPG	Total phosphoglycerides	Lipid and lipoprotein	X	X	X		
TG/PG	Ratio of triglycerides to phosphoglycerides	Lipid and lipoprotein	X	X	X	X	
PC	Phosphatidylcholine and other cholines	Lipid and lipoprotein	X	X	X		
SM	Sphingomyelins	Lipid and lipoprotein	X	X	X		
TotCho	Total cholines	Lipid and lipoprotein	X	X	X		
ApoA1	Apolipoprotein A--I *	Lipid and lipoprotein	X	X	X		
ApoB	Apolipoprotein B *	Lipid and	X	X	X		

Traits	Description	Supergroup	Single point tests	Gene tests	Gene-set tests	Enrichment near GWAS signals tests	Tails tests
		lipoprotein					
ApoB/ApoA1	Ratio of apolipoprotein B to apolipoprotein A--I	Lipid and lipoprotein	X	X	X		
TotFA	Total fatty acids	Lipid and lipoprotein	X	X	X		
FALen	Estimated description of fatty acid chain length, not actual carbon number	Lipid and lipoprotein	X	X	X		
UnsatDeg	Estimated degree of unsaturation	Lipid and lipoprotein	X	X	X		
DHA	22:6, docosahexaenoic acid	Lipid and lipoprotein	X	X	X		
LA	18:2, linoleic acid	Lipid and lipoprotein	X	X	X		
CLA	Conjugated linoleic acid	Lipid and lipoprotein	X	X	X		
FAw3	Omega--3 fatty acids	Lipid and lipoprotein	X	X	X		
FAw6	Omega--6 fatty acids	Lipid and lipoprotein	X	X	X		
PUFA	Polyunsaturated fatty acids	Lipid and lipoprotein	X	X	X		
MUFA	Monounsaturated fatty acids; 16:1, 18:1	Lipid and lipoprotein	X	X	X		
SFA	Saturated fatty acids	Lipid and lipoprotein	X	X	X		
DHA/FA	Ratio of 22:6 docosahexaenoic acid to total fatty acids	Lipid and lipoprotein	X	X	X		
LA/FA	Ratio of 18:2 linoleic acid to total fatty acids	Lipid and lipoprotein	X	X	X		
CLA/FA	Ratio of conjugated linoleic acid to total fatty acids	Lipid and lipoprotein	X	X	X		
FAw3/FA	Ratio of omega--3 fatty acids to total fatty acids	Lipid and lipoprotein	X	X	X		
FAw6/FA	Ratio of omega--6 fatty acids to total fatty acids	Lipid and lipoprotein	X	X	X		
PUFA/FA	Ratio of polyunsaturated fatty acids to total fatty acids	Lipid and lipoprotein	X	X	X		
MUFA/FA	Ratio of monounsaturated fatty acids to total fatty acids	Lipid and lipoprotein	X	X	X		
SFA/FA	Ratio of saturated fatty acids to total fatty acids	Lipid and lipoprotein	X	X	X		
Ala	Alanine	Aminoacid	X	X	X		
Gln	Glutamine	Aminoacid	X	X	X		
Gly	Glycine	Aminoacid	X	X	X		
His	Histidine	Aminoacid	X	X	X		
Ile	Isoleucine	Aminoacid	X	X	X		
Leu	Leucine	Aminoacid	X	X	X		
Val	Valine	Aminoacid	X	X	X		
Phe	Phenylalanine	Aminoacid	X	X	X		
Tyr	Tyrosine	Aminoacid	X	X	X		
AcAce	Acetoacetate	Ketone bodies	X	X	X		
Crea	Creatinine	Fluid balance	X	X	X		
Alb	Albumin	Fluid balance	X	X	X		
Gp	Glycoprotein acetyls, mainly a1-acid glycoprotein	Inflammation	X	X	X		

Table 3.1: List of traits and analyses where they were used

3.3.8 Gene-set analyses

To perform gene set analysis I obtained a curated gene-disease list from DisGeNET [299, 300] and gene lists of metabolic pathways from KEGG [301-303] and Reactome [304, 305]. The gene-disease list obtained from DisGeNET, combines expert curated gene-disease associations from the following databases: a) CTD (Comparative Toxicogenomics Database); b) UNIPROT; c) ORPHANET (an online rare disease and orphan drug data base); d) PSYGENET (Psychiatric disorders Gene association NETwork); and e) HPO (Human Phenotype Ontology). I limited analysis to gene sets with more than three genes resulting in 7,150 total gene sets to test. Finally, I extracted loss-of-function variants from genes in the gene sets and ran SKAT-O (method="optimal.adj") for each of the traits. Similarly to the gene-based analysis, I used WES data as discovery, and took signals forward for validation in WGS if $p < 0.01$. Covariate selection for correlated traits was performed as described in the gene-based analyses (**Methods 3.3.7**). The gene-set-wide significance threshold was calculated by first estimating the effective number of gene sets tested given the high overlap amongst them. Using PCA I estimated that 1094 PCs explain > 95% of the variance in gene sets. The significance threshold was therefore calculated as: $0.05/(1094*19)=2.41\text{x}10^{-6}$ where 19 corresponds to the effective number of phenotypes tested as described above. A signal was considered to replicate if after meta-analysis: i) it met the defined gene-set-wide significance threshold ($p_{meta} < 2.41\text{x}10^{-6}$); ii) >2 variants were tested; iii) it was nominally significant ($p_{validation}<0.05$) in the unadjusted test for WGS (i.e without adjusting for correlated traits).

3.3.9 Genes near GWAS signals

GWAS catalog data files (release 27-09-2017) were downloaded from https://www.ebi.ac.uk/gwas/docs/file-downloads [79]. I focused on GWAS loci associated with HDL cholesterol, LDL cholesterol, total cholesterol and triglycerides. I extracted all reported genes for GWAS loci that were associated at genome-wide significance ($p<5x10^{-8}$) excluding cases where the "REPORTED GENE" value was: i) NR (not reported); ii) intergenic; iii) APO(APOE) cluster; iv) HLA-area (**Table 3.2**). For this analysis, I ran SKAT-O using the optimal unified approach (method="optimal.adj") on the four gene sets (HDLC reported, LDLC reported, TC reported, TG reported, **Table 3.2**). The list of genes known to be involved in conditions leading to abnormal lipid levels was created extracting relevant genes from the DisGeNET and Reactome gene lists. Afterwards, I conducted a manual review of the published literature to remove genes where functional work in mouse or human has revealed a direct role of the gene in HDL metabolism (**Table 3.2**). The search terms used were "[*gene name*] loss of function HDL" and "[*gene name*] knockout HDL". Significance threshold ($p < 0.005$) was determined by correcting for 10 PCs explaining >95% of the variance of the traits used in this analysis.

HDLC reported*	HDLC reported (known removed)**	LDLC reported*	TC reported*	TG reported*	Known genes
ABCA1	*ACAD11*	*ABCG5*	*ABCA1*	*AFF1*	*ABCA1*
ABCA8	*ADH5*	*ABCG8*	*ABCB11*	*AKR1C4*	*ABCA8*
AC016735.2	*ALDH1A2*	*ABO*	*ABCG5*	*ALDH2*	*AC016735.2*
ACAD11	*ANGPTL1*	*ACAD11*	*ABCG8*	*ANGPTL3*	*ANGPTL4*
ADH5	*ATG7*	*ANGPTL3*	*ABO*	*ANGPTL4*	*ANGPTL8*
ALDH1A2	*CITED2*	*APOA1*	*ADAMTS3*	*APOA1*	*APOA1*
ANGPTL1	*CMIP*	*APOB*	*ANGPTL3*	*APOA5*	*APOA5*
ANGPTL4	*COBLL1*	*APOC1*	*APOA1*	*APOB*	*APOB*
ANGPTL8	*COPB1*	*APOE*	*APOB*	*APOC1*	*APOC3*

HDLC reported*	HDLC reported (known removed)**	LDLC reported*	TC reported*	TG reported*	Known genes
APOA1	*CPS1*	*BRAP*	*APOE*	*APOE*	*APOE*
APOA5	*DAGLB*	*BRCA2*	*ASAP3*	*BAI3*	*ARL15*
APOB	*FADS1*	*CELSR2*	*BRAP*	*LMBRD1*	*C12orf51*
APOC3	*FAM13A*	*CETP*	*C6orf106*	*CAPN3*	*C6orf106*
APOE	*GPAM*	*CILP2*	*CELSR2*	*CCR6*	*CD300LG*
ARL15	*GSK3B*	*CMTM6*	*CETP*	*CEP68*	*CD36*
ATG7	*HAS1*	*CSNK1G3*	*CILP2*	*CETP*	*CETP*
C12orf51	*IKZF1*	*CYP7A1*	*CMTM6*	*CILP2*	*FTO*
C6orf106	*KAT5*	*DLG4*	*CSNK1G3*	*CITED2*	*GALNT2*
CD300LG	*LACTB*	*DNAH11*	*CYP7A1*	*COBLL1*	*HNF4A*
CD36	*LRP4*	*EHBP1*	*DLG4*	*CTF1*	*IGHVII-33-1*
CETP	*LRRC29*	*FAM117B*	*DNAH11*	*CYP26A1*	*IRS1*
CITED2	*MADD*	*FN1*	*DOCK7*	*DNAH17*	*KLF14*
CMIP	*MC4R*	*FRK*	*ERGIC3*	*DOCK7*	*LCAT*
COBLL1	*MLXIPL*	*GATA6*	*EVI5*	*ERGIC3*	*LILRA3*
COPB1	*MVK*	*GPAM*	*FAM117B*	*FADS1*	*LIPC*
CPS1	*MYL2*	*HFE*	*FN1*	*FRMD5*	*LIPG*
DAGLB	*OR4C46*	*HLA*	*FRK*	*FTO*	*LOC100996634*
FADS1	*PDE3A*	*HLA-C*	*GCKR*	*GALNT2*	*LOC55908*
FAM13A	*PEPD*	*HMGCR*	*GPAM*	*GCKR*	*LPA*
FTO	*PGS1*	*HNF1A*	*GPR146*	*GPR85*	*LPL*
GALNT2	*RBM5*	*HPR*	*HBS1L*	*HLA*	*LRP1*
GPAM	*RSPO3*	*IDOL*	*HFE*	*INSR*	*MSL2L1*
GSK3B	*SBNO1*	*INSIG2*	*HLA*	*IRS1*	*PABPC4*
HAS1	*SEMA3C*	*IRF2BP2*	*HLA-C*	*JMJD1C*	*PLTP*
HNF4A	*SETD2*	*LDLR*	*HMGCR*	*KLHL8*	*PPP1R3B*
IGHVII-33-1	*SLC39A8*	*LDLRAP1*	*HNF1A*	*LIPC*	*PRKAG3*
IKZF1	*SNX13*	*LOC84931*	*HNF4A*	*LPA*	*RMI2*
IRS1	*STAB1*	*LPA*	*HPR*	*LPL*	*RP-11-115*
KAT5	*STARD3*	*LRPAP1*	*IDOL*	*LRP1*	*SCARB1*
KLF14	*TMEM176A*	*MAFB*	*INSIG2*	*LRPAP1*	*SIK3*
LACTB	*TRPS1*	*MIR148A*	*IRF2BP2*	*MAP3K1*	*TRIB1*
LCAT	*UBASH3B*	*MOSC1*	*KCNK17*	*MAU2*	*TTC39B*
LILRA3	*ZNF648*	*MTHFD2L*	*LDLR*	*MET*	*UBE2L3*
LIPC		*MTMR3*	*LDLRAP1*	*MIR148A*	*VEGFA*
LIPG		*MYLIP*	*LIPC*	*MLXIPL*	*ZNF664*
LOC100996634		*NCAN*	*LIPG*	*MPP3*	
LOC55908		*NPC1L1*	*LPA*	*MSL2L1*	
LPA		*OSBPL7*	*LRPAP1*	*NAT2*	

HDLC reported*	HDLC reported (known removed)**	LDLC reported*	TC reported*	TG reported*	Known genes
LPL		*PCSK9*	*MAFB*	*PDXDC1*	
LRP1		*PFAS*	*MAMSTR*	*PEPD*	
LRP4		*PLEC1*	*MARCH8*	*PINX1*	
LRRC29		*PPARA*	*MIR148A*	*PLA2G6*	
MADD		*PPARG*	*MOSC1*	*PLTP*	
MC4R		*PPP1R3B*	*MTHFD2L*	*PROX1*	
MLXIPL		*SMARCA4*	*MYLIP*	*RSPO3*	
MSL2L1		*SNX5*	*NAT2*	*SIK3*	
MVK		*SORT1*	*NCAN*	*TIMD4*	
MYL2		*SOX17*	*NPC1L1*	*TM4SF5*	
OR4C46		*SPTLC3*	*OSBPL7*	*TP53BP1*	
PABPC4		*ST3GAL4*	*PCSK9*	*TRIB1*	
PDE3A		*TIMD4*	*PHLDB1*	*TYW1B*	
PEPD		*TOP1*	*PLEC1*	*VEGFA*	
PGS1		*TRIB1*	*PPARA*	*ZNF664*	
PLTP		*VLDLR*	*PPARG*		
PPP1R3B		*ZNF274*	*PPP1R3B*		
PRKAG3			*PXK*		
RBM5			*RAB3GAP1*		
RMI2			*RAF1*		
RP-11-115			*RP11-115*		
J16.1			*J16.1*		
RSPO3			*SAMM50*		
SBNO1			*SNX5*		
SCARB1			*SORT1*		
SEMA3C			*SOX17*		
SETD2			*SPTY2D1*		
SIK3			*ST3GAL4*		
SLC39A8			*TIMD4*		
SNX13			*TMEM57*		
STAB1			*TOP1*		
STARD3			*TRIB1*		
TMEM176A			*TRPS1*		
TRIB1			*TTC39B*		
TRPS1			*UBASH3B*		
TTC39B			*UGT1A1*		
UBASH3B			*VLDLR*		
UBE2L3					
VEGFA					

HDLC reported*	HDLC reported (known removed)**	LDLC reported*	TC reported*	TG reported*	Known genes
ZNF648					
ZNF664					

Table 3.2: Gene sets used for enrichment of genes near GWAS signals analyses. HDL reported -Genes reported associated with "HDL cholesterol" unambiguously ; HDLC reported (known removed) - Genes reported associated with "HDL cholesterol" unambiguously but with known genes involved in HDL metabolism or lipid abnormalities removed; LDLC reported - Genes reported associated with "LDL cholesterol" unambiguously; TC reported - Genes reported associated with "Cholesterol, total" unambiguously; TG reported - Genes reported associated with "Triglycerides" unambiguously; Known genes - Genes removed for sensitivity analysis that are known to be involved in lipid abnormalities or HDL metabolism based on literature review; *Gene sets used in analyses running SKAT-O on gene sets.; **Gene sets used in sensitivity analyses.

3.3.10 Analysis of tails of phenotype distribution

For this analysis, I used all lipoprotein and lipid traits but excluded derived measures (lipid ratios) resulting in 106 traits (**Table 3.1**). I focused on likely deleterious missense and loss-of-function variation in lipid metabolism and disease gene sets (**Table 3.3**) with an allele count <10 in each dataset. I chose an arbitrary cutoff of 10 individuals with the highest and lowest values for the traits to define tails for all 106 traits.

Gene Set	Source
Abnormality_of_lipid_metabolism	DisGeneNet
Dyslipidaemias	DisGeneNet
HDL_assembly	Reactome
HDL_clearance	Reactome
HDL_remodeling	Reactome
Hyperlipidaemia	DisGeneNet
Hypertriglyceridaemia_CTD	DisGeneNet
Hypertriglyceridaemia_HPO	DisGeneNet
LDL_clearance	Reactome
LDL_remodeling	Reactome
Triglyceride_biosynthesis	Reactome
Triglyceride_catabolism	Reactome
Triglyceride_metabolism	Reactome
VLDL_assembly	Reactome
VLDL_clearance	Reactome

Table 3.3: List of gene sets used for tails analyses.

Given the high phenotypic correlation of these traits, there was a high overlap of individuals at the tails of the distributions so I removed traits that shared >=8 individuals with any other trait reducing the number of tested traits to 50. For each trait, total deleterious allele count from each gene set for upper and lower tails was obtained and an empirical *p* was calculated by performing 10,000 permutations extracting 10 random individuals from the phenotype distribution and counting the number of deleterious alleles from the gene set. The significance threshold (p = 0.00037) was chosen by correcting for 9 PCs explaining >95% of the traits variance and 15 pathways. Meta-analysis was done using Stouffer's method [306] as implemented in the metap package [307] in R.

3.4 Results

3.4.1 Single point analyses

I first explored whether I could recapitulate known associations with NMR traits, as well as, potentially identify novel associations with rarer variants not previously tested in GWAS arrays. To this end, I performed single-point association analysis for 226 NMR metabolic biomarkers using WES data from 3,741 healthy blood donors from the INTERVAL cohort as a discovery dataset (**Methods 3.3.6**). Power calculations showed very limited power to detect associations for variants on the rare allele frequency spectrum with this sample size (power=4.6% to find an association with $p<1x10^{-5}$ -threshold to take forward for validation- with beta=1 and variant with MAF=0.1%). I therefore focused on variants with MAF>=0.1%. After association analyses for all traits I took forward for validation 494 variants associated with at least one trait with $p<1x10^{-5}$. I performed validation using whole-genome sequence (WGS) data from 3,401 independent individuals from the same cohort. After meta-analysis,

34 unique loci were associated with at least one trait (**Table 3.4**). All of these associations had already been previously described [38, 173, 308].

Rsid	Gene	most severe consequence	top trait	EA	NEA	discov p	validation p	meta-p	beta	se	EAF	n assoc traits
rs1047891	*CPS1*	missense_variant (Thr1412Asn)	Gly	a	c	$1.48x10^{-68}$	$4.47x10^{-54}$	$2.09x10^{-125}$	0.42	0.02	32.47%	1
rs1077834	*LIPC,ALDH1A2*	intron_variant	L-HDL-TG	t	c	$2.52x10^{-16}$	$6.90x10^{-21}$	$1.11x10^{-35}$	-0.25	0.02	21.41%	35
rs11076176	*CETP*	intron_variant	M-HDL-TG	t	g	$5.82x10^{-7}$	$6.62x10^{-6}$	$1.65x10^{-11}$	-0.15	0.02	16.92%	6
rs11591147	*PCSK9*	missense_variant (Arg46Leu)	IDL-FC	t	g	$7.31x10^{-12}$	$2.20x10^{-5}$	$2.96x10^{-15}$	-0.48	0.06	1.73%	45
rs116843064	*ANGPTL4*	missense_variant (Glu40Lys)	S-VLDL-TG	a	g	$7.81x10^{-7}$	$2.67x10^{-6}$	$9.11x10^{-12}$	-0.40	0.06	1.89%	17
rs1184865	*DOCK7*	intron_variant	M-HDL-TG	a	g	$6.59x10^{-6}$	$5.66x10^{-5}$	$1.45x10^{-9}$	-0.10	0.02	36.13%	1
rs12191266	*SLC16A10*	intron_variant	Tyr	t	c	$4.68x10^{-6}$	$2.42x10^{-5}$	$4.48x10^{-10}$	-0.15	0.02	14.43%	1
rs1260326	*GCKR*	missense_variant (Leu446Pro)	MUFA	t	c	$1.20x10^{-6}$	$5.31x10^{-6}$	$2.61x10^{-11}$	0.12	0.02	39.85%	17
rs138326449	*APOC3*	splice_donor_variant (2^{nd} exon)	S-VLDL-TG	a	g	$7.91x10^{-6}$	$8.80x10^{-6}$	$2.90x10^{-10}$	-1.10	0.17	0.23%	6
rs17231506	*CETP*	upstream_gene_variant	HDL2-C	t	c	$6.73x10^{-17}$	$4.65x10^{-18}$	$1.35x10^{-33}$	0.21	0.02	31.83%	38
rs174476	*RAB3IL1*	intron_variant	UnsatDeg	t	c	$2.05x10^{-9}$	$1.48x10^{-5}$	$1.95x10^{-13}$	0.12	0.02	41.71%	1
rs174547	*FADS1,FADS2*	intron_variant	UnsatDeg	t	c	$1.03x10^{-41}$	$5.96x10^{-38}$	$9.02x10^{-80}$	0.33	0.02	33.71%	8
rs174602	*FADS2*	non_coding_transcript_exon_variant	UnsatDeg	t	c	$1.21x10^{-11}$	$5.64x10^{-7}$	$4.97x10^{-17}$	0.17	0.02	20.16%	2
rs1912826	*KLKB1*	intron_variant	His	a	g	$7.80x10^{-11}$	$5.54x10^{-9}$	$2.04x10^{-18}$	0.15	0.02	48.89%	2
rs2072560	*APOA5*	intron_variant	XS-VLDL-TG_%	t	c	$1.15x10^{-8}$	$2.06x10^{-7}$	$1.07x10^{-14}$	0.27	0.04	5.90%	30
rs2228671	*LDLR*	non_coding_transcript_exon_variant	IDL-FC	t	c	$2.04x10^{-7}$	$6.27x10^{-7}$	$5.55x10^{-13}$	-0.18	0.03	12.26%	38
rs2295601	*ELOVL2*	synonymous_variant	DHA/FA	a	g	$1.54x10^{-10}$	$6.61x10^{-9}$	$4.69x10^{-18}$	-0.17	0.02	22.90%	2
rs2575876	*ABCA1*	intron_variant	HDL3-C	a	g	$1.92x10^{-6}$	$8.30x10^{-8}$	$8.12x10^{-13}$	-0.14	0.02	24.65%	1
rs2657879	*GLS2*	3_prime_UTR_variant	Gln	a	g	$1.16x10^{-11}$	$1.72x10^{-15}$	$1.50x10^{-25}$	0.23	0.02	18.07%	1
rs283813	*PVRL2*	intron_variant	S-LDL-C_%	a	t	$3.08x10^{-8}$	$1.20x10^{-5}$	$2.20x10^{-12}$	-0.23	0.03	6.90%	22
rs28399637	*BCAM*	intron_variant	S-LDL-CE_%	a	g	$4.95x10^{-9}$	$8.59x10^{-7}$	$2.02x10^{-14}$	0.14	0.02	31.77%	25
rs28399654	*BCAM*	missense_variant (Val196Ile)	S-LDL-C_%	a	g	$1.38x10^{-11}$	$8.80x10^{-8}$	$8.29x10^{-18}$	-0.40	0.05	3.37%	34
rs328	*LPL*	stop_gained (Ser474Ter)	TG/PG	c	g	$1.08x10^{-8}$	$1.44x10^{-7}$	$7.00x10^{-15}$	0.22	0.03	10.09%	19
rs3798220	*LPA*	missense_variant (Ile1891Met)	XL-VLDL-CE	t	c	$3.04x10^{-6}$	$4.55x10^{-13}$	$6.15x10^{-17}$	0.55	0.07	1.76%	16
rs386606006	*APOB*	synonymous_variant	ApoB	a	g	$9.37x10^{-6}$	$2.97x10^{-6}$	$1.17x10^{-10}$	0.11	0.02	48.80%	1

Rsid	Gene	most severe consequence	top trait	EA	NEA	discov p	validation p	meta-p	beta	se	EAF	n assoc traits
rs429358	*APOE*	missense_variant (Cys130Arg)	S-LDL-PL_%	t	c	$9.37x10^{-17}$	$1.20x10^{-17}$	$4.69x10^{-33}$	0.27	0.02	15.07%	61
rs435306	*PLTP*	intron_variant	L-HDL-PL_%	t	g	$4.90x10^{-7}$	$4.17x10^{-7}$	$8.84x10^{-13}$	0.14	0.02	25.50%	1
rs4804573	*KANK2*	3_prime_UTR_variant	S-LDL-PL_%	a	g	$1.49x10^{-7}$	$6.26x10^{-5}$	$4.66x10^{-11}$	0.11	0.02	47.05%	9
rs5880	*CETP*	missense_variant (Ala390Pro)	HDL-C	c	g	$7.97x10^{-7}$	$3.05x10^{-8}$	$1.17x10^{-13}$	-0.28	0.04	4.87%	8
rs61937878	*HAL*	missense_variant (Val549Met)	His	t	c	$7.41x10^{-14}$	$3.75x10^{-8}$	$2.01x10^{-20}$	0.95	0.10	0.66%	1
rs693672	*FADS3*	intron_variant	UnsatDeg	t	c	$1.44x10^{-10}$	$1.36x10^{-9}$	$8.97x10^{-19}$	-0.19	0.02	16.76%	1
rs7412	*APOE*	missense_variant (Arg176Cys)	S-LDL-CE_%	t	c	$8.55x10^{-63}$	$1.82x10^{-58}$	$5.97x10^{-124}$	-0.71	0.03	7.80%	89
rs76075198	*CEACAM19*	synonymous_variant	S-LDL-CE_%	t	c	$6.76x10^{-7}$	$5.25x10^{-8}$	$1.72x10^{-13}$	-0.41	0.06	2.20%	10
rs7679	*PCIF1*	3_prime_UTR_variant	L-HDL-PL_%	t	c	$5.43x10^{-18}$	$1.14x10^{-19}$	$2.23x10^{-36}$	-0.27	0.02	18.05%	19

Table 3.4: Single point association analyses results. Most severe consequence=most severe consequence predicted by VEP on CANONICAL transcript. top trait=trait with the lowest p-value. EA=effect allele. NEA=non-effect allele discov p=p-value for top trait in discovery cohort (WES). validation p=p-value for top trait in validation cohort (WGS). meta-p= p-value for top trait. beta=beta for top trait after meta-analysis. se=se for top trait after meta-analysis. EAF=effect allele frequency. n assoc traits=number of associated traits.

3.4.2 Gene-based analyses

I next sought to discover new gene-trait associations using rare-variant aggregate tests. After running association tests using two nested approaches to group rare variants (LoF and MCAP+LoF, **Methods 3.3.7**), genes were taken forward for validation if they reached the arbitrary threshold of p <5x10^{-3} (**Supplementary Tables 1-2 of Riveros-Mckay et al (in preparation, Appendix B)**). A burden test was used when testing only LoF whereas the optimal unified approach was used when adding predicted deleterious missense variants (MCAP+LoF). This is because I expected most high confidence LoF variants to influence a trait with the same direction of effect and therefore the burden test should be better powered than the optimal unified approach to detect an association. When including missense variants one could expect different directions of effect and therefore the optimal unified approach should be better powered. As previously suggested, to boost discovery power I adjusted for correlated metabolic biomarkers [309, 310]. However, to minimise the possible collider bias this could incur, I only did this at the validation stage. This was to ensure there was at least suggestive evidence for association in the discovery stage without adjusting for any metabolite (**Methods 3.3.7**). After meta-analysis, five genes (*APOB*, *APOC3*, *PCSK9*, *PAH*, *HAL*) associated with 92 different traits with p < 1.32x10^{-7}, which is the stringent significance threshold after correcting for the effective number of tested phenotypes (**Table 3.5, Methods 3.3.7**). All five genes have been previously associated with their respective traits [38, 308, 311]. As expected, I found that there was a significant increase in the strength of association signal for traits for which I used other correlated traits as covariates when compared to the unadjusted tests [309, 310]**,** with the most notable example being a >30 order of magnitude increase in association strength for *PAH*

and phenylalanine (**Table 3.5**). In total, 32 of the 92 known gene-trait associations met the stringent significance threshold (p<1.32x10^{-7}) only after adjusting for correlated traits (**Supplementary Tables 1-2 of Riveros-Mckay et al (in preparation, Appendix B)**).

LoF								
Gene	**Top trait**	**p-value (covs)**	**p-value (raw)**	**N WES**	**N WGS**	**N overlap**	**N traits associated**	**Driven by single variant?**
APOB	**IDL-TG**	**3.20x10^{-13}**	**1.72x10^{-10}**	**6**	**5**	**0**	**45 (57)**	**No**
APOC3	**XS-VLDL-TG**	**6.10x10^{-13}**	**3.58x10^{-12}**	**3**	**2**	**2**	**46 (56)**	**No**
PAH	**Phe**	**5.82x10^{-11}**	**8.25x10^{-3}**	**4**	**3**	**1**	**1 (1)**	**Yes**
MCAP+LoF								
Gene	**Top trait**	**p-value (covs)**	**p-value (raw)**	**N WES**	**N WGS**	**N overlap**	**N traits associated**	**Driven by single variant?**
PAH	**Phe**	**8.33x10^{-63}**	**1.67x10^{-28}**	**39**	**41**	**18**	**1 (1)**	**No**
HAL	**His**	**NA**	**3.72x10^{-42}**	**48**	**37**	**22**	**1 (1)**	**No**
APOC3	**XS-VLDL-TG**	**5.46x10^{-11}**	**2.15x10^{-10}**	**6**	**6**	**3**	**26 (40)**	**No**
PCSK9	**IDL-FC**	**2.39x10^{-10}**	**1.11x10^{-7}**	**15**	**17**	**3**	**29 (34)**	**No**
ACSL1	IDL-P	1.82x10^{-7}	1.76x10^{-4}	4	6	2	0 (1)	Yes
MYCN	M-VLDL-L	6.20x10^{-7}	3.97x10^{-6}	8	8	3	0 (5)	No
ALDH1L1	Gly	NA	4.56x10^{-7}	39	38	19	0 (1)	No
SCARB1	XL-HDL-FC	NA	6.93x10^{-7}	25	18	10	0 (6)	No
FBXO36	IDL-CE_%	NA	1.98x10^{-6}	5	2	1	0 (1)	Yes
B4GALNT3	L-VLDL-FC_%	NA	7.59x10^{-7}	28	22	13	0 (1)	No
LIPC	XXL-VLDL-C_%	NA	9.03x10^{-7}	28	29	11	0 (2)	No

Table 3.5:Genes significantly associated (p<2.5x10-6) with at least one trait in gene-based analyses focusing on loss-of-function (LoF) or predicted deleterious missense by M-CAP plus loss-of-function (MCAP+LoF). Genes that meet gene-level significance after adjusting for multiple phenotypes (p<1.32x10^{-7}) are highlighted in bold. Top trait: trait with the smallest p-value after meta-analysis adjusting for correlated metabolites. p-value (covs): p-value of meta-analysis after adjusting for correlated metabolites for top trait. If NA, this analysis was not performed for this trait due to no metabolic biomarkers meeting the criteria to be included as covariates in meta-analysis. p-value (raw): p-value of meta-analysis without adjusting for correlated metabolites for top trait. N WES: number of tested variants in WES. N WGS: number of tested variants in WGS. N overlap: number of variants present in both WES and WGS. N traits associated: number of traits that meet gene-level significance after adjusting for multiple phenotypes (p<1.32x10^{-7}), traits meeting standard gene-level significance (2.5x10^{-6}) in parenthesis. Driven by single variant?: Yes if after conditioning on top associated variant the meta-analysis association disappears (p>0.05). IDL-TG: Triglycerides in IDL. XS-VLDL-TG: Triglycerides in very small VLDL. Phe: Phenylalanine. His: Histidine. IDL-FC: Free cholesterol in IDL. IDL-P: Concentration of IDL particles. M-VLDL-L: Total lipids in medium VLDL. Gly:Glycine. XL-HDL-FC: Free cholesterol in very large HDL. IDL-CE_%: Cholesterol esters to total lipids ratio in IDL. L-VLDL-FC%: Free cholesterol to total lipids ratio in large VLDL. XXL-VLDL-C_%: Total cholesterol to total lipids ratio in extremely large VLDL.

In addition to established genes, I found 15 gene-trait associations in seven genes meeting standard gene-level significance before adjusting for multiple traits ($p< 2.5x10^{-6}$) which also had nominal evidence of association in the validation cohort ($p< 0.05$). Nine of these were gene-trait associations in three established genes (*ALDH1L1*, *SCARB1*, *LIPC*, **Table 3.5**), suggesting that other results achieving this significance threshold may warrant being prioritised for additional follow-up to establish their validity. In particular amongst the remaining four genes, the association between IDL particle concentration (IDL-P) and *ACSL1* ($p = 1.82x10^{-7}$), as well as, the associations of multiple very-low-density lipoprotein (VLDL) traits to *MYCN* (min $p = 6.20x10^{-7}$) merit further exploration as both genes have been previously linked to lipid metabolism in mouse studies [312-314].

3.4.3 Gene set analyses

To find links between predicted loss-of-function rare variants and metabolic biomarker biology, I next explored associations of these variants in 7,150 gene sets. To this end, I used two biological pathway databases (Reactome, KEGG) and one database that contains expert curated disease associated genes (DisGeNET) (**Methods3.3.8**). Gene set analysis yielded 163 gene-set-trait associations with 14 unique gene sets (**Supplementary Table 4 of Riveros-Mckay et al (in preparation, Appendix B)**). Given that 143 gene-set-trait associations were with 13 gene sets that included two genes with a well-established role in lipid biology (*APOB* and *APOC3*), I repeated the test removing variants in these genes. After removal, there is residual evidence of association ($p<0.05$) in 102 of 143 gene-set-trait signals representing 12 of 13 gene sets. Of the 163 gene-set-trait associations, the remaining 20 gene-set-trait associations (in gene sets not containing either *APOB* or *APOC3)* represent associations of

various lipoprotein related metabolic biomarkers with the "regulation of pyruvate dehydrogenase (PDH) complex" pathway in REACTOME (R-HSA-204174, min p= 7.85x10^{-7}, trait=phospholipids in intermediate density lipoproteins (IDL-PL), **Table 3.6**). These associations encompassed 12 LoF variants in WES and four in WGS (**Figure 3.1**). Upon further inspection, I found that most variants in this pathway were contributing to the association suggesting the signal was not driven by a single gene, in addition they all have the same direction of effect (i.e. the rho(ρ) value in the SKAT-O test was one in both the WES and the WGS analyses). Two variants were of particular interest as they were present in both WES and WGS datasets, rs113309941 in Pyruvate Dehydrogenase Complex Component X (*PDHX*), and rs201013643 in Pyruvate Dehydrogenase Phosphatase Regulatory Subunit (*PDPR*). In *PDHX, rs*113309941 leads to a premature stop mutation (Gln248Ter), it has an allele count (AC) of one in each of WES and WGS, and is very rare in gnomAD[1]. rs201013643 in PDRP also leads to a premature stop (Arg714Ter) and is present in a single heterozygous individual in the WES dataset and two heterozygous in the WGS. This variant is also rare in gnomAD[2]. The five individuals with these two variants had higher than average values (upper percentile range from 44.1% to 0.03%) for measurements that are CVD risk factors such as cholesterol in intermediate-density lipoproteins (IDL-C) and low-density lipoproteins (LDL-C) suggesting these variants may have a deleterious impact on lipid metabolism and cardiovascular risk. Notably, one of the carriers of the *PDHX* Gln248Ter variant was in the top 0.03% for LDL-C in INTERVAL (4.086 mmol/l) and had no predicted deleterious missense mutations in known hypercholesterolaemia genes *PCSK9, APOB* or *LDLR* suggesting this novel protein truncating variant may be contributing to their high LDL-

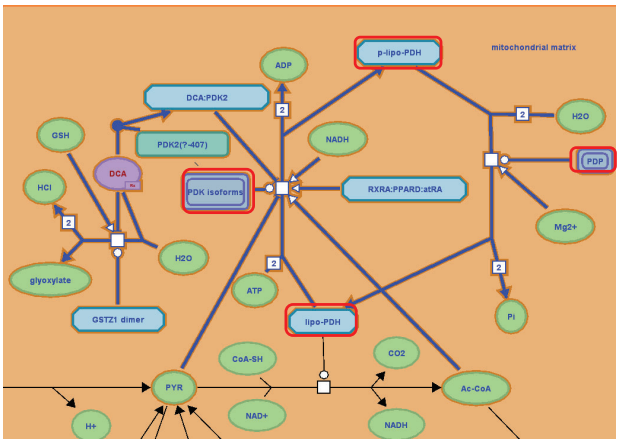
[1] AC (all gnomAD)=3, allele number (AN) (all gnomAD)=246,116, AC (Non-Finnish European (NFE))=2 AN (NFE)=116,604.
[2] AC (all gnomAD)=141, AN (all gnomAD)=275,988, AC (NFE) =8, AN (NFE)=126,382.

C levels. The other carrier was in the top 19.3% percentile of the cohort. None of the genes in this pathway have been previously associated to these traits and therefore this study links these genes collectively to intermediate and low density lipoprotein metabolism and circulating cholesterol for the first time.

Gene set id	Trait	WES p	N WES	WGS p	N WGS	Meta-p	Description	Source
R-HSA-204174	IDL-PL	0.005939	12	0.000503	4	7.85×10^{-7}	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	M-LDL-PL	0.002671	12	0.000594	4	1.01×10^{-6}	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	EstC	0.004754	12	0.001175	4	1.09×10^{-6}	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	IDL-P	0.003992	12	0.000593	4	1.17×10^{-6}	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	L-LDL-P	0.004822	12	0.000258	4	1.20×10^{-6}	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	L-LDL-PL	0.004853	12	0.000423	4	1.21×10^{-6}	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	IDL-L	0.004313	12	0.000574	4	1.21×10^{-6}	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	SerumC	0.005999	12	0.001071	4	1.24×10^{-6}	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	L-LDL-L	0.005082	12	0.000275	4	1.35×10^{-6}	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	IDL-C	0.00475	12	0.001019	4	1.40×10^{-6}	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	L-LDL-FC	0.00681	12	0.0003	4	1.46×10^{-6}	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	L-LDL-C	0.006489	12	0.000275	4	1.87×10^{-6}	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	M-LDL-P	0.006409	12	0.000132	4	1.96×10^{-6}	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	L-LDL-CE	0.006486	12	0.000277	4	2.01×10^{-6}	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	S-LDL-L	0.006413	12	0.000115	4	2.13×10^{-6}	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	S-LDL-P	0.005994	12	0.000113	4	2.13×10^{-6}	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	M-LDL-L	0.006416	12	0.000164	4	2.13×10^{-6}	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	LDL-C	0.007809	12	0.000177	4	2.17×10^{-6}	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	ApoB	0.00504	12	0.000803	4	2.20×10^{-6}	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	IDL-FC	0.009798	12	0.000399	4	2.22×10^{-6}	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome

Table 3.6: Gene set analyses results. WES p = p-value in WES dataset. N WES = number of variants tested in WES dataset. WGS p = p-value in WGS dataset. N WGS = number of variants tested in WGS dataset. Meta-p = Meta-analysis p-value after removing APO genes from gene sets (APOB and APOC3).

a)



b)

Gene	Consequence	AC
Pyruvate dehydrogenase (PDH) complex		
DLAT	Splice acceptor (2nd exon)	WES=1
DLD	Frameshift (Val212SerfsTer32)	WES=1
PDHA2	Stop gain (Tyr28Ter)	WES=1
PDHA2	Frameshift (Val297GlnfsTer14)	WES=1
PDHA2	Stop gain (Gln78Ter)	WES=1
PDHA2	Frameshift (Lys83IlefsTer20)	WES=1
PDHA2	Stop gain (Tyr80Ter)	WES=1
PDHX	Splice donor (2nd exon)	WES=1
PDHX	Stop gain (Gln248Ter)	WES=1 WGS=1
Pyruvate dehydrogenase phosphatase (PDP)		
PDP2	Frameshift (Asn33IlefsTer5)	WES=1
PDP2	Stop gain (Gln352Ter)	WES=1
PDPR	Stop gain (Trp402Ter)	WES=1
PDPR	Stop gain (Arg714Ter)	WES=1 WGS=2
Pyruvate dehydrogenase kinase (PDK)_		
PDK1	Stop gain (Arg66Ter*)	WGS=1

Figure 3.1: Loss-of-function (LoF) variants in regulation of pyruvate dehydrogenase (PDH) complex pathway. a) Figure adapted from REACTOME pathway browser (https://reactome.org/PathwayBrowser/) [315]. Highlighted in red are protein complexes that carry LoF variants in INTERVAL WES or WGS. b) List of genes, consequences and allele count (AC) of LoF variants in the different protein complexes in the pathway.

3.4.4 Enrichment of rare variant associations in genes near established GWAS signals in lipoprotein related metabolic biomarkers

Next, I conducted analyses to investigate whether genes near GWAS index variants associated with traditional lipid traits (HDL-C, LDL-C, TC and TG) were enriched for rare variant associations computationally predicted to affect protein sequence and function with high resolution lipoprotein measurements, which could suggest enrichment of effector transcripts (i.e. transcripts/genes likely to be causal of the original association) in the gene set. Given that this was a hypothesis driven approach using established signals, to boost discovery power I pooled together both WES and WGS data into a single dataset of 7,179 individuals. First, I extracted from the GWAS catalog (release 27-09-2017) the "reported genes" near signals that have been associated with HDL-C, LDL-C, TC or TG and created four gene sets (**Table 3.2**). I only focused on genes that were reported unambiguously (i.e where only one gene is reported) since for associations where more than one gene is reported, it is possible that only one will be the effector gene and rare variants from the non-effector genes will only add noise to the analysis and therefore reduce power. I grouped rare coding variants in the gene set using two nested approaches (LoF and MCAP+LoF) and ran SKAT-O on the gene sets for 157 lipoprotein and lipid traits. Using this approach I found associations ($p < 0.005$, correcting for effective number of tests, **Methods 3.3.9**) for genes near HDL GWAS signals with 18 HDL-related traits (**Table 3.7**), the strongest association being with esterified cholesterol in extra-large HDL (XL-HDL-CE, $p=2.83 \times 10^{-5}$, MCAP+LoF). Associations ($p < 0.005$, **Methods 3.3.9**) in two extra-large HDL cholesterol related traits remained after removing variants in genes known to be involved in conditions leading to abnormal lipid levels or genes where functional work has shown an effect on HDL-C (**Table 3.7**) suggesting there is a contribution to the phenotypic variance of these traits by rare

coding variants in genes, near GWAS signals, without a known role in HDL metabolism, which may represent novel effector transcripts.

Trait	GWAS signal gene set	LoF p-value	MCAP+LoF p-value	LoF p-value (known removed)	MCAP+LoF p-value (known removed)
HDL2-C	HDL-C	$9.03x10^{-3}$	$4.72x10^{-3}$	$4.73x10^{-1}$	$1.41x10^{-1}$
HDL-D	HDL-C	$6.29x10^{-3}$	$2.55x10^{-3}$	$6.88x10^{-1}$	$3.46x10^{-1}$
L-HDL-C_%	HDL-C	$1.49x10^{-3}$	$6.04x10^{-2}$	$4.45x10^{-1}$	$8.78x10^{-1}$
L-HDL-FC_%	HDL-C	$1.67x10^{-4}$	$5.40x10^{-4}$	$1.40x10^{-1}$	$3.52x10^{-1}$
L-HDL-FC	HDL-C	$9.21x10^{-3}$	$3.14x10^{-3}$	$3.95x10^{-1}$	$2.63x10^{-1}$
L-HDL-TG_%	HDL-C	$2.27x10^{-3}$	$1.30x10^{-1}$	$3.40x10^{-1}$	$7.25x10^{-1}$
M-HDL-TG_%	HDL-C	$6.76x10^{-4}$	$1.18x10^{-3}$	$9.98x10^{-2}$	$7.19x10^{-1}$
S-HDL-TG_%	HDL-C	$4.68x10^{-3}$	$4.37x10^{-3}$	$4.37x10^{-1}$	$7.76x10^{-1}$
S-HDL-TG	HDL-C	$1.61x10^{-3}$	$5.47x10^{-3}$	$3.47x10^{-1}$	$3.73x10^{-1}$
XL-HDL-CE	HDL-C	$2.86x10^{-2}$	$2.83x10^{-5}$	1.00	$3.69x10^{-4}$
XL-HDL-C	HDL-C	$1.85x10^{-2}$	$4.43x10^{-5}$	$8.48x10^{-1}$	$9.03x10^{-4}$
XL-HDL-FC	HDL-C	$6.41x10^{-3}$	$2.44x10^{-4}$	$7.43x10^{-1}$	$1.11x10^{-2}$
XL-HDL-L	HDL-C	$1.14x10^{-2}$	$1.75x10^{-4}$	$7.00x10^{-1}$	$7.07x10^{-3}$
XL-HDL-P	HDL-C	$1.17x10^{-2}$	$1.91x10^{-4}$	$6.92x10^{-1}$	$7.56x10^{-3}$
XL-HDL-PL	HDL-C	$8.07x10^{-3}$	$9.94x10^{-4}$	$5.12x10^{-1}$	$1.11x10^{-1}$

Table 3.7:Significant results (p<0.005) in SKAT-O analysis on gene sets built from lists of genes near established GWAS loci. LoF p-value: SKAT-O results for analysis focusing on loss-of-function variants in gene set. MCAP+LoF p-value: SKAT-O results for analysis focusing on rare missense variants (MAF <1%) predicted to be likely deleterious (M-CAP score >0.025) and loss-of-function variants in gene set. LoF p-value (known removed) = SKAT-O results for LoF approach after removing genes known to be involved in lipoprotein metabolism. MCAP+LoF p-value (known removed) = SKAT-O results for MCAP+LoF approach after removing genes known to be involved in lipoprotein metabolism.

3.4.5 Enrichment of rare variation in tails of the phenotypic distribution of lipoprotein and glyceride related traits

Finally, I aimed to investigate whether individuals at the extreme tail of the phenotype distribution for 106 lipoprotein and lipid traits harboured rare coding variants likely to be contributing to their phenotype. I used the WES dataset as a discovery dataset and the WGS dataset as validation. An arbitrary cutoff of 10 individuals at each tail was used to define the tails for all of the 106 traits (**Methods 3.3.10**). After meta-analysis, I found an enrichment of predicted deleterious rare variation ($p < 0.00037$, **Methods 3.3.10, Table 3.8,**

Supplementary Table 9 of Riveros-Mckay et al (in preparation, Appendix B)) in hyperlipidaemia related genes on the lower tail of cholesterol in small VLDL (S-VLDL-C), esterified cholesterol in small VLDL (S-VLDL-CE) and concentration of extra small VLDL particles (XS-VLDL-P), and rare variation on HDL remodelling related genes on the lower tail of concentration of small HDL particles (S-HDL-P). I still observed nominal evidence of association in the WES and WGS datasets for the S-VLDL-C and XS-VLDL-P results using a 0.5% percentile cut-off for the tails but no evidence of association was found when using a 1% percentile cut-off (**Supplementary Table 10 of Riveros-Mckay et al (in preparation, Appendix B)**). This is likely due to the fact that by extending the number of individuals taken from the tails, we are decreasing the average distance to the mean and diluting signal coming from true extreme values.

Upper tails				
Trait	**WES P**	**WGS P**	**Meta-P**	**Gene Set**
S-VLDL-FC	$3.3x10^{-2}$	$2.37x10^{-2}$	$3.45x10^{-3}$	Hypertriglyceridemia_HPO
XS-VLDL-C	$3.3x10^{-2}$	$2.37x10^{-2}$	$3.45x10^{-3}$	Hypertriglyceridemia_HPO
Lower tails				
Trait	**WES P**	**WGS P**	**Meta-P**	**Gene Set**
S-VLDL-C	**$5.8x10^{-3}$**	**$2.31x10^{-3}$**	**$7.61x10^{-5}$**	**Hyperlipidaemia**
XS-VLDL-P	**$1.85x10^{-2}$**	**$7x10^{-4}$**	**$9.42x10^{-5}$**	**Hyperlipidaemia**
S-VLDL-CE	**$5.8x10^{-3}$**	**$6.75x10^{-3}$**	**$2.07x10^{-4}$**	**Hyperlipidaemia**
S-HDL-P	**$2.72x10^{-3}$**	**$1.84x10^{-2}$**	**$2.89x10^{-4}$**	**HDL_remodeling**
S-HDL-P	$4.10x10^{-2}$	$3.92x10^{-2}$	$8.x24x10^{-3}$	Hypertriglyceridemia_CTD

Table 3.8:Gene sets where there is a nominally significant enrichment of rare variation in the tails of a lipid or lipoprotein measurement (p<0.05) in both WES and WGS. Highlighted in bold are gene sets that are significant after meta-analysis using Stouffer's method [306] and after adjusting for multiple traits (p<=0.00037). WES P: permutation p in WES. WGS P: permutation p in WGS. Meta-P: p after meta-analysis using Stouffer's method. S-VLDL-FC: Free cholesterol in small VLDL. XS-VLDL-C : Cholesterol in very small VLDL. S-VLDL-C: Cholesterol in small VLDL. XS-VLDL-P: Concentration of very small VLDL particles. S-VLDL-CE: Cholesterol esters in small VLDL. S-HDL-P: Concentration of small HDL particles.

3.5 Discussion

Exploring rare coding variation provides an opportunity to gain insights into biological processes regulating the circulating levels of metabolic biomarkers. Here I took advantage of the combination of sequencing data and high-resolution NMR measurements to elucidate how this variation influences multiple metabolic measurements in a healthy cohort of UK blood donors.

To identify variants, genes and gene sets associated with metabolic biomarkers, I used a two-stage approach using WES data in discovery ($N_{discovery}$=3,741), and WGS data for validation ($N_{validation}$=3,401). I first performed single-point association analysis to assess whether I was able to recapitulate established associations with metabolic biomarkers, and potentially identify novel associated rare variants. This yielded associations at 34 previously established loci. The lack of novel findings was expected given the smaller sample size compared to similar studies using the same NMR platform (INTERVAL N=7,142, Kettunen et al. (2016) [173] N=24,925) and the limited power to detect associations with rare variants. As an example, for 7,142 individuals, I only had 2.5% power to detect a significant association (p<9.51x10^{-9} in a combined analysis, **Methods 3.3.6**) with an effect size of 1 for variants with MAF 0.1%. This study was part of a collaboration with Dr Adam Butterworth's group in the University of Cambridge. As such, array based genotype data for the full INTERVAL cohort was analysed by them and will form part of a large-scale meta-analysis collaborative effort. For this reason, I did not explore these results further.

Rare-variant aggregation tests were used to identify genes harbouring multiple rare coding variants associated with metabolic biomarkers. To gain power at the validation stage I adjusted analyses for correlated traits, an approach previously described for single-point analysis [310]. This yielded significant power gains, namely at the known *PAH* association with phenylalanine levels, where adjusting for 71 phenotypically correlated traits resulted in a greater than 30-fold magnitude change in the statistical evidence of association after meta-analysis. This approach therefore can benefit similar studies with multiple phenotypes measured in the same individuals. And, in future efforts, use of association data from these traits in the INTERVAL cohort, instead of publicly available summary statistics, to determine which traits are not genetically correlated could also be used to increase power for many of the measurements that had no publicly available summary statistics, including all derived lipid ratios. Overall, this approach yielded 4,114 gene-trait associations taken forward for validation ($p_{discovery}<5x10^{-3}$). After meta-analysis besides recapitulating previous associations in eight known genes (*APOB, APOC3*, *PAH*, *HAL*, *PCSK9*, *ALDH1L1*, *SCARB1* and *LIPC,* **Table 3.5**), this method also identified four genes (*ACSL1*, *MYCN, B4GALNT3*, *FBXO36*) that met standard gene-level significance ($p<2.5x10^{-6}$, **Table 3.5**) in at least one gene-trait association test. Of these, *ACSL1* and *MYCN* have been previously linked to lipid metabolism [312-314], suggesting that among the gene-level significant findings there may be additional true positives which will merit additional follow-up.

ACSL1, which encodes long-chain-fatty-acid—CoA ligase 1, is the predominant isoform of *ACSL* in the liver. The gene was associated with concentration of IDL particles in this study ($p = 6.20x10^{-7}$), and its deficiency in the liver has been shown to reduce synthesis of triglycerides and beta oxidation, and alter the fatty acid composition of major phospholipids

[316]. An intronic variant (rs60780116) in *ACSL1* has been associated with T2D [317] and elevated expression of *ACSL1* has been shown to be an independent risk factor for acute myocardial infraction [318].

MYCN encodes N-myc proto-oncogene protein and its amplification can lead to tumorigenesis [319, 320]. Previous animal studies have shown that inhibition of *MYCN* can lead to accumulation of intracellular lipid droplets in tumour cells [314]. Here I find association between *MYCN* and concentration of lipids, phospholipids and triglycerides in medium VLDL, total particle concentration of medium VLDL and triglycerides in small VLDL (min $p = 6.20 \times 10^{-7,}$ **Table 3.5, Supplementary Table 2 of Riveros-Mckay et al (in preparation, Appendix B)**).

The other two genes do not have any obvious link to lipid metabolism. *B4GALNT3* encodes beta-1,4-N-acetyl-galactosaminyl transferase 3. This protein mediates the N,N'-diacetyllactosediamine formation on gastric mucosa [321]. Mouse knockouts have been associated with abnormal tail movements, abnormal retinal pigmentation and increased circulating alkaline phosphatase levels [322] and variants near the gene have been associated with height and hip circumference adjusted for BMI in human GWAS [94, 323]. *FBXO36* is a member of the F-box protein family. F-box proteins are known to be involved in protein ubiquitination [324]. Replication of these signals in additional studies would represent a novel link between these genes and lipid metabolism.

In gene set analysis, the “regulation of pyruvate dehydrogenase (PDH) complex” pathway was newly associated with 20 traits, mostly related to IDL and LDL lipoproteins. None of the genes in this pathway have been previously linked to any of these phenotypes, and this data suggests the signal arises from a cumulative effect of predicted loss-of-function variants in

different genes in the pathway (**Figure 3.1**), which represents a novel link between this pathway and lipoprotein metabolism. Of note, a carrier of a rare stop gain mutation (Gln248Ter) in *PDHX* had very high levels of LDL-C (4.086 mmol/l, top 0.03% of full INTERVAL cohort) with no other rare mutation in genes known to harbour rare mutations causative of hypercholesterolaemia (*PCSK9*, *APOB*, *LDLR*). The other carrier of this variant had slightly increased LDL-C levels but within normal clinical range (1.823 mmol/l, top 19.3% of the full INTERVAL cohort). Since we lack information on medication, specifically, lipid lowering medication, the degree to which this variant influences the observed LDL-C levels is difficult to assess. The PDH complex has been shown to be crucial for metabolic flexibility, i.e. the capacity to adjust fuel oxidation based on nutrient availability, which itself has been shown to play a role in cardiovascular disease [325].

In analyses aiming at identifying effector transcripts at established GWAS loci associated with traditional lipid measurements (HDL-C, LDL-C, TC and TG), I established that reported genes mapping near HDL-C associated loci were enriched for rare coding variants associated with multiple HDL-related measurements. The results remained significant ($p<0.005$) after removing genes known to be directly involved in HDL metabolism, suggesting rare coding variants in this gene set contribute to variation in these traits, and that this gene set is potentially enriched for additional effector transcripts, though common variants in the same haplotype as these rare variants could also account for some of the signal we observe. One of the major limitations of this approach is that most of the times, the reported gene is the closest gene and we may miss the true causal gene if the GWAS signal is regulating a more distant gene. It is also important to note that an enrichment of rare variant associations

near reported genes does not necessarily mean that they solely explain the GWAS non-coding association and other genes might also be contributing to the signal.

Finally, I showed that one can detect enrichment of rare variation in genes involved in lipoprotein metabolism in phenotypic extremes of some of these NMR measurements. Specifically, I showed enrichment of rare variants in hyperlipidaemia related genes in individuals with very low levels of cholesterol and esterified cholesterol in small VLDL, total small VLDL particle concentration, and enrichment of rare variants in HDL remodelling genes in individuals with very low levels of small HDL particles. Given that high levels of small HDL particles have been previously associated with higher incidence of ischemic stroke (IS) [326] some of these variants could have protective effects. These results are in agreement with previous work on LDL-C [285] and HDL-C [327] that show that common polygenic signals seem to have a higher impact on the higher extremes of lipid traits whereas there is evidence for a higher prevalence of rare variation on the lower extremes [327]. This is also expected since the INTERVAL cohort consists predominantly of healthy blood donors and therefore the distribution of many of these traits might be truncated and depleted of individuals with rare "damaging" variants. Another factor that could contribute to the observed results is that each trait will have a different distribution and given the fact I am choosing an arbitrary number of participants at the top and bottom of the distribution, these participants will not represent equivalent "extremes".

A major limitation of rare variant association analyses to date is that, despite the advances in computational methods predicting the pathogenicity of rare variants, many of these predicted deleterious variants appear to exert little to no effect as evidenced by the non-significant associations with known positive controls where one should be well powered to

detect association if most of these variants were sufficiently deleterious. Some reported gene-based associations may be due to a few population specific variants, making those findings hard to replicate. As an example, a study using the same NMR platform and performing gene-based analysis using exome-chip data found a significant association of *LIPG* with many HDL subclass traits (min p=3.8x10^{-17}, all protein-truncating and missense variants, $N_{variants}$=5 in a Finnish population [288] whereas in this study the same gene was only nominally significant in triglycerides in medium HDL (p=0.049) querying 19 missense and LoF variants predicted to be deleterious. Power in our study was ~ 82% to find an association at $p < 0.001$ if 50% of the variants included in the test were causal and had the same direction and maximum beta is 1.1, this dropped to ~75% power if 20% of those variants had opposite directions of effect. Upon further inspection, the burden in the original study is mostly driven by one LoF variant (rs200435657, p=4x10^{-6}), and one missense variant (rs201922257, p=8.6x10^{-9}) that are almost monomorphic in Non-Finnish Europeans (gnomAD AC=1 and 7 respectively, AN= 126,228 and 126,712) but have an increased AC in Finnish populations (gnomAD AC=43 and 44 respectively, AN= 25,782 and 25,784). Another missense variant contributing to the association (rs77960347, p=4.8x10^{-6}) is low frequency in NFE (INTERVAL MAF=1.6%) and therefore was not included in our analysis, but it is worth noting that this variant is predicted to be tolerated by SIFT and only possible damaging by PolyPhen. Another study using the same platform but focusing on amino acids [289] found a burden of rare variants in *BCAT2* (p=7.4x10^{-7}, all protein-truncating and missense variants, $N_{variants}$=3) affecting valine levels where one of the two variants driving the association (rs199999090, p=5.36x10^{-4}) was monomorphic in our data and the other variant (rs117048185, p=4.12x10^{-4}) was also similarly associated in my dataset (p=3.x89x10^{-3}) but was not predicted to be deleterious by MCAP (or other similar algorithms

like PolyPhen and SIFT) and therefore was not included in the burden test that included eight variants (p_{burden}=0.76). Other examples of non-significant associations from traditionally measured lipid traits include a *PNPLA5* association with LDL-C [328] and a *TEAD2* association with HDL-C [284]. In the case of *PNPLA5* we tested 10 predicted deleterious variants and found no association p=0.59. However, the reported association with *PNPLA5*, was driven by an African American signal [283]. In the case of *TEAD2* the SNP driving the signal, rs142665148, is monomorphic in the European population and was found in a Chinese population, although unlike *LIPG*, *BCAT2* and *PNPLA5,* this gene is not a known effector transcript and might represent a false positive.

Further work on the INTERVAL cohort incorporating proteomics data could help better understand the potential functional consequences of rare coding variation and help bridge the gap between the rare variant analyses associations presented in this chapter and the observed consequences to circulating metabolic biomarkers.Altogether, my results showed that focusing on rare variation and deep metabolic phenotyping provides new insights into circulating metabolic biomarker biology. This argues for the expansion of deeper molecular phenotyping as part of large cohort sequencing efforts to gain further understanding on the role of rare coding variation on circulating metabolic biomarkers which may potentially lead to novel drug target discovery and/or provide additional genetic validation for specific targets.