Lipo-proteomic and metabolic fine-mapping of genetic lipid risk loci

Karsten Suhre[1,2,], Raghad Al-Ishaq[1,2], Aziz Belkadi[1,2], Tanwir Habib[1,2], Anna Halama[1,2], Nisha Stephan[1,2], Gaurav Thareja[1,2], Shaza Zaghlool[1,2], Eric B. Fauman[3], S. Hani Najafi-Shoushtari[4,5*]*

[1] Bioinformatics Core, Weill Cornell Medicine-Qatar, Education City, 24144 Doha, Qatar

[2] Department of Biophysics and Physiology, Weill Cornell Medicine, New York, NY, U.S.A.

[3] Internal Medicine Research Unit, Pfizer Worldwide Research, Development and Medical, Cambridge, MA, 02139, USA

[4] Division of Research, Weill Cornell Medicine-Qatar, Education City, 24144 Doha, Qatar

[5] Department of Cell and Developmental Biology, Weill Cornell Medicine, New York, NY, U.S.A.

[*] Correspondence to K.S. (kas2049@qatar-med.cornell.edu) and S.H.N-Sh. (han2005@qatar-med.cornell.edu)

Abstract

Dysregulated blood lipid levels sit at the nexus of cardiometabolic disorders and are major predictors of human cardiovascular health. Using five major lipid traits (HDL-C, LDL-C, non-HDL-C, TC, and TG), a recent genome-wide association study (GWAS) in 1.65 million individuals identified and fine-mapped almost 900 genetic loci that may be implicated in the etiology of dyslipidemia and related cardiovascular disease. However, a deeper functional understanding of these associations is needed to assess their therapeutic potential as druggable targets. Here we leveraged data from over 98,000 participants of UK Biobank for deep molecular phenotypic refinement to identify loci associated with 168 distinct NMR-derived lipo-proteomics and metabolomics traits. We found, irrespective of the higher multiple-testing burden, 215 lipid risk variants out of 892 to be associated at a combined trait- and genome-wide significance level with at least one NMR trait, while only 115 variants were found using the five "classical" lipid traits. Hypothesis-free testing of >14,000 ratios between metabolite pairs significantly increased statistical power (p-gain) at 74% of the loci, revealing distinct groups of variants primarily affecting the ratio of lipid content in TG-rich remnants and HDLs, and impacting lipoprotein particle composition and remodeling more generally. Moreover, we demonstrate that NMR-association profiles can serve as surrogate endpoints for future drug trials and Mendelian randomization studies, as shown for lipid-lowering drug usage and incident myocardial infarction. Our NMR association profiles provide a resource for the functional interpretation of lipid risk loci and their evaluation as potential drug targets in the prevention and treatment of ASCVD.

Keywords: Lipid and cholesterol metabolism; Lipoprotein particle composition and remodeling; Atherosclerotic cardiovascular disease; Genetic association; Drug target evaluation

INTRODUCTION

The trafficking of plasma cholesterol and other lipids in lipoprotein particles lies at the heart of metabolic dysregulation and is associated with atherosclerotic cardiovascular disease (ASCVD). Patients suffering from ASCVD often display with elevated LDL-cholesterol (LDL-C) and triglyceride-rich lipoproteins (TRLs) levels accompanied by reduced concentrations of HDL-Cholesterol (HDL-C) [1-4]. While LDL-C levels are efficiently controlled by statins and the recently developed PCSK9 inhibitors, newer reports indicate that a large burden of ASCVD remains, due mainly to a lack of targeted therapies to alleviate persistent dyslipidemia caused by impaired hepatic clearance of TLRs (aka cholesterol remnants) from the arteries, and reduced cholesterol efflux capacity from foam cells, two major pro-atherogenic events that directly contribute to vascular inflammation and atherosclerotic plaque formation [5-10].

Moreover, while LDL-C levels are recognized as a major causal risk factor for ASCVD, Mendelian randomization studies ruled out a causal effect of clinical HDL-C levels on risk of ASCVD, suggesting that HDL exerts its anti-atherogenic activities and function through specific lipids and protein components other than its core cholesterol content [11,12]. Indeed, emerging evidence suggests that variations in HDL particle size and composition of lipids, proteins, and small non-coding RNAs affect HDL function under healthy and diseased conditions, underscoring the importance of identifying gene targets that contribute to the cardioprotective function of HDL subclasses [13,14].

Over the last decade, GWAS have shed much light on common gene variants associated with major lipid traits that were found to be implicated in key regulatory pathways underlying cholesterol and triglyceride metabolism, thereby identifying new modes of metabolic control in the pathogenesis of ASCVD, highlighting potential therapeutic targets in the reverse cholesterol transport pathway [15]. Most recently, Graham et al. [16] reported the largest GWAS with lipid levels in a multi-ethnic study with 1.65 million individuals, including 350,000 participants of non-European ancestries. They analyzed five blood lipid traits, low-density lipoprotein cholesterol, high-density lipoprotein cholesterol, triglycerides, total cholesterol, and non-high-density lipoprotein cholesterol and reported 1,765 index variants at 773 genomic regions with genome-wide significance ($5x10^{-8}$).

While the genetic architecture of these lipid-risk loci was mapped out in detail in this study, much less is known about the metabolic pleiotropy of these loci and their potential link to CVD risk. We, therefore, used the recently published Nightingale lipo-proteomics data that became recently available for ~100,00 participants of the UK Biobank to refine the metabolic architecture of these genetic risk loci.

Nightingale Inc. developed and operates an NMR-based metabolomics platform that has generated the most comprehensive plasma dataset of 168 metabolite and lipoprotein traits [17-19]. A detailed description of the metabolic pathways covered by this platform is provided in **Supplementary Table 1** and https://biobank.ndph.ox.ac.uk/showcase/label.cgi?id=220. In brief, the platform readouts include several amino acids (alanine, glutamine, glycine, histidine, isoleucine, leucine, valine, phenylalanine, tyrosine), glycolysis related metabolites (glucose, lactate, pyruvate, citrate), ketone bodies (3-hydroxybutyrate, acetate, acetoacetate, acetone), creatinine, and two lipid species (linoleic acid, docosahexaenoic acid). The Nightingale platform further reports aggregated lipid traits, specifically, total fatty acids and their degree of unsaturation, omega-3 and omega-6 fatty acids, polyunsaturated fatty acids (PUFA), monounsaturated fatty acids (MUFAs), saturated fatty acids, phosphoglycerates, total cholines, phosphatidylcholines (PCs), and sphingomyelins (SMs). Moreover, the platform also provides readouts of protein glycosylation (glycoprotein acetyls), apolipoproteins B (ApoB) and A1 (ApoA1), and albumin. However, the largest part of the platform's traits are quantitative measures of total lipids, free and total cholesterol (FC, TC), cholesteryl esters (CE), phospholipids (PL), and triglycerides (TGs), all segregated by lipid particle size, together with the concentrations of these particles. Lipoprotein particle classes cover fourteen size ranges, from small HDL particles with an average diameter of 8.7 nm to chylomicrons and extremely large VLDL, with particle diameters from 75 nm upwards. In addition to the individual traits, Nightingale suggests the computation of 81 ratios between these directly measured traits as additional endpoints, including the ratios of TGs to phosphoglycerides, apoB to apoA1, PUFAs to MUFAs, omega-6 fatty acids to omega-3 fatty acids, and the percentages of different lipid content in the lipoprotein size classes, bringing the total number of features up to 249 (**Supplementary Table 2**).

We previously showed that the hypothesis-free approach of testing all-against-all ratios between metabolite levels can increase the statistical power of metabolomics studies, well beyond what is lost by the increased level of multiple testing [20]. Ratios between metabolic traits as independent variables in statistical hypothesis tests have been used for a long time. However, they are often considered mere normalization techniques, such as dividing metabolite levels in urine by creatinine to account for variability in urine concentrations or normalizing metabolite levels in cell extracts by "Bradford" total protein content to account for variability in metabolite extraction. Nevertheless, several GWAS with ratios between metabolite pairs identified associations with exceptionally large gains in the strength of their association that reflected biochemical conversions between the implicated metabolite pairs [21-24]. Using this approach, we inferred FADS1 gene as a desaturase of C20:3 to C20:4 fatty acids, while others found reduced enzyme activity of alkyl-DHAP in smokers [25], identified metabolic markers for drug screening [26], and unraveled genetic variance in blood-type related genes associated with fibrinogen protein phosphorylation [23]. Interestingly, Nightingale already suggests 81 ratios between the directly measured traits as additional endpoints for analysis. We included these ratios (referred to as NightRatios hereafter) as a benchmark, representing a targeted ratio-metric approach.

Our analysis comprises three steps of genetic association analyses. First, we attempt a replication of the Graham et al. [16] lipid GWAS using the same five lipid traits as used in the original study but measured on the Nightingale platform. Then, we proceed to a deep molecular fine mapping of the replicated loci using the 168 NMR traits as endpoints. Finally, we conduct a hypothesis-free all-against-all ratio-metric association study at these loci, testing over 14,000 ratios at each genetic locus. To interpret our findings, we explore groups of key genetic variants and novel risk loci implicated in HDL and TRL remnant metabolism to show that these NMR trait- and ratio-association profiles reflect the biology of well-known lipid regulating pathways. Using Mendelian randomization with myocardial infarction (MI) as outcome and NMR-based metabolic and lipo-protein traits as exposures, we show that these profiles can predict the intervention outcome of presently less-well studied genetic risk variants as potential drug targets. The overarching goal of our study is to further the concept of phenotypic fine-mapping by getting more precise biochemical measures to better understand the biology of the loci and the

function of the encoded genes. We demonstrate that (at least for metabolites) the extra multiple-testing burden of testing ratios is well worth the increase in information.

RESULTS

Extending the number of endpoints from five "classical" lipid traits to 168 NMR traits almost doubles the number of significant lipid risk loci associations.

Graham *et al.* [16] reported 2,624 genetic associations with five major lipid traits (Supplementary Table 3 in Graham *et al.*). These associations included conditional analyses with up to 46 variants in the model and were conducted in multiple ancestries. However, not all variants were available in UK Biobank (UKB). For 1,906 associations, all genetic variants used in the conditioned models could be retrieved from the UKB research analysis platform (RAP) (**Supplementary Table 3**). We computed the associations of all 168 NMR metabolites and all 14,196 ratios between them in linear models with the 1,906 lead variants, age, age^2, sex, and all respective conditional variants as covariates. While the conditioned models for the five lipid traits were sometimes different, their lead variants could be identical or in high LD. We identified 1,091 primary signals where the lead variants of the models were not in strong LD (R2 < 0.7). When multiple traits and models were available, we kept the model with the strongest association signal in Graham *et al.* To avoid spurious associations with rare variants, we only retained the 963 variants that had a minor allele frequency of MAF>1% and further limited the analysis to 892 variants that had genome-wide significant p-values ($p<10^{-8}$) in the conditioned analysis of Graham *et al.* We used these 892 variants (MAF>1%, LD<0.7, $p<10^{-8}$ in Graham *et al.*) as a reference and starting set for the following analyses (**Figure 1**).

Although we only conducted association tests at a limited number of loci, we nevertheless chose to apply a genome-wide Bonferroni significance cutoff of $p_{Bonf} = 5x10^{-8}$ so that our conclusions would also hold if conducted in a wider GWAS setting. We further corrected the number of traits tested as appropriate by dividing p_{Bonf} by 5 (p_{ref}, for testing five lipids), 168 (p_{NMR}, for testing 168 NMR traits), 249 ($p_{NightRatios}$, for testing 168 NMR traits and 81

ratios), or 14,196 ($p_{AllRatios}$, for testing all possible NMR ratios). For convenience and to avoid rounding small p-values to zero, we report negative log10-scaled p-values throughout the paper.

115 of the 892 primary associations were significant at p_{ref} for at least one of the five lipid traits reported by Graham *et al.* (TC, TG, LDL-C, HDL-C, non-HDL-C) and can therefore be considered replicated (and even discoverable in a GWAS setting) in UKB using the Nightingale platform. 777 of the 892 primary associations that did not replicate with any of these five "classical" lipid traits, 101 were significant at p_{NMR} with at least one of the 168 Nightingale traits, and 119 were significant at $p_{NightRatios}$ with one of the 249 NightRatios. 215 of the 892 associations were significant at p_{NMR} with at least one of the 168 Nightingale traits. Hence, despite the increased multiple-testing burden, using more traits led to almost twice as many significant associations (215 versus 115) at loci that are known to be associated with lipid traits (**Figure 2**). This means that the NMR association profiles contain a considerable amount of additional information and can be used to functionally fine-map the "classical" lipid risk associations.

NMR association profiles represent distinct HDL and TLRs gene clusters

To compare the NMR-association profiles between the lipid risk loci, we visualized the association's statistics ($-\log_{10}(p)$ and effect size, aka beta) as heatmaps with the lipid risk loci as one dimension and the Nightingale traits as the other dimension (**Figure 2**). To avoid counting similar signals at a single locus with several associated variants multiple times, we limited the analysis in the following to 137 sentinel variants located in independent regions, based on a distance cutoff of 500kb between regions and retaining only the variant with the strongest association in each region.

Clusters of loci with similar profiles emerged. It is tempting to speculate that each cluster corresponds to a particular set of biological pathways that are perturbed similarly by each of the associated variants. This would link lipid risk genes with shared profiles into groups of potentially co-acting partners on same pathways. We, therefore, proceeded to a k-means clustering of the -log10(p) association profiles. Based on gap statistics, an optimal number of seven clusters was chosen (**Supplementary Figures 1&2**). Inspection of genes that share the

same cluster or that are close on the heatmaps in **Figure 2** revealed that many well studied genes with established biological relationships between them were close or shared the same clusters.

For instance, lecithin:cholesterol acyltransferase (LCAT, sentinel SNP rs4986970) and cholesteryl ester transfer protein (CETP, sentinel SNP rs821840), two key HDL modulating genes associated with coronary heart disease, play a crucial role in the esterification of free cholesterol, and the distribution of cholesteryl ester among lipoproteins in plasma [27]. Notably, we found that both variants represent similar profiles in the same cluster (cluster 4 in **Figure 2d**) irrespective of the counteracting effect of their respective loss-of-function (LoF) variants on HDL-C levels.

Similarly, angiopoietin-like-4 (ANGPTL4, sentinel SNP rs116843064), a secretory protein that restricts lipoprotein lipase (LPL) activity to control TRL remnant uptake and clearance during fasting and feeding, was found to share a very similar profile shared with LPL [28] (cluster 6 in **Figure 2d)** but clustered independently of ANGPTL3 (cluster 1). Despite being functionally inversely correlated, our data shows both genetic variants to be most significantly associated with VLDL lipid content (+beta), and to a lesser extent with HDL (-beta) in a reciprocal manner, but not LDL (see **Supplementary Table 5**), highlighting the ANGPTL4-LPL axis as a strong determinant of HDL and TRLs metabolism in agreement with previous findings [29].

Loss of function in the ANGPTL3 gene, a homolog of ANGPTL4, is associated with a marked reduction of plasma LDL-C and TG and significantly reduced risk of CAD via activation of lipolysis [30]. Currently, therapeutic inactivation of ANGPTL3 is being carried out in intensive clinical trials to treat homozygous familial hypercholesterolemia [31]. Notably, we found genetic variance at the ANGPTL3 locus (rs598253, annotated as USP1 in Graham et al.) to cluster independently of ANGPTL4 and LPL and to be mainly associated with VLDL, large HDL-TG and LDL-C according to the TG and LDL-C lowering effect of ANGPTL3 inhibition in recent clinical trials, but not cholesterol in large and medium sized HDL[31,32] (**Supplementary Table 4&5**). Furthermore, we found SNP rs598253 near ANGPTL3 to have a very similar profile to SNP rs35246381 near N-Acetyltransferase 2 (NAT2), a recently identified gene to be associated with insulin resistance, increased hepatic de novo lipogenesis, and blood TRL remnants levels,

and SNP rs1260326 near the GCKR, the latter found to be significantly associated with nonalcoholic fatty liver disease (NAFLD) [33-35].

Intriguingly, NMR associations also revealed highly similar profiles of a set of corroborated genetic variants associated with systemic and familial hypercholesterolemia. SNP rs73015024 near LDLR, a central factor of hepatic LDL a TLRs uptake and cholesterol transport pathway, SNP rs12916 near 3-hydroxy-3-methylglutaryl–coenzyme A reductase (HMGCR), the rate-limiting enzyme of cholesterol biosynthesis pathway and the target of statin, SNPs rs11591147 near PCSK9 and rs2235215 near MYLIP (aka IDOL), two major factors facilitating post-translational LDLR degradation (cluster 2 in **Figure 2d**) [36,37]. Loss-of-function variants in PCSK9 have been shown to instigate approximately the same effect as variants in HMGCR on the risk of cardiovascular events and type 2 diabetes through their LDL-C lowering effect leading to a MI resistant status in individuals [38].

Intriguingly, NMR associations also revealed highly similar profiles of a set of corroborated genetic variants associated with systemic and familial hypercholesterolemia. SNP rs73015024 near LDLR, a central factor of hepatic LDL-C uptake and cholesterol transport pathway, SNP rs12916 near 3-hydroxy-3-methylglutaryl–coenzyme A reductase (HMGCR), the rate-limiting enzyme of cholesterol biosynthesis pathway and the target of statin, SNPs rs11591147 near PCSK9 and rs2235215 near Mylip (known as IDOL), two major factors facilitating post-translational LDLR degradation (cluster 2 in Figure 2d) [36,37]. We found that all four genetic variants are most strongly associated with CE content in medium VLDL among other LDL, IDL and VLDL associations involved in LDLR pathwayFurthero. These, phenotypic similarities found at these LDL loci reflects functional links between PCSK9, MYLIP and LDLR confirms experimental findings that were not otherwise concluded from previous GWAS.

Furthermore, we also gain further insight into potential therapeutic targeting of these loci given its close clustering with HMGCR. Indeed, loss-of-function variants in PCSK9 have been shown to instigate approximately the same effect as variants in HMGCR on the risk of cardiovascular events and type 2 diabetes through their LDL-C lowering effect leading to a MI resistant status in individuals [38] (**Supplementary Figure 3**).

NMR association profiles can predict therapeutic intervention outcome

Based on these findings, we then sought to explore whether the similarity of lipid risk association profiles can be linked to other traits of biomedical or clinical relevance. For this purpose, we computed the association of the NMR traits with statin usage and with incident myocardial infarction (MI) (**Supplementary Table 6**). We further conducted Mendelian randomization with MI as an outcome and the 168 NMR traits as exposures using MR-Base (**Figure 3**) [39]. For this analysis, GWAS summary statistics from CARDIoGRAMplusC4D [40] were accessed through MR-Base (trait ieu-a-798, 43,676 cases, 128,199 controls). We used our summary statistics for the associations of 215 uncorrelated lipid-risk SNPs with the 168 NMR traits to generate exposure instruments for the MR. Only strong instruments (p-value $< 10^{-10}$) were retained. 157 of the 168 NMR traits had at least three strong instruments and were included. As a side-product of our study, this exposure data can in the future be used for MR analyses with other outcomes of interest, such as cardiovascular disease [41]. The instruments are provided as Supplementary Data on Figshare (https://doi.org/10.6084/m9.figshare.19728991).

The MR Egger method was used to compute MR estimates (**Supplementary Table 7**). In total, 40 NMR traits had a significant ($p<0.05/168$) MR estimate, 23 of which were LDL-related, eight VLDL, five IDL, and four cholesterol. ApoB was one of the strongest associations, which is in agreement with a previous similar analysis that prioritized ApoB as key lipid risk factor for coronary artery disease [41]. Consistent with previous studies [12], no HDL-related traits were significant in the MR analysis.

For the association analysis with statin usage, we excluded all individuals who did not use statins but took other cholesterol-lowering medication, and for the analysis with incident MI we excluded individuals who used any cholesterol-lowering medication at the time of recruitment. Of the 100,000 UKB participants for whom Nightingale data is available, 80,791 UKB participants did not use any cholesterol-lowering drugs, while 15,050 used statins. Among the UKB participants who did not use any cholesterol-lowering drugs at baseline, 1,100 subsequently had a MI event. We found that the impact on each lipid species due to statin usage was highly correlated with the effect size of the genetic variant at HMGCR on that same lipid

(**Figure 2**), as may be expected due to the pharmacological mechanism of statins acting on the HMGCR gene product.

The relationship between statin usage and incident MI is more complex but generally reflects the effects of the drug (**Figure 3**), that is, that statins have a strong effect on lowering LDL-related traits but little on HDL. Further, as is known from epidemiological studies, higher LDL and lower HDL correlate with incident MI. However, many VLDL-related traits have a stronger effect on MI risk but are less strongly influenced by statin use. The strongest risk factor not influenced by statin usage is Glycoprotein Acetyls (GlycA). GlycA is an inflammatory biomarker that captures distinct sources of inflammation and provides a measure of cardiovascular risk [42].

Taken together, this NMR-profile analysis sheds new light on the potential that statins could have played in preventing MI in these UKB participants, knowing that participants with incident MI included here did not use cholesterol-lowering drugs. In addition, using a similar approach, we can now ask what the effect of a drug targeting for example CETP, ANGPTL3, and ANGPTL4 could be compared to that of statins (**Supplementary Figure 4**).

Ratios between NMR traits corroborate underlying biological pathways through related trait pairs.

Next, we analyzed the associations of all 14,196 possible ratios between the 168 NMR traits with the lipid risk loci. We used the p-gain statistic to identify significant ratios [20]. The p-gain of an association of a metabolite ratio is defined as the smaller value of the two p-values for a pair of metabolite associations divided by the p-value of the association for their ratio. A p-gain of 10 has been shown to be the equivalent of an alpha level of significance of 0.05 for a single test. Out of the 215 variants that were associated with at least one NMR-trait, 159 (74%) had a significant p-gain on a genome-wide level (p-gain > $10 * 10^6 * 14{,}196$, or $\log10(\text{p-gain}_{\text{AllRatios}}) > 11.2$), indicating that in over two-thirds of the cases a ratio added significant information to the associations. Of 677 associations not significantly associated with at least one of the NMR traits, 20 were significant at $p_{\text{NightRatios}}$ with at least one of the 249 NightRatios, and 35 were significant at $p_{\text{AllRatios}}$ with at least one of all possible ratios. Six loci that were significant at p_{NMR} did not

reach the significance level required for using ratios. Hence, testing all possible ratios rather than only the NMR traits allowed us to identify 35 additional loci, while six loci with weaker signals were lost due to the increased multiple testing burden. Ratios are therefore more likely to provide new knowledge at already known loci (74% of the cases) than helping discover entirely new loci (13% increase in loci).

Ratios that show exceptionally strong increases in the significance of their associations with genetic variants compared to the individual traits involved in the ratios (p-gain) can indicate new biologically relevant pathways [20]. We limited our analysis to 66 loci that displayed a significant p-gain (>p-gain$_{AllRatios}$) and, in addition, discarded multiple signals in the same region (defined by a distance cut-off of 500kb) by keeping only the strongest signal. We scaled the $\log_{10}$(p-gain) values by the largest $\log_{10}$(p-gain) at each locus and only retained those 580 ratios that had at least one scaled $\log_{10}$(p-gain) > 0.8. We then clustered the 66x580 matrix of scaled log10pgain values and applied the same clustering to the corresponding effect sizes (betas) (**Figure 3**).

To identify ratio association profiles of highest interest and that represent clear true positive associations, we focused on ratios that had a substantial p-gain by requiring *ad hoc* that the strength of the association doubles compared to testing the traits alone, and this on a logarithmic scale, that is, we required that the $\log_{10}$(p-gain) be larger than the maximum -$\log_{10}$(p-value) observed with non-ratios alone. A total of 44 of the 137 unique genetic loci met this criterion. In addition, we selected those ratios that added substantially to what was already covered by the NightRatios by requiring that the p-gain be at least 11.2 orders of magnitude larger (the genome-wide Bonferroni significance level for ratios) than the largest association obtained with NightRatios. Following these criteria, a total of 17 loci with 16 unique ratios were identified (**Table 2**). We then filtered the log10pgain matrix for these 16 ratios (**Figure 4**) and identified ratios that shared strong p-gains between multiple genes.

For instance, the top ratios shared by the LCAT and CETP locus were the ratio of cholesteryl esters and of phospholipids in small VLDL particles and the ratio of triglycerides in small HDL and in small LDL particles (**Table 3**). LCAT converts cholesterol and phosphatidylcholines to cholesteryl esters on the surface of high and low-density lipoproteins [43-

45]. These ratios hence directly reflect the known function of LCAT. CETP, on the other hand, reduces HDL neutral lipid core content by transferring CE in exchange for TG, which leads to the conversation of TG rich VLDL remnant into cholesterol loaded-LDL particles. Thus, both ratios found are in accord with CETP activity.

Intriguingly, in addition to the key HDL genes LCAT and CETP, two additional loci that harbor SNPs near TTC39B and PPP1R3B were clustered with LCAT and CETP, suggesting both have a similar effect on these two specific NMR-ratios. Indeed, both variants were initially found as GWAS-implicated genes associated with HDL-C [46]. Subsequent studies have shown intestinal and hepatic mouse homolog of TTC39B to be linked to the stability of LXR and subsequent ABCA1-mediated cholesterol efflux and increase in expression of enzymes mediating PC biosynthesis and incorporation of PUFAs into phospholipids [47-49]. LXR, in turn, is a transcription factor important for the transcription of CETP [50]. Thus, the phenotypic similarity at these two loci increases confidence that TTC39B is the causal gene at rs686030.

On the other, no direct link has been established for the protein phosphatase 1 regulatory subunit 3B (PPP1R3B) in HDL metabolism, suggesting a potential role for its nearby non-coding gene RP11-115J16.1, yet to be identified. This information may, in the future, be used to gain insight into the function of the respective causal genes at these two loci, which in turn – due to their shared NMR-ratio profiles – could be considered potential drug targets with properties similar to LCAT and CETP.

A somewhat more complex example is the gene group constituted by LPL, ANGPTL4, APOB, the APOA1-APOA4-APOA5-APOC3 locus, MLXIPL (also known as ChREBP), CD300LG, GALNT2, and three not yet identified genes for a role in lipid metabolism (**Table 3**). These loci are all strongly associated with five biochemically interpretable ratios, including the ratio of cholesterol in IDL particles and remnant cholesterol and of total esterified cholesterol and total cholesterol. The already known genes in this cluster are functionally related and linked to these ratios, such as LPL, which catalyzes the hydrolysis of triglycerides from circulating chylomicrons and very-low-density lipoproteins (VLDL) [51] and its regulator ANGPTL4. GALNT2 encodes GalNAc-T2 and is a regulator of HDL metabolism through direct targeting and activation of phospholipid transfer protein (PLTP) that acts to transfer PLs from TRL to

HDL and is found to be accordingly associated with PL content in large LDL [52]. The NMR ratio association also corroborated apolipoprotein C3 (APOC3), a potential therapeutic target for lowering TRL and HDL to be associated with overall major lipid ratio content in both TG-rich remnants and HDLs [10,53].

Pairs of genes with a shared strong ratio association, and where one of the genes is well characterized while the other is not, can be used to transfer knowledge between loci. For instance, the association of rs174564 near FADS1/FADS2 with the ratio of linoleic acid and polyunsaturated fatty acids is one of the ratios with the strongest p-gains in this study (p-gain > $10^{2,000}$). It reflects the function of FADS2 as the rate-limiting step at the beginning of the omega-6 polyunsaturated fatty acid (PUFA) pathway. However, the association of rs12928099 near RRN3, PDXDC1, and NTAN1 with the classical lipid traits is not clear from that context considering these three genes as candidates. However, given the strong association of this locus with a ratio that involves the composition of PUFAs, phospholipase A2 group X (PLA2G10) emerges as the likely causal gene since PLA2G10 encodes a phospholipase that converts phosphatidylethanolamines and phosphatidylcholines to the corresponding lysolipid and preferentially releases sn-2 bound PUFAs over saturated fatty acids [54,55], plus, the FADS1/FADS2 locus is well known to also associated with phospholipids that harbor PUFAs in their side chains [21].

DISCUSSION

Although the grade of ASCVD has globally declined, its trend is threatened by persistent dyslipidemia mainly caused by the obesity epidemic and the conjoined increase in Type 2 diabetes. New treatments are needed as recent studies indicate that many patients suffering from ASCVD fail to achieve optimal LDL-C reduction, while existing clinical therapies to increase HDL function and lower TLRs and cholesterol remnants are modestly effective and remain elusive. Identification of genes that contribute to changes in lipid and protein constituents of lipoprotein subpopulations under physiologic and pathologic conditions is a challenge and an unmet need to develop novel precision therapeutics to improve ASCVD and MI outcomes.

Over the last decade, only a few novel lipid gene candidates implicated from GWAS have been functionally validated, often hampered by long and tedious process and complicated by using mouse models that have significant differences in lipoprotein profile and hepatic lipid metabolism as compared to primates, leading to controversies and contradictory outcome. At the example of GALNT2 and T39, two novel lipid loci associated with HDL levels, our data suggests that while T39 is mainly associated with HDL, the GALNT2 locus shows additional correlations with both HDL and VLDL particles. Using NMR traits ratios, we then revealed more precisely the link between GALNT2 and phospholipid content in large LDL particles, in line with the findings reported by Rader and colleagues showing GALNT2 as a major activator of PLTP, an enzyme that acts between VLDL and HDL particles by transferring VLDL membrane PL to HDL [52].

Interesting to note in the case of ratio interpretation is the fact that among all possible trait combinations, most often matching pairs were among the strongest associations, that is, two specific metabolite classes from the same particle type (e.g. cholesteryl esters versus phospholipids in small VLDL particles) or the same metabolite class from two-particle types (e.g. triglycerides in small HDL versus small LDL particles). This is a recurring feature of many of the top-ranking ratios that further supports the non-randomness of these associations. Due to their strong increase in the strength of association, these ratios may constitute possible surrogate endpoints for use in clinical trials of drugs targeting the genes in this group.

Our study has of course also its limitations – the biggest being the curse of the "too-many results". We clearly could not discuss all associations to the degree that they deserve. We therefore share all associations in different formats and degrees of condensation as **Supplementary Data** on Figshare (https://doi.org/10.6084/m9.figshare.19728991). These data can (among others) be used as instruments in future Mendelian randomization studies. Another limitation are choices that had to be made regarding the scaling of the traits, the covariates used in the models, significance cut-offs etc. We tried to follow as closely as possible previous work (i.e. following Graham et al. where possible), and be as conservative as possible (i.e. always using Bonferroni correction and not analyzing rare variants). However, some choices, like the number of clusters, distance measures etc. used in the NMR-association profile clustering are to a certain degree *ad hoc*. We therefore did not put too much weight on the interpretation of the

exact cluster boundaries, but rather see this as indicating helper tools in the interpretation and comparison of the NMR association profiles.

We also made some *ad hoc* choices by focussing only on the strongest associations with ratios. There are clearly many more biologically relevant associations with ratio pairs that are worthy of investigation. We there make the full set of ratios associations available in such a form that the interested reader can scrutinize the data down to the individual ratio level. The ratios presented in **Table 2** are certainly of high value to be added to the set of ratios that are analyzed by default when using the Nightingale platform in other studies with clinical outcomes.

In summary, in this study we present the first metabolic and lipo-proteomic fine-mapping of all currently known lipid risk loci using NMR measurements from the Nightingale platform, which is a dataset that has only recently been made available for over 98,000 samples from the UK Biobank. We show that by extending the number of traits from five "classical" lipid traits to the 168 Nightingale NMR traits we almost double the number of risk loci that can be identified in the 98,000 UKB samples, even when imposing genome-wide significance levels. This observation indicates that the NMR association profiles contain a considerable amount of additional information compared to the "classical" lipid traits, possibly reflecting more intricate features related to the composition of lipoprotein particles. At examples we showed that the distinct clusters of NMR association profiles often reflect known underlying biological processes that can be linked to the mode of action of lipid lowering drugs, such as statins, and further through Mendelian randomization to clinical endpoints, such as MI. We further analyzed all possible ratios between the NMR traits for association with the lipid risk variants and found that specific ratios can corroborate the underlying biological pathways through their related trait pairs, revealing for instance distinct groups of variants that are primarily affecting the ratio of lipid content in TG-rich remnants and HDLs. Taken together, our paper provides a functional interpretation of many of the lipid risk loci reported Graham et al. and has the potential to serve as a resource to guide future translation of this knowledge into novel drug targets and therapy approaches.

METHODS

Data sources. All data was obtained using the UKB RAP system on DNAnexus (data dispensed on December 14, 2021; application id 43418). Samples for analysis were selected using the following criteria: "Ethnic background | Instance 0" is "White", "Genetic ethnic grouping" is "Caucasian", "Spectrometer |Instance 0" is not NULL and "Measurement Quality Flag | Instance 0" is NULL, yielding NMR data for 98,698 out of 502,414 UKB participants. Imputed genotypes for 1,906 variants were extracted from UK RAP BGEN files (https://biobank.ndph.ox.ac.uk/showcase/label.cgi?id=100319) using bgenix [56] and reformatted to text format using plink [57]. NMR data (https://biobank.ndph.ox.ac.uk/showcase/label.cgi?id=220) and additional phenotype data (age, sex, use of cholesterol lowering medication, statin usage, clinical biochemistry, …) were extracted using the DNAnexus cohort browser and the table downloader app (https://ukbiobank.dnanexus.com/landing). Incident myocardial infarction was defined when the reported "Date of myocardial infarction" (https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=42000) was after the "Date of attending assessment centre" (https://biobank.ndph.ox.ac.uk/ukb/field.cgi?id=53). When multiple instances were available, "Instance 0" (baseline) was used.

Statistical analysis. The NMR data was log-scaled and ratios were computed using the identity log(A/B) = log(A) - log(B), which renders the association statistics invariant to inversion of nominator and denominator. NMR values equal to zero were treated as missing. The data was inverse-normal scaled after log-scaling to avoid overly strong effects of outliers that may have resulted from division by small values.

Linear models were computed using R (version 4.1.0) using the R-package maplet (version 1.1.1) [58], with the NMR data as dependent variables and all genetic variants reported by Graham et al., age, age^2, sex, and use of lipid lowering drugs to the model (binary) as independent variables. For ratios between two NMR traits, the p-gain was computed as the p-value for the association of the genetic variant of interest with the ratio divided by the smaller of the two p-values for the individual traits.

Assuming a genome-wide multiple testing level of ~10^6 tests and requiring nominal significance ($p<0.05$, p-gain>10 in case of ratios [20]), the following Bonferroni levels of significance were applied, depending on the context:

- $p_{ref} = 10^{-8} = -\log10(5x10^{-8}/5)$ for testing the five lipid traits (also used by Graham *et al.*)
- $p_{NMR} = 10^{-9.5} = -\log10(5x10^{-8} / 168)$ for testing the NMR traits
- $p_{NightRatios} = 10^{-9.7} = -\log10(5x10^{-8} / 249)$ for testing NightRatios
- $p_{AllRatios} = 10^{-11.5} = -\log10(5x10^{-8} / (168*169/2))$ for testing all ratios & traits
- $p\text{-}gain_{AllRatios} = 10^{11.2} = \log10(10^6 * 10 * (168*169/2))$ p-gain for testing all ratios & traits

K-means clustering was performed using the R package factoextra (version 1.0.7).

Association with statin usage and incident MI. The number of non-statin users was 80,791 individuals, defined as individuals not using any cholesterol lowering drugs, the number of statin users was 15,050 individuals. Individuals using cholesterol lowering drugs but not statins were excluded. Covariates used in the model were age, age2, sex, HbA1c and glucose. Association with incident MI was limited to 80,791 samples not using any cholesterol lowering medication, leaving 79,691 controls and 1,100 incident MI cases (date of reported MI > date of visit of the UKB assessment center).

Mendelian Randomization. MR was conducted using the R-interface to MR-Base with MI as an outcome and NMR traits as exposures [39]. GWAS summary statistics from CARDIoGRAMplusC4D [40] were accessed through MR-Base (trait ieu-a-798, 43,676 cases, 128,199 controls). We used the association summary statistics of 215 uncorrelated SNPs with 168 NMR traits to generate exposures. Only strong instruments (p-value < 10^{-10}) were retained. 157 of the 168 NMR traits had at least three strong instruments. The MR Egger method was used to compute MR estimates (**Supplementary Table 7**). Summary statistics for the exposures are provided as **Supplementary Data** in MR-Base compatible format and can be used with the MR-Base online version (http://app.mrbase.org) to generate detailed MR reports for all traits.

AUTHOR STATEMENTS

Data availability statement

All analyzed UK Biobank data was obtained through the UKB RAP system under application reference number 43418. All generated data are provided with the paper.

Ethics statement

All data and samples were collected by UK Biobank protocols following all relevant guidelines and procedures and were shared under rules reviewed by the UK Biobank ethics committee and board.

Code availability statement

Only publicly available software was used in the data analysis (R and Rstudio).

Acknowledgements

The authors wish to thank all UK Biobank participants for their contribution.

Funding

K.S. is supported by the Biomedical Research Program at Weill Cornell Medicine in Qatar, a program funded by the Qatar Foundation. K.S. is also supported by Qatar National Research Fund (QNRF) grant NPRP11C-0115-180010. The statements made herein are solely the responsibility of the authors.

Author contributions

Study design and primary data analysis: K.S.; manuscript writing: K.S. and S.H.N-Sh.; data analysis and interpretation: R.A., A.B., T.H., A.H., N.S., G.T., S.Z., E.B.F.; All authors contributed to the interpretation of results and critically reviewed the manuscript.

Competing interests

E.B.F. is an employee and stockholder of Pfizer. The other authors declare no competing interests.

REFERENCES

1. Stamler, J.; Vaccaro, O.; Neaton, J.D.; Wentworth, D. Diabetes, other risk factors, and 12-yr cardiovascular mortality for men screened in the Multiple Risk Factor Intervention Trial. *Diabetes care* **1993**, *16*, 434-444, doi:10.2337/diacare.16.2.434.
2. Siddiqi, H.K.; Kiss, D.; Rader, D. HDL-cholesterol and cardiovascular disease: rethinking our approach. *Current opinion in cardiology* **2015**, *30*, 536-542, doi:10.1097/hco.0000000000000211.
3. Do, R.; Willer, C.J.; Schmidt, E.M.; Sengupta, S.; Gao, C.; Peloso, G.M.; Gustafsson, S.; Kanoni, S.; Ganna, A.; Chen, J.; et al. Common variants associated with plasma triglycerides and risk for coronary artery disease. *Nature genetics* **2013**, *45*, 1345-1352, doi:10.1038/ng.2795.
4. Klempfner, R.; Erez, A.; Sagit, B.Z.; Goldenberg, I.; Fisman, E.; Kopel, E.; Shlomo, N.; Israel, A.; Tenenbaum, A. Elevated Triglyceride Level Is Independently Associated With Increased All-Cause Mortality in Patients With Established Coronary Heart Disease: Twenty-Two-Year Follow-Up of the Bezafibrate Infarction Prevention Study and Registry. *Circulation. Cardiovascular quality and outcomes* **2016**, *9*, 100-108, doi:10.1161/circoutcomes.115.002104.
5. Nicholls, S.J.; Puri, R.; Anderson, T.; Ballantyne, C.M.; Cho, L.; Kastelein, J.J.; Koenig, W.; Somaratne, R.; Kassahun, H.; Yang, J.; et al. Effect of Evolocumab on Progression of Coronary Disease in Statin-Treated Patients: The GLAGOV Randomized Clinical Trial. *Jama* **2016**, *316*, 2373-2384, doi:10.1001/jama.2016.16951.
6. Groenen, A.G.; Halmos, B.; Tall, A.R.; Westerterp, M. Cholesterol efflux pathways, inflammation, and atherosclerosis. *Critical reviews in biochemistry and molecular biology* **2021**, *56*, 426-439, doi:10.1080/10409238.2021.1925217.
7. deGoma, E.M.; Ahmad, Z.S.; O'Brien, E.C.; Kindt, I.; Shrader, P.; Newman, C.B.; Pokharel, Y.; Baum, S.J.; Hemphill, L.C.; Hudgins, L.C.; et al. Treatment Gaps in Adults With Heterozygous Familial Hypercholesterolemia in the United States: Data From the CASCADE-FH Registry. *Circulation. Cardiovascular genetics* **2016**, *9*, 240-249, doi:10.1161/circgenetics.116.001381.
8. Cohen, J.C.; Boerwinkle, E.; Mosley, T.H., Jr.; Hobbs, H.H. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *The New England journal of medicine* **2006**, *354*, 1264-1272, doi:10.1056/NEJMoa054013.
9. Rohatgi, A.; Khera, A.; Berry, J.D.; Givens, E.G.; Ayers, C.R.; Wedin, K.E.; Neeland, I.J.; Yuhanna, I.S.; Rader, D.R.; de Lemos, J.A.; et al. HDL cholesterol efflux capacity and incident cardiovascular events. *The New England journal of medicine* **2014**, *371*, 2383-2393, doi:10.1056/NEJMoa1409065.
10. Tall, A.R.; Thomas, D.G.; Gonzalez-Cabodevilla, A.G.; Goldberg, I.J. Addressing dyslipidemic risk beyond LDL-cholesterol. *The Journal of clinical investigation* **2022**, *132*, doi:10.1172/jci148559.
11. Barrett, T.J.; Distel, E.; Murphy, A.J.; Hu, J.; Garshick, M.S.; Ogando, Y.; Liu, J.; Vaisar, T.; Heinecke, J.W.; Berger, J.S.; et al. Apolipoprotein AI) Promotes Atherosclerosis Regression in Diabetic Mice by Suppressing Myelopoiesis and Plaque Inflammation. *Circulation* **2019**, *140*, 1170-1184, doi:10.1161/circulationaha.119.039476.
12. Voight, B.F.; Peloso, G.M.; Orho-Melander, M.; Frikke-Schmidt, R.; Barbalic, M.; Jensen, M.K.; Hindy, G.; Hólm, H.; Ding, E.L.; Johnson, T.; et al. Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. *Lancet (London, England)* **2012**, *380*, 572-580, doi:10.1016/s0140-6736(12)60312-2.
13. Zhao, Q.; Wang, J.; Miao, Z.; Zhang, N.R.; Hennessy, S.; Small, D.S.; Rader, D.J. A Mendelian randomization study of the role of lipoprotein subfractions in coronary artery disease. *eLife* **2021**, *10*, doi:10.7554/eLife.58361.

14. Rohatgi, A.; Westerterp, M.; von Eckardstein, A.; Remaley, A.; Rye, K.A. HDL in the 21st Century: A Multifunctional Roadmap for Future HDL Research. *Circulation* **2021**, *143*, 2293-2309, doi:10.1161/circulationaha.120.044221.
15. Nurnberg, S.T.; Zhang, H.; Hand, N.J.; Bauer, R.C.; Saleheen, D.; Reilly, M.P.; Rader, D.J. From Loci to Biology: Functional Genomics of Genome-Wide Association for Coronary Disease. *Circulation research* **2016**, *118*, 586-606, doi:10.1161/circresaha.115.306464.
16. Graham, S.E.; Clarke, S.L.; Wu, K.H.; Kanoni, S.; Zajac, G.J.M.; Ramdas, S.; Surakka, I.; Ntalla, I.; Vedantam, S.; Winkler, T.W.; et al. The power of genetic diversity in genome-wide association studies of lipids. *Nature* **2021**, *600*, 675-679, doi:10.1038/s41586-021-04064-3.
17. Soininen, P.; Kangas, A.J.; Würtz, P.; Suna, T.; Ala-Korpela, M. Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics. *Circulation. Cardiovascular genetics* **2015**, *8*, 192-206, doi:10.1161/circgenetics.114.000216.
18. Ahola-Olli, A.V.; Mustelin, L.; Kalimeri, M.; Kettunen, J.; Jokelainen, J.; Auvinen, J.; Puukka, K.; Havulinna, A.S.; Lehtimäki, T.; Kähönen, M.; et al. Circulating metabolites and the risk of type 2 diabetes: a prospective study of 11,896 young adults from four Finnish cohorts. *Diabetologia* **2019**, *62*, 2298-2309, doi:10.1007/s00125-019-05001-w.
19. Tikkanen, E.; Jägerroos, V.; Holmes, M.V.; Sattar, N.; Ala-Korpela, M.; Jousilahti, P.; Lundqvist, A.; Perola, M.; Salomaa, V.; Würtz, P. Metabolic Biomarker Discovery for Risk of Peripheral Artery Disease Compared With Coronary Artery Disease: Lipoprotein and Metabolite Profiling of 31 657 Individuals From 5 Prospective Cohorts. *Journal of the American Heart Association* **2021**, *10*, e021995, doi:10.1161/jaha.121.021995.
20. Petersen, A.K.; Krumsiek, J.; Wagele, B.; Theis, F.J.; Wichmann, H.E.; Gieger, C.; Suhre, K. On the hypothesis-free testing of metabolite ratios in genome-wide and metabolome-wide association studies. *BMC bioinformatics* **2012**, *13*, 120, doi:10.1186/1471-2105-13-120.
21. Gieger, C.; Geistlinger, L.; Altmaier, E.; Hrabé de Angelis, M.; Kronenberg, F.; Meitinger, T.; Mewes, H.-W.; Wichmann, H.E.; Weinberger, K.M.; Adamski, J.; et al. Genetics Meets Metabolomics: A Genome-Wide Association Study of Metabolite Profiles in Human Serum. *PLOS Genetics* **2008**, *4*, e1000282, doi:10.1371/journal.pgen.1000282.
22. Illig, T.; Gieger, C.; Zhai, G.; Römisch-Margl, W.; Wang-Sattler, R.; Prehn, C.; Altmaier, E.; Kastenmüller, G.; Kato, B.S.; Mewes, H.-W.; et al. A genome-wide perspective of genetic variation in human metabolism. *Nature Genetics* **2010**, *42*, 137-141, doi:10.1038/ng.507.
23. Suhre, K.; Shin, S.-Y.; Petersen, A.-K.; Mohney, R.P.; Meredith, D.; Wägele, B.; Altmaier, E.; Deloukas, P.; Erdmann, J.; Grundberg, E.; et al. Human metabolic individuality in biomedical and pharmaceutical research. *Nature* **2011**, *477*, 54-60, doi:10.1038/nature10354.
24. Suhre, K.; Wallaschofski, H.; Raffler, J.; Friedrich, N.; Haring, R.; Michael, K.; Wasner, C.; Krebs, A.; Kronenberg, F.; Chang, D.; et al. A genome-wide association study of metabolic traits in human urine. *Nature Genetics* **2011**, *43*, 565-569, doi:10.1038/ng.837.
25. Wang-Sattler, R.; Yu, Y.; Mittelstrass, K.; Lattka, E.; Altmaier, E.; Gieger, C.; Ladwig, K.H.; Dahmen, N.; Weinberger, K.M.; Hao, P.; et al. Metabolic profiling reveals distinct variations linked to nicotine consumption in humans--first results from the KORA study. *PloS one* **2008**, *3*, e3863, doi:10.1371/journal.pone.0003863.
26. Suhre, K.; Römisch-Margl, W.; de Angelis, M.H.; Adamski, J.; Luippold, G.; Augustin, R. Identification of a potential biomarker for FABP4 inhibition: the power of lipidomics in preclinical drug testing. *Journal of biomolecular screening* **2011**, *16*, 467-475, doi:10.1177/1087057111402200.
27. Eisenberg, S. High density lipoprotein metabolism. *Journal of Lipid Research* **1984**, *25*, 1017-1058, doi:https://doi.org/10.1016/S0022-2275(20)37713-0.

28. Musunuru, K.; Kathiresan, S. Surprises From Genetic Analyses of Lipid Risk Factors for Atherosclerosis. *Circulation research* **2016**, *118*, 579-585, doi:10.1161/circresaha.115.306398.
29. Dewey, F.E.; Gusarova, V.; O'Dushlaine, C.; Gottesman, O.; Trejos, J.; Hunt, C.; Van Hout, C.V.; Habegger, L.; Buckler, D.; Lai, K.M.; et al. Inactivating Variants in ANGPTL4 and Risk of Coronary Artery Disease. *The New England journal of medicine* **2016**, *374*, 1123-1133, doi:10.1056/NEJMoa1510926.
30. Romeo, S.; Yin, W.; Kozlitina, J.; Pennacchio, L.A.; Boerwinkle, E.; Hobbs, H.H.; Cohen, J.C. Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans. *The Journal of clinical investigation* **2009**, *119*, 70-79, doi:10.1172/jci37118.
31. Gaudet, D.; Gipe, D.A.; Pordy, R.; Ahmad, Z.; Cuchel, M.; Shah, P.K.; Chyu, K.Y.; Sasiela, W.J.; Chan, K.C.; Brisson, D.; et al. ANGPTL3 Inhibition in Homozygous Familial Hypercholesterolemia. *The New England journal of medicine* **2017**, *377*, 296-297, doi:10.1056/NEJMc1705994.
32. Dewey, F.E.; Gusarova, V.; Dunbar, R.L.; O'Dushlaine, C.; Schurmann, C.; Gottesman, O.; McCarthy, S.; Van Hout, C.V.; Bruse, S.; Dansky, H.M.; et al. Genetic and Pharmacologic Inactivation of ANGPTL3 and Cardiovascular Disease. *The New England journal of medicine* **2017**, *377*, 211-221, doi:10.1056/NEJMoa1612790.
33. Camporez, J.P.; Wang, Y.; Faarkrog, K.; Chukijrungroat, N.; Petersen, K.F.; Shulman, G.I. Mechanism by which arylamine N-acetyltransferase 1 ablation causes insulin resistance in mice. *Proceedings of the National Academy of Sciences of the United States of America* **2017**, *114*, E11285-e11292, doi:10.1073/pnas.1716990115.
34. Knowles, J.W.; Xie, W.; Zhang, Z.; Chennamsetty, I.; Assimes, T.L.; Paananen, J.; Hansson, O.; Pankow, J.; Goodarzi, M.O.; Carcamo-Orive, I.; et al. Identification and validation of N-acetyltransferase 2 as an insulin sensitivity gene. *The Journal of clinical investigation* **2016**, *126*, 403, doi:10.1172/jci85921.
35. Stender, S.; Kozlitina, J.; Nordestgaard, B.G.; Tybjærg-Hansen, A.; Hobbs, H.H.; Cohen, J.C. Adiposity amplifies the genetic risk of fatty liver disease conferred by multiple loci. *Nature genetics* **2017**, *49*, 842-847, doi:10.1038/ng.3855.
36. Zelcer, N.; Hong, C.; Boyadjian, R.; Tontonoz, P. LXR regulates cholesterol uptake through Idol-dependent ubiquitination of the LDL receptor. *Science (New York, N.Y.)* **2009**, *325*, 100-104, doi:10.1126/science.1168974.
37. Horton, J.D.; Cohen, J.C.; Hobbs, H.H. PCSK9: a convertase that coordinates LDL catabolism. *Journal of lipid research* **2009**, *50 Suppl*, S172-177, doi:10.1194/jlr.R800091-JLR200.
38. Ference, B.A.; Robinson, J.G.; Brook, R.D.; Catapano, A.L.; Chapman, M.J.; Neff, D.R.; Voros, S.; Giugliano, R.P.; Davey Smith, G.; Fazio, S.; et al. Variation in PCSK9 and HMGCR and Risk of Cardiovascular Disease and Diabetes. *The New England journal of medicine* **2016**, *375*, 2144-2153, doi:10.1056/NEJMoa1604304.
39. Hemani, G.; Zheng, J.; Elsworth, B.; Wade, K.H.; Haberland, V.; Baird, D.; Laurin, C.; Burgess, S.; Bowden, J.; Langdon, R.; et al. The MR-Base platform supports systematic causal inference across the human phenome. *eLife* **2018**, *7*, doi:10.7554/eLife.34408.
40. Nikpay, M.; Goel, A.; Won, H.H.; Hall, L.M.; Willenborg, C.; Kanoni, S.; Saleheen, D.; Kyriakou, T.; Nelson, C.P.; Hopewell, J.C.; et al. A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat Genet* **2015**, *47*, 1121-1130, doi:10.1038/ng.3396.
41. Zuber, V.; Gill, D.; Ala-Korpela, M.; Langenberg, C.; Butterworth, A.; Bottolo, L.; Burgess, S. High-throughput multivariable Mendelian randomization analysis prioritizes apolipoprotein B as key lipid risk factor for coronary artery disease. *International journal of epidemiology* **2021**, *50*, 893-901, doi:10.1093/ije/dyaa216.

42. Chiesa, S.T.; Charakida, M.; Georgiopoulos, G.; Roberts, J.D.; Stafford, S.J.; Park, C.; Mykkänen, J.; Kähönen, M.; Lehtimäki, T.; Ala-Korpela, M.; et al. Glycoprotein Acetyls: A Novel Inflammatory Biomarker of Early Cardiovascular Risk in the Young. *Journal of the American Heart Association* **2022**, *11*, e024380, doi:doi:10.1161/JAHA.121.024380.
43. Kosek, A.B.; Durbin, D.; Jonas, A. Binding affinity and reactivity of lecithin cholesterol acyltransferase with native lipoproteins. *Biochemical and biophysical research communications* **1999**, *258*, 548-551, doi:10.1006/bbrc.1999.0690.
44. Hirsch-Reinshagen, V.; Donkin, J.; Stukas, S.; Chan, J.; Wilkinson, A.; Fan, J.; Parks, J.S.; Kuivenhoven, J.A.; Lütjohann, D.; Pritchard, H.; et al. LCAT synthesized by primary astrocytes esterifies cholesterol on glia-derived lipoproteins. *J Lipid Res* **2009**, *50*, 885-893, doi:10.1194/jlr.M800584-JLR200.
45. Piper, D.E.; Romanow, W.G.; Gunawardane, R.N.; Fordstrom, P.; Masterman, S.; Pan, O.; Thibault, S.T.; Zhang, R.; Meininger, D.; Schwarz, M.; et al. The high-resolution crystal structure of human LCAT. *J Lipid Res* **2015**, *56*, 1711-1719, doi:10.1194/jlr.M059873.
46. Teslovich, T.M.; Musunuru, K.; Smith, A.V.; Edmondson, A.C.; Stylianou, I.M.; Koseki, M.; Pirruccello, J.P.; Ripatti, S.; Chasman, D.I.; Willer, C.J.; et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **2010**, *466*, 707-713, doi:10.1038/nature09270.
47. Rong, X.; Wang, B.; Dunham, M.M.; Hedde, P.N.; Wong, J.S.; Gratton, E.; Young, S.G.; Ford, D.A.; Tontonoz, P. Lpcat3-dependent production of arachidonoyl phospholipids is a key determinant of triglyceride secretion. *eLife* **2015**, *4*, doi:10.7554/eLife.06557.
48. Wang, B.; Rong, X.; Duerr, M.A.; Hermanson, D.J.; Hedde, P.N.; Wong, J.S.; Vallim, T.Q.; Cravatt, B.F.; Gratton, E.; Ford, D.A.; et al. Intestinal Phospholipid Remodeling Is Required for Dietary-Lipid Uptake and Survival on a High-Fat Diet. *Cell metabolism* **2016**, *23*, 492-504, doi:10.1016/j.cmet.2016.01.001.
49. Hsieh, J.; Koseki, M.; Molusky, M.M.; Yakushiji, E.; Ichi, I.; Westerterp, M.; Iqbal, J.; Chan, R.B.; Abramowicz, S.; Tascau, L.; et al. TTC39B deficiency stabilizes LXR reducing both atherosclerosis and steatohepatitis. *Nature* **2016**, *535*, 303-307, doi:10.1038/nature18628.
50. Honzumi, S.; Shima, A.; Hiroshima, A.; Koieyama, T.; Ubukata, N.; Terasaka, N. LXRalpha regulates human CETP expression in vitro and in transgenic mice. *Atherosclerosis* **2010**, *212*, 139-145, doi:10.1016/j.atherosclerosis.2010.04.025.
51. Benson, M.D.; Yang, Q.; Ngo, D.; Zhu, Y.; Shen, D.; Farrell, L.A.; Sinha, S.; Keyes, M.J.; Vasan, R.S.; Larson, M.G.; et al. Genetic Architecture of the Cardiovascular Risk Proteome. *Circulation* **2018**, *137*, 1158-1172, doi:10.1161/circulationaha.117.029536.
52. Khetarpal, S.A.; Schjoldager, K.T.; Christoffersen, C.; Raghavan, A.; Edmondson, A.C.; Reutter, H.M.; Ahmed, B.; Ouazzani, R.; Peloso, G.M.; Vitali, C.; et al. Loss of Function of GALNT2 Lowers High-Density Lipoproteins in Humans, Nonhuman Primates, and Rodents. *Cell metabolism* **2016**, *24*, 234-245, doi:10.1016/j.cmet.2016.07.012.
53. Pollin, T.I.; Damcott, C.M.; Shen, H.; Ott, S.H.; Shelton, J.; Horenstein, R.B.; Post, W.; McLenithan, J.C.; Bielak, L.F.; Peyser, P.A.; et al. A null mutation in human APOC3 confers a favorable plasma lipid profile and apparent cardioprotection. *Science (New York, N.Y.)* **2008**, *322*, 1702-1705, doi:10.1126/science.1161524.
54. Singer, A.G.; Ghomashchi, F.; Le Calvez, C.; Bollinger, J.; Bezzine, S.; Rouault, M.; Sadilek, M.; Nguyen, E.; Lazdunski, M.; Lambeau, G.; et al. Interfacial kinetic and binding properties of the complete set of human and mouse groups I, II, V, X, and XII secreted phospholipases A2. *The Journal of biological chemistry* **2002**, *277*, 48535-48549, doi:10.1074/jbc.M205855200.
55. Hanasaki, K.; Yamada, K.; Yamamoto, S.; Ishimoto, Y.; Saiga, A.; Ono, T.; Ikeda, M.; Notoya, M.; Kamitani, S.; Arita, H. Potent modification of low density lipoprotein by group X secretory

phospholipase A2 is linked to macrophage foam cell formation. *The Journal of biological chemistry* **2002**, *277*, 29116-29124, doi:10.1074/jbc.M202867200.

56. Band, G.; Marchini, J. BGEN: a binary file format for imputed genotype and haplotype data. *bioRxiv* **2018**, 308296, doi:10.1101/308296.
57. Purcell, S.; Neale, B.; Todd-Brown, K.; Thomas, L.; Ferreira, M.A.; Bender, D.; Maller, J.; Sklar, P.; de Bakker, P.I.; Daly, M.J.; et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* **2007**, *81*, 559-575, doi:10.1086/519795.
58. Chetnik, K.; Benedetti, E.; Gomari, D.P.; Schweickart, A.; Batra, R.; Buyukozkan, M.; Wang, Z.; Arnold, M.; Zierer, J.; Suhre, K.; et al. maplet: An extensible R toolbox for modular and reproducible metabolomics pipelines. *Bioinformatics (Oxford, England)* **2021**, *38*, 1168-1170, doi:10.1093/bioinformatics/btab741.

FIGURES

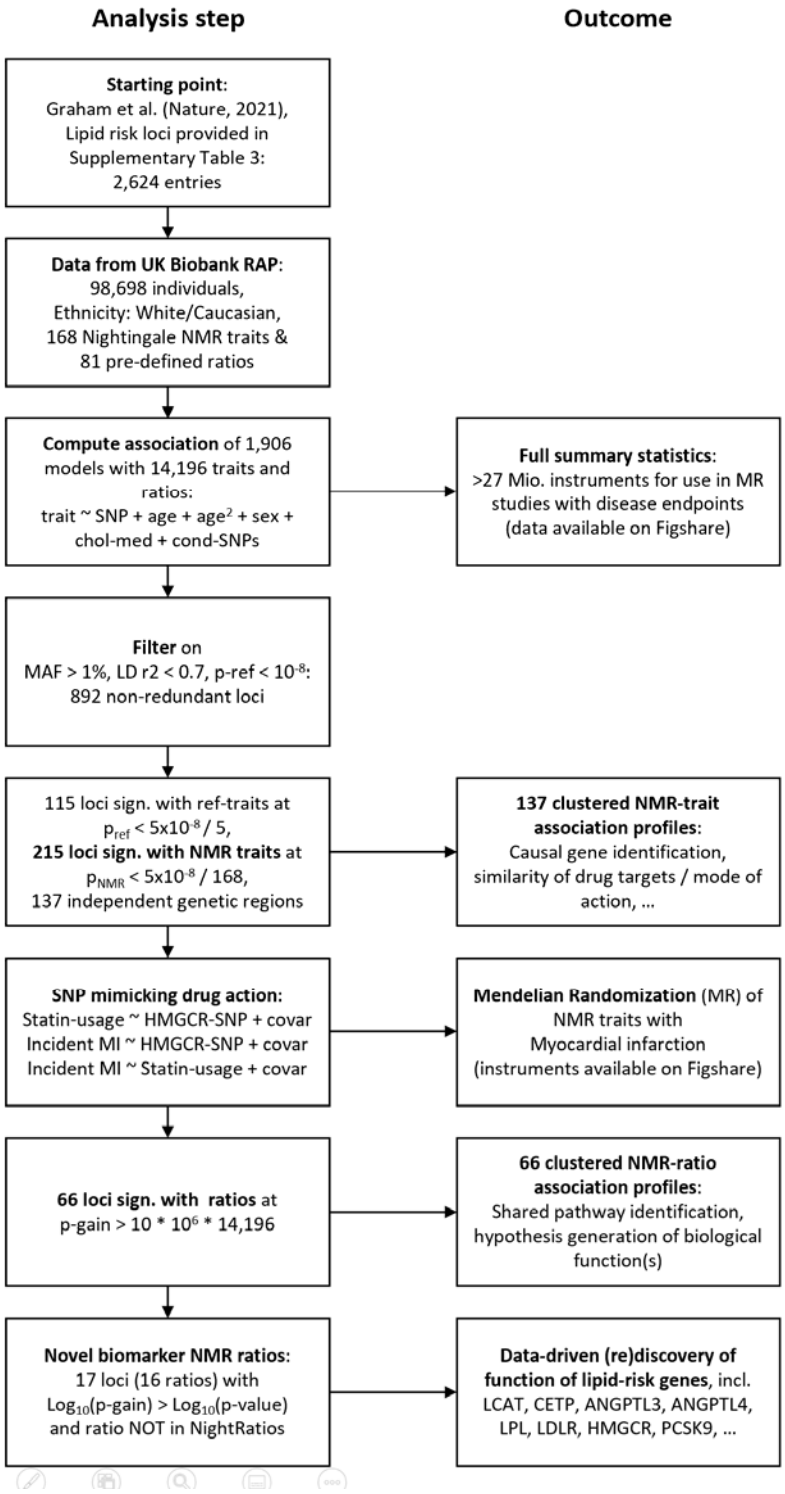


Figure 1. Flow chart of the study.

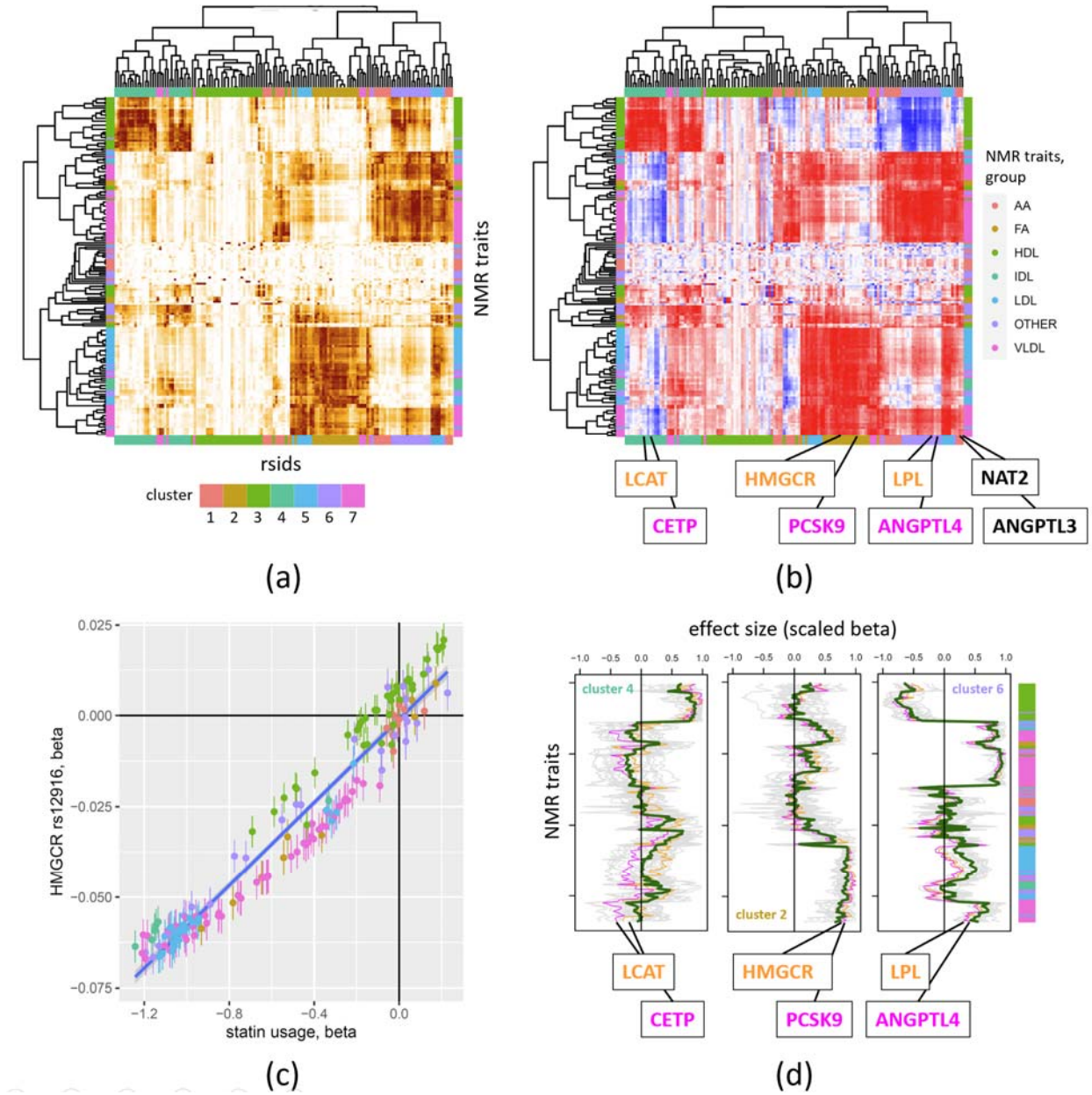


Figure 2. (a) Heatmap of -log10(p-value) for the association of 168 NMR traits at the 137 independent genetic loci (rsids), scaled by division by the largest -log10(p-value) at each locus, darker colors represent stronger associations; (b) As (a), but for effect estimates (beta), scaled by division by the largest absolute beta value and the sign aligned such that the strongest beta at each locus has a positive association with the respective effect allele of that variant (blue colors represent negative associations and red positive associations), heatmaps are also provided in Excel format as **Supplementary Tables 4&5**; (c) Scatterplot of the effect estimates for the association of rs12916 near HMGCR with 168 NMR traits and the effect of the association of statin usage with the 168 NMR traits, colored by NMR trait group (see legend in (b) and **Supplementary Table 1**), error bars represent the standard errors of the effect estimates; (d) Three out of seven clusters of association profiles, individual profiles are in grey and correspond to columns of the heatmap presented in (b), average profiles for each cluster in green, data and order of NMR traits as in (b), details on k-means clustering are in **Supplementary Figures 1&2**; the following selected examples are highlighted: HMGCR (orange, cluster 2), PCSK9 (magenta, cluster 2), LCAT (orange, cluster 4), CETP (magenta, cluster 4), LPL (orange, cluster 6), ANGPTL4 (magenta, cluster 6).

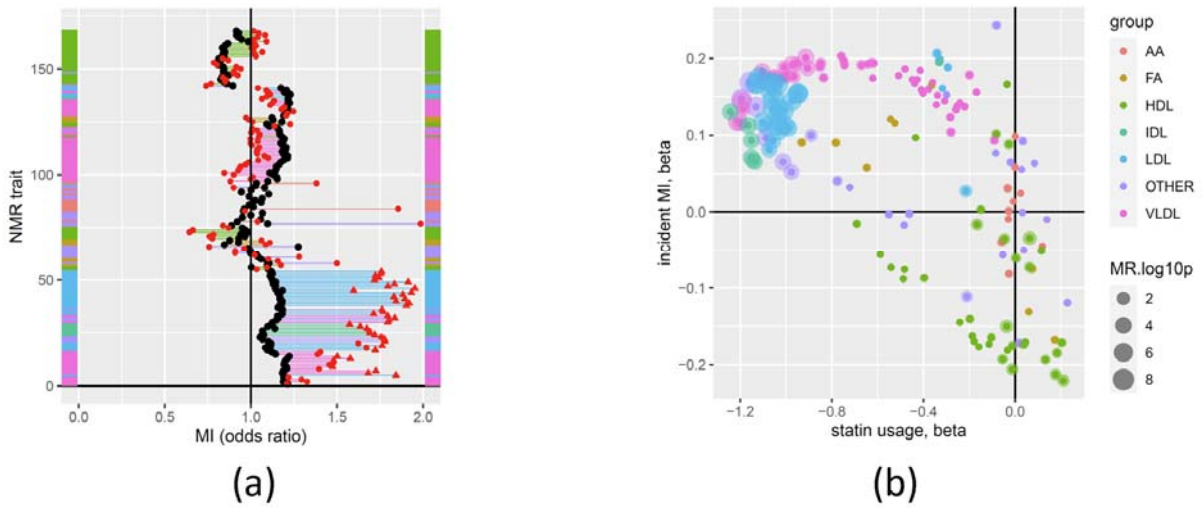


Figure 3. (a) Odds ratios for the association of incident MI with the NMR traits (black) and estimates using Mendelian randomization (red), triangles indicate Bonferroni significant MR estimates, points for the same traits are connected by lines that are colored by the NMR trait group, traits are ordered as in **Figure 2**, (b) scatterplots of the effect size for the association of statin usage (x-axes) and incident MI (y-axes), dot sizes represent the log10(p-value) of the MR estimates (MR Egger).

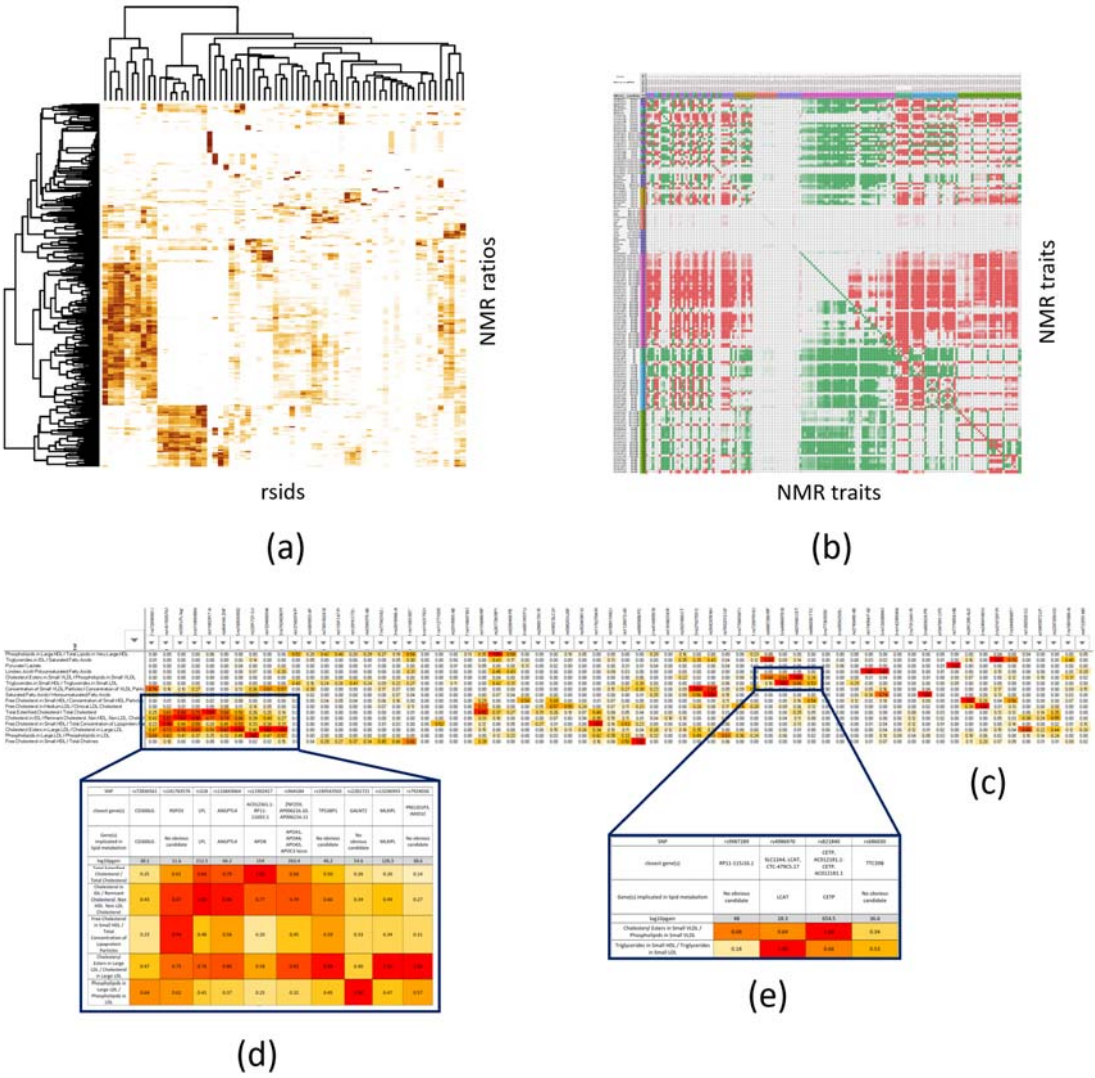


Figure 4. (a) Heatmap of log10(p-gain) values for the association of 580 ratios between NMR traits at 66 independent genetic loci (rsids), scaled by division by the largest log10(p-gain) value at each locus, darker colors represent higher p-gains. The x-axis represents 66 variants that have at least one genome- and metabolome-wide significant association (p-gain > $10^{11.2}$) with at least one of the NMR ratios, the y-axis is limited to 580 ratios that have a scaled p-gain larger than 0.8 in at least one association with the 66 variants (see **Supplementary Tables 7&8** for association data); (b) example of a 168x168 matrix of all log10(p-gain) values for the association of one variant (rs116843064 near ANGPTL4) with all 14,196 traits and ratios, the diagonal contains the -log10(p-value) and the off-diagonal cells contain the log10(p-gain) values, where the sign indicates the direction of the association, data for this example is in Supplementary Table 10, similar data for all loci is available on FigShare under doi:10.6084/m9.figshare.19728991 (private link for review purposes: https://figshare.com/s/426e6dea2172ca463526); (c) log10(p-gain) as in (a), filtered on the ratios listed in **Table 2** to

reveal clusters of ratio-locus pairs that indicate potentially shared biological pathways; (d&e) selected examples that are further detailed in **Table 3** and discussed in the text.

Table 1: NMR-ratio associations for which the p-gain more than doubles the strength of the association on a log-scale. Associations that were already covered by NightRatios were not included here; "log10p Graham et al." is the -$\log_{10}$(p-value) obtained in 1.6 million individuals for "ref trait"; "log10p ref trait" is the -$\log_{10}$(p-value) on the corresponding NMR-trait obtained in this study for ~98,000 individuals, "max log10p no ratios" is the largest -$\log_{10}$(p-value) obtained for 168 NMR-traits, "max log10pgain" is the largest log10(p-gain) obtained when testing all ratio pairs; Complete association data are summarized in **Supplementary Table 3** and available as **Supplementary Data** at https://doi.org/10.6084/m9.figshare.19728991.

Closest gene(s)	rsid	chr	pos	log10p Graham et al.	ref trait	log10p ref trait	max log10p no ratios	max log10 pgain	max log10pgain trait
CERS2	rs267738	1	150,940,625	34.4	HDL	5.7	16.7	18.5	Phospholipids in Large HDL / Total Lipids in Very Large HDL
GALNT2	rs2281721	1	230,297,136	200.1	logTG	16.2	35.7	54.6	Phospholipids in Large LDL / Phospholipids in LDL
APOB	rs11902417	2	21,198,900	12.1	HDL	36.5	135.3	154.0	Total Esterified Cholesterol / Total Cholesterol
HBS1L	rs7776054	6	135,418,916	11.4	nonHDL	3.9	12.2	23.7	Pyruvate / Lactate
MLXIPL	rs13246993	7	73,022,746	53.4	logTG	85.2	110.9	128.3	Cholesteryl Esters in Large LDL / Cholesterol in Large LDL
KLF14, miR-29?	rs11762784	7	130,424,646	88.6	HDL	7.4	11.2	14.6	Free Cholesterol in Small HDL / Total Concentration of Lipoprotein Particles
LPL	rs328	8	19,819,724	882.7	logTG	164.0	194.6	212.5	Cholesterol in IDL / Remnant Cholesterol (Non HDL-C, Non LDL -C)
PRELID1P3, JMJD1C	rs7924036	10	65,191,645	14.9	HDL	5.1	28.3	38.6	Cholesteryl Esters in Large LDL / Cholesterol in Large LDL
PKD2L1	rs603424	10	102,075,479	14.5	LDL	0.4	10.3	42.8	Saturated Fatty Acids /

									Monounsaturated Fatty Acids
GPAM	rs2792703	10	113,931,690	17.6	HDL	18.6	21.4	33.2	Concentration of Small VLDL Particles / Concentration of VLDL Particles
FADS2, FADS1	rs174564	11	61,588,305	205.6	HDL	57.8	1338.4	2047.8	Linoleic Acid / Polyunsaturated Fatty Acids
HNF1A	rs2464190	12	121,415,390	15.7	TC	14.0	35.5	40.2	Free Cholesterol in Medium LDL / Clinical LDL Cholesterol
ALDH1A2	rs261290	15	58,678,720	678.6	HDL	127.7	884.3	1021.4	Free Cholesterol in Small HDL / Concentration of Small HDL Particles
CETP	rs821840	16	56,993,886	223.3	HDL	508.2	568.9	654.5	Cholesteryl Esters in Small VLDL / Phospholipids in Small VLDL
SLC12A4, LCAT	rs4986970	16	67,976,320	15.3	TC	4.9	17.1	28.3	Triglycerides in Small HDL / Triglycerides in Small LDL
PCIF1	rs6065908	20	44,569,930	231.7	HDL	45.1	414.9	423.0	Free Cholesterol in Small HDL / Total Cholines
PNPLA3	rs3747207	22	44,324,855	11.1	logTG	0.1	13.9	47.1	Triglycerides in IDL / Saturated Fatty Acids

Table 2: Examples of clusters of genes that associated with multiple ratios from Table 2. Table entries are log10(p-gain) values scaled by the largest log10pgain value for each locus; data extracted from **Supplementary Table 6**.

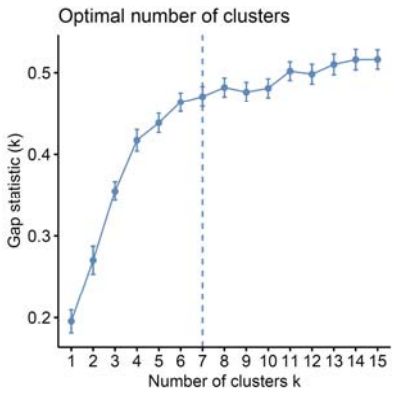
SNP	rs9987289	rs4986970	rs821840	rs686030
closest gene(s)	RP11-115J16.1/ PPP1R3B	SLC12A4, LCAT, CTC-479C5.17	CETP, AC012181.1-CETP, AC012181.1	TTC39B
Gene(s) implicated in lipid metabolism	PPP1R3B	LCAT	CETP	TTC39B
log10pgain	48	28.3	654.5	36.6
Cholesteryl Esters in Small VLDL / Phospholipids in Small VLDL	0.68	0.64	1.00	0.34
Triglycerides in Small HDL / Triglycerides in Small LDL	0.18	1.00	0.66	0.53

(b)

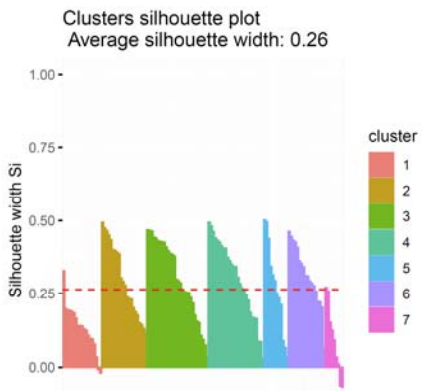
SNP	rs72836561	rs141783576	rs328	rs116843064	rs11902417	rs964184	rs190543502	rs2281721	rs13246993	rs7924036
closest gene(s)	CD300LG	RSPO3	LPL	ANGPTL4	AC012361.1-RP11-116D2.1	ZNF259, AP006216.10, AP006216.11	TP53BP1	GALNT2	MLXIPL	PRELID1P3, JMJD1C
Gene(s) implicated in lipid metabolism	CD300LG	No obvious candidate	LPL	ANGPTL4	APOB	APOA1, APOA4, APOA5, APOC3 locus	No obvious candidate	GALNT2	MLXIPL	No obvious candidate
log10pgain	38.1	11.6	212.5	66.2	154	263.4	46.2	54.6	128.3	38.6
Total Esterified Cholesterol / Total Cholesterol	0.25	0.61	0.84	0.79	1.00	0.68	0.50	0.26	0.20	0.14
Cholesterol in IDL / Remnant Cholesterol (Non-HDL, Non-LDL - Cholesterol)	0.43	0.87	1.00	0.96	0.77	0.79	0.66	0.39	0.49	0.27
Free Cholesterol in Small HDL / Total Concentration of Lipoprotein Particles	0.23	0.94	0.48	0.58	0.20	0.45	0.59	0.33	0.34	0.31
Cholesteryl Esters in Large LDL / Cholesterol in Large LDL	0.47	0.73	0.76	0.80	0.58	0.82	0.98	0.40	1.00	1.00
Phospholipids in Large LDL / Phospholipids in LDL	0.64	0.62	0.41	0.37	0.25	0.32	0.45	1.00	0.47	0.57

(b)

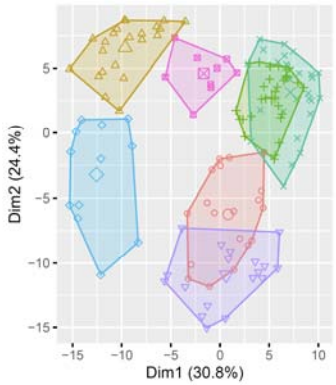
SUPPLEMENTARY FIRGURES



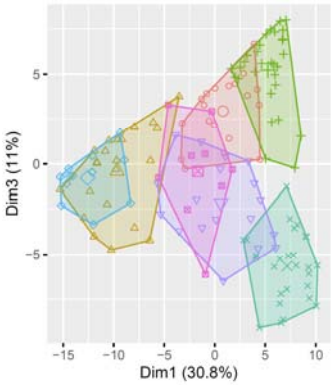
(a)



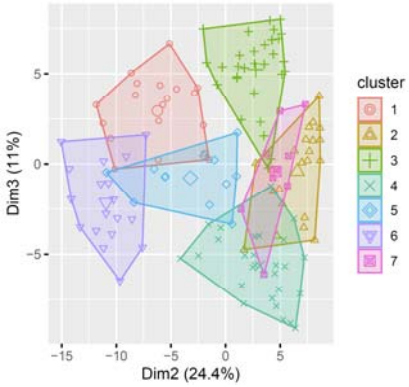
(b)



(c)

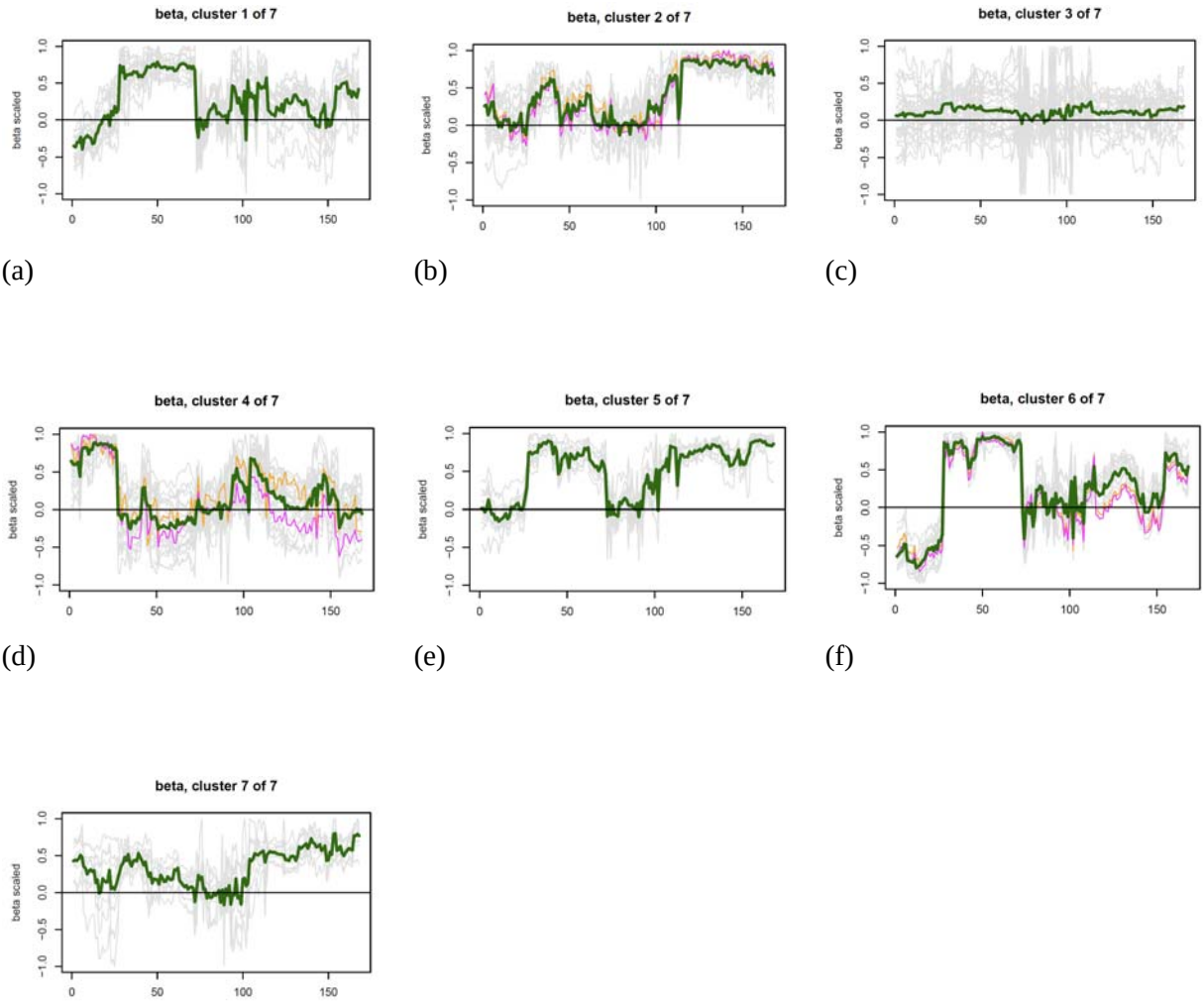


(d)

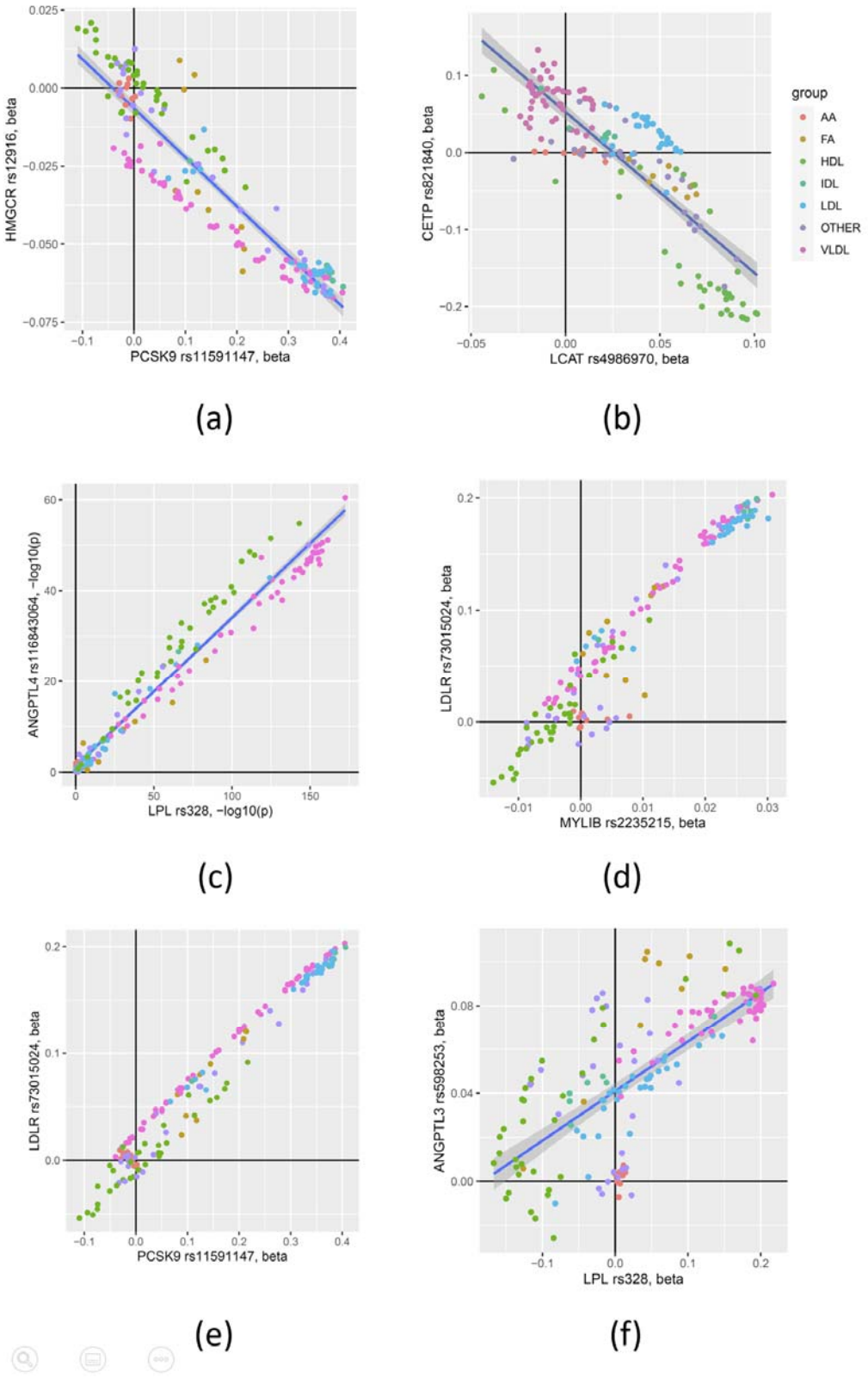


(e)

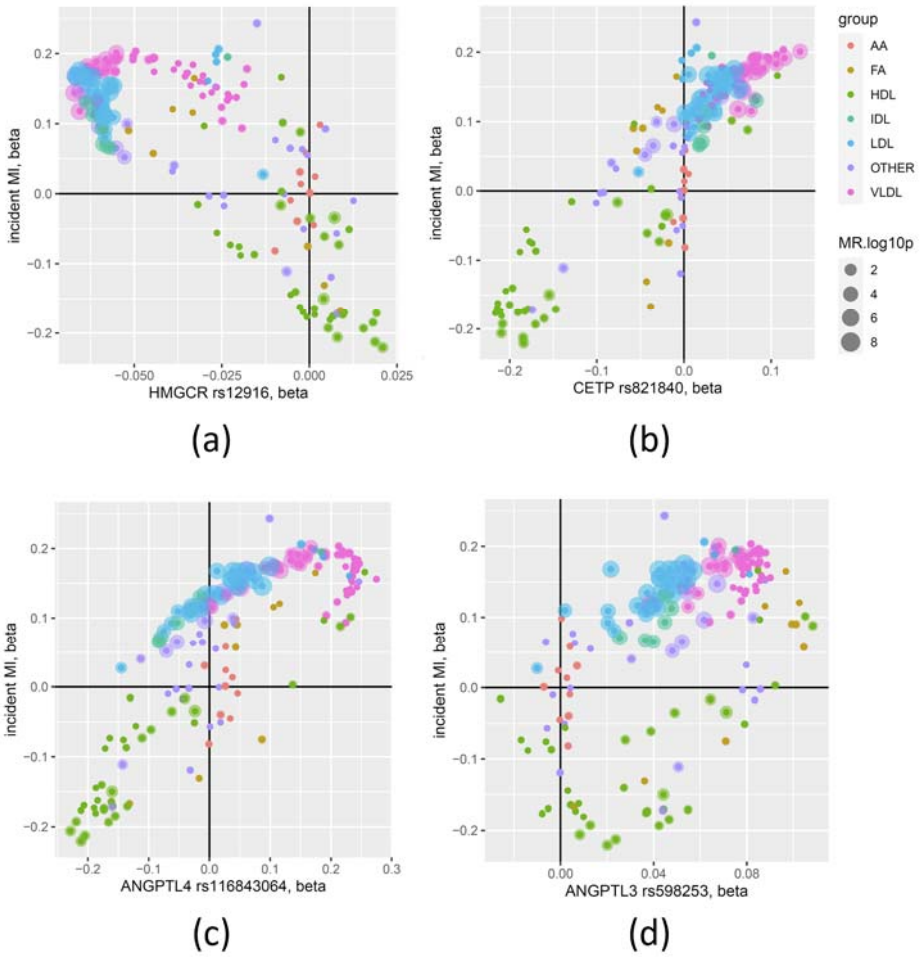
Supplementary Figure 1. (a) Gap statistics as a function of cluster number, used to determine the optimal number clusters, (e) cluster silhouette plot to evaluate cluster specificity, (a-c) principal component analysis (PCA) of the 137 association profiles, colored by cluster membership and showing k-means cluster centers and envelopes.



Supplementary Figure 2. Association profiles for the 137 independent genetic loci by cluster, shown are the effect sizes (beta) of the associations of the lead SNPs with the respective 168 NMR traits, normalized by the largest absolute beta value, and grouped by cluster, NMR traits on the x-axis are ordered as in the heatmaps shown in **Figure 2**, individual profiles are in grey, averaged (reference) profiles per cluster in green, the following loci that are discussed in the paper are highlighted here as follows: HMGCR (orange, cluster 2), PCSK9 (magenta, cluster 2), LCAT (orange, cluster 4), CETP (magenta, cluster 4), LPL (orange, cluster 6), ANGPTL4 (magenta, cluster 6), individual data points can be found in **Supplementary Tables 4&5**.



Supplementary Figure 3: Scatterplots of the effect sizes for the association of selected SNPs with the 168 NMR traits.



Supplementary Figure 4: Scatterplots of the association effect of genetic variance in HMGCR, CETP, ANGPTL3, and ANGPTL4 with incident MI for the 168 NMR traits.

SUPPLEMENTARY TABLES

ST1: List of 168 NMR traits

ST2: List of 81 ratios proposed by Nightingale (NightRatios)

ST3: Association data for 1,906 genetic models (full summary stats provided on FigShare)

ST4: Scaled log10(p-value) for 137 genetic loci and 168 traits (data used in Figure 2a)

ST5: Scaled effect sizes (beta) for 137 genetic loci and 168 traits (data used in Figure 2b)

ST6: Association data for plots shown in Figure 4

ST7: Mendelian Randomization on MI

ST8: Scaled log10(p-gain) for 66 genetic loci and 580 ratios (data used in Figure 5a)

ST9: Scaled effect sizes (beta) for 66 genetic loci and 580 ratios (data used in Figure 5b)

ST10: Example of a formatted p-gain matrix (data provided for 66 genetic loci on FigShare)

SUPPLEMENTARY DATA

Summary statistics for all NMR-trait and NMR-ratio associations at all 1,906 loci, ratio association matrices similar to **Supplementary Table 10** in tab-separated text format, as well as all MR instruments in MRBase format are available on FigShare https://doi.org/10.6084/m9.figshare.19728991.