University of Groningen

Genome-Wide Association Study and Identification of a Protective Missense Variant on Lipoprotein(a) Concentration Protective Missense Variant on Lipoprotein(a) Concentration- Brief Report

Said, M Abdullah; Yeung, Ming Wai; van de Vegte, Yordi J; Benjamins, Jan Walter; Dullaart, Robin P F; Ruotsalainen, Sanni; Ripatti, Samuli; Natarajan, Pradeep; Juarez-Orozco, Luis Eduardo; Verweij, Niek

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2021

Link to publication in University of Groningen/UMCG research database

Arteriosclerosis, Thrombosis, and Vascular Biology

CLINICAL AND POPULATION STUDIES

Genome-Wide Association Study and Identification of a Protective Missense Variant on Lipoprotein(a) Concentration

Protective Missense Variant on Lipoprotein(a) Concentration–Brief Report

M. Abdullah Said,* Ming Wai Yeung,* Yordi J. van de Vegte, Jan Walter Benjamins©, Robin P.F. Dullaart, Sanni Ruotsalainen, Samuli Ripatti©, Pradeep Natarajan©, Luis Eduardo Juarez-Orozco©, Niek Verweij, P. van der Harst©

OBJECTIVE: Lipoprotein(a) (Lp[a]) is associated with coronary artery disease (CAD) but also to LDL (low-density lipoprotein) cholesterol. The genetic architecture of Lp(a) remains incompletely understood, as well as its independence of LDL cholesterol in its association to CAD. We investigated the genetic determinants of Lp(a) concentrations in a large prospective multiethnic cohort. We tested the association for potential causality between genetically determined higher Lp(a) concentrations and CAD using a multivariable Mendelian randomization strategy.

APPROACH AND RESULTS: We studied 371 212 participants of the UK Biobank with available Lp(a) and genome-wide genetic data. Genome-wide association analyses confirmed 2 known and identified 37 novel loci ($P<5×10^{-8}$) associated with Lp(a). Testing these loci as instrumental variables in an independent cohort with 60 801 cases and 123 504 controls, each SD genetically elevated Lp(a) conferred a 1.30 ([95% CI, 1.20–1.41] $P=5.53×10^{-11}$) higher odds of CAD. Importantly, this association was independent of LDL cholesterol. Genetic fine-mapping in the *LPA* gene region identified 15 potential causal variants. This included a rare missense variant (rs41267813[A]) associated with lower Lp(a) concentration. We observed a strong interaction between rs41267813 and rs10455872 on Lp(a) concentrations, indicating a protective effect of rs41267813(A).

CONCLUSIONS: This study supports an LDL cholesterol–independent causal link between Lp(a) and CAD. A rare missense variant in the *LPA* gene locus appears to be protective in people with the Lp(a) increasing variant of rs10455872. In the search for therapeutic targets of Lp(a), future work should focus on understanding the functional consequences of this missense variant.

GRAPHIC ABSTRACT: A graphic abstract is available for this article.

Key Words: causality ■ coronary artery disease ■ genetics ■ lipoproteins ■ polymorphism, single nucleotide

L ipoprotein(a) (Lp[a]) is a macromolecular complex that consists of an LDL (low-density lipoprotein) particle to which a glycoprotein—called apo(a) (apolipoprotein[a])—is linked via a disulfide bond to the constitutional apolipoprotein-B(100)—the principal protein that is carried on LDL particles. Elevated levels of Lp(a) have been associated with an increased risk of coronary artery disease (CAD) in epidemiological studies and meta-analyses.[1] More recently, Mendelian randomization (MR) studies using genetic determinants of Lp(a)

Nonstandard Abbreviations and Acronyms

apo(a)	apolipoprotein(a)
CAD	coronary artery disease
CARDIoGRAM plusC4D	Coronary Artery Disease Genome Wide Replication and Meta-Analysis Plus the Coronary Artery Disease Genetics
GWAS	genome-wide association study
HDL	high-density lipoprotein
KIV	kringle IV
KIV2	kringle IV subtype 2
LD	linkage disequilibrium
LDL	low-density lipoprotein
LDL-C	low-density lipoprotein cholesterol
LDLR	low-density lipoprotein receptor
Lp(a)	lipoprotein(a)
MR	Mendelian randomization
MR-PRESSO	Mendelian Randomization Pleiotropy Residual Sum and Outlier
MVMR	multivariable Mendelian randomization
PCSK9	proprotein convertase subtilisin-kexin type 9
SNP	single-nucleotide polymorphism
SuSiE	Sum of Single Effects
WES	whole-exome sequencing

Highlights

- Leveraging data from over 370 000 UK Biobank participants, we found 37 novel genetic variants associated with lipoprotein(a) values.
- We identified a missense variant that was associated with strong lipoprotein(a)-lowering effects in carriers of the lipoprotein(a) increasing rs10455872 variant.
- Mendelian randomization analyses in over 60 000 cases and 123 000 controls provide evidence for a causal link between genetically determined higher lipoprotein(a) and increased risk of coronary artery disease, independently of LDL (low-density lipoprotein) cholesterol within multivariable Mendelian randomization analyses.

have suggested a causal link between elevated Lp(a) levels and CAD.[2–6] MR analyses use genetic variants, which are randomly distributed and fixed at conception, as instrumental variables for a risk factor of interest to minimize confounding and reversed causality bias and, therefore, have the potential to provide evidence on the putative causal links with a disease.[7]

The *LPA* gene (6q25.3-q26) encodes the apo(a) component, which has evolved from the plasminogen gene and contains a variable number of KIV (kringle IV) repeats. This includes the highly polymorphic KIV$_2$ (KIV-subtype 2), which can have multiple repeats ranging from 1 to over 40 copies.[8] The number of KIV$_2$ repeats have been inversely associated with Lp(a) concentrations.[8] Twin studies have suggested that >90% of variance is genetically determined.[9] Genome-wide association studies (GWAS) of Lp(a) concentrations have thus far identified single-nucleotide polymorphisms (SNPs) outside KIV$_2$, explaining 21% to 63% of the variance in Lp(a).[2,3,6,10] Yet, several aspects of the genetic architecture and causal relationships of Lp(a) remain to be better understood. Not only 30% of variance remains to be explained but also the causal mechanism marked by the strongest associated SNP (the intronic rs10455872) is not well understood,

and few independent genetic variants affecting Lp(a) concentrations outside the *LPA* gene region are known. Finally, although the association between Lp(a) and CAD has been suggested not to be affected by LDL cholesterol (LDL-C)–lowering therapies,[3,11,12] additional lines of evidence could be obtained by multivariable MR (MVMR) taking LDL-C into account in this association.

We aimed to better characterize the genetic architecture underlying Lp(a) concentrations across the whole genome in a large prospective observational study. Further, to increase our understanding of the functional variants in the *LPA* gene, we applied genetic fine-mapping by incorporating exome sequencing data. Finally, we applied a 2-sample MVMR approach to investigate whether the genetic variants associated with Lp(a) influence CAD independently from LDL-C.

METHODS

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Study Population

The UK Biobank study has been described in detail previously.[13] The UK Biobank study is a population-based prospective cohort in the United Kingdom in which >500 000 individuals aged between 40 and 69 years were included from 2006 to 2010. All participants have given informed consent.[14] The study has approval from the North West Multi-Centre Research Ethics Committee for the United Kingdom, from the National Information Governance Board for Health and Social Care for England and Wales, and from the Community Health Index Advisory Group for Scotland.[15] Ethnic background was determined using self-reported data at the assessment center (field ID 2100).

Lp(a) and LDL-C Measurement

Lp(a), in nmol/L, was measured using an immunoturbidimetric assay (Randox Bioscience, United Kingdom). LDL-C (in

mmol/L) was measured using an enzymatic selective protection assay. Both lipoproteins were measured on a Beckman Coulter AU5800 (Beckman Coulter, Ltd, United Kingdom).

Whole-Exome Sequencing

Full details of the whole-exome sequencing (WES) in the UK Biobank have been reported previously.[16] In short, WES was performed using IDT xGen Exome Research Panel v1.0, targeting 38 997 831 bases in 19 396 genes. Exomes were captured including 100 bp flanking regions. Coverage exceeded 20× at 94.6% of sites on average in all samples and among targeted bases. All variants passed quality control criteria, had <10% individual and variant missingness, and Hardy Weinberg $P > 10^{-15}$. Compared with the imputed sequence data, the WES data contain over 7× more coding variants and 20× more loss-of-function variants. A total of 4.7M variants within targeted regions and 9.7M across all covered bases were identified and mapped to a full CRCh38 reference in the used functionally equivalent pipeline. At the time of writing, 7554 targets were incorrectly mapped and removed from the analysis as recommended by the UK Biobank. This resulted in a 0.48% loss of variants overall. The *LPA* gene region was not affected. At the time of analysis, WES data were available for 49 960 participants.

Genotyping and Imputation

The UK Biobank participants were genotyped using custom Affymetrix Axiom (UK Biobank Lung Exome Variant Evaluation or UK Biobank) arrays with >95% common content. The genotyping methods, arrays, and quality control procedures have been extensively described previously.[17]

Functional Annotation of Variants

Candidate genes at each locus were prioritized based on proximity by selecting the nearest protein coding gene and any additional gene within 10 kb of the sentinel SNP. Variants identified using fine-mapping were annotated using Ensembl Variant Effect Predictor[18] for allele frequency, variant consequences from dbNSFP and MaxEntScan, and effect prediction by various tools including CADD (Combined Annotation Dependent Deletion), SIFT (Sorting Intolerant From Tolerant), PolyPhen, Condel (CONsensus DELeteriousness), and LoFtool.

MR Assumptions

SNPs were considered valid instrumental variables for the MR analyses if (1) they were strongly associated with the risk factor of interest, (2) the SNPs were not associated with confounders of the association between risk factor and outcome, and (3) the SNPs affected the outcome exclusively through their effect on the risk factor being studied (Figure I in the Data Supplement).

Statistical Analysis

All genetic analyses were adjusted for age at inclusion, squared age at inclusion, genotyping array, the first 30 principal components to adjust for population stratification (provided by the UK Biobank), and lipid-lowering drug usage at inclusion. The study design is depicted schematically in Figure 1.

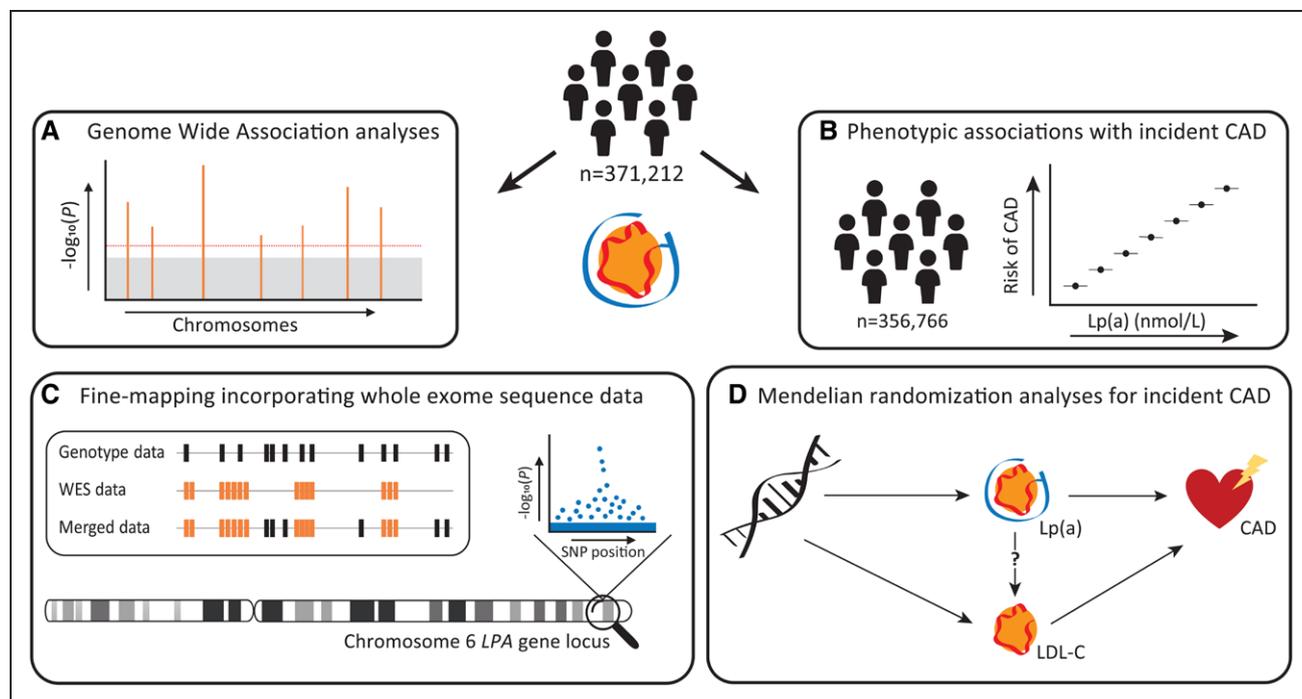### Genome-Wide Association Studies

We performed GWAS for inverse rank normalized serum Lp(a) concentrations. GWAS using the genotyped and imputed data were performed using BOLT-LMM v2.3.1[19] and included 19M SNPs. To obtain a set of independent SNPs associated with Lp(a), SNPs that passed the genome-wide significance threshold of $P < 5 \times 10^{-8}$ were clumped together based on linkage disequilibrium (LD) $r^2 > 0.005$ and 2.5-Mb distance using the clumping procedure in PLINK 1.9. A locus was defined as a 1-Mb region surrounding the most significant SNP. SNPs with minor allele frequencies <0.005 or INFO scores <0.3 were excluded. The proportion of additive variance explained by the top variants was estimated by fitting a multivariable linear regression model on Lp(a) concentration, assuming an additive genetic model for the genetic variants and using the covariates as described above.

Genetic Fine-Mapping

To allow statistical fine-mapping with a higher variant density, the WES data were overlaid with the genotyped data, using the WES data when a variant was present in both sources (Figure 1). LiftOver[20] was used to convert the genotype data from GRCh37 to GRCh38. Genetic fine-mapping in the merged data was subsequently performed using 2 Bayesian fine-mapping methods to identify putative causal variants, namely the Sum of Single Effects[21] (SuSiE) model and FINEMAP.[22] SuSiE implements an iterative Bayesian stepwise selection procedure, which creates a number of credible sets with independent or highly correlated variables of which one has a nonzero effect, while all are associated with Lp(a). FINEMAP performs a shotgun stochastic search to efficiently evaluate possible causal configurations of SNPs. Fine-mapped variants were annotated using Variant Effect Predictor for the variant's primary effect prediction. Genetic fine-mapping was performed across a 1-Mb region surrounding the sentinel SNP rs10455872 in the *LPA* region. After selecting individuals with WES data who were also included in the GWAS, PLINK 2.0 was used to perform a linear regression analysis on the inverse rank normalized Lp(a) concentrations. Variants with minor allele frequencies >0.0005 and genotype missingness <0.1 were included, amounting to 7173 variants for fine-mapping. Pairwise LD estimates were calculated from the genotype dosages for individuals included in the GWAS, rather than using external reference panels, which may be inaccurate when scaled to large sample sizes.[23] In the scenario in which multiple SNPs were in a credible set identified by SuSiE, the SNP with the highest posterior inclusion probability was taken as the most likely causal variant for that set. SuSiE was performed first, and the number of credible sets was taken forward as the maximum number of allowed causal SNPs in FINEMAP. SNPs in the top causal configuration in FINEMAP were taken as likely causal variants. SNPs identified by both SuSiE and FINEMAP based on identical rsID or $r^2 > 0.8$ were prioritized and considered more likely to be causal. In the case of selection based on r^2, the SNP with the highest posterior inclusion probability as indicated by SuSiE was selected.

MR Analyses

We performed univariable and multivariable 2-sample MR analyses to investigate evidence for causal links between genetically determined elevated Lp(a) and CAD. MR analyses were performed using summary statistics data from the Coronary Artery

Figure 1. Schematic study design.
Schematic overview of the study design including (**A**) genome-wide association analyses, (**B**) phenotypic associations between lipoprotein(a) (Lp[a]) and coronary artery disease (CAD), (**C**) genetic fine-mapping analyses, and (**D**) Mendelian randomization analyses using the genetic variants associated with Lp(a) as instrumental variables to investigate the association with coronary artery disease. LDL-C indicates low-density lipoprotein cholesterol; and WES, whole-exome sequencing.

Disease Genome Wide Replication and Meta-Analysis Plus the Coronary Artery Disease Genetics (CARDIoGRAMplusC4D) consortium (123 504 controls and 60 801 [33.0%] cases).[24] Sentinel SNPs identified in the GWAS on Lp(a) were used as instrumental variables for Lp(a). SNPs that were not available in CARDIoGRAMplusC4D were replaced with proxies in LD of $r^2 > 0.8$ or excluded from the MR if no eligible proxies were available. Harmonization of SNP effects was performed using the built-in feature of the TwoSampleMR package in R. The framework used for the MR and corresponding heterogeneity and sensitivity analyses is depicted in Figure II in the Data Supplement. F statistics[25] and $I2_{GX}$[26] were calculated to assess potential weak instrument bias. I^2 index,[27] Cochran Q, Rücker Q′, and Q-Q′[28] were used as heterogeneity tests. Univariable MR analyses to investigate evidence for a potential causal association between genetically determined Lp(a) and CAD included fixed and random-effects inverse-variance weighted (IVW) MR, MR-Egger,[28] MR-Steiger,[29] MR Pleiotropy Residual Sum and Outlier[30] (MR-PRESSO), and median- and mode-based estimator MR analyses[31] as outlined in the Data Supplement. MVMR-IVW analyses were performed to estimate the direct effect of Lp(a) on CAD not mediated by the effect of LDL-C on CAD and the direct effect of LDL-C on CAD not mediated by the effect of Lp(a) on CAD (Figure I in the Data Supplement). By conditioning the effects of each SNP on LDL-C in the UK Biobank, the direct effect of Lp(a) on CAD in CARDIoGRAMplusC4D can be estimated. Q_{x1} and Q_{x2} were calculated to test for weak instrument bias and Q_a to test for pleiotropy.[32] MVMR-Egger and MR-PRESSO were performed as sensitivity analyses. Odds ratios with 95% CIs are presented for the MR outcomes. We considered a conservative α of 0.005[33] instead of 0.05 to

provide evidence for a significant causal association. MR analyses were performed using the TwoSampleMR (version 0.4.26), MR-PRESSO (version 1.0), MendelianRandomization (version 0.4.1), and MVMR (version 0.1) packages in R, version 3.5.1.

RESULTS

Cohort Characteristics

A total of 371 212 individuals were included in the analyses (Figure III in the Data Supplement). Baseline characteristics are shown in Table I in the Data Supplement. Compared with women, men were more often diagnosed with hyperlipidemia and CAD, next to having a higher blood pressure and BMI. Lp(a) concentrations ranged between 3.8 and 189 nmol/L, with a median value of 21.1 nmol/L (interquartile range, 9.58–61.9). A total of 723 individuals with Lp(a) concentrations had no LDL-C values available. Among the 370 489 individuals with LDL-C values, Lp(a) was correlated with LDL-C with a Pearson ρ of 0.081 (Figure IV in the Data Supplement).

GWAS on Lp(a)

The GWAS identified 177 genome-wide significant SNPs in 39 loci associated with Lp(a) (Figure V in the Data Supplement; Table II in the Data Supplement). Notably, 37 of these have not been reported previously. Two variants in *CHKA* and *PEMT* were not associated with other

lipid traits (LDL-C, apolipoprotein-B, HDL [high-density lipoprotein] cholesterol, triglycerides, and cholesterol) at $P<0.05$ (Table III in the Data Supplement). We identified 2 sentinel SNPs with extreme P, namely the previously identified intronic rs10455872 ($P=4.3\times10^{-19,380}$) and rs1065853 ($P=1.3\times10^{-542}$), near the *APOE* gene. Regional association plots for each locus are presented in Figure VI in the Data Supplement and the QQ-plot in Figure VII in the Data Supplement. We replicated the novel variants in 7044 participants of the FINRISK study from Zekavat et al[4] and found that 2 SNPs were nominally significant (Table IV in the Data Supplement). The 39 top variants explained 24.9% of the phenotypic variance of Lp(a) level, the vast majority (24.4%) of which was attributable to rs10455872. Because of this, fine-mapping was performed solely for the *LPA* region centered around rs10455872.

Genetically Determined Lp(a) and CAD

The sentinel SNPs were tested for their association with CAD using a 2-sample MR approach in the CARDIoGRAMplusC4D data. In total, 38 variants or their LD buddies were available in CARDIoGRAMplusC4D (Table V in the Data Supplement). Heterogeneity and pleiotropy test results are presented in Table VI in the Data Supplement. In the univariable MR setting, there was no evidence for weak instrument bias and low chances of measurement error in MR-Egger. In the fixed-effects MR-IVW model, there was potential balanced horizontal pleiotropy based on the I^2 and significant Cochran Q but no evidence for unbalanced horizontal pleiotropy (MR-Egger intercept $P>0.05$). Therefore, despite the significant Q-Q′, the random-effects MR-IVW estimate was considered the causal estimate. Using this model, an SD increase in genetically determined Lp(a) was associated with a 1.42 ([95% CI, 1.26–1.59] $P=5.62\times10^{-9}$; Figure VIII in the Data Supplement) higher odds of CAD. MR-Steiger filtering, MR-PRESSO, and weighted and mode-based estimator MR analyses resulted in similar estimates (Table VII in the Data Supplement). Heterogeneity test results for the MVMR are presented in Table VIII in the Data Supplement. Q_{x1} and Q_{x2} were both higher than the critical value, indicating the SNPs could predict both Lp(a) and LDL-C. There was no evidence for unbalanced horizontal pleiotropy. The value for the adjusted Q statistic, however, was higher than the critical value. The null hypothesis that there is no heterogeneity could, therefore, not be rejected. Using the IVW method, each SD of genetically determined higher Lp(a) remained associated with a 1.30 ([95% CI, 1.20–1.41] $P=5.53\times10^{-11}$) higher odds of developing CAD, independently of the effects of LDL-C (Table IX in the Data Supplement; Figure IX in the Data Supplement). After filtering one outlier (rs3785549), MR-PRESSO indicated an odds ratio of 1.42 ([95% CI, 1.26–1.60] $P=1.03\times10^{-6}$).

Phenotypic Lp(a) and CAD

We tested the association between the measured Lp(a) concentrations with new-onset CAD in the UK Biobank. Of 356766 individuals with no history of CAD, 14710 individuals were diagnosed with CAD during a median (interquartile range) 8.1 (7.5–8.6) years of follow-up. There was a linear increase in risk of CAD per decile of Lp(a) when compared with people in the lowest decile, with a hazard ratio of 1.34 ([95% CI, 1.25–1.44] $P=3.64\times10^{-16}$) in the highest decile after adjusting for age, sex, and lipid-lowering drug usage at inclusion (Table X in the Data Supplement).

Genetic Fine-Mapping Analyses

Genetic fine-mapping analyses in the merged WES and genotype data were performed in a subset of 36773 individuals who were also included in the GWAS. The linear regression performed in the *LPA* region resulted in 3313 variants associated at a significant $P<6.9\times10^{-7}$ (Bonferroni-corrected $P=0.005/7283$ analyzed variants). In the merged data, rs10455872 remained the strongest associated variant ($\beta=1.44$; SE=0.01; $P=2.02\times10^{-2,290}$). Genetic fine-mapping using SuSiE yielded 30 credible sets (Table XI in the Data Supplement) of which 20 contained single SNPs and 9 contained <5 SNPs. Together, SuSiE and FINEMAP identified 47 variants, but only 15 variants were identified through both methods and were prioritized (Table XII in the Data Supplement). This included rs118039278 ($P_{GWAS}=1.1\times10^{-19,360}$), which is in perfect LD with rs10455872 in Europeans. The variants identified by SuSiE, FINEMAP, SuSiE and FINEMAP, and all 47 variants together explained, respectively, 51.4%, 49.6%, 46.6% and 52.1% of variance in Lp(a). Variant Effect Predictor annotation of the prioritized SNPs indicated 2 missense variants that were both reported as deleterious by SIFT and Condel and probably damaging by PolyPhen (Table XIII in the Data Supplement).

Protective Variant on Lp(a) Concentration

When plotting the Lp(a) distribution per Lp(a) increasing G allele of rs10455872, a small group of individuals (n=2314 [4.8%]) had low Lp(a) concentrations (median [interquartile range], 8.6 [5.4–13.6] nmol/L) despite having 1 or 2 Lp(a) increasing G alleles (Figure 2). To investigate whether the low Lp(a) concentrations observed in heterozygous carriers of rs10455872 were the result of protective effects of variants within or outside the *LPA* gene region, we compared the distribution of the weighted GRS for low (<25 nmol/L) and elevated (>50 nmol/L) concentrations using the 39 GWAS variants across the whole genome and the 15 fine-mapped variants in the *LPA* region. The distribution of the GRS for both the low and elevated concentrations deviated using the fine-mapped variants but not the GWAS variants.
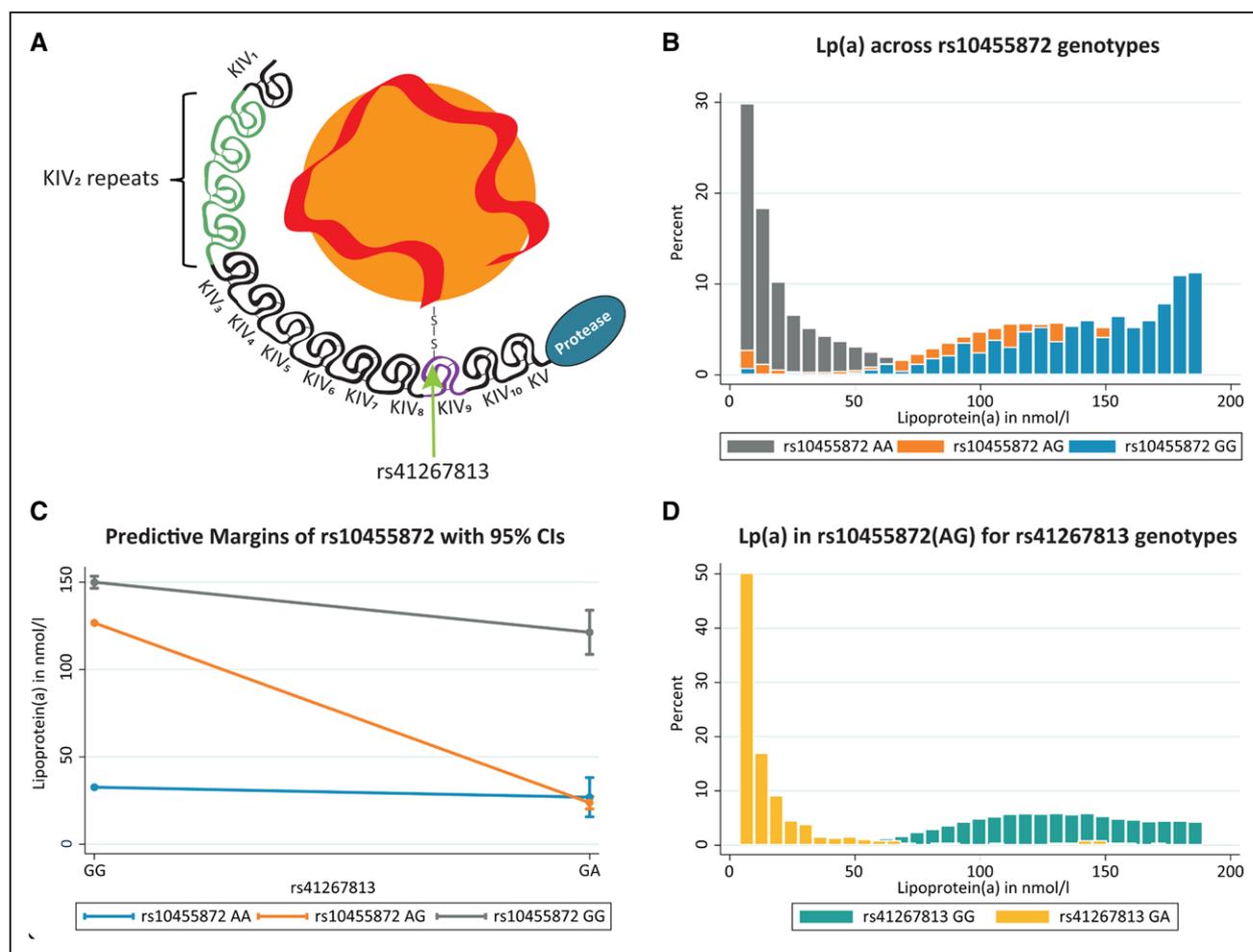
Figure 2. Missense variant rs41267813 in KIV9 (kringle IV subtype 9).
A, The schematic position of rs41267813 in KIV$_9$ in the apolipoprotein(a) tail of lipoprotein(a) (Lp[a]). **B**, Distribution of Lp(a) for different genotypes of rs10455872. **C**, Interaction between rs10455872 and rs41367813. No individuals with 2 A alleles of rs41267813 were available. **D**, Distribution of Lp(a) among heterozygous carriers of rs10455872 for carriers and noncarriers of the rs41267813 A allele.

This suggests that the protective variant resides in the *LPA* region. Hence, we prioritized the fine-mapped variants with functional consequences. Because of the small percentage of individuals with low Lp(a) values despite the G allele(s) of rs10455872, we disregarded SNPs with minor allele frequencies >5%. Two missense variants, rs41267807 and rs41267813, remained, and their effects on Lp(a) independently of rs10455872 were estimated (Table XIV in the Data Supplement). No interaction existed between rs41267807 with rs10455872, and there was only a minor shift toward lower values in people with 1 or 2 Lp(a)-lowering C alleles of rs41267807. Much stronger effects were observed for rs41267813. Not only was the minor allele (A) of this variant associated with a strong Lp(a)-lowering effect but there was also a strong interaction between rs41267813 and rs10455872 (P=3.75×10^{-19}; Figure 2). Among individuals who were heterozygous for rs10455872, as well as for rs41267813, the median Lp(a) concentration (9.44 [5.41−19.41] nmol/L) was over 13× lower compared with individuals

with rs41267813(GG) (median, 127.94 [100.7−155.15] nmol/L; Figure 2). Median Lp(a) concentrations for each genotype of rs10455872 and rs41267813 are provided in Table XV in the Data Supplement. When the model with all 15 overlapping fine-mapped SNPs was refitted using a multivariable regression model, each A allele of rs41267813 had a β of −1.99 (SE, 0.023; P=3.66×10^{-48}), which was larger than the effect of rs118039278 (β=1.38 [SE=0.005]; P=1.41×10^{-82}; Table XVI in the Data Supplement), which is in almost perfect LD with rs10455872. Among participants heterozygous for rs10455872, LDL-C values were comparable in noncarriers of rs41267813 (mean, 3.57 [SD=0.85]) and heterozygous carriers of rs41267813 (mean, 3.60 [SD=0.86]). The proportion of individuals in the UK Biobank heterozygous for rs10455872 with CAD in their history or during follow-up was lower among individuals with the missense variant compared with those without (6.5% versus 9.1%; 1-sided Fisher exact, P=0.03). However, we did not find evidence for an interaction between

rs10455872 and rs41267813 on CAD risk in logistic regression analyses in the UK Biobank. When looking primarily at the effect of rs41267813(A) and not taking into account the interaction with the rs10455872 genotype, individuals with rs41267813(A) had a lower prevalence of CAD compared with noncarriers of the A allele, although this difference was not statistically significant (6.7% versus 7.9%; 1-sided Fisher exact P=0.16). Unfortunately, rs41267813 was not available in the CARDIoGRAMplusC4D cohort and could, therefore, not be looked up to assess the effect on CAD. In our sample, no individual had 2 A alleles of rs41267813. Among individuals with no A alleles of rs41267813 but with 1 or 2 G alleles of rs10455872, a total of 1072 (2.4%) remained with Lp(a) concentrations <25 nmol/L.

DISCUSSION

We explored the genetic architecture of Lp(a) in over 370 000 individuals through GWAS. Genetic fine-mapping analyses identified a rare missense variant in the *LPA* locus with protective effects in individuals with Lp(a) increasing G alleles of the well-established Lp(a) SNP rs10455872.[2] We provide a novel line of evidence supporting an LDL-C independent causal link between Lp(a) biology and the development of CAD.

Comparison With Previous Studies

We identified 39 variants that were strongly associated with Lp(a) across the genome. Notably, the majority of these variants (37 of 39) had not been reported previously; only 1 variant in the *LPA* locus and 1 variant in proximity of the *APOE* locus have been published before.[2,4,6] Two of the novel variants are located in *CHKA* and *PEMT* (phosphatidylethanolamine N-methyltransferase), which both play important roles in phospholipid biosynthesis pathways, were not associated with other lipid traits in the UK Biobank. *CHKA* plays an important role in the cytidine diphosphate-choline pathway, which is the major pathway for the biosynthesis of phosphatidylcholine.[34] Phosphatidylcholine is a major membrane phospholipid of all lipoproteins and plays essential roles in membrane structure and permeability. *PEMT* is responsible for the alternative pathway (PEMT pathway) for phosphatidylcholine biosynthesis in the liver and contributes to ≈30% of phosphatidylcholine biosynthesis.[34] Neither gene has been previously reported in GWAS on Lp(a). *PEMT*, next to *APOH*, *PGS1*, *APOE*, *PGS1*, *LDLR* (LDL receptor), *PCSK9* (proprotein convertase subtilisin-kexin type 9), *APOB*, *ABCA6*, *PPP1R3B*, and *LPA*, has, however, been associated to LDL-C.[35,36] Of these, *LDLR* and *PCSK9* have a role in the reduction of Lp(a).[37] PCSK9 plays a role in the modulation of Lp(a) concentrations, and its inhibition leads to a reduction in Lp(a) concentrations. This is likely due to an increased expression

of LDLR, allowing more internalization of LDL particles. In turn, more LDLRs become available and with higher affinity ligands that can more easily bind Lp(a) particles. This is important, as Lp(a) particles have a lower affinity to LDLR compared with LDL particles.[37] One previous GWAS on Lp(a) reported a variant on chromosome 11 in the *APOC3* locus,[38] which was not available in the UK Biobank and could, therefore, not be replicated. Substantial advantages of the present study are the much larger sample size in comparison with previous GWASs on Lp(a)[2,4,6] and the utilization of a single sample study design that overcomes potential drawbacks of meta-analyses in terms of power and heterogeneity among studies. We found that the variants outside the LPA locus had a smaller effect on the variance in Lp(a) levels, which, for the vast majority, was accounted for by rs10455872. Fine-mapping around rs10455872 resulted in credible sets, which explained 46.6% to 52.1% of variance in Lp(a), which is comparable to previous estimates.[3,6]

Since evidence of the association between Lp(a) and CAD has been described previously in multiple studies,[2,4,6] we proceeded sequentially to validate prior findings and further disentangle the relationships between Lp(a), LDL-C, and CAD. We found a similar estimate as in previous studies, with each SD increase (corresponding to an 83.83 nmol/L change in Lp[a]) in genetically determined Lp(a) translating in a 42% ([95% CI, 26%–59%] P=5.6×10⁻⁹) increased risk of CAD. A previous MR study found that the causal estimate of Lp(a) and CAD was independent of LDL-C levels using rs12916 to mimic the effect of statins.[3] However, this method does not allow for investigation of the direct effect of Lp(a) on CAD (ie, the effect not driven by LDL) and differentiation between different scenarios encountered within epidemiological studies (confounding, collider, pleiotropy, and mediation). Here, using an MVMR approach, we found that Lp(a) was robustly associated with CAD risk (odds ratio, 1.30 [95% CI, 1.20–1.41]; P=5.5×10⁻¹¹) independently from LDL-C. The present study provides strong supportive evidence for a direct causal role of Lp(a) in the development of CAD.

Protective Missense Variant in the *LPA* Gene

Although rs10455872 is the major genetic determinant for Lp(a) concentrations (explaining 24.4% of its variance), we found individuals that carried the increasing allele still have low Lp(a) concentrations, not consistent with an additive model. We found that rs41267813—one of the fine-mapped candidate causal variants—could explain this phenomenon and showed strong interaction effects with rs10455872 on Lp(a) concentrations. rs41267813 causes a change from a histidine residue to tyrosine on exon 28. Exons 28 and 29 together form KIV₉ of the apo(a) tail, which contains an extra unpaired cysteine residue responsible for the linkage between apo(a) and the apolipoprotein-B(100) via a disulfide

bond.[39] rs41267813 was significantly more associated with lower Lp(a) values among individuals carrying at least 1 G allele of rs10455872. This effect-modifying SNP was previously reported to be associated with Lp(a) in an independent cohort (n=48333) but not further investigated.[3] A separate study also reported the association between rs41267813(A) and lower LDL-C levels but could not confirm the association with Lp(a) concentration as this measure was unavailable.[40] We could not test the association between rs41267813 and Lp(a) isoform size, as this was not measured in the UK Biobank. A potential explanation for the current findings is that individuals with the missense variant and rs10455872(G) have large isoform sizes and, therefore, low Lp(a) concentrations.[41] This should be studied in a sufficiently large separate cohort with rs41267813 and isoform size data available. The deep dive of this variant in the present study, however, highlights it as a potential protective variant that may be of special interest to therapeutic developments aimed at lowering Lp(a) concentrations.

Clinical Perspectives

This study provides further evidence for the causal association between Lp(a) and CAD. Screening for patients with high Lp(a) values is not a common practice, as Lp(a) is relatively refractory to both lifestyle and drug interventions. High Lp(a) values may, however, identify high-risk individuals that could benefit from early treatment. Future therapies include antisense oligonucleotide therapy, which shows promise in clinical trials.[42] This study further highlights the importance of finding Lp(a)-lowering therapies. The missense variant reported in this study may be a potential drug target.

Strengths and Limitations

Major strengths of the present study are the large sample size of the UK Biobank, fine-mapping of the *LPA* gene region using 2 Bayesian fine-mapping approaches, identification of a protective missense variant, and MVMR strategy to investigate causal links between Lp(a) with CAD in an independent cohort with over 60000 cases and 120000 controls. Bayesian fine-mapping approaches are superior to conditional analyses used in previous reports and simulations,[3,4,6,43] as the latter fail to provide probabilistic measures of causality for variants. There are also limitations. We found evidence for potential heterogeneity in the MR analyses, meaning pleiotropy cannot be ruled out. However, we, therefore, provided a framework for the MR analyses to report the correct estimate per degree of pleiotropy. In addition, sensitivity analyses showed consistent results with respect to the main analyses. We could not analyze the KIV_2 copy number nor test variants in the KIV_2 region as these data were not available in the UK Biobank. We,

however, aimed to provide some insight into the association between rs41267813 and the KIV_2 copy number using rs10455872, which has been reported to be tagging the number of KIV_2 copies.[2]

Conclusions

In conclusion, this study determined genetic variants associated with Lp(a) and found additional strong support for a causal link with CAD, independent of LDL-C. Furthermore, we identified a novel rare missense variant, rs41267813(A), in the *LPA* gene locus with protective effects in people with Lp(a) increasing G alleles of rs10455872. In the continuing search for therapeutic targets of Lp(a), this missense variant may be of interest.

REFERENCES

1. Emerging Risk Factors Collaboration; Erqou S, Kaptoge S, Perry PL, Di Angelantonio E, Thompson A, White IR, Marcovina SM, Collins R, Thompson SG, Danesh J. Lipoprotein(a) concentration and the risk of coronary heart disease, stroke, and nonvascular mortality. *JAMA*. 2009;302:412–423. doi: 10.1001/jama.2009.1063

2. Clarke R, Peden JF, Hopewell JC, Kyriakou T, Goel A, Heath SC, Parish S, Barlera S, Franzosi MG, Rust S, et al; PROCARDIS Consortium. Genetic variants associated with Lp(a) lipoprotein level and coronary disease. *N Engl J Med*. 2009;361:2518–2528. doi: 10.1056/NEJMoa0902604

3. Burgess S, Ference BA, Staley JR, Freitag DF, Mason AM, Nielsen SF, Willeit P, Young R, Surendran P, Karthikeyan S, et al; European Prospective Investigation Into Cancer and Nutrition–Cardiovascular Disease (EPIC-CVD) Consortium. Association of LPA variants with risk of coronary disease and the implications for lipoprotein(a)-lowering therapies: a

mendelian randomization analysis. *JAMA Cardiol.* 2018;3:619–627. doi: 10.1001/jamacardio.2018.1470

4. Zekavat SM, Ruotsalainen S, Handsaker RE, Alver M, Bloom J, Poterba T, Seed C, Ernst J, Chaffin M, Engreitz J, et al; NHLBI TOPMed Lipids Working Group. Deep coverage whole genome sequences and plasma lipoprotein(a) in individuals of European and African ancestries. *Nat Commun.* 2018;9:2606. doi: 10.1038/s41467-018-04668-w

5. Saleheen D, Haycock PC, Zhao W, Rasheed A, Taleb A, Imran A, Abbas S, Majeed F, Akhtar S, Qamar N, et al. Apolipoprotein(a) isoform size, lipoprotein(a) concentration, and coronary artery disease: a mendelian randomisation analysis. *Lancet Diabetes Endocrinol.* 2017;5:524–533. doi: 10.1016/S2213-8587(17)30088-8

6. Mack S, Coassin S, Rueedi R, Yousri NA, Seppälä I, Gieger C, Schönherr S, Forer L, Erhart G, Marques-Vidal P, et al; KORA-Study Group. A genome-wide association meta-analysis on lipoprotein (a) concentrations adjusted for apolipoprotein (a) isoforms. *J Lipid Res.* 2017;58:1834–1844. doi: 10.1194/jlr.M076232

7. Davies NM, Holmes MV, Davey Smith G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *BMJ.* 2018;362:k601. doi: 10.1136/bmj.k601

8. Tsimikas S. A test in context: lipoprotein(a): diagnosis, prognosis, controversies, and emerging therapies. *J Am Coll Cardiol.* 2017;69:692–711. doi: 10.1016/j.jacc.2016.11.042

9. Kronenberg F, Utermann G. Lipoprotein(a): resurrected by genetics. *J Intern Med.* 2013;273:6–30. doi: 10.1111/j.1365-2796.2012.02592.x

10. Li J, Lange LA, Sabourin J, Duan Q, Valdar W, Willis MS, Li Y, Wilson JG, Lange EM. Genome- and exome-wide association study of serum lipoprotein (a) in the Jackson Heart Study. *J Hum Genet.* 2015;60:755–761. doi: 10.1038/jhg.2015.107

11. Wei WQ, Li X, Feng Q, Kubo M, Kullo IJ, Peissig PL, Karlson EW, Jarvik GP, Lee MTM, Shang N, et al. LPA variants are associated with residual cardiovascular risk in patients receiving statins. *Circulation.* 2018;138:1839–1849. doi: 10.1161/CIRCULATIONAHA.117.031356

12. Dullaart RPF. Lipoprotein(a): the renaissance of an enigmatic lipoprotein. *J Clin Endocrinol Metab.* 2020;105:dgz111. doi: 10.1210/clinem/dgz111

13. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* 2015;12:e1001779. doi: 10.1371/journal.pmed.1001779

14. UK Biobank. UK Biobank: protocol for a large-scale prospective epidemiological resource. Updated 2007. Accessed December 15, 2015. https://www.ukbiobank.ac.uk/media/gnkeyh2q/study-rationale.pdf.

15. UK Biobank. UK biobank ethics and governance framework. Updated 2007. Accessed December 15, 2015. https://www.ukbiobank.ac.uk/media/0xsbmfmw/egf.pdf.

16. Van Hout C, Tachmazidou I, Backman J, et al. Whole exome sequencing and characterization of coding variation in 49,960 individuals in the UK biobank. *Nature.* 2020;586:749–756.

17. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature.* 2018;562:203–209. doi: 10.1038/s41586-018-0579-z

18. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. The ensembl variant effect predictor. *Genome Biol.* 2016;17:122. doi: 10.1186/s13059-016-0974-4

19. Loh PR, Tucker G, Bulik-Sullivan BK, Vilhjálmsson BJ, Finucane HK, Salem RM, Chasman DI, Ridker PM, Neale BM, Berger B, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet.* 2015;47:284–290. doi: 10.1038/ng.3190

20. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F, et al. The UCSC genome browser database: update 2006. *Nucleic Acids Res.* 2006;34(Database issue): D590–D598. doi: 10.1093/nar/gkj144

21. Wang G, Sarkar A, Carbonetto P, Stephens M. A simple new approach to variable selection in regression, with application to genetic fine-mapping. BioRxiv. 2018.

22. Benner C, Spencer CC, Havulinna AS, Salomaa V, Ripatti S, Pirinen M. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics.* 2016;32:1493–1501. doi: 10.1093/bioinformatics/btw018

23. Benner C, Havulinna AS, Järvelin MR, Salomaa V, Ripatti S, Pirinen M. Prospects of fine-mapping trait-associated genomic regions by using summary statistics from genome-wide association studies. *Am J Hum Genet.* 2017;101:539–551. doi: 10.1016/j.ajhg.2017.08.012

24. Nikpay M, Goel A, Won HH, Hall LM, Willenborg C, Kanoni S, Saleheen D, Kyriakou T, Nelson CP, Hopewell JC, 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

25. Palmer TM, Lawlor DA, Harbord RM, Sheehan NA, Tobias JH, Timpson NJ, Davey Smith G, Sterne JA. Using multiple genetic variants as instrumental variables for modifiable risk factors. *Stat Methods Med Res.* 2012;21:223–242. doi: 10.1177/0962280210394459

26. Bowden J, Del Greco M F, Minelli C, Davey Smith G, Sheehan NA, Thompson JR. Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I2 statistic. *Int J Epidemiol.* 2016;45:1961–1974. doi: 10.1093/ije/dyw220

27. Greco M FD, Minelli C, Sheehan NA, Thompson JR. Detecting pleiotropy in Mendelian randomisation studies with summary data and a continuous outcome. *Stat Med.* 2015;34:2926–2940. doi: 10.1002/sim.6522

28. Bowden J, Del Greco M F, Minelli C, Davey Smith G, Sheehan N, Thompson J. A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Stat Med.* 2017;36:1783–1802. doi: 10.1002/sim.7221

29. Hemani G, Bowden J, Davey Smith G. Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Hum Mol Genet.* 2018;27(R2):R195–R208. doi: 10.1093/hmg/ddy163

30. Verbanck M, Chen CY, Neale B, Do R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat Genet.* 2018;50:693–698. doi: 10.1038/s41588-018-0099-7

31. Hartwig FP, Davey Smith G, Bowden J. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *Int J Epidemiol.* 2017;46:1985–1998. doi: 10.1093/ije/dyx102

32. Sanderson E, Davey Smith G, Windmeijer F, Bowden J. An examination of multivariable Mendelian randomization in the single-sample and two-sample summary data settings. *Int J Epidemiol.* 2019;48:713–727. doi: 10.1093/ije/dyy262

33. Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, Berk R, Bollen KA, Brembs B, Brown L, Camerer C, et al. Redefine statistical significance. *Nat Hum Behav.* 2018;2:6–10. doi: 10.1038/s41562-017-0189-z

34. Vance DE. Phospholipid methylation in mammals: from biochemistry to physiological function. *Biochim Biophys Acta.* 2014;1838:1477–1487. doi: 10.1016/j.bbamem.2013.10.018

35. Klarin D, Damrauer SM, Cho K, Sun YV, Teslovich TM, Honerlaw J, Gagnon DR, DuVall SL, Li J, Peloso GM, et al; Global Lipids Genetics Consortium; Myocardial Infarction Genetics (MIGen) Consortium; Geisinger-Regeneron DiscovEHR Collaboration; VA Million Veteran Program. Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. *Nat Genet.* 2018;50:1514–1523. doi: 10.1038/s41588-018-0222-9

36. van Leeuwen EM, Karssen LC, Deelen J, Isaacs A, Medina-Gomez C, Mbarek H, Kanterakis A, Trompet S, Postmus I, Verweij N, et al; Genome of the Netherlands Consortium. Genome of the Netherlands population-specific imputations identify an ABCA6 variant associated with cholesterol levels. *Nat Commun.* 2015;6:6065. doi: 10.1038/ncomms7065

37. Pirillo A, Catapano AL. PCSK9 inhibition and Lp(a) reduction: another piece of the puzzle? *Eur Heart J.* 2018;39:2586–2588. doi: 10.1093/eurheartj/ehy311

38. Lu W, Cheng YC, Chen K, Wang H, Gerhard GS, Still CD, Chu X, Yang R, Parihar A, O'Connell JR, et al. Evidence for several independent genetic variants affecting lipoprotein (a) cholesterol levels. *Hum Mol Genet.* 2015;24:2390–2400. doi: 10.1093/hmg/ddu731

39. Schmidt K, Noureen A, Kronenberg F, Utermann G. Structure, function, and genetics of lipoprotein (a). *J Lipid Res.* 2016;57:1339–1359. doi: 10.1194/jlr.R067314

40. Wang Z, Chen H, Bartz TM, Bielak LF, Chasman DI, Feitosa MF, Franceschini N, Guo X, Lim E, Noordam R, et al; CHARGE Gene-Lifestyle Interactions Working Group. Role of rare and low-frequency variants in gene-alcohol interactions on plasma lipid levels. *Circ Genom Precis Med.* 2020;13:e002772. doi: 10.1161/CIRCGEN.119.002772

41. Kronenberg F. Prediction of cardiovascular risk by Lp(a) concentrations or genetic variants within the LPA gene region. *Clin Res Cardiol Suppl.* 2019;14(suppl 1):5–12. doi: 10.1007/s11789-019-00093-5

42. Zabaneh D, Kumari M, Sandhu M, Wareham N, Wainwright N, Papamarkou T, Hopewell J, Clarke R, Li K, Palmen J, et al. Meta analysis of candidate gene variants outside the LPA locus with Lp(a) plasma levels in 14,500 participants of six White European cohorts. *Atherosclerosis.* 2011;217:447–451. doi: 10.1016/j.atherosclerosis.2011.04.015

43. Schaid DJ, Chen W, Larson NB. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet.* 2018;19:491–504. doi: 10.1038/s41576-018-0016-z