Whole genome sequence analysis of blood lipid levels in >66,000 individuals

Blood lipids are heritable modifiable causal factors for coronary artery disease. Despite well-described monogenic and polygenic bases of dyslipidemia, limitations remain in discovery of lipid-associated alleles using whole genome sequencing (WGS), partly due to limited sample sizes, ancestral diversity, and interpretation of clinical significance. Among 66,329 ancestrally diverse (56% non-European) participants, we associate 428M variants from deep-coverage WGS with lipid levels; ~400M variants were not assessed in prior lipids genetic analyses. We find multiple lipid-related genes strongly associated with blood lipids through analysis of common and rare coding variants. We discover several associated rare non-coding variants, largely at Mendelian lipid genes. Notably, we observe rare \(LDLR\) intronic variants associated with markedly increased LDL-C, similar to rare \(LDLR\) exonic variants. In conclusion, we conducted a systematic whole genome scan for blood lipids expanding the alleles linked to lipids for multiple ancestries and characterize a clinically-relevant rare non-coding variant model for lipids.

The discovery of rare alleles linked to plasma lipids (i.e., low-density lipoprotein cholesterol \([LDL-C]\), high-density lipoprotein cholesterol \([HDL-C]\), total cholesterol \([TC]\), and triglycerides \([TG]\)) continue to yield important translational insights toward coronary artery disease (CAD), including \(PCSK9\) and \(ANGPTL3\) inhibitors now available in clinical practice\(^1\)–\(^5\). The monogenic and polygenic bases of plasma lipids are well-suited to population-based discovery analyses and confer broader insights for genetic analyses of complex traits. We now evaluate numerous newly catalogued, largely rare, alleles never previously systematically analyzed with lipids.

Analyses of imputed array-derived genome-wide genotypes and whole exome sequences in hundreds of thousands of increasingly diverse individuals continue to uncover low-frequency protein-coding variants linked to lipids. Due to purifying selection, causal variants conferring large effects tend to occur relatively more recently, and are thus rare and often specific to families or communities\(^6\). Most discovery analyses for large-effect rare alleles have focused on the analysis of disruptive protein-coding variants given (1) well-recognized constraint in coding regions, (2) incomplete genotyping of rare non-coding sequence given relative sparsity of deep-coverage (i.e., \(>30X\)) whole genome sequencing (WGS), and (3) better prediction of coding versus non-coding sequence variation consequence\(^1\)–\(^12\). We recently described a statistical framework incorporating multi-dimensional reference datasets paired with genomic data to improve rare coding and non-coding variant analyses for WGS analysis of lipids and other complex traits\(^13\)–\(^14\). Furthermore, including individuals of non-European ancestry facilitates the discovery of both novel alleles at established loci as well as novel loci\(^15\)–\(^16\).

Here, we examine the full allelic spectrum with plasma lipids using whole genome sequences and harmonized lipids from the National Heart, Lung, and Blood Institute (NHLBI) Trans-Omics for Precision Medicine (TOPMed) program\(^17\)–\(^18\). We studied 66,329 participants and 428 million variants across multiple ancestry groups—44.48% European, 25.60% Black, 21.02% Hispanic, 7.11% Asian, and 1.78% Samoan. We identified robust allelic heterogeneity at known loci with several novel variants at these loci; we additionally identified novel loci and pursued replication in independent cohorts. We then explored the association of genome-wide rare variants with lipids, with detailed explorations of rare coding and non-coding variant models at known Mendelian dyslipidemia genes. Our systemic effort yields new insights for plasma lipids and provides a framework for population-based WGS analysis of complex traits.
Results
Overview
We studied the TOPMed Freeze8 dataset of 66,329 samples from 21 studies and performed genome-wide association studies (GWAS) separately for the four plasma lipid phenotypes (i.e., LDL-C, HDL-C, TC, and TG) using 28 M individual autosomal variants (minor allele count [MAC] >20) and aggregated rare autosomal variant (minor allele frequency [MAF] <1%) association testing for 417 M variants (Fig. 1, Supplementary Fig. 1). Secondarily, we associated individual variants with minor allele frequencies (MAF) >0.01% within each ancestry group to detect ancestry-specific lipid-associated alleles. We intersected our results with currently published array-based GWAS results15 to identify novel associations with lipids. We performed replication analyses for the putative novel associations identified, in up to ~45,000 independent samples with array-based genotyping imputed to TOPMed and 400 K samples from UK Biobank (UKB) imputed genotypes. Finally, we conducted rare variant association studies as multiple aggregate tests across the genome to identify gene-specific functional categories and non-coding genomic regions influencing plasma lipid concentrations. We replicated the significant rare variant aggregates in ~130 K whole genomes from UKB.

TOPMed baseline characteristics
The TOPMed Informatics Research Center (IRC) and TOPMed Data Coordinating Center (DCC) performed quality control, variant calling, and calculated the relatedness of population structures of Freeze 8 data17. We studied 66,329 samples across 21 cohorts, and 41,182 (62%) were female. The ancestry distribution was 29,502 (44.46%) White, 16,983 (25.60%) Black, 13,943 (21.02%) Hispanic, 4719 (7.11%) Asian, and 1182 (1.78%) Samoan (Supplementary Data 1). The mean (standard deviation [SD]) age of the full cohort was 53 (15.00) years which varied by cohort from 25 (3.56) years for Coronary Artery Risk Development in Young Adults (CARDIA) to 73 (5.38) years for Cardiovascular Health Study (CHS). The Amish cohort had a higher-than-average concentration of LDL-C (140 [SD 43] mg/dL) and HDL-C (56 [SD 16] mg/dL) as well as lower TG (median 63 [IQR 50] mg/dL) consistent with the known founder mutations in APOB and APOC37,8,14. In the Women's Health Initiative (WHI) cohort, the TC (230 [SD 41] mg/dL) and TG (median 129 [IQR 87] mg/dL) concentrations were higher than for other cohorts as previously described12. We accounted for lipid-lowering medications and fasting status and inverse rank normalized the phenotypes as before12,14 which are further detailed in the Methods. The adjusted normalized lipid concentrations for the four lipids were similar across the cohorts.

![Fig. 1 | Overall study schematic.](image-url)

The analyses were conducted using the multi-ancestral TOPMed freeze8 data to associate whole genome sequence variation with lipid phenotypes (i.e., LDL-C, HDL-C, TC, and TG). A total of 66,329 samples with lipids quantified data from five ancestry groups were analyzed. Single variant GWAS were carried out using SAIGE on the Encore platform using SNPs with MAC >20. Both trans-ancestry and ancestry-specific GWAS were conducted. Genome-wide rare variant (MAF <1%) gene-centric and region-based aggregate tests were grouped and analyzed using STAAR pipeline. Finally, single variant and rare variant associations at Mendelian dyslipidemia genes were investigated in further detail. TOPMed Trans-Omics for Precision Medicine, HDL-C high-density lipoprotein cholesterol, LDL-C low-density lipoprotein cholesterol, TC total cholesterol, TG triglycerides, GWAS genome wide association study, SAIGE Scalable and Accurate Implementation of GEneralized mixed model, MAC minor allele count, MAF minor allele frequency, SNPs single nucleotide polymorphisms.
A total of 428 M variants passed the quality criteria with an average depth >30X in 22 autosomes. 202 M variants were singletons, 417 M were rare variants (MAF <1%), and 11 M were common or low-frequency variants (MAF >1%) with differences by cohort (Supplementary Data 2).

**Individual variant associations with lipids**

We performed single variant analysis of ~28 M variants with a MAC > 20 for four lipid phenotypes. We identified significant genomic risk loci for each lipid level (Supplementary Data 3) and considered a p-value <5 x 10^-9 to claim significance as previously recommended for whole genome sequencing common variant association studies\(^1^4,1^9\). The total numbers of variants that met our significance threshold were 2214, 2314, 2697, and 2442 for LDL-C, HDL-C, TC and TG, respectively, and after clumping\(^2^0\) the numbers of variants were 357, 338, 324, and 289, respectively. Of these variants, 99% were previously demonstrated to be associated with plasma lipids either at the variant- or locus-level\(^1^5\) (Supplementary Data 4, Supplementary Fig. 2).

To identify putative novel variant associations, we compared our results to a recent multi-ethnic lipid GWAS among 312,571 participants of the Million Veteran Program (MVP)\(^1^5\) as well as the GWAS Catalog (All associations(v1.0) file dated 06/04/2020) (Fig. 2). We clumped (window 250 kb, \(r^2\) 0.5) significant variants using Plink\(^2^0\) and queried these in the GWAS Catalog and MVP. Among genome-wide significant variants, we tabulated ‘known-position’ (variant previously associated), ‘known-loci’ (variants not previously significantly associated with the corresponding lipid phenotype but within 500 kb of a known locus, thereby representing additional allelic heterogeneity), and ‘novel’ variants (variants not in a known lipid locus) (Supplementary Data 4).

The novel variants, tabulated in Table 1, are divided into two subsets – ‘novel variants’ or variants at established lipid loci for another lipid phenotype, and ‘novel loci,’ representing new loci associations for any lipid phenotype. For example, the CETP locus is well-known for its link to HDL-C, but we now found that rs183130 (16:56957451:C:T, MAF 28.3%) at the locus is associated with LDL-C. Similarly, the variants rs7140110 (13:113841051:T:C, MAF 27.8%) \(*GAS6*\) and rs73729083 (7:137875053:T:C, MAF 4.5%) \(*CREB3L2*\) are newly associated with TC, while previous studies showed that rs73729083 associates with LDL-C\(^2^1\) and rs7140110 associates with LDL-C\(^2^2\) and TG\(^2^3\). Index variants at novel loci were typically low-frequency variants often observed in non-European ancestries, so we also conducted ancestry-specific association analyses for these alleles (Supplementary Data 5). For example, 12q23.1 (12:97352354:T:C, MAF 0.3%) and 4q34.2 (4:176382171:C:T, MAF 0.2%) associations with LDL-C are specific to Hispanic (MAF 1.3%) and Black (MAF 0.6%) populations, respectively and among Asians (MAF 1.5%) alone, 11q13.3 (11:69219641:C:T, MAF 0.2%) was associated with TG. One variant initially passing the novel locus filter for HDL-C \(*RNF111*\)
In-silico analysis to gain mechanic insights from single variant GWAS results

Prioritization and functional enrichment analysis. We first mapped the variants to genes and to functional regions using ANNOVAR. Second, we determined gene tissue specificity, relating tissue-specific gene expression with disease-gene associations, using MAGMA. Significantly associated variants were enriched in intronic and intergenic regions (Supplementary Fig. 3). Using GTEx, tissue-specific gene expression was enriched among liver, stomach, and pancreatic tissues (Supplementary Fig. 4) with top tissue-gene sets tabulated in Supplementary Data 8. Using the STRING protein-protein interaction database examining liver-specific genes, we highlight that the HDL-C protein network uniquely harbored metal-ions related genes (MTIA, MTIB, MIFH, MT1G, MT1H) and anticipated LCAT-CEPT interactions (Supplementary Fig. 5). Enriched pathways from Reactome, GeneOntology and other curated and canonical pathways (Supplementary Data 9) with a p-value < 2.5 × 10^-6 were observed including response to metal ions, lipoprotein assembly, and chylomicron remodeling.

The enrichment analysis was carried out with the full single variant summary statistics, where we identified that most of the prioritized loci genes were previously documented for lipid associations. Next, we specifically investigated the novel variants that we identified from this study. Out of the seven variants documented in Table 1, four were low frequency variants, 12973s2354:T:C (rs89010847) closest to NEDD1, 14763s2171:C:T (rs15489644) closest to SPG5, 1L69219461:C:T (rs74971571) near to SMIM38, are all intergenic variants and 131075s1611:C:T (rs77867061) is an intronic variant in FAM115A. We did not find any information for these variants in the Open Target Genetics database. Finally, two of the common novel-loci variants (rs183130 and rs7140110) were present in eQTL and sQTL databases, therefore, we performed analysis to determine the correlation among effects and the importance of these variants more in detail.

CETP locus, HDL-C, and LDL-C. CETP is a well-recognized Mendelian HDL-C gene and the locus was previously known to be significantly associated with HDL-C, TC, and TG at genome-wide significance. Pharmacologic CETP inhibitors have shown strong associations with increased HDL-C but mixed effects for LDL-C reduction in clinical trials.

We found that the CETP locus variant rs183130 in chr16:56957451:C:T, MAF 28.3%, intergenic variant was associated
with reduced LDL-C concentration (beta = −1.568 mg/dL, SE = 0.264, p-value = 2.88 × 10^{−09}). The lead HDL-C-associated variant at the locus, rs3764261 (chr16:56959412:C:A, MAF 30.3%), was associated with 3.5 mg/dL increased HDL-C (p-value = 8.03 × 10^{−284}) as well. Among the ancestry groups analyzed, rs183130 was most significantly associated with LDL-C among those of African ancestry (beta = −2.762 mg/dL, p-value = 5.71 × 10^{−05}) (Supplementary Data 10). We next investigated variants by their HDL-C and LDL-C effects within this locus (+/−500 kb of rs183130 and rs3764261) (Fig. 3). We identified five variants showing at least suggestive (p-value < 5 × 10^{−05}) association with both HDL-C and LDL-C. Though variants with strong LD (linkage disequilibrium) existed, ancestry-specific analyses showed that the stronger LDL-C effects were among those of African ancestry.

To better understand the mechanisms for HDL-C and LDL-C effects at the CETP locus, we pursued colocalization with eQTLs from three tissues (Liver, Adipose Subcutaneous and Adipose Visceral [Omentum]) from GTEx^9. We analyzed 5 LDL-C and 441 HDL-C associated (p-values < 5 × 10^{−05}) variants. We correlated eQTL effect estimates for genes at the locus with lipid outcome effect estimates. Indeed, CETP gene expression effects were strongly negatively correlated with HDL-C effects (Liver: p = 0.933, p-value 4.01 × 10^{−17}; Adipose Subcutaneous: p = 0.762, p-value 8.87 × 10^{−12}; Adipose Visceral: p = 0.739, p-value 5.52 × 10^{−09}) (Supplementary Fig. 6). However, CETP expression effects were not significantly correlated with LDL-C (Liver: p = 0.007, p-value 0.99; Adipose Subcutaneous: p = 0.344, p-value 0.57; Adipose Visceral: p = 0.59, p-value 0.29). Given the possibility that the observed lack of correlation for LDL-C could be due to reduced power from a limited number of variants attaining a suggestive p-value (<5 × 10^{−05}), we repeated the analysis with a subset of 122 nominally significant (p-value < 0.05) LDL-C associated variants in this locus. Indeed, CETP gene expression effects were strongly positively correlated with LDL-C effects (Liver: p = 0.957, p-value 2.28 × 10^{−06}; Adipose Subcutaneous: p = 0.922, p-value 1.34 × 10^{−10}; Adipose Visceral: p = 0.868, p-value 6.09 × 10^{−06}).

GAS6 locus, LDL-C/TG, and TC. Variants at GAS6 were previously associated with LDL-C and TG^22,23, but in our analysis, rs7140110 was now significantly associated with TC. We performed colocalization analysis of the variants+/−500 kb from rs7140110 in liver and adipose tissues from GTEx. Across the three lipid-related tissues (liver, adipose subcutaneous, and adipose visceral), strong colocalization was observed in liver for all three lipid phenotypes (TG 46.6%; LDL-C 33.3%; TC 28%). The TG and LDL-C-associated variants were eQTLs for the GAS6 gene only. However, the TC-associated eQTLs at this locus influenced the cis expression of multiple genes, including GAS6, anti-sense genes of GAS6 (AS1, AS2) as well as other genes (i.e., TDFPI, CHAMP1, LINCO00565, ADPRHL1, RASA3, UPE3A, GRTPI, AL442125.1, C13orf46, DCUNID2, CDC16, TMEM225B, GRTPI-AS1, ATP4B, TMCO3). In addition to GAS6, the TC-associated rs7140110 is an sQTL for TMEM225B in adipose subcutaneous tissue (p-value 3.6 × 10^{−08}).

![Comparison of effects estimates for HDL-C and LDL-C among variants in the CETP locus.](https://doi.org/10.1038/s41467-022-33510-7) The color scale of the data points was based on −log_{10} p-values from HDL-C association and the size of each data point was based on −log_{10} p-values of LDL-C association. Variants which are genome wide significant with LDL-C are represented as chromosome:position:reference allele:alternate allele.
further support from TC colocalization analysis and was not significant for other lipid levels.

Phenome-wide association with complex traits. We conducted a phenome-wide association (PheWAS) of 1572 binary complex traits using UK Biobank for the three replicated common variants (16:56957451:CT (CETP); 13:13841051:TC (GAS6); 7:137875053:CT (CREB3L2)) adjusting for PC1–10, age, age², sex, and race. We claimed significance at FDR of 0.05 and identified various complex traits significant, including ischemic heart disease for the CETP variant and heart failure/atherosclerosis, hypercholesterolemia traits for GAS6 variant. The summary statistics from PheWAS analysis for the significant complex traits are tabulated in Supplementary Data 11.

Rare variant aggregates associated with lipids Gene-centric associations. We next evaluated the association of aggregated rare (MAF < 1%) variants, linked to protein-coding genes (gene-centric). We employed a Bonferroni-corrected significance threshold of 0.05/20,000 = 2.5 × 10⁻⁶ for coding and non-coding gene-centric rare variant analyses (Supplementary Fig. 7). We identified 102 coding and 160 non-coding gene-centric rare variant aggregates significantly associated with at least one of the four plasma lipid phenotypes in a non-conditional analysis (Supplementary Data 12, 13). We secondarily conditioned our significant aggregate sets on variants individually associated with lipid levels from the GWAS catalog, MPW summary statistics and the TOPMed data. We identified 74 coding and 25 non-coding rare variants aggregates associated with at least one lipid level after conditional analyses (Supplementary Data 14, 15).

Most of the coding gene-centric sets remained significant after secondary conditioning, while a minority of non-coding gene-centric sets remained significant after conditioning. Significant genes identified from coding rare variant analyses included multiple known Mendelian lipid genes including LCAT, LDLR, and APOB (Supplementary Data 13). RFC2 putative loss-of-function mutations (combined allele frequency < 0.002%) were significantly associated with triglycerides (p-value = 2 × 10⁻⁶) representing a putative novel association for triglycerides. The RFC2 aggregate set (plof) was associated with reduced TG (beta = −0.89 for log(TG)). The persistently significant regions identified from non-coding rare variant analyses linked to genes included the UTR (untranslated region) for CETP and promoter-CAGE (CAGE—Cap Analysis of Gene Expression Sites) around APOA1 for HDL-C, and APOE promoter-CAGE, APOE enhancer-DHS (DHS—DNase hypersensitivity sites), and EHDS promoter-DHS for total cholesterol (Supplementary Data 15). Most of the coding aggregates had larger effects compared to non-coding aggregates, and among the non-coding aggregates SPC24 non-coding aggregate (enhancer-CAGE) at the LDLR locus had the strongest effect for LDL-C (beta = 2.320 mg/dL; p-value = 1.75 × 10⁻⁶).

We analyzed the UK Biobank whole genome sequences among ~130 K participants to provide evidence of replication for the significant coding and non-coding aggregate sets. We used a Bonferroni-corrected significance threshold based on the number of genes tested in each type of aggregate-based test. For gene-centric coding aggregates, we conducted replication of 21 genes (p-value < 0.05/21 = 2.38 × 10⁻⁶) and for non-coding aggregates we replicated the findings from 13 genes (p-value < 0.05/13 = 3.85 × 10⁻⁶). At Bonferroni significance, 71% and 62% of genes replicated for at least one coding and non-coding aggregate set, respectively (Supplementary Data 14, 15). We observed that most of the Mendelian lipid genes replicated for coding aggregates including ABCA1, ABCG5, LCAT, APOB, LDLR, PCSK9, and LPL. For the non-coding aggregate set, the most significant replications were observed for the APOB, LDLR (SPC24), and PCSK9 loci, further corroborating the observation that both coding and noncoding rare variant signals contribute to variation in lipid levels at these loci.

Region-based associations. We also performed unbiased region-based rare variant association analyses tiled across the genome with both static and dynamic window sizes. We first evaluated 2.6 M regions statically at 2 kb size and 1 kb window overlap by the sliding window approach. Statistical significance was assigned at 0.05/(2.6 × 1⁻¹⁰⁴) = 1.88 × 10⁻⁵. We identified 28 significantly associated windows with at least one lipid phenotype. After conditioning of variants individually associated with the corresponding lipid phenotype, we identified two windows at LDLR still significantly associated with both total cholesterol and LDL-C, although these regions included both intronic and exonic variants (Supplementary Data 16). LDLR intron 1, which encodes LDLR-ASI (LDLR antisense RNA 1) on the minus strand, had suggestive evidence for association with TC (p-value = 3.17 × 10⁻⁵) with −2.76 mg/dL reduction in TC. A prior study identified that a common variant (rs651720, MAF 0.11) in LDLR intron 1 is associated with increased LDLR expression in a luciferase assay and reduction in LDL-C when adjusting for rs651720, the significance improved (p-value = 1.43 × 10⁻⁶) with −3.35 mg/dL reduction in TC.

For dynamic window scanning of the genome, we implemented the SCANG method. The SCANG procedure accounts for multiple testing by controlling the genome-wide error rate (GWER) at 0.1⁴. In the dynamic window-based workflow, STAAR-O detected 51 regions significantly associated with at least one lipid phenotype after conditioning on known variants (Supplementary Data 17). Most of the regions mapped to known Mendelian lipid genes, including LCAT (8.7 × 10⁻⁶) for HDL-C, and LDLR (2.4 × 10⁻⁵, 7.3 × 10⁻⁶) and PCSK9 (2.9 × 10⁻⁵, 5.5 × 10⁻⁶) for LDL-C and TC, respectively. Exon 4 aggregates of LDLR were specifically associated with 20 mg/dL increase in LDL-C. PCSK9 exon2-Intron2 region spanning chr13:50437382–55045960 had significantly reduced LDL-C by 6 mg/dL (p-value = 3 × 10⁻⁶), and the effect persisted even with only Intron 2 rare variants of PCSK9 (<5 mg/dL; p-value = 2 × 10⁻⁶). Strikingly, in secondary analyses, we found evidence for very large effects for rare variants in LDLR Introns 2 and 3 (±21 mg/dL; p-value = 7 × 10⁻⁴) and LDLR Introns 16 and 17 (±17 mg/dL; p-value = 0.02), similar to rare coding LDLRM mutations. While 32 of the significant dynamic windows also included exonic regions, there were also several dynamic windows significantly independently associated with lipids not containing exonic regions. For example, four non-coding windows (two overlapping) at 2p24.1, which harbors the Mendelian APOE gene, were significantly associated with LDL-C. Intronic non-coding regions were associated with both LDL-C and TC-associated windows at LPAL2-LPA-SLC22A3; for example, LPAL2 Intron 3 was associated with a 3.7 mg/dL increase in TC. Non-coding TC-associated significant dynamic windows were near TOMM40/APOE. One rare variant signal observed was at TOMM40 Intron 6, where the ‘poly-T’ variant in this region is on the APOE4 haplotype and influences expressivity for Alzheimer’s disease age-of-onset⁵³.⁵⁴. For HDL-C, we identified significant non-coding windows at an intergenic region near LPL and CD36 Intron 4. In the generation of the spontaneously hypertensive rat model, the deletion of intron 4 in CD36 with resultant CD36 deficiency has been mapped to defective fatty acid metabolism in this model⁵⁵. Several regions significant in SCANG were not even nominally significant in burden association analyses indicating the likelihood of causal variants with bidirectional effects.

We replicated 28 sliding and 51 dynamic window aggregate sets using UKB whole genomes, at a Bonferroni-corrected alpha threshold of 0.05/no. of regions for each approach separately. At Bonferroni significance, 61% of the regions from each of the sliding window (p-value < 0.05/28 = 1.79 × 10⁻⁴) and dynamic window (p-value < 0.05/51 = 9.80 × 10⁻⁵) approaches significantly replicated (Supplementary Data 16, 17). Multiple regions linked to LDLR, PCSK9, CETP, APOC3, and ABCA1 were highly significant.

Several gene-centric non-coding aggregates associated with lipids near known monogenic lipid genes but mapped to another gene at the...
 locus via annotations. Therefore, we performed downstream condi-
tional analyses adjusting the gene-centric non-coding results for rare
coding variants (MAF < 1%) within known lipid monogenic genes
(Supplementary Data 18). When accounting for both common and rare
coding variants at the nearby familial hypercholesterolemia LDLR
gene, SPC24-enhancer DHS was significantly associated with total
cholesterol (p-value = 3.01 × 10^{-15}) and with suggestive evidence for
LDL-C (p-value = 1.57 × 10^{-14}). In a similarly adjusted model, LDLR-
enhancer-DHS showed a strong association with TC (p-value 5.18 × 10^{-14}). When adjusting for known common variants as well as rare
coding variants in PCSK9, both PCSK9-enhancer DHS and PCSK9-
promoter DHS were significantly associated with total cholesterol
(Fig. 4, Supplementary Fig. 8). Through this procedure, CETP UTR
retained significance for its independent association with HDL-C as
well as the putatively novel gene EHD3-promoter DHS association
with TC. However, the non-coding gene-centric APOC3 and APOE
associations were rendered non-significant for HDL-C and TC,
respectively.

Since we cannot rule out the possibility of reduced power for
geno-mwide rare variant analyses, we leveraged current knowl-
etge of 22 Mendelian lipid genes for more focused exploratory
analyses. We validated most genes in rare variant coding analyses.
The genes with the strongest coding signals typically had at least
nominal evidence of gene-centric non-coding rare variant associa-
tions (Supplementary Data 19, Supplementary Fig. 9). When rare
coding variants were introduced into the model, the evidence for
non-coding rare variant associations were largely unchanged. Our
findings expanding the currently described genetic basis for
hypercholesterolemia to include rare non-coding variation at LDLR
and PCSK9 (Fig. 5).

Heritability contributions from rare variants
To understand the contribution of rare variants towards lipid trait
heritability, we examined heritability of lipids by variant allele fre-
quency across three ancestral samples (White, Black, and Hispanic) in
TOPMed. We calculated trait heritability using Gremi-LDMS following
the steps as implemented by Wainschtein et al. Using the TOPMed
WGS, we grouped the variants into 4 MAF bins for the three ancestral
samples. In each MAF bin, we grouped variants based on the LD scores
into four quartiles and calculated variance contributed by the SNPs (h^2)
for each of the lipids using unrelated individuals from each ancestral
group (Supplementary Fig. 10) and set negative estimate to zero. We
observed that rare variants from the lower MAF bins contributed to
trait heritability but have large standard errors (Supplementary
Data 20). We observed an increase in h^2 values including WGS variants
relative to estimates obtained from array-genotypes as reported by
Cadby et al. for the European samples. We also compared the h^2
estimates from all the variants from WGS TOPMed cohort against
array-genotypes captured in MGB Biobank to understand the differ-
ences contributed by these two sequencing methods. As expected, the
h^2 estimates from array-genotypes were reduced corresponding to
missing heritability from the lower MAF bins captured by WGS. The
heritability estimates from array-genotypes were markedly higher for
European samples relative to African and Hispanic sample sets indi-
cating that WGS better captured heritability for the latter groups.

Discussion
Conducting one of the largest population-based WGS association
analyses, we now simultaneously interrogate and establish a common,
rare coding, and rare non-coding variant model for a complex trait.
Utilizing 66,329 diverse individuals with deep-coverage WGS, we

APOB1, SPC24 (enhancer-CAGE), HBB and APOE signal drops after the conditional
analysis on rare-coding variants. The different colored dots on the plot represents
the conditional STAAR-O p-values when adjusting for known variants (Set1) and
rare-coding variants of the same or near-by gene. The p-values were calculated from
two-sided aggregate testing preformed using STAAR gene-centric model, where
the model was adjusted for all the covariates; see Methods. STAAR variant-Set Test
for Association using annotation information, TC total cholesterol, TG triglyc-
erides, CAGE cap analysis of gene expression, DHS DNase hypersensitivity.

Fig. 4 | Conditional analysis of coding rare-variants from the same gene and a
near-by gene. Non-coding rare variant sets significantly associated with TC and TG
after the conditional analysis on known variants are shown with additional
adjustment on rare-coding variants. The additional adjustment for rare-coding
variants were carried out for the same gene of the aggregate set and for certain
gene aggregates (SPC24) the conditional analysis was carried out with a nearby
Mendelian gene. After adjusting for rare-coding variants and known variants, EHD3
signal drops minimally, whereas signal from PCSK9 (promoter-DHS, enhancer-
DHS), LDLR loci (enhancer-DHS, SPC24 enhancer-DHS) enhances significantly.
interrogated 428 M variants with plasma lipids expanding the allelic series to rare non-coding variants, often within introns, of Mendelian lipid genes with prior robust rare coding variant support. Our observations have important implications for plasma lipids as well as the genetic basis of complex traits more broadly.

WGS of diverse ancestries enables both allelic and locus heterogeneity for complex traits. Population genetic analyses have largely been enriched for individuals of European descent. Genetic association of plasma lipids using arrays or whole exome sequencing among Europeans has yielded several important insights regarding plasma lipids and the causal determinants of CAD.4,5,42 Larger studies among non-Europeans have often yielded new genetic loci and sometimes new genes, such as PCSK9.41,47,57. Such differences have also led to concerns about the use of polygenic risk scores typically larger than common non-coding variants and less than rare coding variants, with the exception of LDLR, consistent with whole genome mutational constraint models.

The detection of independent rare non-coding variant signals has remained elusive largely due to limited sample sizes with requisite WGS and limitations in the interpretation of rare non-coding variation functional consequence. Previously, we used annotated functional variants also yield a polygenic basis for hypercholesterolemia. In the present work, we now identify rare non-coding variants in proximity of Mendelian hypercholesterolemia genes, specifically LDLR and PCSK9, that also contribute to the genetic basis of hypercholesterolemia. Parts of the figure were generated using pictures from Servier Medical Art. Servier Medical Art by Servier is licensed under a Creative Commons Attribution 3.0 Unported License (https://creativecommons.org/licenses/by/3.0/). LDL-C low-density lipoprotein cholesterol.

The presence of familial hypercholesterolemia variants is independently prognostic for CAD, beyond lipids, and merits the consideration of more costly lipid-lowering medications. The small fraction with severe hypercholesterolemia and features consistent with strong genetic predisposition have a familial hypercholesterolemia variant in LDLR, APOB, or PCSK9.6,7,15,16. We now observe that rare LDLR variants in Introns 3, 16, and 17 lead to a standard deviation increase in LDL-C, approximating effects observed with clinically reported exonic familial hypercholesterolemia variants in LDLR. Small studies have indicated the possibility of rare intronic LDLR variants causing familial hypercholesterolemia due to altered splicing, which we now observe in our unbiased population-based WGS study40,48. A WGS approach to lipid disorders, particularly for familial

Nature Communications | (2022)13:5995
hypercholesterolemia, will markedly improve the diagnostic yield beyond existing limited approaches.

Our dynamic window approach may also improve the clinical curation of exonic variants. Among the data used to curate exonic variants is the use of in silico functional prediction tools. Although evolutionary constraint measures are typically employed, such tools are largely agnostic to functional domain. As it relates to lipids, disruptive APOB and PCSK9 exonic variants can lead to strikingly opposing directions with large effects for LDL-C depending on locations. Using SCANc, we detect a significant association with large effect for LDLR Exon 4 itself. This observation supports the pathogenicity of LDLR Exon 4 disruptive variants among patients with severe hypercholesterolemia. The majority of familial hypercholesterolemia variants worldwide occur in Exon 4 of LDLR. Conventional rare coding variant analyses aggregate all exonic variants for transcript. Here, we demonstrate an opportunity for exon-level rare variant association testing.

Our discovery analyses with replication as well as heritability assessment are consistent with the notion that both rare coding and non-coding alleles, not well-captured by genome-wide arrays. Furthermore, we observe that heritability gains relative to genome-wide non-coding alleles, not well-captured by genome-wide arrays. Functional rare coding variant analyses aggregate all exonic variants for limiting power. The striking difference for most STAAR and burden sequence to prioritize causal variants remains a challenge thereby using WGS in diverse ancestries, we can study hundreds of millions GWAS datasets limiting power for novel discovery. Nevertheless, by average, cost may no longer be a downside.

Our study has important limitations. First, while our study is large for a WGS study by contemporary standards, it is dwarfed by existing GWAS datasets limiting power for novel discovery. Nevertheless, by using WGS in diverse ancestries, we can study hundreds of millions new variants. Second, prediction of rare non-coding variation consequence to prioritize causal variants remains a challenge thereby limiting power. The striking difference for most STAAR and burden results also highlights bidirectional effects for rare non-coding variants within the same region and further challenges for clinical utility. Third, given the paucity of multi-ancestral WGS datasets with lipids, our analyses are largely restricted to TOPMed and replication to European rich UK Biobank WGS data. For single variant associations, we pursued TOPMed-imputed GWAS datasets but were limited by the lack of ancestral diversity. As TOPMed is a consortium of multiple different cohorts, we demonstrate consistencies by cohort. Furthermore, rare variant non-coding signals were largely restricted to regions with rare variant coding signals supporting biological plausibility.

In conclusion, using WGS and lipids among 66,329 ancestrally diverse individuals we expand the catalog of alleles associated with lipids, including allelic heterogeneity at known loci and locus heterogeneity by ancestry. We characterize the common, rare coding, and rare non-coding variant model for lipids and replicated the results. Lastly, we now demonstrate a monogenic-equivalent model for rare LDLR intronic variants predisposing to marked alterations in LDL-C, currently not recognized in current population or clinical models for LDL-C.

Methods

Dataset

Contributing studies. The discovery cohort includes the whole genome sequenced (WGS) data of 66,329 samples from 21 studies of the Trans-Omics for Precision Medicine (TOPMed) program with blood lipids available. The overall goal of TOPMed is to generate and use trans-omics, including whole genome sequencing, of large numbers of individuals from diverse ancestral backgrounds with rich phenotypic data to gain novel insights into heart, lung, blood, and sleep disorders. The Freeze 8 data includes 140,306 samples out of which 66,329 samples qualified with lipid phenotype. Freeze 8 dataset passed the central quality control protocol implemented by the TOPMed Informatics Research Core (described below) and was deposited in the dbGaP TOPMed Exchange Area.

The studies included in the current dataset, along with their abbreviations and sample sizes, contains the Old Order Amish (Amish, n = 1083), Atherosclerosis Risk in Communities study (ARIC, n = 8016), Mt Sinai BioMe Biobank (BioMe, n = 9848), Coronary Artery Risk Development in Young Adults (CARDIA, n = 3,056), Cleve-

land Family Study (CSF, n = 579), Cardiovascular Health Study (CHS, n = 3,456), Diabetes Heart Study (DHS, n = 365), Framingham Heart Study (FHS, n = 3992), Genetic Studies of Atherosclerosis Risk (GeneSTAR, n = 1757), Genetic Epidemiology Network of Arterio-
pathy (GENOA, n = 1046), Genetic Epidemiology Network of Salt Sensitivity (GenSalt, n = 1772), Genetics of Lipid Lowering Drugs and Diet Network (GOLDN, n = 926), Hispanic Community Health Study - Study of Latinos (HCHS_SOL, n = 7714), Hypertension Genetic Epidemiology Network and Genetic Epidemiology Network of Arterio-
pathy (HyperGEN, n = 1853), Jackson Heart Study (JHS, n = 2847), Multi-Ethnic Study of Atherosclerosis (MESA, n = 5290), Massachus-
etts General Hospital Atrial Fibrillation Study (MGH_AF, n = 683), San Antonio Family Study (SAFS, n = 619), Samoan Adiposity Study (SAS, n = 1182), Taiwan Study of Hypertension using Rare Variants (THRIV, n = 1982) and Women’s Health Initiative (WHI, n = 8263) (Please see Supplementary Note 1 for additional details). The multi-
national data set included individuals from White (44%), Black (26%), Hispanic (21%), Asian (7%), and Samoan (2%) ancestries. Study partici-
pants granted consent per each study’s Institutional Review Board (IRB) approved protocol. Secondarily, these data were analyzed through a protocol approved by the Massachusetts General Hospital IRB. Supplementary Data 1 details the number of samples across different studies and ancestral group.

The replication cohorts for single variant GWAS include TOPMed-imputed genome-wide array data from the Mass General Brigham (MGB), Penn Medicine Biobanks and UK Biobank (UKB) imputed data which consist of 25,137, 20,079, and 424,955 samples respectively. The replication cohort for rare variant aggregates test include UKB whole genome sequenced data which consists of a subset of 133,360 UKB participants, where we removed unconsented and related individuals. We curated the MGB Biobank and Penn Medi-
cine Biobank phenotype data from the corresponding electronic health record databases in accordance with corresponding institutional IRB approvals. The UKB data included volunteer residents of the UK aged 40–69 and were recruited between 2006 and 2010. Consent was previously obtained from each participant regarding storage of biological specimens, genetic sequencing, access to all available elec-
tronich health record (EHR) data, and permission to recontact for future studies. All UKB participants gave written informed consent per UK primary protocol. The MGB Biobank consists of 54%, Penn Medicine Biobank consist of 52% and UK Biobank imputed data consist of 54% of female samples and average ages were 55.89, 58.35 and 56.55 years, respectively (Supplementary Data 6).

Phenotypes. The primary outcomes in this study included LDL cho-

lesterol (LDL-C), HDL cholesterol (HDL-C), total cholesterol (TC), and triglycerides (TG) phenotypes. LDL-C was either directly measured or calculated by the Friedewald equation when triglycerides were <400 mg/dL. Given the average effect of lipid lowering medicines, when lipid-lowering medicines were present, we adjusted the total cholesterol by dividing by 0.8 and LDL-C by dividing by 0.7, as previously done. Triglycerides remained natural log transformed for analysis. Fasting status was accounted for with an indicator variable.

We harmonized the phenotypes across each cohort and inverse rank normalization of the residuals of each race within each cohort scaled by the standard deviation of the trait and adjusted for covariates. We included covariates such as age, sex, PCI–11, study-groups as well as Mendelian founder lipid variants APOB

Nature Communications| (2022) 13:5995
p.R3527Q and APOC3 p.R19X for the Amish cohort. Supplementary Data 1 provides the distributions of each of the four lipid phenotypes by cohort, ancestral groups, and gender. For the UK Biobank, we curated the first instance of the four lipids (data field numbers: HDL-C: 30760; LDL-C: 30780; TC: 30690; TG: 30870). The lipid measurements from mmol/L were converted to mg/dl by multiplying TG measurements by 88.57 and for other lipids by multiplying by 38.67. We executed similar steps of phenotype harmonization and normalization for the replication cohorts. In addition, we adjusted the MGB Biobank for study-center and array-type, and Penn Medicine Biobank for ancestry and BMI in addition to the other common covariates.

Genotypes. Whole genome sequencing of goal >30X coverage was performed at seven centers (Broad Institute of MIT and Harvard, Northwest Genomics Center, New York Genome Center, Illumina Genomic Services, PSOMAGEN [formerly Macrogen], Baylor College of Medicine Human Genome Sequencing Center, and McDonnell Genome Institute [MGI] at Washington University). In most cases, all samples for a given study within a given Phase were sequenced at the same center (Supplementary Note 1). The reads were aligned to human genome build GRCh38 using a common pipeline across all centers (BWA-MEM+GATK). The TOPMed Informatics Research Core at the University of Michigan performed joint genotype calling on all samples in Freeze 8. The variant calling “GotCloud” pipeline (https://github.com/statgen/topmed_variant_calling) is under continuous development and details on each step can be accessed through TOPMed website for Freeze817. The resulting BCF files were split by study and consent group for distribution to approved dbGaP users. Quality control was performed centrally by the TOPMed IRC and the TOPMed Data Coordinating Center (DCC) as previously described10. Briefly, the two sequence quality criteria used in freeze 8 are: estimated DNA sample contamination below 10%, and 95% or more of the genome covered to 10x or greater. The variant filtering in TOPMed Freeze 8 is performed by (1) first calculating Mendelian consistency scores using known familial relatedness and duplicates, and (2) training a Support Vector Machine (SVM) classifier between known variant sites (positive labels) and Mendelian inconsistent variants. A small number of sex mismatches were detected as annotated females with low X and high Y chromosome depth or annotated males with high X and low Y chromosome depth. These samples were either excluded from the sample set to be released on dbGaP or their sample identities were resolved using information from prior array genotype comparisons and/or pedigree checks. Details regarding WGS data acquisition, processing and quality control vary among the TOPMed data freezers. Freeze-specific methods are described on the TOPMed website (https://www.nhlbiwgs.org/data-sets) and in documents included in each TOPMed accession released on dbGaP. The VCF/BCF files were converted to GDS (Genomic Data Structure) format by the DCC and were deposited into the dbGaP TOPMed Exchange Area.

The genetic relationship matrix (GRM) is an N*N matrix of relatedness information of the samples included in the study and was computed centrally using ‘PC-relate’ R package (version: 1.24.0)12. Using the ‘Genesis’ R package (version:2.20.1)13 we generated sub-setted GRM for the samples with plasma lipid profiles. The GDS files with the variants were annotated internally by curating data from multiple database sources using Functional Annotation of Variant–Online Resource (FAVOR (http://favor.genohub.org))14. This study used the resultant aGDS (annotation GDS) files.

The MGB Biobank replication cohort was genotyped using three different arrays (Multiethnic Exome Global (Meg), Human multi-ethnic array (Mega)), and Expanded multi-ethnic genotyping array (Megex)), and we separately imputed the data using TOPMed imputation server with default parameters15. This study applied the Version-r2 of the imputation panel, it includes 97,256 reference samples and ~300 M genetic variants. The Illumina Global Screening array was used to genotype the Penn Medicine Biobank. Penn Medicine Biobank TOPMed imputation was performed using EAGLE16 and Minimac17 software. For this study, we downloaded variants that passed a min R² threshold of 0.3. The TOPMed imputation panel is robust, built from 97,256 deeply sequenced human genomes and contains 308,107,085 genetic variants from multi-ethnic samples. Imputation was performed in independent non-overlapping samples agnostic to phenotypes. The UKB imputed data was derived using merged UK10K7, 1000 Genomes phase2 reference panels and was combined to the Haplotype reference Consortium (HRC) using IMPUTE 4 program (https://marchini.org/software/). The UKB WGS data consist of whole genomes of 150,119 UKB participants with an average coverage of 32.5X. We used joint called VCFs from GraphTyper, which consist of 710,913,648 variants79. We used VCFs provided on the UK Biobank and conducted all the analysis in UKB Research Analysis Platform (UKB RAP).

Single variant association
We performed genome-wide single variant association analyses for autosomal variants with minor allele frequency (MAF)>0.1% across the dataset with each of the four lipid phenotypes. We implemented the SAIGE-QT method, which employs fast linear mixed models with kinship adjustment, in Encore (https://encore.sph.umich.edu/) for single variant association analyses. We additionally adjusted the model for covariates (PC1-PCII, age, sex, age², and study-groups [cohort-race subgrouping]).

We conducted single variant association replications for putative novel variants. After comparing the results with published lipid GWAS summary statistics, we filtered putative novel GWAS variants based on a stringent whole genome-wide significant threshold (alpha = 5 × 10⁻⁸)90. Replication was performed in the MGB, Penn Medicine Biobanks and UK Biobank where linear regression models were fitted and selected for covariates as indicated above. In addition, we adjusted the MGB Biobank for study recruitment center and array and Penn Medicine Biobank for ancestry and BMI. In the MGB Biobank, we selected lipid concentrations closest to the sample acquisition time point and adjusted for statins if prescribed within one year prior to sample acquisition. In the Penn Biobank, we utilized each participant’s median lipid concentration for replication; statins prescribed prior to lipid concentration used were adjusted in the models. In addition, we carried out meta-analysis using fixed effects model based on inverse-variance-weighted effect size for the two replication cohorts using METASOFT72.

Rare variant association test
We performed rare variant association (RVA) using the Variant-Set Test for Association using Annotation Information (STAAR) pipeline81. STAARpipeline is a regression-based framework that permits adjustment of covariates, population structure, and relatedness by fitting linear and logistic mixed models for quantitative and dichotomous traits85. We chose STAAR to leverage the annotation information and associated scores that were available for TOPMed Freeze 8 data to incorporate the analysis of rare non-coding variants from whole genome sequencing. The method implements genome-wide scanning of rare variants (MAF <0.01) in gene-centric and region-based workflows. For each variant set, STAARpipeline calculates a set-based p-value using the STAAR method, which increases the analysis power by incorporating multiple in silico variant functional annotation scores capturing diverse genomic features and biochemical readouts85. We aggregated rare variants into multiple groups for coding and non-coding analyses. For the coding region, we defined five different aggregate masks of rare variants i) plof (putative loss-of-function), plof-Ds (putative loss-of-function or disruptive missense), missense, disruptive-missense, and synonymous. For the non-coding regions, we used seven rare variant masks: (1) promoter-CAGE (promoter variants
within Cap Analysis of Gene Expression (CAGE) sites\(^\text{29}\), (2) promoter-DHS (promoter variants within DNase hypersensitivity [DHS] sites\(^\text{29}\)), (3) enhancer-CAGE (enhancer within CAGE sites\(^\text{29,30}\)), (4) enhancer-DHS (enhancer variants within DHS sites\(^\text{29,30}\)), (5) UTR (rare variants in 3′ untranscribed region [UTR] and 5′ UTR untranscribed region), (6) upstream, and (7) downstream. Detailed explanations of the regions defined based on these tasks is discussed with STAARpipeline\(^\text{13,83}\).

In the gene-centric workflows, for both coding (within exonic boundaries) and non-coding (promoter: +/-3 kb window of transcription starting site (TSS), enhancer: GeneHancer predicted regions, UTR (both 5′ and 3′ UTR regions)/upstream/downstream: GENCODE Variant Effect Predictor (VEP) categories) regions, we considered only genes with at least two rare variants (i.e., 18,445 genes in all 22 autosomes). In the region-based workflows, we implemented two protocols: (1) a ‘sliding window’ approach, where we aggregated rare variants within 2-kb sliding windows and with 1-kb overlap length, and (2) a ‘dynamic window’ approach, where we executed SCANG\(^\text{44}\) method and aggregated dynamically variant-sets between 40–300 variants per set, where the method systematically scans the whole genome with overlapping windows of varying sizes. The STAARpipeline R-package implements multiple rare-variants aggregate tests including SKAT\(^\text{90}\), SVEP\(^\text{94}\) and LAMS\(^\text{94}\) overlapping windows of varying sizes. The STAARpipeline R-package where the method systematically scans the whole genome with overlapping windows of varying sizes. The STAARpipeline R-package implements multiple rare-variants aggregate tests including SKAT\(^\text{90}\), SVEP\(^\text{94}\) and LAMS\(^\text{94}\).

We completed aggregate tests as three-step process. In the first step, we fitted a null model using glmnmkin() function. The null model was fitted for each of the four lipid phenotypes adjusted for all covariates and relatedness except the genotype of interest. In the second step, we ran genome-wide gene-centric and region-based rare-variant aggregate tests. The third step directed conditional analyses, where the results were adjusted for previously known significantly lipid-associated (i.e., \( p < 5 \times 10^{-8} \) in external datasets) individual variants from GWAS Catalog\(^\text{91}\) and Million Veterans Program (MVP\(^\text{92}\)) GWAS summary statistics. To obtain effect estimates of significant aggregate sets, we associated the cumulative genotypes (binary scores) based on the variants forming the aggregated and used Glimm.Wald test from GMMAT R package\(^\text{83}\) (version 1.3.1). For signficantly associated window-based rare variant aggregations, we trimmed the exonic variants and estimated the effects with only non-coding variants.

For the rare variant replication in UKB WGS data, we curated the rare variant aggregate sets in UKB RAP for the gene-centric coding/ non-coding and region-based significant sets and applied STAAR workflow as demonstrated by the STAARpipeline (https://github.com/xhaoli/STAARpipeline) and describe above.

Computational mining of single variant GWAS

Gene-set enrichment using FUMA. We performed enrichment analysis with single variant GWAS summary stats from the four lipids using FUMA\(^\text{46}\) (version 1.3.7) with default parameters and significance at \(5 \times 10^{-7}\). FUMA is an integrated platform which efficiently facilitates functional mapping and enrichment of GWAS-associated genes using multiple useful resources. The method uses 18 different biological data repositories and tools to process GWAS data. We additionally used MAGMA\(^\text{47}\) (version 1.08) gene-based analysis enrichment workflow within FUMA with the complete GWAS summary data for eQTL based tissue enrichment. The functionally prioritized genes were visualized based on their protein-protein interaction networks using the STRING database\(^\text{48}\).

CEPTP and GAS6 gene expression and lipid trait colocalization. We studied the correlation of LDL-C and HDL-C effects with eQTL effects at chromosome 16q13, which includes C E T P and correlation of LDL-C and TC with eQTLs at rs7140110 of GAS6. We downloaded GTEx eQTL build 38 (version8) data for liver, adipose subcutaneous, and adipose visceral (omentum) tissues from GTEx on 16/APR/2020\(^\text{97}\). For the C E T P variant analysis, we selected eQTLs with nominal significance (\(p\text{-value} < 0.05\)) and utilized the eQTL-gene pairs with the most significant \(p\)-values. Genes with at least 5 eQTLs were selected for the colocalization analysis. We selected variants with a suggestive significance (\(p\text{-value} < 5 \times 10^{-8}\) ) for LDL-C or HDL-C effects within 500 kb of the lead locus variant. For the GAS6 variant analysis, we curated all the GWAS variants within 500 kb of the lead variant with nominal significance (\(p\text{-value} < 0.05\)) and matched them to eQTL data where the transcription starting site of the corresponding gene is within +/-500 kb. We conducted colocalization analysis using the coloc.abf() function\(^\text{98}\) and identified nominally significant (PP.H4 > 1 \( \times 10^{-8}\) ) genes-eQTL pairs. The coloc methodology implements an efficient statistical framework to identify shared variants from two association signals through posteriors probabilities. Finally, we used the colocalized signals and compared the significant genes using STRING\(^\text{49}\), a protein-protein interaction database. All the correlation tests were conducted in R, where we calculated Pearson correlations between the lipid effect estimates and gene expression effects (slope) from GTEx.

Phenome wide association analysis. The complex trait information was curated from UK Biobank resource, where we curated multiple outcome phenotypes for UKB samples into International Classification of Diseases (ICD)-based phecodes based on phecode map (https://phecwascatalog.org) using the PhewAS R package (version PhewAS_0.99.5-4). We conducted a phenome-wide association analysis (PhewAS) using a logistic regression model glm() in R. We adjusted the models for PC1–10, age, age\(^2\), sex, and race.

Calculation of heritability estimates from TOPMed WGS data

We calculated heritabilities estimated for the four lipids using TOPMed WGS data using Greml-LDMS approach\(^\text{39}\), where we binmed the variants into four MAF bins based on minor allele frequency and grouped the variants to four LD quartiles based on LD score calculated by GCTA method\(^\text{37}\). The four MAF bins used in this study includes >=0.05, >=0.01 to <0.05, >=0.001 to <0.01 and >=0.0001 to <0.001. We excluded any variant with MAF < 0.0001 from this analysis. The hereditary estimation was calculated for three ancestral groups (African, European, Hispanic) where only unrelated samples (kinship score < 0.025) were included in the analysis. We excluded the other two ancestral groups (i.e., Asian and Samoan) from this analysis due to insufficient sample sizes. In total we included 9640, 21568 and 10631 in African, European and Hispanic ancestries respectively. For each MAF bin, we implemented certain quality control (QC) measures using PLINK software\(^\text{50}\), which includes; genotype missingness (−geno 0.05), sample missingness (−mind 0.05), Hardy-Weinberg equilibrium (−hwe 10\(^{-1}\) ) and LD pruned variants (−indep-pairwise 50 5 0.1) as implemented by Wainschtein et al.\(^\text{51}\). Next, we implemented Greml-LDMS with LD score region as 200 and GRM cut-off as 0.05 for the four lipid phenotypes. We calculated 20 principal components from the QC passed variants in each MAF bin and implemented GCTA workflow with −reml-no-constrain, −reml-no-lrt and −reml-maxit 10,000 parameters to avoid the non-convergence issues and negative \(h^2\) estimates. For comparing the \(h^2\) estimates between variants from WGS data and array-genotypes, first, we used QC passed WGS variants as mentioned above, second, we curated the variants from MGB Biobank array data and intersected them with WGS variants from TOPMed. Next, we calculated heritability estimates for array-genotype variants and compared with \(h^2\) estimates from WGS variants for the three ancestral groups.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.
Data availability

Individual whole-genome sequence data for TOPMed and harmonized lipids at individual sample level are available through restricted access via the TOPMed dbGaP Exchange area. Summary level genotype data from TOPMed are available through the BRAVO browser (https://bravo.sph.umich.edu/). The UK Biobank (UKB) whole-genome sequence data can be accessed through UKB Research Analysis Platform (RAP), through the UKB approval system (https://www.ukbiobank.ac.uk). The Mass General Brigham Biobank (MGBB) individual-level data are available from https://personalizedmedicine.partners.org/Biobank/Default.aspx, where the data is available through institutional review board (IRB) approval, therefore not publicly available. Individual-level data from Penn Medicine Biobank is not publicly available due to research participants privacy concerns. The summary data captured using whole exome sequencing can be accessed through PMBB Genome Browser (https://pmbb.med.upenn.edu/allele-frequency/). The dbGaP accessions for TOPMed cohorts are as follows: Old Order Amish (Amish) phs000956 and phs00039; Atherosclerosis Risk in Communities study (ARIC) phs001121 and phs000280; Mt Sinai BioMe Biobank (BioMe) phs001644 and phs000925; Coronary Artery Risk Development in Young Adults (CARDIA) phs001612 and phs00258; Cleveland Family Study (CFS) phs000954 and phs00294; Cardiovascular Health Study (CHS) phs001368 and phs000027; Diabetes Heart Study (DHS) phs001412 and phs001012; Framingham Heart Study (FHS) phs000974 and phs000007; Genetic Studies of Atherosclerosis Risk (GeneSTAR) phs001218 and phs000375; Genetic Epidemiology Network of Arteriopathy (GENOA) phs001345 and phs001238; Genetic Epidemiology Network of Salt Sensitivity (GenSalt) phs001217 and phs000784; Genetics of Lipid-Lowering Drugs and Diet Network (GOLDN) phs001339 and phs000741; Hispanic Community Health Study - Study of Latinos (HCHS_SOL) phs001395 and phs000810; Hypertension Genetic Epidemiology Network and Genetic Epidemiology Network of Arteriopathy (HyperGEN) phs001293 and phs001293; Jackson Heart Study (JHS) phs000964 and phs000286; Multi-Ethnic Study of Atherosclerosis (MESA) phs001416 and phs000209; Massachusetts General Hospital Atrial Fibrillation Study (MGH.AF) phs001062 and phs001001; San Antonio Family Study (SAFS) phs001215 and phs000462; Samoan Adiposity Study (SAS) phs000972 and phs000914; Taiwan Study of Hypertension using Rare Variants (THRV) phs001387 and phs001387; Women’s Health Initiative (WHI) phs001237 and phs000200.

Code availability

Codes used to implement STAAR workflows are available at https://github.com/xihaoi/STAAR and https://github.com/xihaoi/STAAR pipeline. Workflows implemented for whole genome heritability calculations are available at https://github.com/CGNSGenomics/Heritability_WGS.

References

1. Cohen, J. C., Boerwinkle, E., Mosley, T. H. & Hobbins, H. H. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. N. Engl. J. Med. 354, 1264–1272 (2006).
2. Cohen, J. et al. Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. Nat. Genet. 37, 161–165 (2005).
3. Musunuru, K. et al. Exome sequencing, ANGPTL3 mutations, and familial combined hyperlipidemia. N. Engl. J. Med. 363, 2220–2227 (2010).
4. Stitziel, N. O. et al. ANGPTL3 deficiency and protection against coronary artery disease. J. Am. Coll. Cardiol. 68, 2054–2063 (2017).
5. Dewey, F. E. et al. Genetic and pharmacologic inactivation of ANGPTL3 and cardiovascular disease. N. Engl. J. Med. 377, 211–221 (2017).
6. Manolio, T. A. et al. Finding the missing heritability of complex diseases. Nature 461, 747–753 (2009).
7. Pollin, T. I. et al. A null mutation in human APOC3 confers a favorable plasma lipid profile and apparent cardioprotection. Science 322, 1702–1705 (2008).
8. Shen, H. et al. Familial defective apolipoprotein B-100 and increased low-density lipoprotein cholesterol and coronary artery calcification in the old order amish. Arch. Intern. Med. 170, 1850–1855 (2010).
9. Salehine, D. et al. Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. Nature 544, 235–239 (2017).
10. Exome Aggregation Consortium. et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285–291 (2016).
11. Samocha, K. E. et al. A framework for the interpretation of de novo mutation in human disease. Nat. Genet. 46, 944–950 (2014).
12. Natarajan, P. et al. Chromosome Xq23 is associated with lower atherogenic lipid concentrations and favorable cardiometabolic indices. Nat. Commun. 12, 2182 (2021).
13. Li, X. et al. Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. Nat. Genet. 52, 969–983 (2020).
14. Natarajan, P. et al. Deep-coverage whole genome sequences and blood lipids among 16,324 individuals. Nat. Commun. 9, 3391 (2018).
15. Klarin, D. et al. Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. Nat. Genet. 50, 1514–1523 (2018).
16. Hu, Y. et al. Minority-centric meta-analyses of blood lipid levels identify novel loci in the Population Architecture using Genomics and Epidemiology (PAGE) study. PLoS Genet. 16, e1008684 (2020).
17. NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature 590, 290–299 (2021).
18. Stilp, A. M. et al. A System for Phenotype Harmonization in the NHLBI Trans-Omics for Precision Medicine (TOPMed) Program. Am. J. Epidemiol. https://doi.org/10.1093/aje/kwab115 (2021).
19. Fadista, J., Manning, A. K., Florez, J. C. & Groop, L. The (in)formative GWAS P-value threshold revisited and updated for low-frequency variants. Eur. J. Hum. Genet. 24, 1202–1205 (2016).
20. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575 (2007).
21. Bentley, A. R. et al. Multi-ancestry genome-wide gene-smoking interaction study of 387,272 individuals identifies new loci associated with serum lipids. Nat. Genet. 51, 636–648 (2019).
22. Ripatti, P. et al. Polygenic hyperlipidemias and coronary artery disease risk. Circ. Genom. Precis. Med. 13, e002725 (2020).
23. van Leeuwen, E. M. et al. Meta-analysis of 49,549 individuals imputed with the 1000 Genomes Project reveals an exonic damaging variant in ANGPTL4 determining fasting TG levels. J. Med. Genet. 53, 441–449 (2016).
24. Nielsen, J. B. et al. Loss-of-function genomic variants highlight potential therapeutic targets for cardiovascular disease. Nat. Commun. 11, 6417 (2020).
25. Aragam, K. G. et al. Limitations of contemporary guidelines for managing patients at high genetic risk of coronary artery disease. J. Am. Coll. Cardiol. 75, 2769–2780 (2020).
26. Park, J. et al. Exome-wide evaluation of rare coding variants using electronic health records identifies new gene-phenotype associations. Nat. Med. 27, 66–72 (2021).
27. Mountjoy, E. et al. An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. Nat. Genet. 53, 1527–1533 (2021).
28. Lonsdale, J. et al. The Genotype-Tissue Expression (GTEx) project. Nat. Genet. 45, 580–585 (2013).
29. Barter, P. J. et al. Effects of torcetrapib in patients at high risk for coronary events. N. Engl. J. Med. 357, 2109–2122 (2007).
30. Schwartz, G. G. et al. Effects of dalcetrapib in patients with a recent acute coronary syndrome. N. Engl. J. Med. 367, 2089–2099 (2012).
31. The ESP3/TIMI55-REVEAL Collaborative Group. Effects of anacetrapib in patients with atherosclerotic vascular disease. N. Engl. J. Med. 377, 1217–1227 (2017).
32. Lincoff, A. M. et al. Evacetrapib and cardiovascular outcomes in high-risk vascular disease. N. Engl. J. Med. 376, 1933–1942 (2017).
33. Fairoosy, R. H., White, J., Palmén, J., Kalea, A. Z. & Humphries, S. E. Identification of the functional variant(s) that explain the low-density lipoprotein receptor (LDLR) GWAS SNP rs6151720 association with lower LDL-C and risk of CHD. PLoS ONE 11, e0167676 (2016).
34. Li, Z. et al. Dynamic scan procedure for detecting rare-variant association regions in whole-genome sequencing studies. Am. J. Hum. Genet. 104, 802–814 (2019).
35. Roses, A. D. et al. A TOMM40 variable-length polymorphism predicts the risk of late-onset Alzheimer’s disease. Pharmacogenomics J. 10, 375–384 (2010).
36. Li, G. et al. TOMM40 intron 6 poly-T length, age at onset, and neuropathology of AD in individuals with APOE ε3/c3. Alzheimers Dement. J. Alzheimers Assoc. 9, 554–561 (2013).
37. Glazier, A. M., Scott, J. & Altman, T. J. Molecular basis of the Cd36 chromosomal deletion underlying SHR defects in insulin action and fatty acid metabolism. Mamm. Genome. J. Int. Mamm. Genome Soc. 13, 108–113 (2002).
38. The LifeLines Cohort Study. et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat. Genet. 47, 1114–1120 (2015).
39. Wainschtein, P. et al. Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. Nat. Genet. 54, 263–273 (2022).
40. Cadby, G. et al. Heritability of 596 lipid species and genetic correlation with cardiovascular traits in the Busselton Family Heart Study. J. Lipid Res. 61, 537–545 (2020).
41. Global Lipids Genetics Consortium. Discovery and refinement of loci associated with lipid levels. Nat. Genet. 45, 1274–1283 (2013).
42. Willer, C. J. et al. Discovery and refinement of loci associated with lipid levels. Nat. Genet. 45, 1274–1283 (2013).
43. ENGAGE Consortium. et al. The impact of low-frequency and rare variants on lipid levels. Nat. Genet. 47, 589–597 (2015).
44. The Myocardial Infarction Genetics Consortium Investigators. Inactivating mutations in NPC1L1 and protection from coronary heart disease. N. Engl. J. Med. 371, 2072–2082 (2014).
45. GLGC Consortium. et al. Exome chip meta-analysis identifies novel loci and East Asian-specific coding variants that contribute to lipid levels and coronary artery disease. Nat. Genet. 49, 1722–1730 (2017).
46. Hoffmann, T. J. et al. A large electronic-health-record-based genome-wide study of serum lipids. Nat. Genet. 50, 401–413 (2018).
47. Martin, A. R. et al. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat. Genet. 51, 584–591 (2019).
48. Peloso, G. M. & Natarajan, P. Insights from population-based analyses of plasma lipids across the allele frequency spectrum. Curr. Opin. Genet. Dev. 50, 1–6 (2018).
49. Kremer, L. S. et al. Genetic diagnosis of Mendelian disorders via RNA sequencing. Nat. Commun. 8, 15824 (2017).
50. Cummings, B. B. et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Sci. Transl. Med. 9, eaal5209 (2017).
51. Genome Aggregation Database Production Team. et al. Transcript expression-aware annotation improves rare variant interpretation. Nature 581, 452–453 (2020).
52. Mendes de Almeida, R. et al. Whole gene sequencing identifies deep-intronic variants with potential functional impact in patients with hypertrophic cardiomyopathy. PLoS ONE 12, e0182946 (2017).
53. Vitsios, D., Dhindsa, R. S., Middleton, L., Gussow, A. B. & Petrovski, S. Prioritizing non-coding regions based on human genomic constraint and sequence context with deep learning. Nat. Commun. 12, 1504 (2021).
54. di Iulio, J. et al. The human noncoding genome defined by genetic diversity. Nat. Genet. 50, 333–337 (2018).
55. Genome Aggregation Database Consortium. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581, 434–443 (2020).
56. Khera, A. V. et al. Diagnostic yield and clinical utility of sequencing familial hypercholesterolemia genes in patients with severe hypercholesterolemia. J. Am. Coll. Cardiol. 67, 2578–2589 (2016).
57. Benn, M., Watts, G. F., Tybjærg-Hansen, A. & Nordestgaard, B. G. Mutations causative of familial hypercholesterolemia: screening of 98,098 individuals from the Copenhagen General Population Study estimated a prevalence of 1 in 217. Eur. Heart J. 37, 1384–1394 (2016).
58. Grundy, S. M. et al. 2018 AHA/ACC/AACVPR/AAPA/ACPM/ADA/AGS/APA/ASPC/NLA/PCNA Guideline on the Management of Blood Cholesterol: Executive Summary. J. Am. Coll. Cardiol. 73, 3168–3209 (2019).
59. Sturm, A. C. et al. Clinical genetic testing for familial hypercholesterolemia. J. Am. Coll. Cardiol. 72, 662–680 (2018).
60. Reekamp, L. F. et al. A Deep intrinsic variant in LDLR in familial hypercholesterolemia: time to widen the scope? Circ. Genomic Precis. Med. 11, e002385 (2018).
61. Calandra, S., Tarugi, P. & Bertolini, S. Altered mRNA splicing in lipoprotein disorders. Curr. Opin. Lipidol. 22, 93–99 (2011).
62. on behalf of the ACMG Laboratory Quality Assurance Committee. et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet. Med. 17, 405–423 (2015).
63. Peloso, G. M. et al. Rare protein-truncating variants in APOB, lower low-density lipoprotein cholesterol, and protection against coronary heart disease. Circ. Genom. Precis. Med. 12, e002376 (2019).
64. Abifadel, M. et al. Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. Nat. Genet. 34, 154–156 (2003).
65. Jiang, L. et al. The distribution and characteristics of LDL receptor mutations in China: a systematic review. Sci. Rep. 5, 17272 (2015).
66. Arráz, N. et al. Novel mutations identification in exon 4 of LDLR gene in patients with moderate hypercholesterolemia in a Venezuelan population. Am. J. Ther. 17, 325–329 (2010).
67. Gudnason, V. et al. Identification of recurrent and novel mutations in exon 4 of the LDLR receptor gene in patients with familial hypercholesterolemia in the United Kingdom. Arterioscler. Thromb. J. Vasc. Biol. 13, 56–63 (1993).
68. Goldmann, R. et al. Genomic characterization of large rearrangements of the LDLR gene in Czech patients with familial hypercholesterolemia. BMC Med. Genet. 11, 115 (2010).
69. Zuk, O. et al. Searching for missing heritability: Designing rare variant association studies. Proc. Natl Acad. Sci. USA 111, E455–E464 (2014).
70. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. Nature 562, 203–209 (2018).
71. Soria, L. F. et al. Association between a specific apolipoprotein B mutation and familial defective apolipoprotein B-100. Proc. Natl Acad. Sci. USA 86, 587–591 (1989).
72. Conomos, M. P., Reiner, A. P., Weir, B. S. & Thornton, T. A. Model-free estimation of recent genetic relatedness. Am. J. Hum. Genet. 98, 127–148 (2016).
73. Gogarten, S. M. et al. Genetic association testing using the GENESIS R/Bioconductor package. Bioinforma. Oxf. Engl. 35, 5346–5348 (2019).
74. Das, S. et al. Next-generation genotype imputation service and methods. Nat. Genet. 48, 1284–1287 (2016).
75. Loh, P.-R. et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat. Genet. 48, 1443–1448 (2016).
76. Fuchsberger, C., Abecasis, G. R. & Hinds, D. A. minimac, faster genotype imputation. Bioinformatics 31, 782–784 (2015).
77. UK10K Consortium, et al. Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. Nat. Commun. 6, 8111 (2015).
78. The Haplotype Reference Consortium. A reference panel of 64,976 haplotypes for genotype imputation. Nat. Genet. 48, 1279–1283 (2016).
79. Halldorsson, B. V. et al. The sequences of 150,119 genomes in the UK biobank. Nature 507, 732–740 (2022).
80. Zhou, W. et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat. Genet. 50, 1335–1341 (2018).
81. Pulit, S. L., de With, S. A. J. & de Bakker, P. I. W. Resetting the bar: statistical significance in whole-genome sequencing-based association studies of global populations. Genet. Epidemiol. 41, 145–151 (2017).
82. Han, B. & Eskin, E. Random-effects model aimed at discovering functional variants. Bioinformatics 37, 587–595 (2021).
83. Li, Z. et al. A framework for detecting noncoding rare variant associations of large-scale whole-genome sequencing studies. bioRxiv (2021). doi: https://doi.org/10.1101/2021.11.05.467531.
84. Chen, H. et al. Control for population structure and relatedness for genome-wide complex trait analysis. Am. J. Hum. Genet. 88, 76–82 (2011).
85. Gogarten, S. M. et al. Genetic association testing using the GENESIS R/Bioconductor package. Bioinforma. Oxf. Engl. 35, 5346–5348 (2019).
86. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012).
87. The Fantom Consortium. et al. An atlas of active enhancers across human cell types and tissues. Nature 507, 455–461 (2014).
88. Fishlievich, S. et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database J. Biol. Databases Curation 2017, (2017).
89. Wu, M. C. et al. Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet. 89, 82–93 (2011).
90. Madsen, B. E. & Browning, S. R. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 5, e1000384 (2009).
91. Liu, Y. et al. ACAT: a fast and powerful p value combination method for rare-variant analysis in sequencing studies. Am. J. Hum. Genet. 104, 410–421 (2019).
92. Buncieli, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 47, D1005–D1012 (2019).
93. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. Nat. Commun. 8, 1826 (2017).
94. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. PLoS Comput. Biol. 11, e1004219 (2015).
95. Szklarczyk, D. et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 47, D607–D613 (2019).
96. GTEx Consortium. Genetic effects on gene expression across human tissues. Nature 550, 204–213 (2017).
97. Giambartolomé, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. 10, e1004383 (2014).
98. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. 88, 76–82 (2011.).
M.E.M. receives funding from Regeneron Pharmaceutical Inc. unrelated to this work. S.A. has employment and equity in 23andMe, Inc. The spouse of C.J.W. works at Regeneron. S.A.L. is a full-time employee of Novartis as of July 18, 2022. S.A.L. has received sponsored research support from Bristol Myers Squibb, Pfizer, Boehringer Ingelheim, Fitbit, Medtronic, Premier, and IBM, and has consulted for Bristol Myers Squibb, Pfizer, Blackstone Life Sciences, and Invitae. X. Lin is a consultant of AbbVie Pharmaceuticals and Verily Life Sciences. The remaining authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-33510-7.

Correspondence and requests for materials should be addressed to Gina M. Peloso or Pradeep Natarajan.

Peer review information Nature Communications thanks David Meyre and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022

Margaret Sunitha Selvaraj 1,2,3, Xihao Li 4, Zilin Li 5, Akhil Pampana 5, David Y. Zhang 5,6, Joseph Park 5,6, Stella Aslibekyan 7, Joshua C. Bis 8, Jennifer A. Brody 8, Brian E. Cade 8, Lee-Ming Chuang 9, Ren-Hua Chung 10, Joanne E. Curran 11, Lisa de las Fuentes 13,14, Paul S. de Vries 15, Ravindranath Duggirala 12, Barry I. Freedman 16, Mariaelisa Graff 17, Xiuxing Guo 18, Nancy Heard-Costa 19, Bertha Hidalgo 17, Chi-Min Hwu 20, Marguerite R. Irvin 17, Tanika N. Kelly 21,22, Brian G. Kral 23, Leslie Lange 24, Xiaohui Li 25, Martin Lisa 28, Steven A. Lubitz 12,16, Ani W. Manichaikul 26, Preuss Michael 28, May E. Montasser 29, Alanna C. Morrison 15, Take Naseri 30, Jeffrey R. O’Connell 29, Nicholette D. Palmer 31, Patricia A. Peyser 32, Muagututia S. Reupena 33, Jennifer A. Smith 32, Xiao Sun 21, Kent D. Taylor 18, Russell P. Tracy 24, Michael Y. Tsai 35, Zhe Wang 28, Yuxuan Wang 36, Wei Bao 37, John T. Wilkins 38, Lisa R. Yanek 23, Wei Zhao 22, Donna K. Arnett 39, John Blangero 12, Eric Boerwinkle 16, Donald W. Bowden 31, Yi-De Ida Chen 40, Adolfo Correa 41, L. Adrienne Cupples 36, Susan K. Dutcher 42, Patrick T. Ellinor 1,2,6, Myriam Fornage 43, Stacey Gabriel 44, Soren Germer 45, Richard Gibbs 46, Jianguo He 21,22, Robert C. Kaplan 47,48, Shany L. Kardia 49, Ryan Kim 49, Charles Kooperberg 48, Ruth J. F. Loos 28,50, Karine A Viazou-Martino 25, Risaka A. Mathias 23, Stephen T. McGarvey 52, Braxton D. Mitchell 29,53, Deborah Nickerson 54, Kari E. North 17, Bruce M. Psaty 8,55,56, Susan Redline 9, Alexander P. Reiner 35,54, Ramachandran S. Vasan 57,58,59, Stephen S. Rich 27, Cristen Willer 50, Jerome I. Rotter 18, Daniel J. Rader 5,6,11, Xihong Lin 2,4,6, NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium*, Gina M. Peloso 36,220 & Pradeep Natarajan 1,2,3,220

1Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA 02114, USA. 2Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA. 3Department of Medicine, Harvard Medical School, Boston, MA 02115, USA. 4Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA. 5Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA. 6Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA. 7Department of Epidemiology, University of Alabama at Birmingham School of Public Health, Birmingham, AL, USA. 8Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA, USA. 9Department of Medicine, Brigham and Women’s Hospital, Harvard Medical School, Boston, MA, USA. 10Department of Internal Medicine, National Taiwan University Hospital, Taipei, Taiwan. 11Institute of Population Health Sciences, National Health Research Institutes, Zhunan 350, Taiwan. 12Department of Human Genetics and South Texas Diabetes and Obesity Institute, University of Texas Rio Grande Valley School of Medicine, Brownsville, TX 78520, USA. 13Department of Medicine, Cardiovascular Division, Washington University School of Medicine, St. Louis, MO, USA. 14Division of Biostatistics, Washington University School of Medicine, St. Louis, MO, USA. 15Human Genetics Center, Department of Epidemiology, Human Genetics, and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, USA. 16Department of Internal Medicine, Section on Nephrology, Wake Forest School of Medicine, Winston-Salem, NC 27157, USA. 17Department of Epidemiology, UNC Chapel Hill, Chapel Hill, NC, USA. 18The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA, USA. 19Department of Neurology, Boston university School of Medicine, Boston, MA, USA. 20Section of Endocrinology and Metabolism, Department of Medicine, Taipe Veterans General Hospital, Taipei, Taiwan. 21Department of Epidemiology, Tulane University School of Public Health and Tropical Medicine, New Orleans, LA 70112, USA. 22Tulane University Translational Science Institute, New Orleans, LA 70112, USA. 23Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA. 24Division of Biomedical
