

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

Margaret Sunitha Selvaraj^{1,2,3}, Xihao Li⁴, Zilin Li⁴, Akhil Pampana², David Y Zhang^{5,6}, Joseph Park^{5,6}, Stella Aslibekyan⁷, Joshua C Bis⁸, Jennifer A Brody⁸, Brian E Cade⁹, Lee-Ming Chuang¹⁰, Ren-Hua Chung¹¹, Joanne E Curran¹², Lisa de las Fuentes^{13,14}, Paul S de Vries¹⁵, Ravindranath Duggirala¹², Barry I Freedman¹⁶, Mariaelisa Graff¹⁷, Xiuqing Guo¹⁸, Nancy Heard-Costa¹⁹, Bertha Hidalgo⁷, Chii-Min Hwu²⁰, Marguerite R Irvin⁷, Tanika N Kelly^{21,22}, Brian G Kral²³, Leslie Lange²⁴, Xiaohui Li¹⁸, Martin Lisa²⁵, Steven A Lubitz^{1,26}, Ani W Manichaikul²⁷, Preuss Michael²⁸, May E Montasser²⁹, Alanna C Morrison¹⁵, Take Naseri³⁰, Jeffrey R O'Connell²⁹, Nicholette D Palmer³¹, Patricia A Peyser³², Muagututia S Reupena³³, Jennifer A Smith³², Xiao Sun²¹, Kent D Taylor¹⁸, Russell P Tracy³⁴, Michael Y Tsai³⁵, Zhe Wang²⁸, Yuxuan Wang³⁶, Bao Wei³⁷, John T Wilkins³⁸, Lisa R Yanek²³, Wei Zhao³², Donna K Arnett³⁹, John Blangero¹², Eric Boerwinkle¹⁵, Donald W Bowden³¹, Yii-Der Ida Chen⁴⁰, Adolfo Correa⁴¹, L Adrienne Cupples³⁶, Susan K Dutcher⁴², Patrick T Ellinor^{1,26}, Myriam Fornage⁴³, Stacey Gabriel⁴⁴, Soren Germer⁴⁵, Richard Gibbs⁴⁶, Jiang He^{21,22}, Robert C Kaplan^{47,48}, Sharon LR Kardia³², Ryan Kim⁴⁹, Charles Kooperberg⁴⁸, Ruth J. F. Loos^{28,50}, Karine Martinez⁵¹, Rasika A Mathias²³, Stephen T McGarvey⁵², Braxton D Mitchell^{29,53}, Deborah Nickerson⁵⁴, Kari E North¹⁷, Bruce M Psaty^{8,55,56}, Susan Redline⁹, Alexander P Reiner^{55,48}, Ramachandran S Vasan^{57,58,59}, Stephen S Rich²⁷, Cristen Willer⁶⁰, Jerome I Rotter¹⁸, Daniel J Rader^{5,6,61}, Xihong Lin^{2,4,62}, NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, Gina M Peloso^{36 # \$}, Pradeep Natarajan^{1,2,3 # \$}

1. Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, USA, 02114
2. Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA, 02142
3. Department of Medicine, Harvard Medical School, Boston, MA, USA, 02115
4. Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA, 02115
5. Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA, 19104
6. Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA, 19104
7. Department of Epidemiology, University of Alabama at Birmingham School of Public Health
8. Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA

- 27 9. Department of Medicine, Brigham and Women's Hospital, Harvard Medical School
- 28 10. Department of Internal Medicine, National Taiwan University Hospital, Taipei, Taiwan
- 29 11. Institute of Population Health Sciences, National Health Research Institutes, Zhunan, 350, Taiwan
- 30 12. Department of Human Genetics and South Texas Diabetes and Obesity Institute, University of Texas Rio Grande
- 31 Valley School of Medicine, Brownsville, TX 78520
- 32 13. Department of Medicine, Cardiovascular Division, Washington University School of Medicine, St. Louis, MO
- 33 14. Division of Biostatistics, Washington University School of Medicine, St. Louis, MO
- 34 15. Human Genetics Center, Department of Epidemiology, Human Genetics, and Environmental Sciences, School of
- 35 Public Health, The University of Texas Health Science Center at Houston, Houston, Texas, USA
- 36 16. Department of Internal Medicine, Section on Nephrology, Wake Forest School of Medicine, Winston-Salem, NC,
- 37 USA, 27157
- 38 17. Department of Epidemiology, UNC Chapel Hill
- 39 18. The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, The Lundquist
- 40 Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA USA
- 41 19. Department of Neurology, Boston university School of Medicine
- 42 20. Section of Endocrinology and Metabolism, Department of Medicine, Taipei Veterans General Hospital
- 43 21. Department of Epidemiology, Tulane University School of Public Health and Tropical Medicine, New Orleans,
- 44 Louisiana, US, 70112
- 45 22. Tulane University Translational Science Institute, New Orleans, Louisiana, US, 70112
- 46 23. Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA 21206
- 47 24. Division of Biomedical Informatics and Personalized Medicine, Department of Medicine
- 48 25. Department of Medicine, George Washington University, Washington DC
- 49 26. Cardiovascular Disease Initiative, The Broad Institute of MIT and Harvard, Cambridge, MA 02124
- 50 27. Department of Public Health Sciences, Center for Public Health Genomics, University of Virginia, Charlottesville,
- 51 VA USA
- 52 28. The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY
- 53 29. Department of Medicine, University of Maryland School of Medicine, Baltimore, MD
- 54 30. Ministry of Health, Government of Samoa, Samoa
- 55 31. Department of Biochemistry, Wake Forest School of Medicine, Winston-Salem, NC, USA, 27157
- 56 32. Department of Epidemiology, University of Michigan, Ann Arbor, MI, USA 48109
- 57 33. Lusia i Puava ae Mapu i Fagalele, Apia, Samoa
- 58 34. Departments of Pathology & Laboratory Medicine and Biochemistry, Larner College of Medicine at the University of
- 59 Vermont, Colchester, VT USA

35. Department of Laboratory Medicine and Pathology, University of Minnesota, Minneapolis, MN USA
36. Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA, 02118
37. Department of Epidemiology, University of Iowa
38. Department of Medicine (Cardiology) and Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL
39. Dean's Office, University of Kentucky College of Public Health
40. Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center
41. Department of Population Health Science, University of Mississippi Medical Center
42. McDonnell Genome Institute
43. Brown Foundation Institute of Molecular Medicine, McGovern Medical School, The University of Texas Health Science Center at Houston, Houston, Texas, 77225
44. Broad Institute, Cambridge, Massachusetts, 02142
45. New York Genome Center, New York, New York, 10013
46. Baylor College of Medicine Human Genome Sequencing Center, Houston, Texas, 77030
47. Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx NY 10461 USA
48. Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle WA 98109
49. Psomagen
50. NNF Center for Basic Metabolic Research, University of Copenhagen, Copenhagen, Denmark
51. Illumina
52. Department of Epidemiology, International Health Institute, Brown University, Providence RI
53. Geriatrics Research and Education Clinical Center, Baltimore Veterans Administration Medical Center, Baltimore, MD
54. University of Washington, Department of Genome Sciences, Seattle, Washington, 98195
55. Department of Epidemiology, University of Washington, Seattle, WA
56. Department of Health Services, University of Washington, Seattle, WA
57. Sections of Preventive medicine and Epidemiology, Cardiovascular medicine, Department of Medicine, Boston University School of Medicine
58. Department of Epidemiology, Boston University School of Public Health
59. Framingham Heart Study
60. University of Michigan, Internal Medicine, Ann Arbor, Michigan, 48109
61. Institute for Translational Medicine and Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA, 19104
62. Department of Statistics, Harvard University, Cambridge, MA, USA, 02138

93
94 # Jointly supervised the work
95 \$ Corresponding authors
96
97
98 Word Count:
99 Abstract: 190
100 Main Text: 3919
101 Online Methods: 1902
102 Main Display Items: Figures-5, Tables-1
103
104 Correspondence:
105 Pradeep Natarajan, MD MMSc
106 185 Cambridge St
107 CPZN 3.184
108 Boston, MA 02114
109 (617) 724-3526
110 pnatarajan@mgh.harvard.edu
111 [@pnatarajanmd](https://twitter.com/pnatarajanmd)
112
113 Gina M. Peloso, PhD
114 801 Massachusetts Ave
115 Boston, MA 02118
116 (617) 358-4466
117 gpeloso@bu.edu

Abstract:

Plasma lipids are heritable modifiable causal factors for coronary artery disease, the leading cause of death globally. Despite the well-described monogenic and polygenic bases of dyslipidemia, limitations remain in discovery of lipid-associated alleles using whole genome sequencing, partly due to limited sample sizes, ancestral diversity, and interpretation of potential clinical significance. Increasingly larger whole genome sequence datasets with plasma lipids coupled with methodologic advances enable us to more fully catalog the allelic spectrum for lipids. Here, among 66,329 ancestrally diverse (56% non-European ancestry) participants, we associate 428M variants from deep-coverage whole genome sequences with plasma lipids. Approximately 400M of these variants were not studied in prior lipids genetic analyses. We find multiple lipid-related genes strongly associated with plasma lipids through analysis of common and rare coding variants. We additionally discover several significantly associated rare non-coding variants largely at Mendelian lipid genes. Notably, we detect 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 plasma lipids expanding the alleles linked to lipids for multiple ancestries and characterize a clinically-relevant rare non-coding variant model for lipids.

Introduction

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^{1,2,3,4,5}. 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 communities⁶. 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^{1,7,8,9,10,11,12}. 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^{14,15,16}.

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 (31.50% non-European samples). 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 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 28M individual autosomal variants (minor allele count [MAC] > 20) and aggregated rare autosomal variant (minor allele

frequency [MAF] < 1%) association testing for 417M variants (**Fig. 1, Supplementary Fig. 1**). Secondly, 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 approximately 45,000 independent samples with array-based genotyping imputed to TOPMed. 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.

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, 4,719 (7.11%) Asian, and 1,182 (1.78%) Samoan (**Supplementary Table 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 *APOC3*^{7,8,14}. 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 before^{12,14} which are further detailed in the **Methods**. The adjusted normalized lipid concentrations for the four lipids were similar across the cohorts.

A total of 428M variants passed the quality criteria with an average depth >30X in 22 autosomes. 202M variants were singletons, 417M were rare variants (MAF<1%), and 11M were common or low frequency variants (MAF>1%) with differences by cohort (**Supplementary Table 2**).

Individual variant associations with lipids

Approximately 28M variants with MAC > 20 were individually associated with LDL-C, HDL-C, TC and TG. We used p-value < 5x10⁻⁹ to claim significance as previously recommended for whole genome sequencing common variant association studies^{14,19}. The total numbers of variants that met our significance threshold were 2,214, 2,314, 2,697 and 2,442 for LDL-C, HDL-C, TC and TG, respectively, and after clumping²⁰ the numbers of variants were 357, 338, 324, and 289, respectively. Of these variants, most were previously demonstrated to be associated with plasma lipids either at the variant- or locus-level¹⁵ (**Supplementary Table 3, 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)¹⁵ as well as the GWAS Catalog (All associations(v1.0) file

198 dated 06/04/2020) (**Fig. 2**). We clumped (window 250 kb, r^2 0.5) significant variants using Plink²⁰ and queried these in the
199 GWAS Catalog and MVP. Among genome-wide significant variants, we tabulated ‘known-position’ (variant previously
200 associated), ‘known-loci’ (variants not previously significantly associated with the corresponding lipid phenotype but within
201 500 kb of a known locus, thereby representing additional allelic heterogeneity), and ‘novel’ variants (variants not in a
202 known lipid locus) (**Supplementary Table 3**).

203 The novel variants, tabulated in **Table 1**, are divided into two subsets – ‘novel variants’ or variants at established
204 lipid loci for another lipid phenotype, and ‘novel loci,’ representing new loci associations for any lipid phenotype. For
205 example, the *CETP* locus is well-known for its link to HDL-C, but we now found that rs183130 (16:56957451:C:T, MAF
206 28.3%) at the locus is associated with LDL-C. Similarly, the variants rs7140110 (13:113841051:T:C, MAF 27.8%) *GAS6*
207 and rs73729083 (7:137875053:T:C, MAF 4.5%) *CREB3L2* are newly associated with TC, while previous studies showed
208 that rs73729083 associates with LDL-C²¹ and rs7140110 associates with LDL-C²² and TG²³. Index variants at novel loci
209 were typically low frequency variants often observed in non-European ancestries, so we also conducted ancestry-specific
210 association analyses for these alleles (**Supplementary Table 4**). For example, 12q23.1 (12:97352354:T:C, MAF 0.3%)
211 and 4q34.2 (4:176382171:C:T, MAF 0.2%) associations with LDL-C are specific to Hispanic (MAF 1.3%) and Black (MAF
212 0.6%) populations, respectively and among Asians (MAF 1.5%) alone, 11q13.3 (11:69219641:C:T, MAF 0.2%) was
213 associated with TG. One variant initially passing the novel locus filter for HDL-C (*RNF111* - rs112147665, beta = 8.664, p-
214 value = 6.51×10^{-10}), was in LD ($r=0.7$) with LIPC p.Thr405Met (rs113298164) which is known to be associated with HDL-

C. The lead variant from MVP was 604 kb away from the *RNF111* variant but the rare *LIPC* missense variant p.Thr405Met was 421 kb away. Conditional analysis accounting for *LIPC* p.Thr405Met rendered the non-coding variant near *RNF111* variant non-significant (beta = 4.351, p-value = 2.47×10^{-02}), therefore we reclassified *RNF111* variant as a known-position variant. Ancestry-specific GWAS did not yield additional novel loci beyond our larger trans-ancestry GWAS. Majority of genome significant single variants were captured by previous lipid GWAS¹⁵, but ancestry specific novel-hits are unique to WGS TOPMed data.

Due to the paucity of available diverse WGS datasets with lipids of comparable size, we pursued replication with two genome-wide array-based genotyped datasets imputed to TOPMed WGS^{17,24}: Mass General Brigham (MGB) Biobank (N=25,137) and Penn Medicine Biobank (N=20,079)^{25,26}, the replication cohorts had diverse ancestry distribution, where non-European samples accounted for 15.77% in MGB Biobank and 51.20% in Penn Medicine Biobank (**Supplementary Table 5**). We brought seven putative novel variants with p-values $< 5 \times 10^{-9}$ forward for replication. The three common variants, rs183130 (*CETP*), rs7140110 (*GAS6*) and rs73729083 (*CREB3L2*), that were associated with both LDL-C and TC in TOPMed also replicated in MGB and two (rs183130, rs73729083) replicated in Penn Biobank at an alpha level of 0.05 and consistent direction of effect (**Table 1**). The two variants that were associated in both replication studies were most significantly associated among African Americans in TOPMed (rs183130: beta = -2.762 mg/dL, p-value = 5.71×10^{-07} ; rs73729083: beta = -3.725 mg/dL, p-value = 5.25×10^{-07}). Low-frequency variants from specific ancestry groups associated with lipids in TOPMed were not replicated but we cannot rule out the possibility of reduced power due

to general underrepresentation of non-white ancestry groups in the replication data. In exploratory analyses, we extended the same approach for variants discovered to have $5 \times 10^{-9} < \text{p-value} < 5 \times 10^{-7}$ but did not observe replication (**Supplementary Table 6**).

***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 significance¹⁵. Pharmacologic *CETP* inhibitors have shown strong associations with increased HDL-C but mixed effects for LDL-C reduction in clinical trials^{27,28,29,30}. We found that the *CETP* locus variant rs183130 (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^{-283}), and rs183130 was associated with 3.9 mg/dL increased HDL-C (p-value $< 1 \times 10^{-284}$) 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^{-07}) (**Supplementary Table 7**). We next investigated variants by their HDL-C and LDL-C effects within this locus (+/- 500kb of rs183130 and rs3764261) (**Fig. 3**). We identified five variants showing at least suggestive (p-value $< 5 \times 10^{-07}$) 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 3 tissues (Liver, Adipose Subcutaneous and Adipose Visceral [Omentum]) from GTEx³¹. We analyzed 5 LDL-C and 441 HDL-C associated (p-values < 5x10⁻⁰⁷) 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: ρ -0.933, p-value 4.01x10⁻¹⁷; Adipose Subcutaneous: ρ -0.762, p-value 8.87x10⁻¹²; Adipose Visceral: ρ -0.739, p-value 5.52x10⁻¹⁰) (**Supplementary Fig. 3**). However, *CETP* expression effects were not significantly correlated with LDL-C (Liver: ρ 0.007, p-value 0.99; Adipose Subcutaneous: ρ 0.344, p-value 0.57; Adipose Visceral: ρ -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 (< 5x10⁻⁰⁷), 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: ρ 0.957, p-value 2.28x10⁻⁰⁸; Adipose Subcutaneous: ρ 0.922, p-value 1.34 x10⁻¹⁵; Adipose Visceral: ρ 0.868, p-value 6.09x10⁻¹¹).

Rare variant aggregates associated with lipids

I) 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=2.5x10⁻⁰⁶ for coding and non-coding

gene-centric rare variant analyses (**Supplementary Fig. 4**). 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 nonconditional analysis (**Supplementary Table 8-9**). We secondarily conditioned our significant aggregate sets on variants individually associated with lipid levels from the GWAS catalog, MVP 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 Table 10-11**).

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 Table 10**). *RFC2* putative loss-of-function mutations (combined allele frequency < 0.002%) were significantly associated with triglycerides (p-value 2×10^{-06}) 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 *EHD3* promoter-DHS for total cholesterol (**Supplementary Table 11**). 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^{-05}).

II) 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 2.6M regions statically at 2 kb size and 1 kb window overlap by the sliding window approach. Statistical significance was assigned at $0.05/(2.6 \times 10^6) = 1.88 \times 10^{-08}$. We identified 28 significantly associated windows with at least one lipid phenotype. After conditioning on variants individually associated with the corresponding lipid phenotype, we identified two regions at *LDLR* still significantly associated with both total cholesterol and LDL-C although these regions included both intronic and exonic variants (**Supplementary Table 12**). *LDLR* intron 1, which encodes *LDLR-AS1* (LDLR antisense RNA 1) on the minus strand, had suggestive evidence for association with TC (p-value 3.17×10^{-6}) with -2.76 mg/dL reduction in TC. A prior study identify that a common variant (rs6511720, MAF 0.11) in *LDLR* intron 1 is associated with increased *LDLR* expression in a luciferase assay and reduction in LDL-C³². When adjusting for rs6511720, the significance improved (p-value 1.43×10^{-8}) with -3.35 mg/dL reduction in TC.

For dynamic window scanning of the genome, we implemented the SCANG method³³. The SCANG procedure accounts for multiple testing by controlling the genome-wide error rate (GWER) at 0.1³³. In the dynamic window-based workflow, STAAR-O detected 51 regions significantly associated with at least one lipid phenotype after conditioning on

299 known variants (**Supplementary Table 13**). Most of the regions mapped to known Mendelian lipid genes, including *LCAT*
300 (8.7×10^{-13}) for HDL-C, and *LDLR* (2.4×10^{-28} , 7.3×10^{-26}) and *PCSK9* (2.9×10^{-12} , 5.5×10^{-12}) for LDL-C and TC, respectively.
301 Exon 4 aggregates of *LDLR* were specifically associated with 20 mg/dL increase in LDL-C. *PCSK9* Exon2-Intron2 region
302 spanning chr1:55043782-55045960 had significantly reduced LDL-C by 6 mg/dL (p-value = 3×10^{-13}), and the effect
303 persisted even with only Intron 2 rare variants of *PCSK9* (-5 mg/dL, p-value = 2×10^{-8}). Strikingly, in secondary analyses,
304 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
305 *LDLR* Introns 16 and 17 (+17 mg/dL, p-value = 0.02), similar to rare coding *LDLR* mutations. While 32 of the significant
306 dynamic windows also included exonic regions, there were also several dynamic windows significantly independently
307 associated with lipids not containing exonic regions. For example, four non-coding windows (two overlapping) at 2p24.1,
308 which harbors the Mendelian *APOB* gene, were significantly associated with LDL-C. Intronic non-coding regions were
309 associated with both LDL-C and TC -associated windows at *LPAL2-LPA-SLC22A3*; for example *LPAL2* Intron 3 was
310 associated with a 3.7 mg/dL increase in TC. Non-coding TC-associated significant dynamic windows were near
311 *TOMM40/APOE*. One rare variant signal observed was at *TOMM40* Intron 6, where the 'poly-T' variant in this region is on
312 the *APOE4* haplotype and influences expressivity for Alzheimer's disease age-of-onset^{34,35}. For HDL-C, we identified
313 significant non-coding windows at an intergenic region near *LPL* and *CD36* Intron 4. In the generation of the
314 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 model³⁶. Several regions significant in SCANG were not even nominally significant in burden association analyses indicating the likelihood of causal variants with bidirectional effects.

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 conditional analyses adjusting the gene-centric non-coding results for rare coding variants (MAF<1%) within known lipid monogenic genes (**Supplementary Table 14**). 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^{-11}) and with suggestive evidence for LDL-C (p-value= 1.57×10^{-06}). In a similarly adjusted model, *LDLR*-enhancer-DHS showed a strong association with TC (p-value 5.18×10^{-12}). 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. 5**). 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 genome-wide rare variant analyses, we leveraged current knowledge of 22 Mendelian lipid genes for more focused exploratory analyses¹⁴. 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 associations (**Supplementary Table 15, Supplementary Fig. 6**). 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).

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 interrogated 428M 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 descent³⁷. Genetic association of plasma lipids using arrays or whole exome sequencing among Europeans have yielded several important insights regarding plasma lipids and the causal determinants of CAD^{5,4,38,39,40}. Similar increasingly larger studies among non-Europeans have often yielded new genetic loci and sometimes new genes, such as *PCSK9*^{1,15,41,42,16}. Such differences have also led to concerns about the use of polygenic risk scores gleaned from much larger European GWAS of complex traits for non-Europeans⁴³. Aided by the availability of WGS data, we identify new putative loci associated with lipids in non-Europeans. Furthermore, our

study enabled the discovery of several novel alleles at known loci, with richly distinct allelic heterogeneity across ancestry groups. For example, HDL-C-raising *CETP* locus variants linked to *CETP* gene expression were only associated with LDL-C reduction among those of African ancestry. While all pharmacologic *CETP* inhibitors increase HDL-C, only those that decrease LDL-C also reduce cardiovascular disease risk^{27,28,30,29}. Given the contribution of genetic differences, clinical trials with more diverse samples would show insights.

Our study now provides increasingly robust evidence for a rare non-coding variant model for complex traits. Our rare non-coding variant associations in both gene-centric and sliding window models were largely restricted to the introns of Mendelian lipid genes with prior robust rare coding variant support consistent with biologic plausibility⁴⁴. Rare intronic variants, often impacting splicing, have been previously implicated in afflicted Mendelian families or small exceptional case series, often through candidate gene approaches^{45,46,47,48}. We discovered one example of a rare non-coding signal without prior rare coding support – i.e., *EHD3*. We obtained estimates of phenotypic effect using burden tests. For most regions, even nominal significance was not detected using burden testing indicating the likelihood of variants with bidirectional effects further complicating clinical interpretation. When burden signals were detected, observed effects were 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^{49,50,51}.

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 non-coding sequence in 16,324 TOPMed participants, and found that rare non-coding gene regions associated with lipid levels, but they were not independent of individually associated single variants¹⁴. Using STAAR, we observed putative rare non-coding variant associations for lipids independent of individual variants associated with lipids in TOPMed.

WGS can improve diagnostic yield beyond the current standard of next-generation gene panel sequencing for dyslipidemias. A very small fraction with severe hypercholesterolemia and features consistent with strong genetic predisposition have a familial hypercholesterolemia variant in *LDLR*, *APOB*, or *PCSK9*^{52,53}. The presence of familial hypercholesterolemia variants is independently prognostic for CAD, beyond lipids, and merits the consideration of more costly lipid-lowering medications^{52,53,54,55}. We now observe that rare *LDLR* variants in Introns 2, 3, 16, and 17 lead to approximately 0.5 standard deviation increase in LDL-C, approximating effects observed with clinically reported exonic familial hypercholesterolemia variants in *LDLR*⁵⁵. 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 study^{56,57}. A WGS approach to lipid disorders, particularly for familial 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 tools⁵⁸. 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^{1,8,59,60}. Using SCANG³³, 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*^{61,62,63,64}. Conventional rare coding variant analyses aggregate all exonic variants for a transcript. Here, we demonstrate an opportunity for exon-level rare variant association testing.

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 power⁶⁵. 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. 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. 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.

Online Methods

Dataset

i) Contributing studies

The discovery cohort includes whole genome sequenced (WGS) data of 66,329 samples from 21 studies of the Trans-Omics for Precision Medicine (TOPMed) program with blood lipids available¹⁷. 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.

415 The studies included in the current dataset, along with their abbreviations and sample sizes, contains the Old Order
 416 Amish (Amish, n=1,083), Atherosclerosis Risk in Communities study (ARIC, n=8,016), Mt Sinai BioMe Biobank (BioMe,
 417 n=9,848), Coronary Artery Risk Development in Young Adults (CARDIA, n=3,056), Cleveland Family Study (CFS, n=579),
 418 Cardiovascular Health Study (CHS, n=3,456), Diabetes Heart Study (DHS, n=365), Framingham Heart Study (FHS,
 419 n=3,992), Genetic Studies of Atherosclerosis Risk (GeneSTAR, n=1,757), Genetic Epidemiology Network of Arteriopathy
 420 (GENOA, n=1,046), Genetic Epidemiology Network of Salt Sensitivity (GenSalt, n=1,772), Genetics of Lipid-Lowering
 421 Drugs and Diet Network (GOLDN, n=926), Hispanic Community Health Study - Study of Latinos (HCHS_SOL, n=7714),
 422 Hypertension Genetic Epidemiology Network and Genetic Epidemiology Network of Arteriopathy (HyperGEN, n=1,853),
 423 Jackson Heart Study (JHS, n=2,847), Multi-Ethnic Study of Atherosclerosis (MESA, n=5,290), Massachusetts General
 424 Hospital Atrial Fibrillation Study (MGH_AF, n=683), San Antonio Family Study (SAFS, n=619), Samoan Adiposity Study
 425 (SAS, n=1,182), Taiwan Study of Hypertension using Rare Variants (THRV, n=1,982) and Women's Health Initiative
 426 (WHI, n=8,263) (Please see **Supplementary Text** for additional details). The multi-ancestral data set included individuals
 427 from White (44%), Black (26%), Hispanic (21%), Asian (7%), and Samoan (2%) ancestries. Study participants granted
 428 consent per each study's Institutional Review Board (IRB) approved protocol. Secondly, these data were analyzed
 429 through a protocol approved by the Massachusetts General Hospital IRB. **Supplementary Table 1** details the number of
 430 samples across different studies and ancestral group.

The replication cohorts include TOPMed-imputed genome-wide array data from the Mass General Brigham (MGB) and Penn Medicine Biobanks which consist of 25,137 samples and 20,079 samples respectively^{26,25}. We curated the MGB Biobank and Penn Medicine Biobank phenotype data from the corresponding electronic health record databases in accordance with corresponding institutional IRB approvals. Consent was previously obtained from each participant regarding storage of biological specimens, genetic sequencing, access to all available electronic health record (EHR) data, and permission to recontact for future studies. The MGB Biobank consists of 54% and Penn Medicine Biobank consist of 52% female samples and average ages were 55.89 years and 58.35 years, respectively (**Supplementary Table 5**).

ii) Phenotypes

The primary outcomes in this study included LDL cholesterol (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 done¹⁴. Triglycerides remained natural log transformed for analysis. Fasting status was accounted for with an indicator variable.

We harmonized the phenotypes across each cohort¹⁸ and inverse rank normalization of the residuals of each race within each cohort scaled by the standard deviation of the trait and adjusted for covariates¹². We included covariates such

as age, age², sex, PC1-11, study-groups as well as Mendelian founder lipid variants *APOB* p.R3527Q and *APOC3* p.R19X for the Amish cohort^{7,66,8}. **Supplementary Table 1** provides the distributions of each of the four lipid phenotypes by cohort, ancestral groups, and gender. We executed similar steps of phenotype harmonization and normalization for the replication cohorts. Additionally, 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.

iii) 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 Text**). The reads were aligned to human genome build GRCh38 using a common pipeline across all centers (BWA-MEM).

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 Freeze8 (<https://www.nhlbiwgs.org/topmed-whole-genome-sequencing-methods-freeze-8>)¹⁷. The resulting BCF files were split by

study and consent group for distribution to approved dbGaP users. Quality control was performed at each stage of the process, poor variant quality was indicated by missing rate >20%, mappability score <0.8, mean depth of coverage >500X, and Ti/Tv ratio, by the Sequencing Centers, the IRC and the TOPMed Data Coordinating Center (DCC). 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)⁶⁷. Using the 'Genesis' R package (version:2.20.1)⁶⁸ we generated subsetted 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>))¹³. 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 parameters^{69,70}. This study applied the Version-r2 of the imputation panel, it includes 97,256 reference samples and ~300M genetic variants. The Illumina Global Screening array was used to genotype the Penn Medicine Biobank. Penn Medicine Biobank TOPMed imputation was performed using EAGLE⁷⁰ and Minimac⁷¹ software. For this study we downloaded variants that passed a min R² threshold of 0.3.

Single Variant Association

We performed genome-wide single variant association analyses for autosomal variants with minor allele frequency (MAF) greater than 0.1% across the dataset with each of the four lipid phenotypes. We implemented the SAIGE-QT⁷² 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-PC11, 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 ($\alpha = 5 \times 10^{-9}$)⁷³. Replication was performed in the MGB and Penn Medicine Biobanks where models were fitted as indicated above. Additionally, 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. Additionally, we carried out meta-analysis using fixed effects model based on inverse-variance-weighted effect size for the two replication cohorts using METASOFT⁷⁴.

Rare variant association test

499 We performed rare variant association (RVA) using the Variant-Set Test for Association using Annotation
500 infoRmation (STAAR) pipeline¹³ from STARtopmed R package. STAARpipeline is a regression-based framework that
501 permits adjustment of covariates, population structure, and relatedness by fitting linear and logistic mixed models for
502 quantitative and dichotomous traits⁷⁵. We chose STAAR to leverage the annotation information and associated scores
503 that were available for TOPMed Freeze 8 data to incorporate the analysis of rare non-coding variants from whole genome
504 sequencing. The method implements genome-wide scanning of rare variants (MAF<0.01) in gene-centric and region-
505 based workflows. For each variant set, STAARpipeline calculates a set-based p-value using the STAAR method, which
506 increases the analysis power by incorporating multiple *in silico* variant functional annotation scores capturing diverse
507 genomic features and biochemical readouts¹³. We aggregated rare variants into multiple groups for coding and non-
508 coding analyses. For the coding region, we defined five different aggregate masks of rare variants 1) plof (putative loss-of-
509 function), plof-Ds (putative loss-of-function or disruptive missense), missense, disruptive-missense, and synonymous. For
510 the non-coding regions, we used seven rare variant masks: 1) promoter-CAGE (promoter variants within Cap Analysis of
511 Gene Expression [CAGE] sites^{77,78}), 2) promoter-DHS (promoter variants within DNase hypersensitivity [DHS] sites⁷⁹), 3)
512 enhancer-CAGE (enhancer within CAGE sites⁷⁸), 4) enhancer-DHS (enhancer variants within DHS sites⁸⁰), 5) UTR (rare
513 variants in 3' untranslated region [UTR] and 5' UTR untranslated region), 6) upstream, and 7) downstream. Detailed
514 explanations of the regions defined based on these masks is discussed within STAARpipeline¹³.

In the gene-centric workflows, for both coding (within exonic boundaries) and non-coding (promoter: +/- 3kb window of transcription starting site (TSS), enhancer: GeneHancer predicted regions) 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 SCANG³³ 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 STAARtopmed R-package implements multiple rare-variants aggregate tests including SKAT⁸¹, Burden⁸² and ACAT⁸³ and integrates them as STAAR-O¹³. We performed gene-centric and region-based rare variant tests using annotated GDS files of TOPMed.

We completed aggregate tests as three-step process. In the first step, we fitted a null model using glmkin() function in STAARtopmed. 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 Catalog⁸⁴ and Million Veterans Program (MVP)¹⁵ GWAS summary statistics. To obtain effect estimates of significant aggregate sets, we associated the cumulative genotypes (binary scores) based on the variants forming the aggregates and used Glmm.Wald

test from GMMAT R package⁷⁵(version 1.3.1). For significantly-associated window-based rare variant aggregations, we trimmed the exonic variants and estimated the effects with only non-coding variants.

***CETP* 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 *CETP*. We downloaded GTEx eQTL build 38 (version8) data for Liver, Adipose Subcutaneous and Adipose Visceral (Omentum) tissues from GTEx Portal on 16/APR/2020⁸⁵. We selected eQTLs with nominal significance (p-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-value < 5×10^{-7}) for LDL-C or HDL-C effects within 500 kb of the lead locus variant. We performed Pearson correlation tests between the lipid effect estimates and gene expression effects (slope) from GTEx.

Acknowledgments

Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). P.N. is supported by grants from the National Heart, Lung, and Blood Institute (R01HL142711, R01HL148050, R01HL151283, R01HL148565, R01HL135242, R01HL151152), Fondation Leducq (TNE-18CVD04), and Massachusetts General Hospital (Paul and Phyllis Fireman Endowed Chair in Vascular Medicine). The Amish studies were supported by NIH grants R01 AG18728, U01 HL072515, R01 HL088119, R01 HL121007, and P30 DK072488. The Atherosclerosis Risk in Communities (ARIC) study has been funded in whole or in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, Department of Health and Human Services (contract numbers HHSN268201700001I, HHSN268201700002I, HHSN268201700003I, HHSN268201700004I and HHSN268201700005I). The authors thank the staff and participants of the ARIC study for their important contributions. The Mount Sinai BioMe Biobank (BioMe) has been supported by The Andrea and Charles Bronfman Philanthropies and in part by Federal funds from the NHLBI and NHGRI (U01HG00638001; U01HG007417; X01HL134588). Coronary Artery Risk Development in Young Adults (CARDIA) Study (phs001612) was performed at the Baylor College of Medicine Human genome Sequencing Center (contract HHSN268201600033I). Core support including centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Core support including phenotype harmonization, data management, sample-identity QC, and general program coordination were

560 provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I).

561 We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed. The

562 Coronary Artery Risk Development in Young Adults Study (CARDIA) is conducted and supported by the National Heart,

563 Lung, and Blood Institute (NHLBI) in collaboration with the University of Alabama at Birmingham (HHSN268201800005I &

564 HHSN268201800007I), Northwestern University (HHSN268201800003I), University of Minnesota (HHSN268201800006I),

565 and Kaiser Foundation Research Institute (HHSN268201800004I). Cleveland Family Study (CFS) is supported by grants

566 from the NHLBI (HL046389, HL113338, and 1R35HL135818). Cardiovascular Health Study (CHS) was supported by

567 contracts HHSN268201200036C, HHSN268200800007C, HHSN268201800001C, N01HC55222, N01HC85079,

568 N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086, 75N92021D00006, and grants U01HL080295

569 and U01HL130114 from the National Heart, Lung, and Blood Institute (NHLBI), with additional contribution from the

570 National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided by R01AG023629 from

571 the National Institute on Aging (NIA). A full list of principal CHS investigators and institutions can be found at CHS-

572 NHLBI.org. The content is solely the responsibility of the authors and does not necessarily represent the official views of

573 the National Institutes of Health. Diabetes Heart Study (DHS) was supported by HL92301, HL67348, NS058700,

574 AR48797, DK071891, AG058921, the General Clinical Research Center of the Wake Forest University School of

575 Medicine (RR07122, HL085989), the American Diabetes Association, and a pilot grant from the Claude Pepper Older

576 Americans Independence Center of Wake Forest University Health Sciences (AG10484). Framingham Heart Study (FHS)

577 acknowledges the support of contracts NO1-HC-25195 and HHSN268201500001I from the National Heart, Lung and
578 Blood Institute and grant supplement R01 HL092577-06S1 for this research. WGS for “NHLBI TOPMed: Whole Genome
579 Sequencing and Related Phenotypes in the Framingham Heart Study” (phs000974) was performed at the Broad Institute
580 of MIT and Harvard (HHSN268201500014C, 3R01HL092577-06S1, and 3U54HG003067-12S2). We also acknowledge
581 the dedication of the FHS study participants without whom this research would not be possible. Genetic Studies of
582 Atherosclerosis Risk (GeneSTAR) was supported by grants from the National Institutes of Health/National Heart, Lung,
583 and Blood Institute (U01 HL72518, HL087698, HL49762, HL58625, HL071025, HL112064), the National Institutes of
584 Health/National Institute of Nursing Research (NR0224103), and by a grant from the National Institutes of Health/National
585 Center for Research Resources (M01-RR000052) to the Johns Hopkins General Clinical Research Center. Genetic
586 Epidemiology Network of Arteriopathy (GENOA) was supported by the National Heart, Lung and Blood Institute
587 (HL054457, HL054464, HL054481, HL087660, and HL119443) of the National Institutes of Health. Genetic Epidemiology
588 Network of Salt Sensitivity (GenSalt) was supported by research grants (U01HL072507, R01HL087263, and
589 R01HL090682) from the National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, MD. Genetics
590 of Lipid-Lowering Drugs and Diet Network (GOLDN) biospecimens, baseline phenotype data, and intervention phenotype
591 data were collected with funding from National Heart, Lung and Blood Institute (NHLBI) grant U01 HL072524. The
592 Hispanic Community Health Study/Study of Latinos (HCHS-SOL) was carried out as a collaborative study supported by
593 contracts from the National Heart, Lung, and Blood Institute (NHLBI) to the University of North Carolina (N01-HC65233),

594 University of Miami (N01-HC65234), Albert Einstein College of Medicine (N01-HC65235), Northwestern University (N01-
595 HC65236), and San Diego State University (N01-HC65237). The Hypertension Genetic Epidemiology Network and
596 Genetic Epidemiology Network of Arteriopathy (HyperGEN) Study is part of the National Heart, Lung, and Blood Institute
597 (NHLBI) Family Blood Pressure Program; collection of the data represented here was supported by grants U01 HL054472
598 (MN Lab), U01 HL054473 (DCC), U01 HL054495 (AL FC), and U01 HL054509 (NC FC). The HyperGEN: Genetics of Left
599 Ventricular Hypertrophy Study was supported by NHLBI grant R01 HL055673 with whole-genome sequencing made
600 possible by supplement -18S1. The Jackson Heart Study (JHS) is supported and conducted in collaboration with Jackson
601 State University (HHSN268201800013I), Tougaloo College (HHSN268201800014I), the Mississippi State
602 Department of Health (HHSN268201800015I) and the University of Mississippi Medical Center
603 (HHSN268201800010I, HHSN268201800011I and HHSN268201800012I) contracts from the National Heart, Lung,
604 and Blood Institute (NHLBI) and the National Institute on Minority Health and Health Disparities (NIMHD). The
605 authors also wish to thank the staffs and participants of the JHS. Multi-Ethnic Study of Atherosclerosis (MESA) and the
606 MESA SHARe projects are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in
607 collaboration with MESA investigators. Support for MESA is provided by contracts 75N92020D00001,
608 HHSN268201500003I, N01-HC-95159, 75N92020D00005, N01-HC-95160, 75N92020D00002, N01-HC-95161,
609 75N92020D00003, N01-HC-95162, 75N92020D00006, N01-HC-95163, 75N92020D00004, N01-HC-95164,
610 75N92020D00007, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-000040,

611 UL1-TR-001079, and UL1-TR-001420. Funding for SHARe genotyping was provided by NHLBI Contract N02-HL-64278.
612 Genotyping was performed at Affymetrix (Santa Clara, California, USA) and the Broad Institute of Harvard and MIT
613 (Boston, Massachusetts, USA) using the Affymetrix Genome-Wide Human SNP Array 6.0. Also supported in part by
614 NHLBI CHARGE Consortium Contract HL105756. The provision of genotyping data was supported in part by the National
615 Center for Advancing Translational Sciences, CTSI grant UL1TR001881, and the National Institute of Diabetes and
616 Digestive and Kidney Disease Diabetes Research Center (DRC) grant DK063491 to the Southern California Diabetes
617 Endocrinology Research Center. Infrastructure for the CHARGE Consortium is supported in part by the National Heart,
618 Lung, and Blood Institute (NHLBI) grant R01HL105756. The Massachusetts General Hospital Atrial Fibrillation Study
619 (MGH-AF) was supported by grants to Dr. Ellinor from the Fondation Leducq (14CVD01), the National Institutes of Health
620 to Dr. Ellinor (1RO1HL092577, R01HL128914, K24HL105780) and Dr. Lubitz (1R01HL139731) and by grants from the
621 American Heart Association to Dr. Ellinor (18SFRN34110082) and to Dr. Lubitz (18SFRN34250007). San Antonio Family
622 Study (SAFS) was supported in part by National Institutes of Health (NIH) grants R01 HL045522, MH078143, MH078111
623 and MH083824; and whole genome sequencing of SAFS subjects was supported by U01 DK085524 and R01 HL113323.
624 We are very grateful to the participants of the San Antonio Family Study for their continued involvement in our research
625 programs. Samoan Adiposity Study (SAS) was funded by NIH grant R01-HL093093. We thank the Samoan participants of
626 the study and local village authorities. We acknowledge the support of the Samoan Ministry of Health and the Samoa
627 Bureau of Statistics for their support of this research. The Rare Variants for Hypertension in Taiwan Chinese (THRV) is

628 supported by the National Heart, Lung, and Blood Institute (NHLBI) grant (R01HL111249) and its participation in TOPMed
 629 is supported by an NHLBI supplement (R01HL111249-04S1). SAPPHiRe was supported by NHLBI grants (U01HL54527,
 630 U01HL54498) and Taiwan funds, and the other cohorts were supported by Taiwan funds. The Women's Health Initiative
 631 (WHI) program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department
 632 of Health and Human Services through contracts HHSN268201600018C, HHSN268201600001C, HHSN268201600002C,
 633 HHSN268201600003C, and HHSN268201600004C. The Centers for Common Disease Genomics (CCDG) program was
 634 supported by NHGRI and NHLBI, and whole genome sequencing was performed at the Baylor College of Medicine
 635 Human Genome Sequencing Center (UM1 HG008898 and R01HL059367). We like to acknowledge all the grants that
 636 supported this study, R01 HL121007, U01 HL072515, R01 AG18728, X01HL134588, HL 046389, HL113338, and
 637 1R35HL135818, K01 HL135405, R03 HL154284, U01HL072507, R01HL087263, R01HL090682, P01HL045522,
 638 R01MH078143, R01MH078111, R01MH083824, U01DK085524, R01HL113323, R01HL093093, R01HL140570,
 639 R01HL142711, R01HL127564, R01HL148050, R01HL148565, and Leducq TNE-18CVD04. The views expressed in this
 640 manuscript are those of the authors and do not necessarily represent the views of the National Heart, Lung, and Blood
 641 Institute; the National Institutes of Health; or the U.S. Department of Health and Human Services.

642

643 **Author's information**

644 **NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium**

645 Namiko Abe⁶³, Gonçalo Abecasis⁶⁴, Francois Aguet⁶⁵, Christine Albert⁶⁶, Laura Almasy⁶⁷, Alvaro Alonso⁶⁸, Seth Ament⁶⁹,
646 Peter Anderson⁷⁰, Pramod Anugu⁷¹, Deborah Applebaum-Bowden⁷², Kristin Ardlie⁶⁵, Dan Arking⁷³, Allison Ashley-Koch⁷⁴,
647 Tim Assimes⁷⁵, Paul Auer⁷⁶, Dimitrios Avramopoulos⁷³, Najib Ayas⁷⁷, Adithya Balasubramanian⁷⁸, John Barnard⁷⁹,
648 Kathleen Barnes⁸⁰, R. Graham Barr⁸¹, Emily Barron-Casella⁷³, Lucas Barwick⁸², Terri Beaty⁷³, Gerald Beck⁸³, Diane
649 Becker⁸⁴, Lewis Becker⁷³, Rebecca Beer⁸⁵, Amber Beitelshes⁶⁹, Emelia Benjamin⁸⁶, Takis Benos⁸⁷, Marcos Bezerra⁸⁸,
650 Larry Bielak⁶⁴, Thomas Blackwell⁶⁴, Russell Bowler⁸⁹, Ulrich Broeckel⁹⁰, Jai Broome⁷⁰, Deborah Brown⁹¹, Karen Bunting⁶³,
651 Esteban Burchard⁹², Carlos Bustamante⁹³, Erin Buth⁹⁴, Jonathan Cardwell⁹⁵, Vincent Carey⁹⁶, Julie Carrier⁹⁷, Cara
652 Carty⁹⁸, Richard Casaburi⁹⁹, Juan P Casas Romero¹⁰⁰, James Casella⁷³, Peter Castaldi¹⁰¹, Mark Chaffin⁶⁵, Christy
653 Chang⁶⁹, Yi-Cheng Chang¹⁰², Daniel Chasman¹⁰³, Sameer Chavan⁹⁵, Bo-Juen Chen⁶³, Wei-Min Chen¹⁰⁴, Yii-Der Ida
654 Chen¹⁰⁵, Michael Cho⁹⁶, Seung Hoan Choi⁶⁵, Mina Chung¹⁰⁶, Clary Clish¹⁰⁷, Suzy Comhair¹⁰⁸, Matthew Conomos⁹⁴,
655 Elaine Cornell¹⁰⁹, Carolyn Crandall⁹⁹, James Crapo¹¹⁰, L. Adrienne Cupples¹¹¹, Jeffrey Curtis⁶⁴, Brian Custer¹¹², Coleen
656 Damcott⁶⁹, Dawood Darbar¹¹³, Sean David¹¹⁴, Colleen Davis⁷⁰, Michelle Daya⁹⁵, Mariza de Andrade¹¹⁵, Michael
657 DeBaun¹¹⁶, Ranjan Deka¹¹⁷, Dawn DeMeo⁹⁶, Scott Devine⁶⁹, Huyen Dinh⁷⁸, Harsha Doddapaneni⁷⁸, Qing Duan¹¹⁸,
658 Shannon Dugan-Perez⁷⁸, Ravi Duggirala¹¹⁹, Jon Peter Durda¹⁰⁹, Charles Eaton¹²⁰, Lynette Ekunwe⁷¹, Adel El Boueiz¹²¹,
659 Leslie Emery⁷⁰, Serpil Erzurum⁷⁹, Charles Farber¹⁰⁴, Jesse Farek⁷⁸, Tasha Fingerlin¹²², Matthew Flickinger⁶⁴, Nora
660 Franceschini¹²³, Chris Frazar⁷⁰, Mao Fu⁶⁹, Stephanie M. Fullerton⁷⁰, Lucinda Fulton¹²⁴, Weiniu Gan⁸⁵, Shanshan Gao⁹⁵,
661 Yan Gao⁷¹, Margery Gass¹²⁵, Heather Geiger¹²⁶, Bruce Gelb¹²⁷, Mark Geraci¹²⁸, Robert Gerszten¹²⁹, Auyon Ghosh⁹⁶,

662 Chris Gignoux⁷⁵, Mark Gladwin⁸⁷, David Glahn¹³⁰, Stephanie Gogarten⁷⁰, Da-Wei Gong⁶⁹, Harald Goring¹³¹, Sharon
663 Graw⁸⁰, Kathryn J. Gray¹³², Daniel Grine⁹⁵, Colin Gross⁶⁴, C. Charles Gu¹²⁴, Yue Guan⁶⁹, Namrata Gupta⁶⁵, David M.
664 Haas¹³³, Jeff Haessler¹²⁵, Michael Hall¹³⁴, Yi Han⁷⁸, Patrick Hanly¹³⁵, Daniel Harris¹³⁶, Nicola L. Hawley¹³⁷, Ben Heavner⁹⁴,
665 Susan Heckbert¹³⁸, Ryan Hernandez⁹², David Herrington¹³⁹, Craig Hersh¹⁴⁰, Bertha Hidalgo¹⁴¹, James Hixson¹⁴², Brian
666 Hobbs⁹⁶, John Hokanson⁹⁵, Elliott Hong⁶⁹, Karin Hoth¹⁴³, Chao (Agnes) Hsiung¹⁴⁴, Jianhong Hu⁷⁸, Yi-Jen Hung¹⁴⁵, Haley
667 Huston¹⁴⁶, Chii Min Hwu¹⁴⁷, Rebecca Jackson¹⁴⁸, Deepti Jain⁷⁰, Cashell Jaquish⁸⁵, Jill Johnsen¹⁴⁹, Andrew Johnson⁸⁵,
668 Craig Johnson⁷⁰, Rich Johnston⁶⁸, Kimberly Jones⁷³, Hyun Min Kang¹⁵⁰, Shannon Kelly¹⁵¹, Eimear Kenny¹²⁷, Michael
669 Kessler⁶⁹, Alyna Khan⁷⁰, Ziad Khan⁷⁸, Wonji Kim¹⁵², John Kimoff¹⁵³, Greg Kinney¹⁵⁴, Barbara Konkle¹⁴⁶, Holly Kramer¹⁵⁵,
670 Christoph Lange¹⁵⁶, Ethan Lange⁹⁵, Cathy Laurie⁷⁰, Cecelia Laurie⁷⁰, Meryl LeBoff⁹⁶, Jiwon Lee⁹⁶, Sandra Lee⁷⁸, Wen-
671 Jane Lee¹⁴⁷, Jonathon LeFaive⁶⁴, David Levine⁷⁰, Dan Levy⁸⁵, Joshua Lewis⁶⁹, Yun Li¹¹⁸, Henry Lin¹⁰⁵, Honghuang Lin¹⁵⁷,
672 Simin Liu¹⁵⁸, Yongmei Liu¹⁵⁹, Yu Liu¹⁶⁰, Kathryn Lunetta¹⁵⁷, James Luo⁸⁵, Ulysses Magalang¹⁶¹, Michael Mahaney¹⁶²,
673 Barry Make⁷³, Alisa Manning¹⁶³, JoAnn Manson⁹⁶, Lisa Martin¹⁶⁴, Melissa Marton¹²⁶, Susan Mathai⁹⁵, Susanne May⁹⁴,
674 Patrick McArdle⁶⁹, Merry-Lynn McDonald¹⁴¹, Sean McFarland¹⁵², Daniel McGoldrick¹⁶⁵, Caitlin McHugh⁹⁴, Becky
675 McNeil¹⁶⁶, Hao Mei⁷¹, James Meigs¹⁶⁷, Vipin Menon⁷⁸, Luisa Mestroni⁸⁰, Ginger Metcalf⁷⁸, Deborah A Meyers¹⁶⁸,
676 Emmanuel Mignot¹⁶⁹, Julie Mikulla⁸⁵, Nancy Min⁷¹, Mollie Minear¹⁷⁰, Ryan L Minster⁸⁷, Matt Moll¹⁰¹, Zeineen Momin⁷⁸,
677 Courtney Montgomery¹⁷¹, Donna Muzny⁷⁸, Josyf C Mychaleckyj¹⁰⁴, Girish Nadkarni¹²⁷, Rakhi Naik⁷³, Sergei Nekhai¹⁷²,
678 Sarah C. Nelson⁹⁴, Bonnie Neltner⁹⁵, Caitlin Nessner⁷⁸, Osuji Nkechinyere⁷⁸, Jeff O'Connell¹⁷³, Tim O'Connor⁶⁹, Heather

679 Ochs-Balcom¹⁷⁴, Geoffrey Okwuonu⁷⁸, Allan Pack¹⁷⁵, David T. Paik¹⁷⁶, James Pankow¹⁷⁷, George Papanicolaou⁸⁵, Cora
680 Parker¹⁷⁸, Juan Manuel Peralta¹¹⁹, Marco Perez⁷⁵, James Perry⁶⁹, Ulrike Peters¹⁷⁹, Lawrence S Phillips⁶⁸, Jacob
681 Pleiness⁶⁴, Toni Pollin⁶⁹, Wendy Post¹⁸⁰, Julia Powers Becker¹⁸¹, Meher Preethi Boorgula⁹⁵, Michael Preuss¹²⁷, Pankaj
682 Qasba⁸⁵, Dandi Qiao⁹⁶, Zhaohui Qin⁶⁸, Nicholas Rafaels¹⁸², Laura Raffield¹⁸³, Mahitha Rajendran⁷⁸, Vasana S.
683 Ramachandran¹⁵⁷, D.C. Rao¹²⁴, Laura Rasmussen-Torvik¹⁸⁴, Aakrosh Ratan¹⁰⁴, Robert Reed⁶⁹, Catherine Reeves¹⁸⁵,
684 Elizabeth Regan¹¹⁰, Alex Reiner¹⁸⁶, Muagututi'a Sefuiva Reupena¹⁸⁷, Ken Rice⁷⁰, Rebecca Robillard¹⁸⁸, Nicolas Robine¹²⁶,
685 Dan Roden¹⁸⁹, Carolina Roselli⁶⁵, Ingo Ruczinski⁷³, Alexi Runnels¹²⁶, Pamela Russell⁹⁵, Sarah Ruuska¹⁴⁶, Kathleen
686 Ryan⁶⁹, Ester Cerdeira Sabino¹⁹⁰, Danish Saleheen¹⁹¹, Shabnam Salimi⁶⁹, Sejal Salvi⁷⁸, Steven Salzberg⁷³, Kevin
687 Sandow¹⁹², Vijay G. Sankaran¹⁹³, Jireh Santibanez⁷⁸, Karen Schwander¹²⁴, David Schwartz⁹⁵, Frank Sciurba⁸⁷, Christine
688 Seidman¹⁹⁴, Jonathan Seidman¹⁹⁵, Frédéric Sériès¹⁹⁶, Vivien Sheehan¹⁹⁷, Stephanie L. Sherman¹⁹⁸, Amol Shetty⁶⁹, Aniket
689 Shetty⁹⁵, Wayne Hui-Heng Sheu¹⁴⁷, M. Benjamin Shoemaker¹⁹⁹, Brian Silver²⁰⁰, Edwin Silverman⁹⁶, Robert Skomro²⁰¹,
690 Albert Vernon Smith²⁰², Josh Smith⁷⁰, Nicholas Smith¹³⁸, Tanja Smith⁶³, Sylvia Smoller²⁰³, Beverly Snively²⁰⁴, Michael
691 Snyder⁷⁵, Tamar Sofer⁹⁶, Nona Sotoodehnia⁷⁰, Adrienne M. Stilp⁷⁰, Garrett Storm²⁰⁵, Elizabeth Streeten⁶⁹, Jessica Lasky
692 Su⁹⁶, Yun Ju Sung¹²⁴, Jody Sylvia⁹⁶, Adam Szpiro⁷⁰, Daniel Taliun⁶⁴, Hua Tang²⁰⁶, Margaret Taub⁷³, Matthew Taylor⁸⁰,
693 Simeon Taylor⁶⁹, Marilyn Telen⁷⁴, Timothy A. Thornton⁷⁰, Machiko Threlkeld²⁰⁷, Lesley Tinker¹²⁵, David Tirschwell⁷⁰,
694 Sarah Tishkoff²⁰⁸, Hemant Tiwari²⁰⁹, Catherine Tong²¹⁰, Dhananjay Vaidya⁷³, David Van Den Berg²¹¹, Peter VandeHaar⁶⁴,
695 Scott Vrieze¹⁷⁷, Tarik Walker⁹⁵, Robert Wallace¹⁴³, Avram Walts⁹⁵, Fei Fei Wang⁷⁰, Heming Wang²¹², Jiongming Wang²⁰²,

696 Karol Watson⁹⁹, Jennifer Watt⁷⁸, Daniel E. Weeks⁸⁷, Joshua Weinstock¹⁵⁰, Bruce Weir⁷⁰, Scott T Weiss²¹³, Lu-Chen
 697 Weng²¹⁴, Jennifer Wessel²¹⁵, Kayleen Williams⁹⁴, L. Keoki Williams²¹⁶, Carla Wilson⁹⁶, James Wilson²¹⁷, Lara
 698 Winterkorn¹²⁶, Quenna Wong⁷⁰, Joseph Wu¹⁷⁶, Huichun Xu⁶⁹, Ivana Yang⁹⁵, Ketian Yu⁶⁴, Seyedeh Maryam Zekavat⁶⁵,
 699 Yingze Zhang²¹⁸, Snow Xueyan Zhao¹¹⁰, Wei Zhao²¹⁹, Xiaofeng Zhu²²⁰, Michael Zody⁶³, Sebastian Zoellner⁶⁴

700

701 63 - New York Genome Center, New York, New York, 10013; 64 - University of Michigan, Ann Arbor, Michigan, 48109; 65
 702 - Broad Institute, Cambridge, Massachusetts, 02142; 66 - Cedars Sinai, Boston, Massachusetts, 02114; 67 - Children's
 703 Hospital of Philadelphia, University of Pennsylvania, Philadelphia, Pennsylvania, 19104; 68 - Emory University, Atlanta,
 704 Georgia, 30322; 69 - University of Maryland, Baltimore, Maryland, 21201; 70 - University of Washington, Seattle,
 705 Washington, 98195; 71 - University of Mississippi, Jackson, Mississippi, 38677; 72 - National Institutes of Health,
 706 Bethesda, Maryland, 20892; 73 - Johns Hopkins University, Baltimore, Maryland, 21218; 74 - Duke University, Durham,
 707 North Carolina, 27708; 75 - Stanford University, Stanford, California, 94305; 76 - University of Wisconsin Milwaukee,
 708 Milwaukee, Wisconsin, 53211; 77 - Providence Health Care, Medicine, Vancouver; 78 - Baylor College of Medicine
 709 Human Genome Sequencing Center, Houston, Texas, 77030; 79 - Cleveland Clinic, Cleveland, Ohio, 44195; 80 -
 710 University of Colorado Anschutz Medical Campus, Aurora, Colorado, 80045; 81 - Columbia University, New York, New
 711 York, 10032; 82 - The Emmes Corporation, LTRC, Rockville, Maryland, 20850; 83 - Cleveland Clinic, Quantitative Health
 712 Sciences, Cleveland, Ohio, 44195; 84 - Johns Hopkins University, Medicine, Baltimore, Maryland, 21218; 85 - National

713 Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, Maryland, 20892; 86 - Boston University,
 714 Massachusetts General Hospital, Boston University School of Medicine, Boston, Massachusetts, 02114; 87 - University of
 715 Pittsburgh, Pittsburgh, Pennsylvania, 15260; 88 - Fundação de Hematologia e Hemoterapia de Pernambuco - Hemope,
 716 Recife, 52011-000; 89 - National Jewish Health, National Jewish Health, Denver, Colorado, 80206; 90 - Medical College
 717 of Wisconsin, Milwaukee, Wisconsin, 53226; 91 - University of Texas Health at Houston, Pediatrics, Houston, Texas,
 718 77030; 92 - University of California, San Francisco, San Francisco, California, 94143; 93 - Stanford University, Biomedical
 719 Data Science, Stanford, California, 94305; 94 - University of Washington, Biostatistics, Seattle, Washington, 98195; 95 -
 720 University of Colorado at Denver, Denver, Colorado, 80204; 96 - Brigham & Women's Hospital, Boston, Massachusetts,
 721 02115; 97 - University of Montreal; 98 - Washington State University, Pullman, Washington, 99164; 99 - University of
 722 California, Los Angeles, Los Angeles, California, 90095; 100 - Brigham & Women's Hospital; 101 - Brigham & Women's
 723 Hospital, Medicine, Boston, Massachusetts, 02115; 102 - National Taiwan University, Taipei, 10617; 103 - Brigham &
 724 Women's Hospital, Division of Preventive Medicine, Boston, Massachusetts, 02215; 104 - University of Virginia,
 725 Charlottesville, Virginia, 22903; 105 - Lundquist Institute, Torrance, California, 90502; 106 - Cleveland Clinic, Cleveland
 726 Clinic, Cleveland, Ohio, 44195; 107 - Broad Institute, Metabolomics Platform, Cambridge, Massachusetts, 02142; 108 -
 727 Cleveland Clinic, Immunity and Immunology, Cleveland, Ohio, 44195; 109 - University of Vermont, Burlington, Vermont,
 728 05405; 110 - National Jewish Health, Denver, Colorado, 80206; 111 - Boston University, Biostatistics, Boston,
 729 Massachusetts, 02115; 112 - Vitalant Research Institute, San Francisco, California, 94118; 113 - University of Illinois at

730 Chicago, Chicago, Illinois, 60607; 114 - University of Chicago, Chicago, Illinois, 60637; 115 - Mayo Clinic, Health
 731 Quantitative Sciences Research, Rochester, Minnesota, 55905; 116 - Vanderbilt University, Nashville, Tennessee, 37235;
 732 117 - University of Cincinnati, Cincinnati, Ohio, 45220; 118 - University of North Carolina, Chapel Hill, North Carolina,
 733 27599; 119 - University of Texas Rio Grande Valley School of Medicine, Edinburg, Texas, 78539; 120 - Brown University,
 734 Providence, Rhode Island, 02912; 121 - Harvard University, Channing Division of Network Medicine, Cambridge,
 735 Massachusetts, 02138; 122 - National Jewish Health, Center for Genes, Environment and Health, Denver, Colorado,
 736 80206; 123 - University of North Carolina, Epidemiology, Chapel Hill, North Carolina, 27599; 124 - Washington University
 737 in St Louis, St Louis, Missouri, 63130; 125 - Fred Hutchinson Cancer Research Center, Seattle, Washington, 98109; 126 -
 738 New York Genome Center, New York City, New York, 10013; 127 - Icahn School of Medicine at Mount Sinai, New York,
 739 New York, 10029; 128 - University of Pittsburgh, Pittsburgh, Pennsylvania; 129 - Beth Israel Deaconess Medical Center,
 740 Boston, Massachusetts, 02215; 130 - Boston Children's Hospital, Harvard Medical School, Department of Psychiatry,
 741 Boston, Massachusetts, 02115; 131 - University of Texas Rio Grande Valley School of Medicine, San Antonio, Texas,
 742 78229; 132 - Mass General Brigham, Obstetrics and Gynecology, Boston, Massachusetts, 02115; 133 - Indiana
 743 University, OB/GYN, Indianapolis, Indiana, 46202; 134 - University of Mississippi, Cardiology, Jackson, Mississippi,
 744 39216; 135 - University of Calgary, Medicine, Calgary; 136 - University of Maryland, Genetics, Philadelphia, Pennsylvania,
 745 19104; 137 - Yale University, Department of Chronic Disease Epidemiology, New Haven, Connecticut, 06520; 138 -
 746 University of Washington, Epidemiology, Seattle, Washington, 98195; 139 - Wake Forest Baptist Health, Winston-Salem,

747 North Carolina, 27157; 140 - Brigham & Women's Hospital, Channing Division of Network Medicine, Boston,
 748 Massachusetts, 02115; 141 - University of Alabama, Birmingham, Alabama, 35487; 142 - University of Texas Health at
 749 Houston, Houston, Texas, 77225; 143 - University of Iowa, Iowa City, Iowa, 52242; 144 - National Health Research
 750 Institute Taiwan, Institute of Population Health Sciences, NHRI, Miaoli County, 350; 145 - Tri-Service General Hospital
 751 National Defense Medical Center; 146 - Blood Works Northwest, Seattle, Washington, 98104; 147 - Taichung Veterans
 752 General Hospital Taiwan, Taichung City, 407; 148 - Oklahoma State University Medical Center, Internal Medicine,
 753 Division of Endocrinology, Diabetes and Metabolism, Columbus, Ohio, 43210; 149 - Blood Works Northwest, Research
 754 Institute, Seattle, Washington, 98104; 150 - University of Michigan, Biostatistics, Ann Arbor, Michigan, 48109; 151 -
 755 University of California, San Francisco, San Francisco, California, 94118; 152 - Harvard University, Cambridge,
 756 Massachusetts, 02138; 153 - McGill University, Montréal, QC H3A 0G4; 154 - University of Colorado at Denver,
 757 Epidemiology, Aurora, Colorado, 80045; 155 - Loyola University, Public Health Sciences, Maywood, Illinois, 60153; 156 -
 758 Harvard School of Public Health, Biostats, Boston, Massachusetts, 02115; 157 - Boston University, Boston,
 759 Massachusetts, 02215; 158 - Brown University, Epidemiology and Medicine, Providence, Rhode Island, 02912; 159 -
 760 Duke University, Cardiology, Durham, North Carolina, 27708; 160 - Stanford University, Cardiovascular Institute, Stanford,
 761 California, 94305; 161 - Ohio State University, Division of Pulmonary, Critical Care and Sleep Medicine, Columbus, Ohio,
 762 43210; 162 - University of Texas Rio Grande Valley School of Medicine, Brownsville, Texas, 78520; 163 - Broad Institute,
 763 Harvard University, Massachusetts General Hospital; 164 - George Washington University, cardiology, Washington,

764 District of Columbia, 20037; 165 - University of Washington, Genome Sciences, Seattle, Washington, 98195; 166 - RTI
765 International; 167 - Massachusetts General Hospital, Medicine, Boston, Massachusetts, 02114; 168 - University of
766 Arizona, Tucson, Arizona, 85721; 169 - Stanford University, Center For Sleep Sciences and Medicine, Palo Alto,
767 California, 94304; 170 - National Institute of Child Health and Human Development, National Institutes of Health,
768 Bethesda, Maryland, 20892; 171 - Oklahoma Medical Research Foundation, Genes and Human Disease, Oklahoma City,
769 Oklahoma, 73104; 172 - Howard University, Washington, District of Columbia, 20059; 173 - University of Maryland,
770 Baltimore, Maryland, 21201; 174 - University at Buffalo, Buffalo, New York, 14260; 175 - University of Pennsylvania,
771 Division of Sleep Medicine/Department of Medicine, Philadelphia, Pennsylvania, 19104-3403; 176 - Stanford University,
772 Stanford Cardiovascular Institute, Stanford, California, 94305; 177 - University of Minnesota, Minneapolis, Minnesota,
773 55455; 178 - RTI International, Biostatistics and Epidemiology Division, Research Triangle Park, North Carolina, 27709-
774 2194; 179 - Fred Hutchinson Cancer Research Center, Fred Hutch and UW, Seattle, Washington, 98109; 180 - Johns
775 Hopkins University, Cardiology/Medicine, Baltimore, Maryland, 21218; 181 - University of Colorado at Denver, Medicine,
776 Denver, Colorado, 80204; 182 - University of Colorado at Denver, Denver, Colorado, 80045; 183 - University of North
777 Carolina, Genetics, Chapel Hill, North Carolina, 27599; 184 - Northwestern University, Chicago, Illinois, 60208; 185 - New
778 York Genome Center, New York Genome Center, New York City, New York, 10013; 186 - Fred Hutchinson Cancer
779 Research Center, University of Washington, Seattle, Washington, 98109; 187 - Lutia I Puava Ae Mapu I Fagalele, Apia;
780 188 - University of Ottawa, Sleep Research Unit, University of Ottawa Institute for Mental Health Research, Ottawa, ON

781 K1Z 7K4; 189 - Vanderbilt University, Medicine, Pharmacology, Biomedical Informatics, Nashville, Tennessee, 37235; 190
 782 - Universidade de Sao Paulo, Faculdade de Medicina, Sao Paulo, 01310000; 191 - Columbia University, New York, New
 783 York, 10027; 192 - Lundquist Institute, TGPS, Torrance, California, 90502; 193 - Harvard University, Division of
 784 Hematology/Oncology, Boston, Massachusetts, 02115; 194 - Harvard Medical School, Genetics, Boston, Massachusetts,
 785 02115; 195 - Harvard Medical School, Boston, Massachusetts, 02115; 196 - Université Laval, Quebec City, G1V 0A6; 197
 786 - Emory University, Pediatrics, Atlanta, Georgia, 30307; 198 - Emory University, Human Genetics, Atlanta, Georgia,
 787 30322; 199 - Vanderbilt University, Medicine/Cardiology, Nashville, Tennessee, 37235; 200 - UMass Memorial Medical
 788 Center, Worcester, Massachusetts, 01655; 201 - University of Saskatchewan, Saskatoon, SK S7N 5C9; 202 - University
 789 of Michigan; 203 - Albert Einstein College of Medicine, New York, New York, 10461; 204 - Wake Forest Baptist Health,
 790 Biostatistical Sciences, Winston-Salem, North Carolina, 27157; 205 - University of Colorado at Denver, Genomic
 791 Cardiology, Aurora, Colorado, 80045; 206 - Stanford University, Genetics, Stanford, California, 94305; 207 - University of
 792 Washington, University of Washington, Department of Genome Sciences, Seattle, Washington, 98195; 208 - University of
 793 Pennsylvania, Genetics, Philadelphia, Pennsylvania, 19104; 209 - University of Alabama, Biostatistics, Birmingham,
 794 Alabama, 35487; 210 - University of Washington, Department of Biostatistics, Seattle, Washington, 98195; 211 -
 795 University of Southern California, USC Methylation Characterization Center, University of Southern California, California,
 796 90033; 212 - Brigham & Women's Hospital, Mass General Brigham, Boston, Massachusetts, 02115; 213 - Brigham &
 797 Women's Hospital, Channing Division of Network Medicine, Department of Medicine, Boston, Massachusetts, 02115; 214

798 - Massachusetts General Hospital, Boston, Massachusetts, 02114; 215 - Indiana University, Epidemiology, Indianapolis,
799 Indiana, 46202; 216 - Henry Ford Health System, Detroit, Michigan, 48202; 217 - Beth Israel Deaconess Medical Center,
800 Cardiology, Cambridge, Massachusetts, 02139; 218 - University of Pittsburgh, Medicine, Pittsburgh, Pennsylvania, 15260;
801 219 - University of Michigan, Department of Epidemiology, Ann Arbor, Michigan, 48109; 220 - Case Western Reserve
802 University, Department of Population and Quantitative Health Sciences, Cleveland, Ohio, 44106
803

804 **Competing interests**

805 P.N. reports investigator-initiated grant support from Amgen, Apple, AstraZeneca, and Boston Scientific, personal fees
806 from Apple, AstraZeneca, Blackstone Life Sciences, Foresite Labs, Genentech, and Novartis, and spousal employment at
807 Vertex, all unrelated to the present work. BP serves on the Steering Committee of the Yale Open Data Access Project
808 funded by Johnson & Johnson. MEM receives funding from Regeneron Pharmaceutical Inc. unrelated to this work. SA
809 has employment and equity in 23andMe, Inc. The spouse of CJW works at Regeneron.

References:

1. Cohen, J. C., Boerwinkle, E., Mosley, T. H. & Hobbs, 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 Hypolipidemia. *N. Engl. J. Med.* **363**, 2220–2227 (2010).
4. Stitzel, N. O. *et al.* *ANGPTL3* Deficiency and Protection Against Coronary Artery Disease. *J. Am. Coll. Cardiol.* **69**, 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).

- 827 9. Saleheen, D. *et al.* Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. *Nature* **544**, 235–239
828 (2017).
- 829 10. Exome Aggregation Consortium *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291
830 (2016).
- 831 11. Samocha, K. E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).
- 832 12. Natarajan, P. *et al.* Chromosome Xq23 is associated with lower atherogenic lipid concentrations and favorable cardiometabolic
833 indices. *Nat. Commun.* **12**, 2182 (2021).
- 834 13. Li, X. *et al.* Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large
835 whole-genome sequencing studies at scale. *Nat. Genet.* **52**, 969–983 (2020).
- 836 14. Natarajan, P. *et al.* Deep-coverage whole genome sequences and blood lipids among 16,324 individuals. *Nat. Commun.* **9**, 3391
837 (2018).
- 838 15. Klarin, D. *et al.* Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. *Nat. Genet.*
839 **50**, 1514–1523 (2018).
- 840 16. Hu, Y. *et al.* Minority-centric meta-analyses of blood lipid levels identify novel loci in the Population Architecture using
841 Genomics and Epidemiology (PAGE) study. *PLoS Genet.* **16**, e1008684 (2020).
- 842 17. NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium *et al.* Sequencing of 53,831 diverse genomes from the
843 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.* (2021) doi:10.1093/aje/kwab115.
19. Fadista, J., Manning, A. K., Florez, J. C. & Groop, L. The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. *Eur. J. Hum. Genet. EJHG* **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. Genomic 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).

- 861 27. Barter, P. J. *et al.* Effects of Torcetrapib in Patients at High Risk for Coronary Events. *N. Engl. J. Med.* **357**, 2109–2122 (2007).
- 862 28. Schwartz, G. G. *et al.* Effects of Dalcetrapib in Patients with a Recent Acute Coronary Syndrome. *N. Engl. J. Med.* **367**, 2089–
863 2099 (2012).
- 864 29. The HPS3/TIMI55–REVEAL Collaborative Group. Effects of Anacetrapib in Patients with Atherosclerotic Vascular Disease. *N.*
865 *Engl. J. Med.* **377**, 1217–1227 (2017).
- 866 30. Lincoff, A. M. *et al.* Evacetrapib and Cardiovascular Outcomes in High-Risk Vascular Disease. *N. Engl. J. Med.* **376**, 1933–1942
867 (2017).
- 868 31. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
- 869 32. Fairwozy, R. H., White, J., Palmen, J., Kalea, A. Z. & Humphries, S. E. Identification of the Functional Variant(s) that Explain the
870 Low-Density Lipoprotein Receptor (LDLR) GWAS SNP rs6511720 Association with Lower LDL-C and Risk of CHD. *PloS One*
871 **11**, e0167676 (2016).
- 872 33. Li, Z. *et al.* Dynamic Scan Procedure for Detecting Rare-Variant Association Regions in Whole-Genome Sequencing Studies. *Am.*
873 *J. Hum. Genet.* **104**, 802–814 (2019).
- 874 34. Roses, A. D. *et al.* A TOMM40 variable-length polymorphism predicts the age of late-onset Alzheimer’s disease.
875 *Pharmacogenomics J.* **10**, 375–384 (2010).
- 876 35. Li, G. *et al.* TOMM40 intron 6 poly-T length, age at onset, and neuropathology of AD in individuals with APOE ε3/ε3.
877 *Alzheimers Dement. J. Alzheimers Assoc.* **9**, 554–561 (2013).

- 878 36. Glazier, A. M., Scott, J. & Aitman, T. J. Molecular basis of the Cd36 chromosomal deletion underlying SHR defects in insulin
879 action and fatty acid metabolism. *Mamm. Genome Off. J. Int. Mamm. Genome Soc.* **13**, 108–113 (2002).
- 880 37. Global Lipids Genetics Consortium. Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274–1283
881 (2013).
- 882 38. Willer, C. J. *et al.* Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274–1283 (2013).
- 883 39. ENGAGE Consortium *et al.* The impact of low-frequency and rare variants on lipid levels. *Nat. Genet.* **47**, 589–597 (2015).
- 884 40. The Myocardial Infarction Genetics Consortium Investigators. Inactivating Mutations in *NPC1L1* and Protection from Coronary
885 Heart Disease. *N. Engl. J. Med.* **371**, 2072–2082 (2014).
- 886 41. GLGC Consortium *et al.* Exome chip meta-analysis identifies novel loci and East Asian-specific coding variants that contribute to
887 lipid levels and coronary artery disease. *Nat. Genet.* **49**, 1722–1730 (2017).
- 888 42. Hoffmann, T. J. *et al.* A large electronic-health-record-based genome-wide study of serum lipids. *Nat. Genet.* **50**, 401–413 (2018).
- 889 43. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591
890 (2019).
- 891 44. Peloso, G. M. & Natarajan, P. Insights from population-based analyses of plasma lipids across the allele frequency spectrum.
892 *Curr. Opin. Genet. Dev.* **50**, 1–6 (2018).
- 893 45. Kremer, L. S. *et al.* Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nat. Commun.* **8**, 15824 (2017).

- 894 46. Cummings, B. B. *et al.* Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci. Transl. Med.* **9**,
895 eaal5209 (2017).
- 896 47. Genome Aggregation Database Production Team *et al.* Transcript expression-aware annotation improves rare variant
897 interpretation. *Nature* **581**, 452–458 (2020).
- 898 48. Mendes de Almeida, R. *et al.* Whole gene sequencing identifies deep-intronic variants with potential functional impact in patients
899 with hypertrophic cardiomyopathy. *PLOS ONE* **12**, e0182946 (2017).
- 900 49. Vitsios, D., Dhindsa, R. S., Middleton, L., Gussow, A. B. & Petrovski, S. Prioritizing non-coding regions based on human
901 genomic constraint and sequence context with deep learning. *Nat. Commun.* **12**, 1504 (2021).
- 902 50. di Iulio, J. *et al.* The human noncoding genome defined by genetic diversity. *Nat. Genet.* **50**, 333–337 (2018).
- 903 51. Genome Aggregation Database Consortium *et al.* The mutational constraint spectrum quantified from variation in 141,456
904 humans. *Nature* **581**, 434–443 (2020).
- 905 52. Khera, A. V. *et al.* Diagnostic Yield and Clinical Utility of Sequencing Familial Hypercholesterolemia Genes in Patients With
906 Severe Hypercholesterolemia. *J. Am. Coll. Cardiol.* **67**, 2578–2589 (2016).
- 907 53. Benn, M., Watts, G. F., Tybjaerg-Hansen, A. & Nordestgaard, B. G. Mutations causative of familial hypercholesterolaemia:
908 screening of 98 098 individuals from the Copenhagen General Population Study estimated a prevalence of 1 in 217. *Eur. Heart J.*
909 **37**, 1384–1394 (2016).

54. Grundy, S. M. *et al.* 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA Guideline on the Management of Blood Cholesterol: Executive Summary. *J. Am. Coll. Cardiol.* **73**, 3168–3209 (2019).
55. Sturm, A. C. *et al.* Clinical Genetic Testing for Familial Hypercholesterolemia. *J. Am. Coll. Cardiol.* **72**, 662–680 (2018).
56. Reeskamp, L. F. *et al.* A Deep Intronic Variant in *LDLR* in Familial Hypercholesterolemia: Time to Widen the Scope? *Circ. Genomic Precis. Med.* **11**, (2018).
57. Calandra, S., Tarugi, P. & Bertolini, S. Altered mRNA splicing in lipoprotein disorders: *Curr. Opin. Lipidol.* **22**, 93–99 (2011).
- 58.; 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).
59. Peloso, G. M. *et al.* Rare Protein-Truncating Variants in APOB, Lower Low-Density Lipoprotein Cholesterol, and Protection Against Coronary Heart Disease. *Circ. Genomic Precis. Med.* **12**, e002376 (2019).
60. Abifadel, M. *et al.* Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. *Nat. Genet.* **34**, 154–156 (2003).
61. Jiang, L. *et al.* The distribution and characteristics of LDL receptor mutations in China: A systematic review. *Sci. Rep.* **5**, 17272 (2015).
62. Arráiz, 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).

- 926 63. Gudnason, V. *et al.* Identification of recurrent and novel mutations in exon 4 of the LDL receptor gene in patients with familial
927 hypercholesterolemia in the United Kingdom. *Arterioscler. Thromb. J. Vasc. Biol.* **13**, 56–63 (1993).
- 928 64. Goldmann, R. *et al.* Genomic characterization of large rearrangements of the LDLR gene in Czech patients with familial
929 hypercholesterolemia. *BMC Med. Genet.* **11**, 115 (2010).
- 930 65. Zuk, O. *et al.* Searching for missing heritability: Designing rare variant association studies. *Proc. Natl. Acad. Sci.* **111**, E455–E464
931 (2014).
- 932 66. Soria, L. F. *et al.* Association between a specific apolipoprotein B mutation and familial defective apolipoprotein B-100. *Proc.*
933 *Natl. Acad. Sci. U. S. A.* **86**, 587–591 (1989).
- 934 67. Conomos, M. P., Reiner, A. P., Weir, B. S. & Thornton, T. A. Model-free Estimation of Recent Genetic Relatedness. *Am. J. Hum.*
935 *Genet.* **98**, 127–148 (2016).
- 936 68. Gogarten, S. M. *et al.* Genetic association testing using the GENESIS R/Bioconductor package. *Bioinforma. Oxf. Engl.* **35**, 5346–
937 5348 (2019).
- 938 69. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
- 939 70. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).
- 940 71. Fuchsberger, C., Abecasis, G. R. & Hinds, D. A. minimac2: faster genotype imputation. *Bioinformatics* **31**, 782–784 (2015).
- 941 72. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies.
942 *Nat. Genet.* **50**, 1335–1341 (2018).

- 943 73. Pulit, S. L., de With, S. A. J. & de Bakker, P. I. W. Resetting the bar: Statistical significance in whole-genome sequencing-based
944 association studies of global populations. *Genet. Epidemiol.* **41**, 145–151 (2017).
- 945 74. Han, B. & Eskin, E. Random-Effects Model Aimed at Discovering Associations in Meta-Analysis of Genome-wide Association
946 Studies. *Am. J. Hum. Genet.* **88**, 586–598 (2011).
- 947 75. Chen, H. *et al.* Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic
948 Mixed Models. *Am. J. Hum. Genet.* **98**, 653–666 (2016).
- 949 76. Chen, H. *et al.* Efficient Variant Set Mixed Model Association Tests for Continuous and Binary Traits in Large-Scale Whole-
950 Genome Sequencing Studies. *Am. J. Hum. Genet.* **104**, 260–274 (2019).
- 951 77. The FANTOM Consortium and the RIKEN PMI and CLST (DGT). A promoter-level mammalian expression atlas. *Nature* **507**,
952 462–470 (2014).
- 953 78. The FANTOM Consortium *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
- 954 79. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74
955 (2012).
- 956 80. Fishilevich, S. *et al.* GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database J. Biol.*
957 *Databases Curation* **2017**, (2017).
- 958 81. Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.*
959 **89**, 82–93 (2011).

- 960 82. Madsen, B. E. & Browning, S. R. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* **5**,
961 e1000384 (2009).
- 962 83. Liu, Y. *et al.* ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. *Am. J.*
963 *Hum. Genet.* **104**, 410–421 (2019).
- 964 84. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary
965 statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
- 966 85. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).

967
968

Figure Legends

Fig1: Overall study schematic. 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 STAARtopmed. 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.

Fig2: Summary of single variant genome wide association. Representation of the single variant GWAS results from TOPMed Freeze 8 whole genome sequenced data of 66,329 samples. Each quarter represents a different lipid phenotype, and dots extending in clock-wise fashion represent variants with increasing evidence of association as noted by $-\log_{10}(\text{p-value})$, which was truncated at 200. The outer three circles show the GWAS data from TOPMed freeze8 where variants binned to nominally significant (p-value 0.05 - 5×10^{-7}), suggestive significant (p-value 5×10^{-7} - 5×10^{-9})

and genome wide significant ($p\text{-value} < 5 \times 10^{-09}$). The inner three circles compare our TOPMed results with known significantly associated lipid loci and variants from the MVP summary statistics and GWAS catalog to the identified novel variants and loci that are genome-wide significant from the current study, respectively.

TOPMed – Trans-Omics for Precision Medicine; GWAS – Genome Wide Association Study; MVP – Million Veteran Program.

Fig3: Comparison of effects estimates for HDL-C and LDL-C among variants in the *CETP* locus. 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.

HDL-C – High-Density Lipoprotein Cholesterol; LDL-C – Low-Density Lipoprotein Cholesterol.

Fig4: 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. *APOB1*, *SPC24* (enhancer-CAGE), *HBB* and *APOE* signal drops after the conditional analysis on rare-coding

1003 variants. The different colored dots on the plot represents the conditional STAAR-O p-values when adjusting for known
1004 variants (Set1) and rare-coding variants of the same or near-by gene.

1005 STAAR – variant-Set Test for Association using Annotation information; TC – Total Cholesterol; TG – Triglycerides; CAGE
1006 – Cap Analysis of Gene Expression; DHS – DNase hypersensitivity.

1007 **Fig5: Influence of common and rare variants with hypercholesterolemia.** In addition to monogenic contributions from
1008 rare variants in Mendelian hypercholesterolemia genes, multiple genome-wide significant LDL-C-associated common
1009 variants also yield a polygenic basis for hypercholesterolemia. In the present work, we now identify rare non-coding
1010 variants in proximity of Mendelian hypercholesterolemia genes, specifically *LDLR* and *PCSK9*, that also contribute to the
1011 genetic basis of hypercholesterolemia.

1012 LDL-C – Low-Density Lipoprotein Cholesterol

1013

Associated Lipid Phenotype	Novel variant class	Variants (Gene)	TOPMed Freeze8 (N=66329)			MGB Biobank (N=25137)			Penn Medicine Biobank (N=20079)			Meta Analysis (METASOFT)	
			TOPMed	TOPMed	TOPMed	MGB Biobank	MGB Biobank	MGB Biobank	Penn Medicine Biobank	Penn Medicine Biobank	Penn Medicine Biobank	Beta	P-value
			Effect Estimate	P-value	MAF	Effect Estimate	P-value	MAF	Effect Estimate	P-value	MAF		
LDL-C	Novel locus	12:97352354:T:C	-12.439	4.88x10 ⁻⁰⁹	0.003	1.055	8.08x10 ⁻⁰¹	0.002	11.441	3.19x10 ⁻⁰¹	0.001	2.357	5.62 x10 ⁻⁰¹
LDL-C	Novel variant	16:56957451:C:T (CETP)	-1.568	2.88x10 ⁻⁰⁹	0.283	-1.375	1.53x10 ⁻⁰⁴	0.309	-2.35	1.54x10 ⁻⁰⁴	0.578	-1.624	2.21 x10 ⁻⁰⁷
LDL-C	Novel locus	4:176382171:C:T	-16.086	2.82x10 ⁻⁰⁹	0.002	-13.340	1.71x10 ⁻⁰¹	0.001	4.716	3.52x10 ⁻⁰¹	0.005	0.882	8.44 x10 ⁻⁰¹
TC	Novel variant	13:113841051:T:C (GAS6)	1.731	1.12x10 ⁻⁰⁹	0.278	0.890	3.94x10 ⁻⁰²	0.304	0.416	5.50x10 ⁻⁰¹	0.563	0.758	3.89 x10 ⁻⁰²
TC	Novel variant	7:137875053:T:C (CREB3L2)	-4.106	7.54x10 ⁻¹¹	0.045	-4.755	1.06x10 ⁻⁰²	0.013	-3.365	7.62x10 ⁻⁰³	0.118	-3.803	2.69 x10 ⁻⁰⁴
TG	Novel locus	11:69219641:C:T	0.232	1.98x10 ⁻⁰⁹	0.002	-0.047	7.33x10 ⁻⁰¹	0.000	0.202	7.82x10 ⁻⁰²	0.001	0.101	2.53 x10 ⁻⁰¹
TG	Novel variant	13:107551611:C:T (FAM155A)	0.052	6.78x10 ⁻¹⁰	0.045	0.016	4.68x10 ⁻⁰¹	0.014	0.039	3.26x10 ⁻⁰¹	0.016	0.021	2.62 x10 ⁻⁰¹

1015 **Table 1. Putative novel variants identified in TOPMed and evidence for replication.** Variants identified as novel after
 1016 comparing with the GWAS catalog and MVP summary statistics for associations with lipid phenotypes, including LDL-C,
 1017 TC, and TG. All effect estimates are in mg/dL units, except for TG which was log-transformed in analysis thereby
 1018 representing fractional change. Variants are categorized as novel loci or novel variant (i.e., known locus associated with
 1019 another lipid phenotype) and the genes assigned to the variants per TOPMed whole genome sequence annotations
 1020 (WGSA) are listed. Data is provided for the discovery (TOPMed freeze8) and replication cohorts (MGB Biobank and Penn
 1021 Medicine Biobank). Meta-analysis with the replication cohorts was carried out and the corresponding beta and p-values
 1022 are provided.

1023 GWAS – Genome Wide Association Study; MVP – Million Veteran Program; LDL-C – Low-Density Lipoprotein
 1024 Cholesterol; TC – Total Cholesterol; TG – Triglycerides; TOPMed – Trans-Omics for Precision Medicine; WGSA – Whole
 1025 Genome Sequence Annotations.

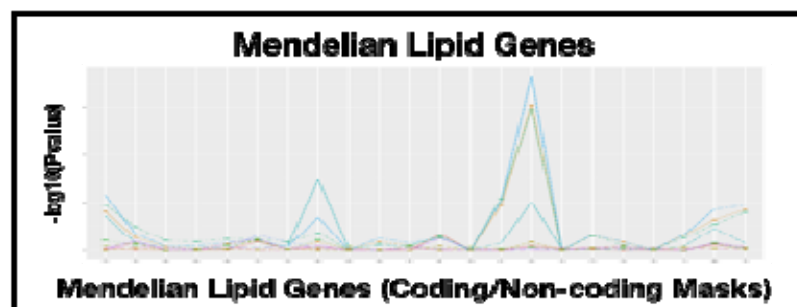
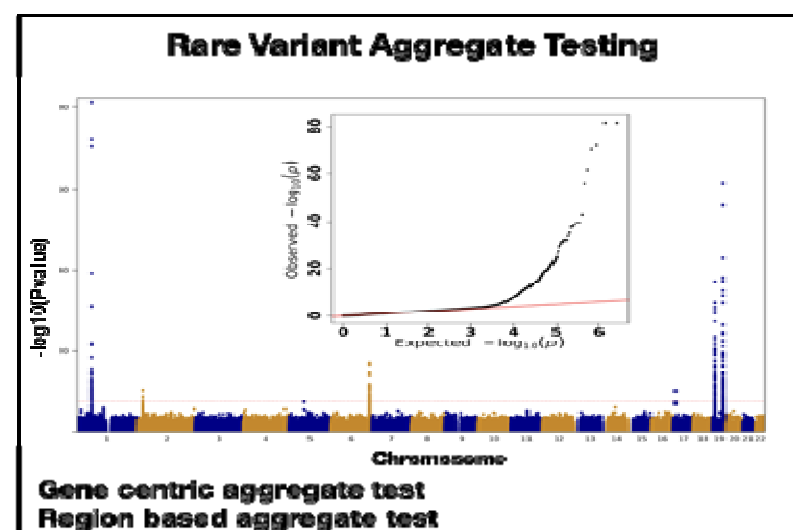
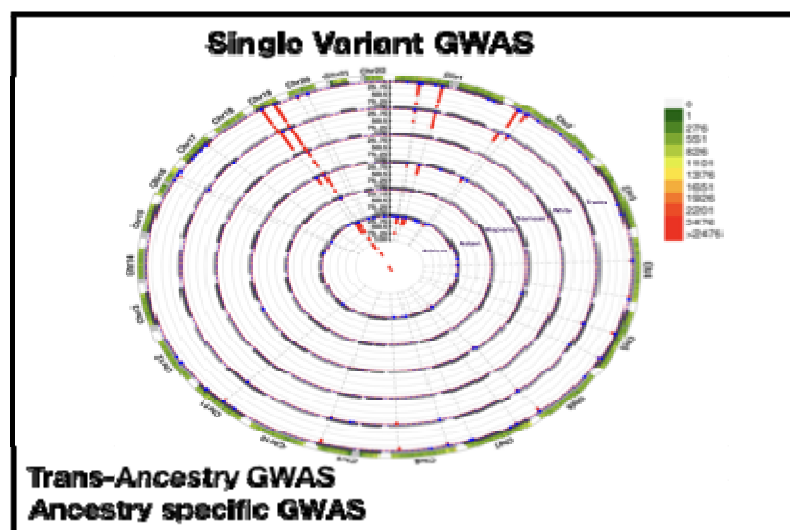
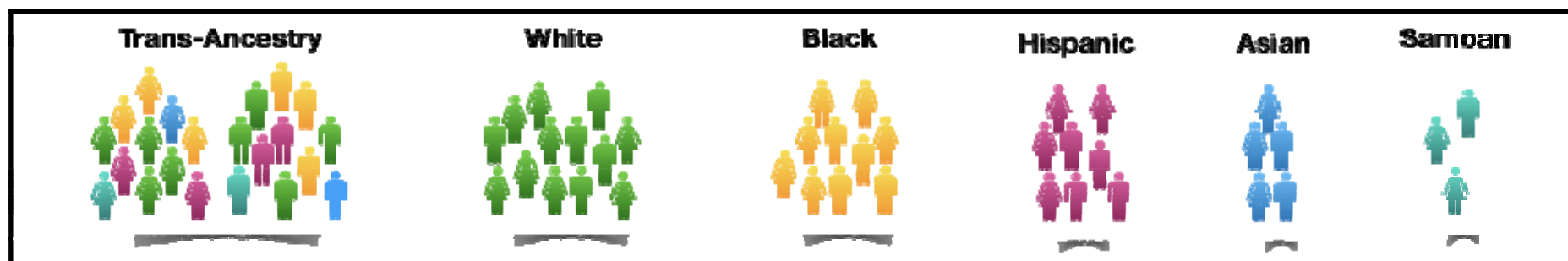


Fig. 1

Overall study schematic. 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 STAARtopmed. 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.

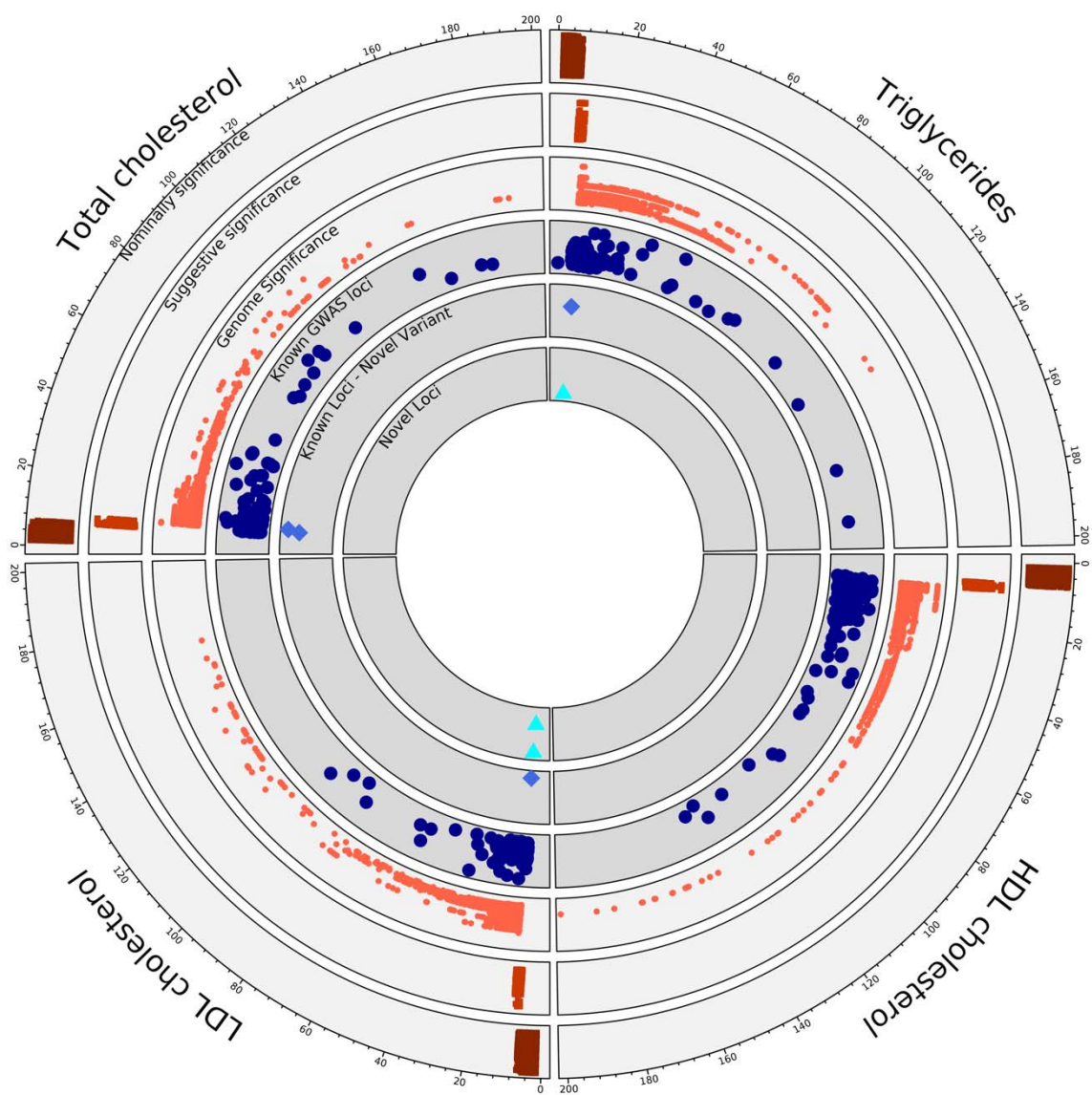


Fig. 2

Summary of single variant genome wide association. Representation of the single variant GWAS results from TOPMed Freeze 8 whole genome sequenced data of 66,329 samples. Each quarter represents a different lipid phenotype, and dots extending in clock-wise fashion represent variants with increasing evidence of association as noted by $-\log_{10}(\text{p-value})$, which was truncated at 200. The outer three circles show the GWAS data from TOPMed freeze8 where variants binned to nominally significant ($\text{p-value } 0.05 - 5 \times 10^{-07}$), suggestive significant ($\text{p-value } 5 \times 10^{-07} - 5 \times 10^{-09}$) and genome wide significant ($\text{p-value } < 5 \times 10^{-09}$). The inner three circles compare our TOPMed results with known significantly associated lipid loci and variants from the MVP summary statistics and GWAS catalog to the identified novel variants and loci that are genome-wide significant from the current study, respectively.

TOPMed – Trans-Omics for Precision Medicine; GWAS – Genome Wide Association Study; MVP – Million Veteran Program.

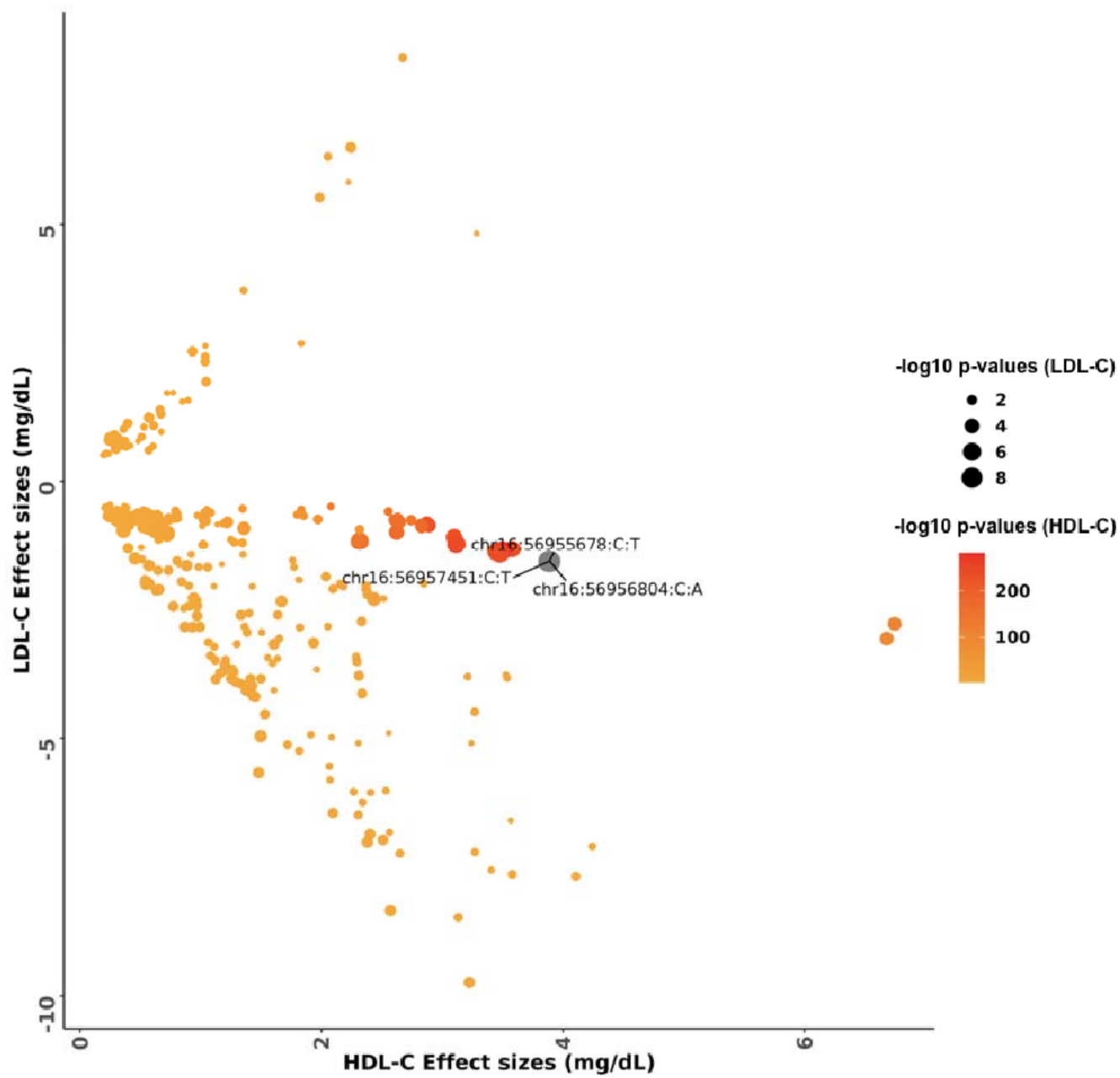
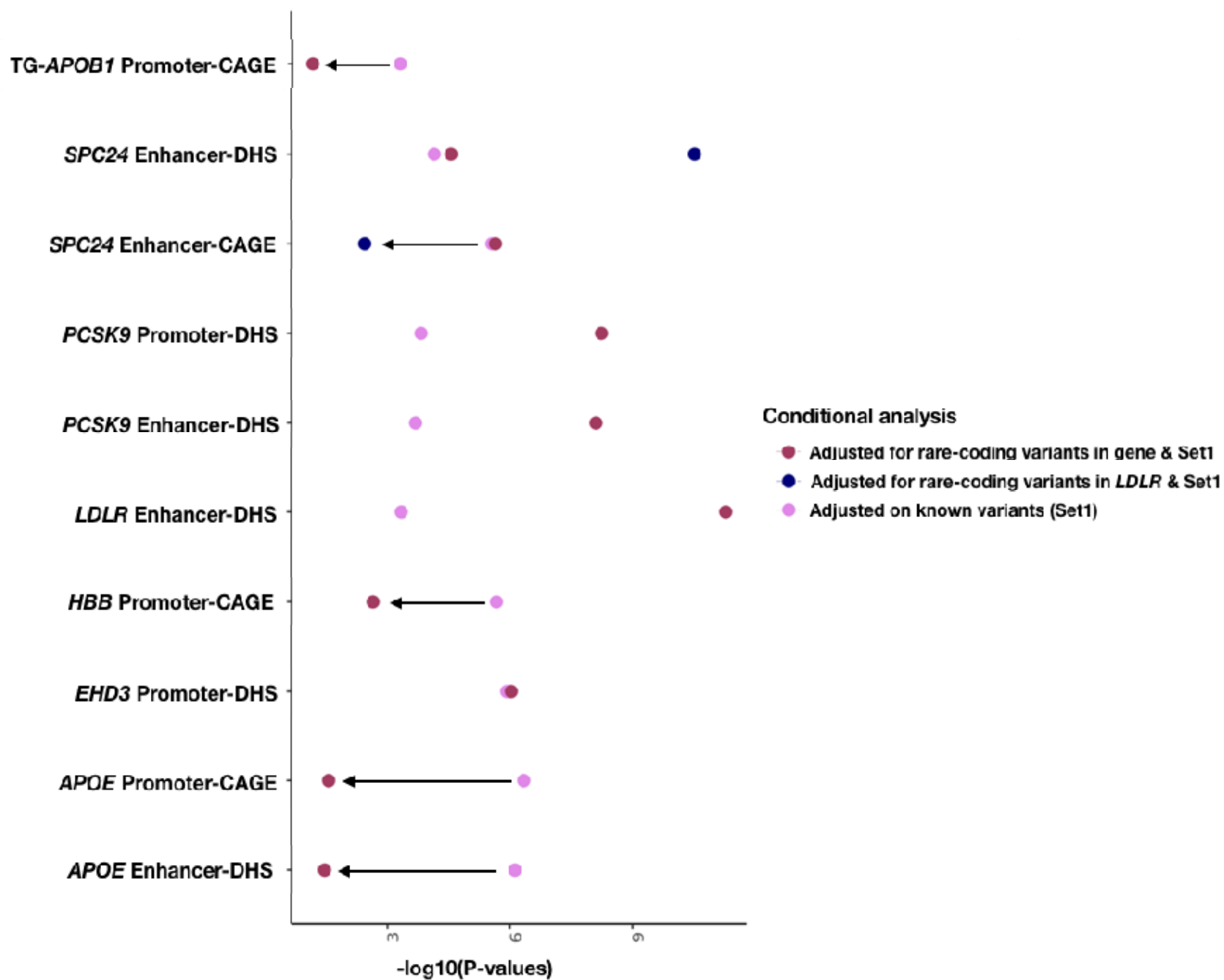


Fig. 3

Comparison of effects estimates for HDL-C and LDL-C among variants in the *CETP* locus. 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.

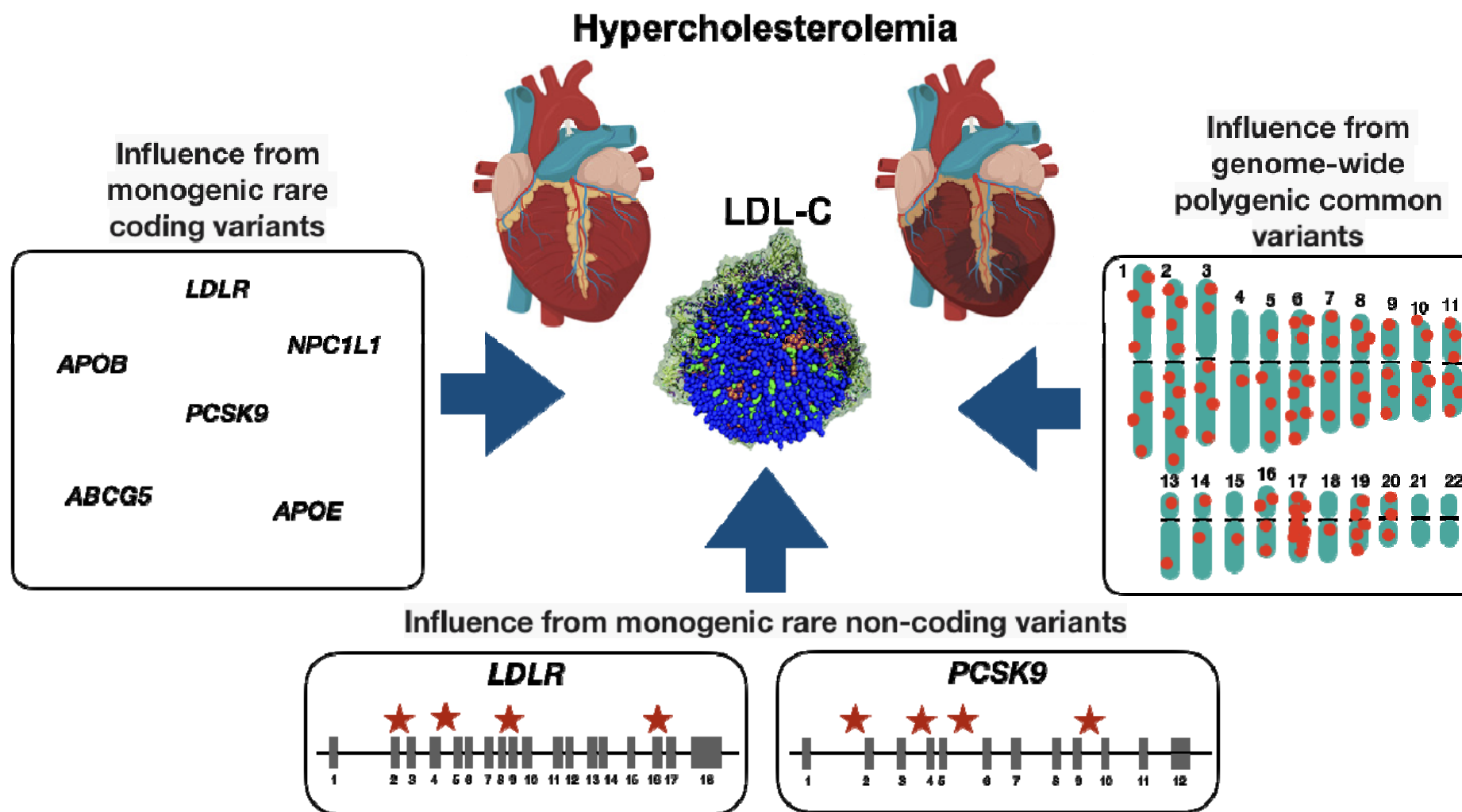
HDL-C – High-Density Lipoprotein Cholesterol; LDL-C – Low-Density Lipoprotein Cholesterol.



1063 **Fig. 4**

1064 **Conditional analysis of coding rare-variants from the same gene and a near-by gene.** Non-coding rare variant sets
 1065 significantly associated with TC and TG after the conditional analysis on known variants are shown with additional
 1066 adjustment on rare-coding variants. The additional adjustment for rare-coding variants were carried out for the same gene
 1067 of the aggregate set and for certain gene aggregates (*SPC24*) the conditional analysis was carried out with a nearby
 1068 Mendelian gene. After adjusting for rare-coding variants and known variants, *EHD3* signal drops minimally, whereas
 1069 signal from *PCSK9* (promoter-DHS, enhancer-DHS), *LDLR*-loci (enhancer-DHS, *SPC24* enhancer-DHS) enhances
 1070 significantly. *APOB1*, *SPC24* (enhancer-CAGE), *HBB* and *APOE* signal drops after the conditional analysis on rare-coding
 1071 variants. The different colored dots on the plot represents the conditional STAAR-O p-values when adjusting for known
 1072 variants (Set1) and rare-coding variants of the same or near-by gene.
 1073 STAAR – variant-Set Test for Association using Annotation information; TC – Total Cholesterol; TG – Triglycerides; CAGE
 1074 – Cap Analysis of Gene Expression; DHS – DNase hypersensitivity.

1075



1076

1077

1078 **Fig. 5**

1079 **Influence of common and rare variants with hypercholesterolemia.** In addition to monogenic contributions from rare
1080 variants in Mendelian hypercholesterolemia genes, multiple genome-wide significant LDL-C-associated common variants
1081 also yield a polygenic basis for hypercholesterolemia. In the present work, we now identify rare non-coding variants in
1082 proximity of Mendelian hypercholesterolemia genes, specifically *LDLR* and *PCSK9*, that also contribute to the genetic
1083 basis of hypercholesterolemia.

1084 LDL-C – Low-Density Lipoprotein Cholesterol

