medRxiv preprint doi: https://doi.org/10.1101/2023.06.28.23291966; this version posted June 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Rare variants in long non-coding RNAs are associated with blood lipid levels in the TOPMed Whole Genome Sequencing Study

Yuxuan Wang[1], Margaret Sunitha Selvaraj[2,3,4], Xihao Li[5], Zilin Li[5,6,7], Jacob A. Holdcraft[1], Donna K. Arnett[8,9], Joshua C. Bis[10], John Blangero[11], Eric Boerwinkle[12], Donald W. Bowden[13], Brian E. Cade[14,15], Jenna C. Carlson[16,17], April P. Carson[18], Yii-Der Ida Chen[19], Joanne E. Curran[11], Paul S. de Vries[12], Susan K. Dutcher[20], Patrick T. Ellinor[21,22], James S. Floyd[10,23], Myriam Fornage[24], Barry I. Freedman[25], Stacey Gabriel[26], Soren Germer[27], Richard A. Gibbs[28], Xiuqing Guo[19], Jiang He[29,30], Nancy Heard-Costa[31,32], Bertha Hildalgo[33], Lifang Hou[34], Marguerite R. Irvin[33], Roby Joehanes[35], Robert C. Kaplan[36,37], Sharon LR. Kardia[38], Tanika N. Kelly[39], Ryan Kim[40], Charles Kooperberg[37], Brian G. Kral[41], Daniel Levy[31,35], Changwei Li[30,29], Chunyu Liu[1,31], Don Lloyd-Jone[34], Ruth JF. Loos[42,43], Michael C. Mahaney[11], Lisa W. Martin[44], Rasika A. Mathias[41], Ryan L. Minster[45], Braxton D. Mitchell[46], May E. Montasser[46], Alanna C. Morrison[12], Joanne M. Murabito[31,47], Take Naseri[48], Jeffrey R. O'Connell[46], Nicholette D. Palmer[13], Michael H. Preuss[42], Bruce M. Psaty[10,23,49], Laura M. Raffield[50], Dabeeru C. Rao[51], Susan Redline[52], Alexander P. Reiner[23], Stephen S. Rich[53], Muagututi'a Sefuiva Ruepena[54], Wayne H-H. Sheu[55], Jennifer A. Smith[38], Albert Smith[56], Hemant K. Tiwari[57], Michael Y. Tsai[58], Karine A. Viaud-Martinez[59], Zhe Wang[42], Lisa R. Yanek[41], Wei Zhao[38], NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, Jerome I. Rotter[19], Xihong Lin[3,5,60], Pradeep Natarajan[2,3,4], Gina M. Peloso[1,*]

[1]Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA, [2]Cardiovascular Research Center and Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA, [3]Program in Medical and Population Genetics, Broad Institute of

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2023.06.28.23291966; this version posted June 29, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Harvard and MIT, Cambridge, MA, USA, [4]Department of Medicine, Harvard Medical School, Boston, MA, USA, [5]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA, [6]Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, IN, USA, [7]Center for Computational Biology & Bioinformatics, Indiana University School of Medicine, Indianapolis, IN, USA, [8]Provost Office, University of South Carolina, Columbia, SC, USA, [9]Department of Epidemiology and Biostatistics, University of South Carolina Arnold School of Public Health, Columbia, SC, USA, [10]Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA, USA, [11]Department of Human Genetics and South Texas Diabetes and Obesity Institute, University of Texas Rio Grande Valley School of Medicine, Brownsville, TX, USA, [12]Human Genetics Center, Department of Epidemiology, Human Genetics, and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, USA, [13]Department of Biochemistry, Wake Forest University School of Medicine, Winston-Salem, NC, USA, [14]Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA, [15]Division of Sleep Medicine, Harvard Medical School, Boston, MA, USA, [16]Department of Human Genetics, School of Public Health, University of Pittsburgh, Pittsburgh, PA, USA, [17]Department of Biostatistics, School of Public Health, University of Pittsburgh, Pittsburgh, PA, USA, [18]Department of Medicine, University of Mississippi Medical Center, Jackson, MS, USA, [19]The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA, USA, [20]The McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO, USA, [21]Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, USA, [22]Cardiovascular Disease Initiative, The

medRxiv preprint doi: https://doi.org/10.1101/2023.06.28.23291966; this version posted June 29, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Broad Institute of MIT and Harvard, Cambridge, MA, USA, [23]Department of Epidemiology, University of Washington, Seattle, WA, USA, [24]Center for Human Genetics, University of Texas Health at Houston, Houston, TX, USA, [25]Department of Internal Medicine, Nephrology, Wake Forest University School of Medicine, Winston-Salem, NC, USA, [26]Broad Institute of Harvard and MIT, Cambridge, MA, USA, [27]New York Genome Center, New York, NY, USA, [28]Baylor College of Medicine Human Genome Sequencing Center, Houston, TX, USA, [29]Department of Epidemiology, Tulane University School of Public Health and Tropical Medicine, New Orleans, LA, USA, [30]Tulane University Translational Science Institute, New Orleans, LA, USA, [31]Framingham Heart Study, Framingham, MA, USA, [32]Department of Neurology, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, USA, [33]Department of Epidemiology, University of Alabama at Birmingham School of Public Health, Birmingham, AL, USA, [34]Department of Preventive Medicine, Northwestern University, Chicago, IL, USA, [35]Population Sciences Branch, Division of Intramural Research, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD, USA, [36]Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY, USA, [37]Division of Public Health Sciences, Fred Hutchinson Cancer Center, Seattle, WA, USA, [38]Department of Epidemiology, University of Michigan, Ann Arbor, MI, USA, [39]Department of Medicine, Division of Nephrology, University of Illinois Chicago, Chicago, IL, USA, [40]Psomagen, Inc. (formerly Macrogen USA), Rockville, MD, USA, [41]GeneSTAR Research Program, Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA, [42]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA, [43]NNF Center for Basic Metabolic Research, University of Copenhagen, Cophenhagen, Denmark, [44]George Washington University School of Medicine and Health

medRxiv preprint doi: https://doi.org/10.1101/2023.06.28.23291966; this version posted June 29, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Sciences, Washington, DC, USA, [45]Department of Human Genetics and Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA, USA, [46]Department of Medicine, University of Maryland School of Medicine, Baltimore, MD, USA, [47]Department of Medicine, Boston Medical Center, Boston University Chobanian and Avedisian School of Medicine, Boston, MA, USA, [48]Ministry of Health, Apia, Samoa, [49]Department of Health Systems and Population Health, University of Washington, Seattle, WA, USA, [50]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, [51]Division of Biostatistics, Washington University School of Medicine, St. Louis, MO, USA, [52]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA, [53]Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA, [54]Lutia i Puava ae Mapu i Fagalele, Apia, Samoa, [55]National Health Research Institute (NHRI), Taiwan, [56]Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA, [57]Department of Biostatistics, University of Alabama, Birmingham, AL, USA, [58]Department of Laboratory Medicine and Pathology, University of Minnesota, Minneapolis, MN, USA, [59]Illumina Laboratory Services, Illumina Inc., San Diego, CA, USA, [60]Department of Statistics, Harvard University, Cambridge, MA, USA

medRxiv preprint doi: https://doi.org/10.1101/2023.06.28.23291966; this version posted June 29, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

*Corresponding author:

Gina M. Peloso, PhD

Department of Biostatistics

Boston University School of Public Health

Boston, MA 02118

gpeloso@bu.edu

Abstract word counts: 240

Text word counts: 4787

No. of references: 56

No. of figures: 3

No. of tables: 1

medRxiv preprint doi: https://doi.org/10.1101/2023.06.28.23291966; this version posted June 29, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Abstract

Long non-coding RNAs (lncRNAs) are known to perform important regulatory functions. Large-scale whole genome sequencing (WGS) studies and new statistical methods for variant set tests now provide an opportunity to assess the associations between rare variants in lncRNA genes and complex traits across the genome. In this study, we used high-coverage WGS from 66,329 participants of diverse ancestries with blood lipid levels (LDL-C, HDL-C, TC, and TG) in the National Heart, Lung, and Blood Institute (NHLBI) Trans-Omics for Precision Medicine (TOPMed) program to investigate the role of lncRNAs in lipid variability. We aggregated rare variants for 165,375 lncRNA genes based on their genomic locations and conducted rare variant aggregate association tests using the STAAR (variant-Set Test for Association using Annotation infoRmation) framework. We performed STAAR conditional analysis adjusting for common variants in known lipid GWAS loci and rare coding variants in nearby protein coding genes. Our analyses revealed 83 rare lncRNA variant sets significantly associated with blood lipid levels, all of which were located in known lipid GWAS loci (in a ±500 kb window of a Global Lipids Genetics Consortium index variant). Notably, 61 out of 83 signals (73%) were conditionally independent of common regulatory variations and rare protein coding variations at the same loci. We replicated 34 out of 61 (56%) conditionally independent associations using the independent UK Biobank WGS data. Our results expand the genetic architecture of blood lipids to rare variants in lncRNA, implicating new therapeutic opportunities.

medRxiv preprint doi: https://doi.org/10.1101/2023.06.28.23291966; this version posted June 29, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Introduction

Blood lipid levels, including low-density lipoprotein cholesterol (LDL-C), total cholesterol (TC), triglyceride (TG), high-density lipoprotein cholesterol (HDL-C), are quantitative clinically important traits with well-described monogenic and polygenic bases[1–19]. Abnormal blood lipid levels contribute to risk of coronary heart disease (CHD) and, in clinical practice, several treatments, including statins, PCSK9 and ANGPTL3 inhibitors[20–22], are available to reduce the risk of developing CHD. Each of these therapeutics has supporting evidence of their efficacy from human genetic analysis of blood lipid levels[21–23].

Long non-coding RNAs (lncRNAs) are broadly defined as transcripts greater than 200 nucleotides in length that biochemically resemble mRNAs but do not code for proteins[24]. lncRNAs are known to perform important regulatory functions in lipid metabolism[25–27]. Rare variants (RVs) in lncRNAs have not been systematically explored for their impact on blood lipid levels as they are not comprehensively genotyped or imputed on non-WGS platforms. In addition, there are difficulties in defining testing units and selecting qualifying variants[28]. Rapidly growing knowledge about the regulatory elements of the non-coding genome[29–33], large-scale WGS studies [34–36], and new statistical methods [37–39] for variant set tests provide the possibility to assess the associations between plasma lipid traits and the genome-wide impact of lncRNAs.

We examined the associations of rare variants in lncRNA genes from high-coverage WGS of 66,329 participants from diverse ancestry who have blood lipid traits (LDL-C, HDL-C, TC and TG) in the National Heart, Lung, and Blood Institute (NHLBI) Trans-omics for Precision

medRxiv preprint doi: https://doi.org/10.1101/2023.06.28.23291966; this version posted June 29, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Medicine (TOPMed) program freeze 8 data[34]. We show that the rare noncoding variants in lncRNA genes located near known Mendelian dyslipidemia genes contribute to phenotypic variation in lipid levels among unselected individuals from population-based cohorts biobanks independently of common variants associated with blood lipid levels.

Results

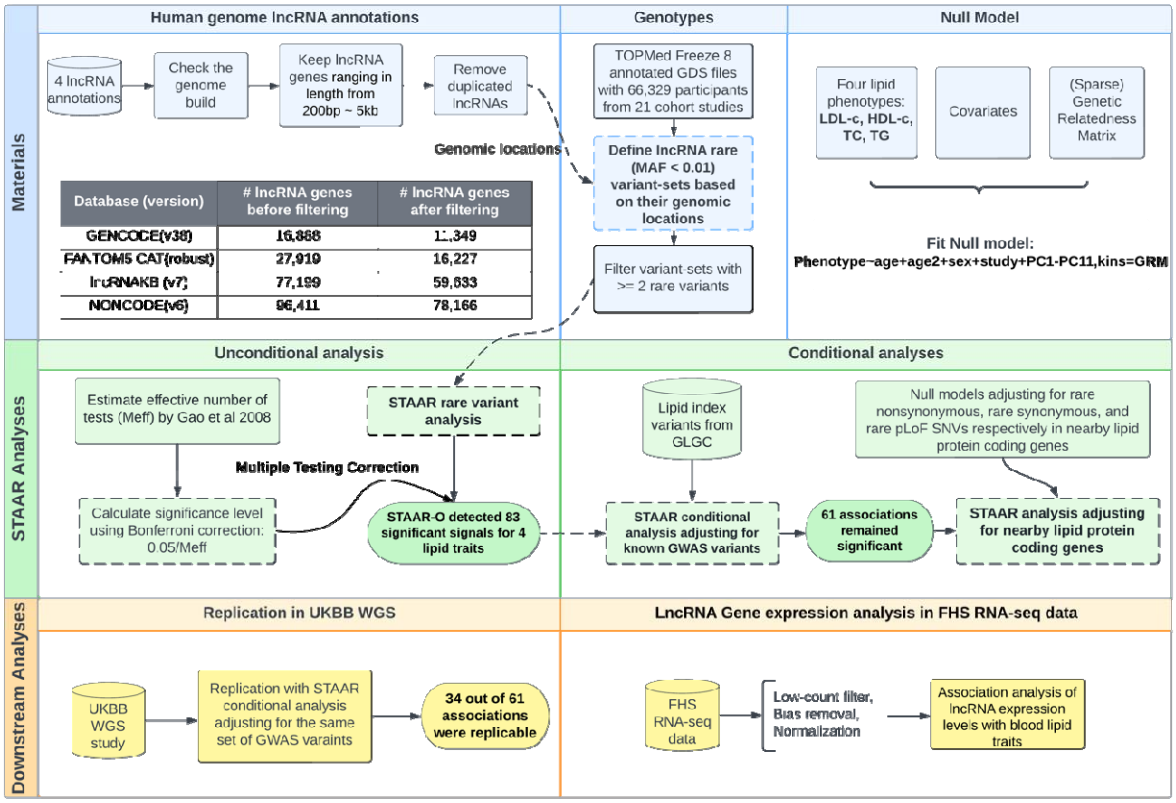
Overview

We performed a comprehensive evaluation of the association between quantitative blood lipid traits and rare variants in lncRNA genes across the genome (**Figure 1**). We systematically curated more than 165k lncRNA genes from the union of four human genome lncRNA annotations, including GENCODE [29,30], FANTOM5 CAT[31], NONCODE[32] and lncRNAKB[33]. We utilized the TOPMed Freeze 8 dataset of 66,329 participants from 21 studies with WGS and measured blood lipid levels and performed the rare variant (MAF <1%) association tests of curated lncRNA genes with four blood lipid phenotypes: LDL-C, HDL-C, TC, and TG. We further conducted the conditional analysis adjusting for known genome-wide association study (GWAS) variants from the Global Lipids Genetics Consortium (GLGC)[18]. Associations between lncRNA genes and lipids that were conditionally independent from the GWAS variants (conditional *P* value < 6.0e-04) were then tested using STAAR procedure for conditional analysis adjusting for rare nonsynonymous variants (MAF < 1%) within the closest protein coding gene and the nearby known lipid monogenic genes in the region. We performed replication in ~140□K genomes from UK Biobank[40]. We intersected our results with the gene expression signatures of lipid traits in 1,505 participants from the Framingham Heart Study

medRxiv preprint doi: https://doi.org/10.1101/2023.06.28.23291966; this version posted June 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

(FHS)[41] with RNA-seq data and blood lipid levels and observed evidence that the lncRNA RVs may both influence their gene expression levels and impact lipid traits.

Figure 1. A schematic illustration of the study.



Characteristics of TOPMed participants

We included 66,329 diverse participants from 21 cohort studies in the NHLBI TOPMed consortium with blood lipid levels. The discovery cohorts consisted of 29,502 (44.5%) self-reported White, 16,983 (25.6%) self-reported Black, 13,943 (21.0%) self-reported Hispanic, 4,719 (7.1%) self-reported Asian, and 1,182 (1.8%) self-reported Samoan participants (**Supplementary Table 1, Supplementary Text**). Among the 66,329 participants, 41,182 (62%)

medRxiv preprint doi: https://doi.org/10.1101/2023.06.28.23291966; this version posted June 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

were female. The mean age of the 66,329 participants was 53 years (SD = 15). The mean ages at lipid measurement varied across 21 cohorts from 25 years (SD = 3.56) for the Coronary Artery Risk Development in Young Adults (CARDIA) to 73 years (SD = 5.38) for the Cardiovascular Health Study (CHS). We observed that the Amish cohort had a higher 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*[35].

Identification of rare lncRNA variants associated with blood lipid traits

We defined lncRNA testing units using the available genomic positions in four genome annotation projects described in the **Methods**. There were 11,349 lncRNA genes obtained from GENCODE[29,30], 16,227 from FANTOM5 CAT[31], 78,166 from NONCODE[32] and 59,633 from lncRNAKB[33]. In total, we tested 165,375 lncRNA genes, among which, the average number of rare variants in each lncRNA was 483 (SD = 572). The minimum and the maximum number of rare variants among the lncRNAs being tested are 2 and 2947, respectively.

Our aggregation of lncRNAs across four lncRNA resources led to an overlap in the lncRNA units, leading to non-independent tests of association of the lncRNAs with blood lipid levels. We estimated the effective number of tests (M_{eff}) using a principal component analysis (PCA) based approach[42] since the traditional Bonferroni correction would be too conservative and reduce power to detect association with blood lipid levels[28]. M_{eff} was estimated as 111,550, providing a significance threshold of $\alpha = 0.05/111{,}550 = 4.5 \times 10^{-7}$.

medRxiv preprint doi: https://doi.org/10.1101/2023.06.28.23291966; this version posted June 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Table 1. Summary of significant lncRNA associations for unconditional analysis, conditional analyses, and replication.

Method	LDL-C	TC	HDL-C	TG	Total No.
STAAR Unconditional analysis*	28	20	19	16	83
Conditioning on known lipid-associated variants **	20	14	15	12	61
Conditioning on rare nonsynonymous variants within the closest gene and nearby lipid monogenic genes ***	18	13	15	12	58
Conditioning on rare synonymous variants within the closest gene and nearby lipid monogenic genes ***	20	14	15	12	61
Conditioning on rare pLoF variants within the closest gene and nearby lipid monogenic genes ***	20	14	15	12	61
Replication in UKBB WGS ***	13	7	8	6	34

* Bonferroni correction level of 0.05/111,550 = 4.5e-07

**Bonferroni correction level of 0.05/83 = 6.0e-04

***Bonferroni correction level of 0.05/61 = 8.2e-04

We applied STAAR (variant-Set Test for Association using Annotation infoRmation) framework[37,38] to identify the lncRNA rare variant (RV) sets that associated with quantitative lipid traits (LDL-C, HDL-C, TC and TG) using TOPMed WGS data. STAAR-O identified 83 genome-wide significant associations (28 with LDL-C, 20 with TC, 19 with HDL-C, and 16 with

medRxiv preprint doi: https://doi.org/10.1101/2023.06.28.23291966; this version posted June 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

TG) (**Table 1, Supplementary Table 2)**. Among the 83 genome-wide significant associations, there are 54 unique lncRNAs. We observed that all the significant associations in the unconditional analysis were in the known lipid GWAS loci (defined as a ±500 kb window beyond a Global Lipids Genetics Consortium index variant)[18]. We performed a sensitivity analysis aggregating only exonic and splicing variants in lncRNA genes and observed consistent results to our primary analysis results (**Supplementary Figure 1**).

Conditional analyses of trait-associated lncRNAs adjusting for known GWAS variants and nonsynonymous variants within the nearby lipid monogenic genes

After conditioning on known lipid-associated variants in a ±500 kb window beyond a variant set[18], 61 out of 83 associations (73%) remained significant (20 with LDL-C, 14 with TC, 15 with HDL-C, and 12 with TG) at the Bonferroni corrected level of $0.05/83 = 6.0 \times 10^{-4}$, indicating that the associations between the lncRNA genes and lipid levels are distinct from the known GWAS variants. The most significant association for LDL-C and TC was the lncRNA NONHSAG026007.2 (chr19:44,892,420-44,903,056) near the *APOE-APOC1* region. NONHSAG026007.2 remained significantly associated with LDL-C (*P* value = 2.44×10^{-15}) and TC (*P* value =2.17×10^{-27}) after adjusting for nearby known lipid-associated variants (**Figure 2**). The most significant associations for HDL-C and TG were NONHSAG063125.1 (chr11:116,790,241-116,805,983) and NONHSAG09700.3 (chr11: 116,773,068-116,779,841), respectively, both near *APOA5-APOC3-APOA1* region. NONHSAG063125.1 remained similarly associated after conditioning on known lipid GWAS variants, while NONHSAG09700.3 became even more significant **(Figure 2**). We then conditioned the GWAS-distinct associations on the rare nonsynonymous variants within the closest protein coding gene and nearby lipid monogenic

medRxiv preprint doi: https://doi.org/10.1101/2023.06.28.23291966; this version posted June 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

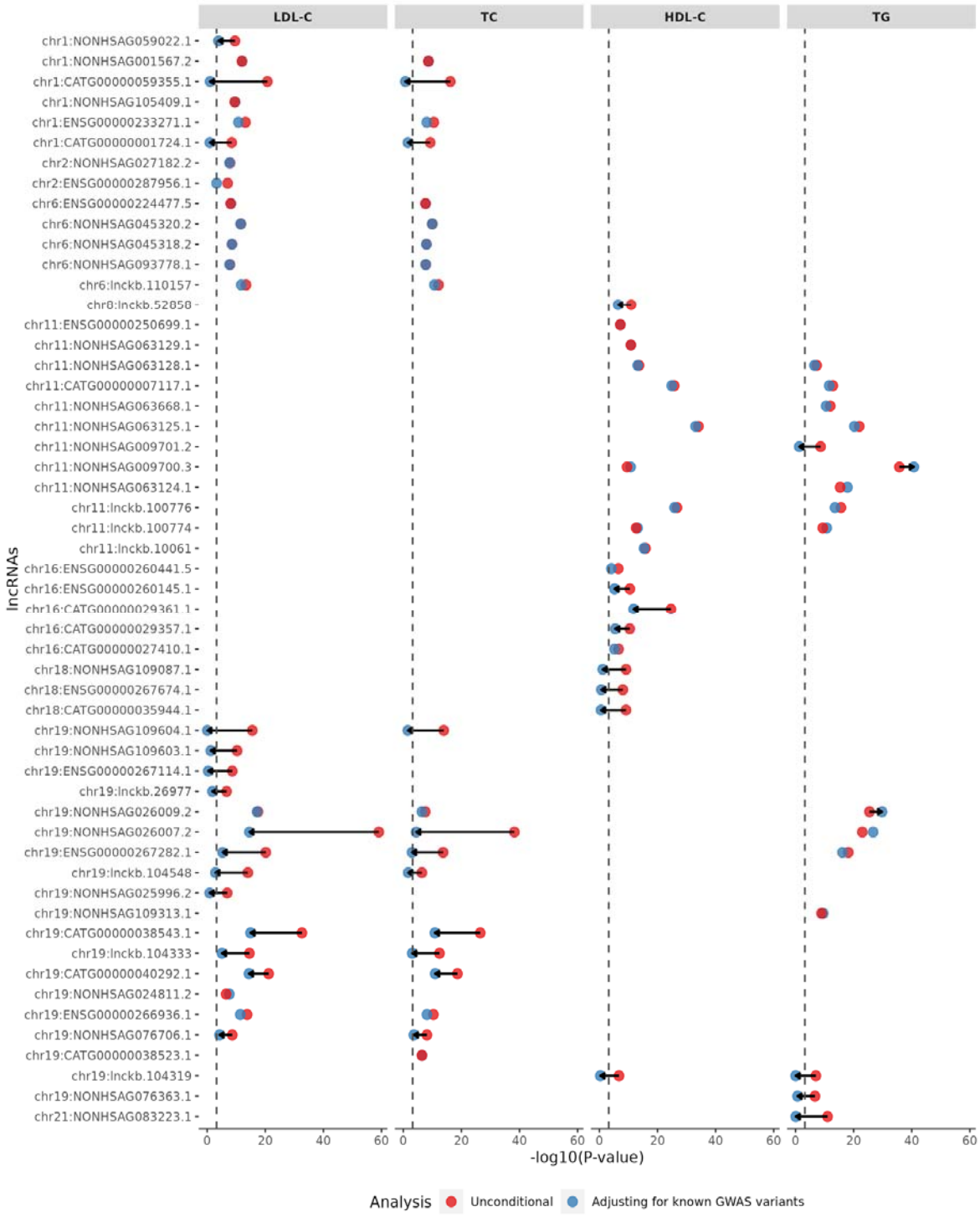
genes and observed that most (94.9%) of the lncRNA associations with lipid levels remained significant (**Table 1; Supplementary Figure 2**). Additionally, when conditioned on the rare synonymous variants or rare pLoF variants within the closest protein coding gene and nearby lipid monogenic genes, the number of associations remained as same as those GWAS-distinct associations (**Table 1; Supplementary Figure 3)**.

Replication of significant lncRNA-blood lipid trait associations

Replication of 61 lncRNAs associated with blood lipid levels was evaluated in 139,849 UK Biobank individuals with WGS and blood lipid levels (**Supplementary Table 3)**. We replicated 34 out of 61 (56%) lncRNA associations with blood lipid levels at a Bonferroni-corrected threshold of 0.05/61 = 8.2e-04 (**Supplementary Table 2**). The most significant associations in the UK Biobank replication were NONHSAG025996.2 (chr19: 44,694,720-44,696,054) near *APOE-APOC1* region for LDL-C, NONHSAG109604.1 near *APOE-APOC1* region for TC, NONHSAG009700.3 near *APOA5-APOC3-APOA1* region for both HDL-C and TG (**Supplementary Table 2**), which were consistent with the results from TOPMed.

Figure 2. Significantly associated lncRNAs with four blood lipid traits (STAAR-O *P* value < 4.5e-07). The lncRNA genes are ordered by chromosome, followed by genomic positions. Dots in red and blue represent the $-\log_{10}$(STAAR-O *P* value) of the STAAR unconditional and conditional analysis adjusting for known lipid-associated GWAS variants, respectively. The black dashed line is the Bonferroni correction level of 0.05/83 = 6.0e-04. Arrows indicate at least 10^4 fold change of STAAR-O *P* values comparing the unconditional analysis and conditional analysis adjusting for known lipid-associated GWAS variants.

medRxiv preprint doi: https://doi.org/10.1101/2023.06.28.23291966; this version posted June 29, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .



259

260

261

medRxiv preprint doi: https://doi.org/10.1101/2023.06.28.23291966; this version posted June 29, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

lncRNA gene expression analysis in FHS RNA-seq data

We overlapped the significant lipid-associated lncRNA genes with the lncRNA genes available in the Framingham Heart Study (FHS) RNA-seq data generated by TOPMed[43]. Since the gene-level expression data in FHS is annotated by GENCODE v30, we limited the lncRNA genes to those presented in GENCODE. Among the 54 unique lncRNA genes that are significantly associated with either one of the lipid traits using TOPMed WGS data, 10 lncRNA genes are annotated by GENCODE, and 8 out of 10 can be found in the FHS data. We performed association analyses of expression levels of those 8 significant lipid-associated lncRNA genes with blood lipid levels (LDL-C, TC, HDL-C, TG) (**Supplementary Text, Supplementary Table 4**). In total, we tested 12 associations of lncRNA gene expression with blood lipid level (**Supplementary Table 4**). The small proportion of overlapping was partially due to lncRNA genes' generally lower expression. The lowly expressed genes were filtered out when processing the gene expression data.

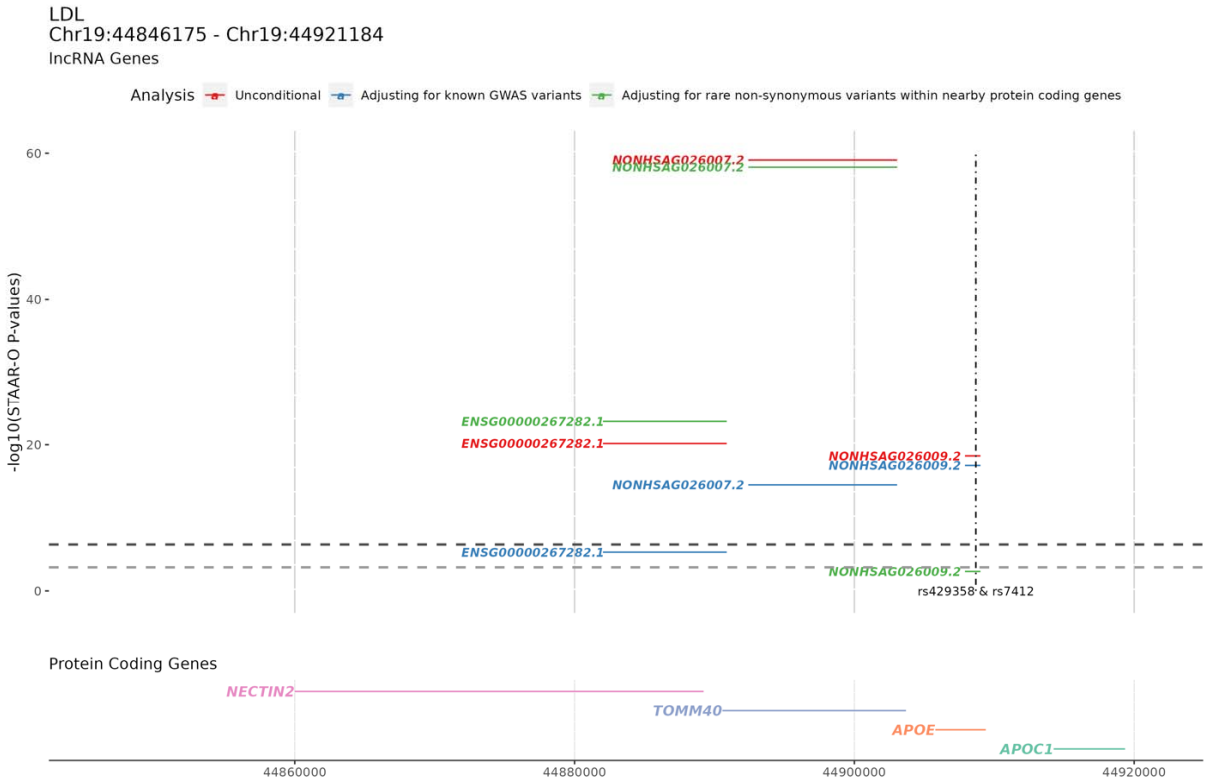
Four associations achieved Bonferroni-adjusted significance, including the gene expression level of ENSG00000267282.1 (chr19:44,881,088-44,890,922) associated with LDL-C, TC, and TG, and the gene expression level of ENSG00000266936.1 (chr19:11,010,917-11,016,011) associated with TC. ENSG00000267282.1 is an antisense of *NECTIN2* (also known as *PVRL2*) (**Figure 3**). The nectin cell adhesion molecule 2 (*NECTIN2*) protein is a cell adhesion molecule involved in lipid metabolism[44]. Additionally, ENSG00000267282.1 was one of the lncRNA associations that we replicated in the independent UK Biobank (**Supplementary Table 2**). We also queried whether the RVs in this lipid-associated lncRNA led to an alteration of the corresponding lncRNA levels in the blood. However, due to the small number of overlapping

medRxiv preprint doi: https://doi.org/10.1101/2023.06.28.23291966; this version posted June 29, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

individuals between FHS RNA-seq data and TOPMed WGS data (N = 512), the number of RVs tested in ENSG00000267282.1 for the association of its gene expression level was only 59. Compared with the original analysis using all 66,329 individuals for the association with lipid levels, the number of RVs tested in ENSG00000267282.1 is 1417. As a result, the association of the RVs in the ENSG00000267282.1 with ENSG00000267282.1 gene expression levels in blood was not significant (STAAR-O *P* value = 0.68).

Figure 3. lncRNAs in the APOE region associated with LDL-C. Upper panel shows the -$\log_{10}$(STAAR-O *P* value) of the STAAR unconditional analysis, STAAR conditional analysis adjusting on known lipid GWAS variants, and STAAR conditional analysis adjusting for rare non-synonymous variants within the closest protein-coding gene and nearby lipid monogenic genes. The bottom panel is the nearby protein coding genes with the genomic coordinates. The vertical dashed line is the position of the known GWAS variants that were conditioned on. The black horizontal dashed line is the Bonferroni correction level of 0.05/111,550 = 4.5e-07, and the gray horizontal dashed line is the Bonferroni correction level of 0.05/83 = 6.0e-04.

medRxiv preprint doi: https://doi.org/10.1101/2023.06.28.23291966; this version posted June 29, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .



300

medRxiv preprint doi: https://doi.org/10.1101/2023.06.28.23291966; this version posted June 29, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Discussion

In this study, we conducted genome-wide rare-variant associations of 165K lncRNAs in ancestrally diverse TOPMed participants (N = 66,329) with measured blood lipid levels. Using rare-variant association tests, we observed 83 rare lncRNAs significantly associated with blood lipid levels, and of these, 61 (73%) were conditionally distinct from common regulatory variation and rare protein coding variation at the same loci. Notably, most of these association signals were replicated in an independent WGS dataset, UK Biobank. We also highlighted one trait-associated lncRNA, ENSG00000267282.1(chr19:44,881,088-44,890,922), whose gene expression level was also shown to be associated with lipid levels using RNA-seq data from the FHS. Together, this systematic assessment of rare lncRNA variants suggests an additional genomic element in known lipid gene regions that is distinct from the known lipid genes.

Genetic variation for blood lipids levels has been observed across the allelic spectrum with common, rare coding, and rare non-coding variants being associated with blood lipids levels[36]. Blood lipids have been associated with non-coding regulatory variants and coding variation in genes, and now also associated with lncRNAs. We show that all the trait-associated lncRNAs are in genomic regions previously associated with blood lipid traits, leading to the plausibility of these results. About 75% of the associations are conditionally distinct from common regulatory variation and rare protein coding variation at the same loci previously identified through GWAS and whole exome sequencing studies. This indicates that the regulatory variants through lncRNAs additionally contribute to the variation of blood lipid levels.

medRxiv preprint doi: https://doi.org/10.1101/2023.06.28.23291966; this version posted June 29, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Despite numerous reports indicating the potential regulatory role of long non-coding RNAs (lncRNAs), only a small proportion of them have substantial evidence to support such claims[25,26,45]. The fraction of lncRNAs that are functional remains unknown. Through a comprehensive study of over 165,000 lncRNAs, we found that the majority of lncRNAs are not associated with a lipid trait, which aligns with the argument made previously that only a few human lncRNAs contribute centrally to human physiology[45]. However, there are still some lncRNAs that harbor variants that predispose individuals to phenotypic differences in blood lipid levels. Our results suggest that investigators should first prioritize individual lncRNAs near the known trait-associated loci for analysis, which is more likely to yield robust experimental observations.

We further investigated one lncRNA, liver-expressed liver X receptor-induced sequence (*LeXis*), which is a mediator of the complex effects of liver X receptor (LXR) signaling on hepatic lipid metabolism to maintain hepatic sterol content and serum cholesterol levels[46,47]. A potential orthologue of *LeXis* in humans, TCONS_00016452 (chr9:104,990,086-104,991,780), is found in a region adjacent to the human *ABCA1* gene. It didn't stand out as a significant signal for any lipid trait in our study, which might suggest that it was not a functional orthologue of *LeXis*. However, the rapid evolutionary turnover of lncRNAs still hinders the functional identification between species[45,47].

Several limitations of our study should be noted. First, our RNA-seq analyses were restricted to GENCODE annotation. The small proportion of overlapping RNA-seq data and WGS data limits the ability to test rare lncRNA variants with their gene expression. Second, we did not correct for

medRxiv preprint doi: https://doi.org/10.1101/2023.06.28.23291966; this version posted June 29, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

the number of tested lipid traits however, there is a moderate to high correlation among the blood lipid levels and therefore this would lead to over correction. Third, to assess a causal role of the rare lncRNA variants, we need to further show that they are correlated with lncRNA expression but not correlated with altered expression or function of other genes nearby.

In summary, our results from a large ancestrally diverse participants add further evidence that lncRNA is an additional genomic element in known lipid gene regions that is distinct from the known genes. We comprehensively evaluated 165K lncRNAs for their association with variation in lipid traits and replicated most of the signals in an independent UKB WGS cohort.

medRxiv preprint doi: https://doi.org/10.1101/2023.06.28.23291966; this version posted June 29, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Methods

Discovery and replication cohorts

Discovery cohorts. The discovery cohort included 66,329 participants in the NHLBI Trans-Omics for Precision Medicine (TOPMed) from 21 cohort studies with Freeze 8 whole genome sequencing (WGS) and blood lipid levels available: Old Order Amish (Amish; n = 1083), Atherosclerosis Risk in Communities study (ARIC; n = 8016), Mt Sinai BioMe Biobank (BioMe; n = 9848), Coronary Artery Risk Development in Young Adults (CARDIA; n = 3,056), Cleveland Family Study (CFS; n = 579), Cardiovascular Health Study (CHS; n = 3,456), Diabetes Heart Study (DHS; n = 365), Framingham Heart Study (FHS; n = 3992), Genetic Studies of Atherosclerosis Risk (GeneSTAR; n = 1757), Genetic Epidemiology Network of Arteriopathy (GENOA; n = 1046), Genetic Epidemiology Network of Salt Sensitivity (GenSalt; n = 1772), Genetics of Lipid-Lowering Drugs and Diet Network (GOLDN; n = 926), Hispanic Community Health Study - Study of Latinos (HCHS-SOL; n = 7714), Hypertension Genetic Epidemiology Network and Genetic Epidemiology Network of Arteriopathy (HyperGEN; n = 1853), Jackson Heart Study (JHS; n = 2847), Multi-Ethnic Study of Atherosclerosis (MESA; n = 5290), Massachusetts General Hospital Atrial Fibrillation Study (MGH_AF; n = 683), San Antonio Family Study (SAFS; n = 619), Samoan Adiposity Study (Samoan; n = 1182), Taiwan Study of Hypertension using Rare Variants (THRV; n = 1982) and Women's Health Initiative (WHI; n = 8263). The discovery cohorts consisted of 29,502 (44.5%) White, 16,983 (25.6%) Black, 13,943 (21.0%) Hispanic, 4719 (7.1%) Asian, and 1182 (1.8%) Samoan. More information for study descriptions can be found in **Supplementary Table 1**.

medRxiv preprint doi: https://doi.org/10.1101/2023.06.28.23291966; this version posted June 29, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Replication cohorts. We sought to replicate the findings using the UK Biobank WGS data for 139,849 genomes with blood lipid traits[40]. The UK Biobank is a large, population-based prospective cohort of half a million United Kingdom residents aged 40–69 years. The replication cohorts consisted of 116, 335 White, and 23,335 others **(Supplementary Table 3).**

Ethical regulations. Participants from each of the studies contributing to the NHLBI TOPMed consortium provided informed consent, and all studies were approved by IRBs in each of the participating institutions.

TOPMed WGS Freeze 8 data

Phenotype data. We included four conventionally measured blood lipids in this study: low-density lipoprotein cholesterol (LDL-C), total cholesterol (TC), triglyceride (TG), high-density lipoprotein cholesterol (HDL-C). Detailed phenotype calculation and harmonization were described elsewhere[36]. Briefly, LDL-C was either directly measured or calculated by the Friedewald equation when triglycerides were <400 mg/dL. We adjusted the total cholesterol by dividing by 0.8 and LDL-C by dividing by 0.7 when statins were present[10,35]. For triglycerides, we additionally performed the natural log transformation for analysis, since triglycerides were skewed. We then fitted a linear regression model for each phenotype to get the residuals after adjusting for age, age2, sex, race/ethnicity, study and the first 11 ancestral PCs (as recommended by the TOPMed DCC). For Amish participants, we additionally adjusted for APOB p.Arg3527Gln in LDL-C and TC, and adjusted for APOC3 p.Arg19Ter in HDL-C and TG[48–50]. The residuals were inverse rank normalized and rescaled by the standard deviation of the original phenotype within each group[36].

medRxiv preprint doi: https://doi.org/10.1101/2023.06.28.23291966; this version posted June 29, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Genotype data. Whole genome sequencing data were accessed from the TOPMed Freeze 8 release. DNA samples were sequenced at the >30× target coverage 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)[34]. The reads were aligned to human genome build GRCh38 using the BWA-MEM algorithm. The genotype calling was performed using the TOPMed variant calling pipeline (https://github.com/statgen/topmed_variant_calling). The resulting BCF files were converted to SeqArray GDS format and annotated were annotated internally by curating data from multiple database sources using Functional Annotation of Variant–Online Resource (FAVOR (http://favor.genohub.org)[37,39]. The resulting annotated GDS (aGDS) files were used in this study. We computed the genetic relationship matrix (GRM) using R package *PC-relate* and subtracted GRM of those samples with lipid phenotypes using R package *GENESIS*.

Human reference genome annotations for long non-coding RNA genes

Multiple lncRNA annotations are available. We obtained four long non-coding RNAs (lncRNAs)annotation resources with different qualities and sizes and merged them to improve comprehensiveness. They included GENCODE [29,30], FANTOM5 CAT[31], NONCODE[32] and lncRNAKB[33].

GENCODE. GENCODE is the default human reference genome annotation for both Ensembl and UCSC genome browsers. It is also widely adopted by many large-scale genomic consortiums including TOPMed. GENCODE gene sets cover lncRNAs, pseudogenes and small RNAs in addition to protein-coding genes. The lncRNA annotation in GENCODE is almost entirely

medRxiv preprint doi: https://doi.org/10.1101/2023.06.28.23291966; this version posted June 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

manual, which ensures the quality and consistency of the data. We downloaded the GENCODE v38 (December 2020) human release from https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_38/gencode.v38.long_noncoding_RNAs.gtf.gz, and kept 17,944 lncRNAs genes with a stable identifier and the genomic location information.

FANTOM CAT. The Functional Annotation of the Mammalian genome (FANTOM) CAGE-associated transcriptome (CAT) meta-assembly combines both published sources and in-house short-read assemblies. It utilized CAGE tags, which mark transcription start sites (TSSs), to identify human lncRNA genes with high-confidence 5' ends. We acquired the FANTOM CAT (lv3 robust) lncRNAs assembly from https://fantom.gsc.riken.jp/5/suppl/Hon_et_al_2016/data/assembly/lv3_robust/FANTOM_CAT.lv3_robust.only_lncRNA.gtf.gz. Since the FANTOM5 annotations were on genome version hg19 (GRCh37), we lifted over to genome version hg38 (GRCh38) using the UCSC liftOver tool[51].

lncRNAKB. Long non-coding RNA Knowledgebase (lncRNAKB) is an integrated resource for exploring lncRNA biology in the context of tissue-specificity and disease association. A systematic integration of annotations using a cumulative stepwise intersection method from six independent databases resulted in 77,199 human lncRNA. We downloaded the lncRNAKB v7 from http://lncrnakb.org.

NONCODE. NONCODE database integrated annotations from both literature searches and other public databases. The latest version, NONCODE version 6, is the single largest collection of lncRNAs, describing 96,422 lncRNA genes in humans. Each lncRNA gene in the NONCODE database had been assigned a unique NONCODE ID. We download the whole NONCODE v6

medRxiv preprint doi: https://doi.org/10.1101/2023.06.28.23291966; this version posted June 29, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

459 human data from

460 http://www.noncode.org/datadownload/NONCODEv6_hg38.lncAndGene.bed.gz.

461 **Integration across the lncRNA annotations.** We kept only those lncRNA genes ranging in
462 length from 200 nucleotides (nt) to 5 kilobases (kb). We limited the maximum length of a
463 lncRNA gene to 5kb to control for the computational complexity[52]. Overlapping lncRNA genes
464 between FANTOM and GENCODE using the Ensembl stable identifier were removed. We split
465 each annotation file into individual files by chromosome with the start and end coordinates of the
466 lncRNA genes. All duplicated lncRNAs between annotation files were removed by checking
467 whether they have the same start and end coordinates. We then used the following intersection
468 order based on experimental validation to merge the four lncRNA annotations: 1. GENCODE, 2.
469 FANTOM5 CAT, 3. NONCODE and 4. lncRNAKB. Approximately 165k lncRNA genes were
470 left for further analysis.

471

472 *LncRNA rare variant association test*

473 **lncRNA rare variant sets.** We obtained the start and end genomic coordinates (human genome
474 build GRCh38) of the lncRNA genomic regions from our previously curated lncRNA gene list.
475 We then defined aggregation units by using all the rare variants (MAF <0.01) based on their
476 genomic locations with respect to the start and end genomic coordinates of the lncRNA genes.
477 We removed lncRNA rare variant sets that had less than two rare variants. For sensitivity
478 analysis, we only aggregated exonic and splicing variants in lncRNA genes provided by
479 GENCODE v29, for which is the default genome annotation employed by TOPMed
480 consortium[34].

medRxiv preprint doi: https://doi.org/10.1101/2023.06.28.23291966; this version posted June 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

STAAR unconditional analysis. We applied the STAAR (variant-set test for association using annotation information) framework to identify rare variants in the lncRNA variant sets that are associated with four quantitative lipid traits (LDL-C, HDL-C, TG and TC). STAAR is a scalable and powerful variant-set test that uses an omnibus multi-dimensional weighting scheme to incorporate both qualitative functional categories and multiple in silico variant annotation scores for genetic variants. STAAR accounts for population structure and relatedness and is scalable for analyzing large WGS studies of continuous and dichotomous traits by fitting linear and logistic mixed models[37,38]. To perform the STAAR unconditional analysis, we first fitted a STAAR null model using *fit_null_glmmkin()* function to account for sample relatedness with phenotypic data, covariates and (sparse) genetic relatedness matrix as input. For each of the four lipid phenotypes, we adjusted for age, age2, sex, study and PC1-PC11. We calculated the *P* value for each lncRNA rare variant set using STAAR-O, an omnibus test in the STAAR framework that combines *P* values from multiple annotation-weighted burden tests, SKAT and ACAT-V using the ACAT method. A total of 13 aggregated variant functional annotations were incorporated in STAAR-O, including three integrative scores (CADD[53], LINSIGHT[54] and FATHMM-XF[55]) and 10 annotation principal components (aPCs) (**Supplementary Table 5**)[38]. All analyses were performed using R packages *STAAR* (version 0.9.6) and *STAARpipeline* (version 0.9.6).

STAAR conditional analysis adjusting for known GLGC GWAS variants. We performed conditional analysis to identify lncRNA rare variant association independent of known lipid-associated variants. We obtained a list of 1,750 significant index variants (**Supplementary Table 6)** associated with one or more lipid levels from The Global Lipids Genetics Consortium (GLGC) latest lipid GWAS results[18,19,56]. The positions of SNV were lifted over to genome build 38. We adjusted for known lipid variants in a ±500 kb window beyond a variant set.

medRxiv preprint doi: https://doi.org/10.1101/2023.06.28.23291966; this version posted June 29, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

STAAR rare variant association test adjusting for nearby protein coding genes. The unconditional analysis showed that most lncRNA genes associated with lipids are near known monogenic lipid genes. We sought to perform conditional analyses adjusting lncRNA rare variant sets for nearby protein coding genes. The adjusted nearby protein coding genes can be divided into two categories: the closest protein coding genes and those nearby known lipid monogenic genes, including *ANGPTL8, APOA1, APOA5, APOB, APOC1, APOC3, APOE, CETP, LDLR, LPA, LPL, PCSK7, PCSK9, PLA2G15, TM6SF2*[19]. Our primary analysis was to adjust for only rare nonsynonymous variants (MAF < 1%) within nearby protein coding genes. We did two sensitivity analyses, one adjusted for rare synonymous variants (MAF < 1%) within nearby protein coding genes, and another adjusted for rare predicted loss-of-function (pLoF) variants (MAF < 1%) within nearby protein coding genes. For each participant, we created three burden scores separately by combining the minor allele counts of nonsynonymous, synonymous, and pLoF variants with a MAF < 1% carried within the closest gene and the nearby lipid monogenic genes in a 250kb window. We re-fitted null models similar to the unconditional analysis and added all the burden scores of the closest gene and the nearby lipid monogenic genes (if any) as additional covariates for each lipid phenotype. We then repeated the STAAR procedures to calculate the STAAR-O *P* values after adjusting for rare nonsynonymous, rare synonymous, and rare pLoF variants.

Effective number of independent tests. Although we removed redundant lncRNAs, the remaining lncRNAs can still have overlapping regions across different genome annotations. Therefore, we adopted a principal component analysis (PCA) based approach, the simple*M* method to calculate the effective number of independent tests[42]. For each chromosome, suppose we had tested K lncRNA rare variant set ($lncRNA_1$, $lncRNA_2$, …, $lncRNA_K$) for N individuals

medRxiv preprint doi: https://doi.org/10.1101/2023.06.28.23291966; this version posted June 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

(1, 2, …, N), we first found the minor allele counts of rare variants (MAF < 1%) carried by each individual within each lncRNA rare variant set that were tested by STAAR and constructed a $N \times K$ matrix. We then derived the pairwise lncRNA correlation matrix R_{KxK} that reflected the correlation structure among the tests from the constructed $N \times K$ matrix. We calculated the eigenvalues, $\{\lambda_i : \lambda_1 \geq \lambda_1 \geq \cdots \geq \lambda_K\}$, from the pairwise lncRNA correlation matrix R_{KxK}. The effective number of tests (M_{eff}) for each chromosome was estimated as $M_{eff} = min(x)\ s.t. \frac{\sum_{i=1}^{x} \lambda_i}{\sum_{i=1}^{K} \lambda_i} > c$, where c was a pre-defined parameter which was set to 0.95. We added up the effective number of tests (M_{eff}) by each chromosome assuming independence between chromosomes. The Bonferroni correction formula was then used to calculate the adjusted significance level as 0.05/ M_{eff} as used for unconditional analysis.

LncRNA gene expression analysis

Framingham Heart Study (FHS) RNA-seq data. We utilized FHS RNA sequencing data to perform the association analyses of lncRNA expression levels with blood lipid traits. This study included 1505 participants from the FHS Third Generation cohort[41]. Blood samples for RNA seq were collected from Third Generation participants who attended the second examination cycle (2008–2011). Protocols for participant examinations and collection of genetic materials were approved by the Institutional Review Board at Boston Medical Center. All participants provided written, informed consent for genetic studies. All research was performed in accordance with relevant guidelines/regulations. The technical details for the blood draw and RNA sequencing can be found elsewhere[43]. For the association analyses (**Supplementary Text**), we first processed the RNASeq Data with following steps: 1. Sample QC by removing misidentified samples and sentinel control samples. 2. TMM normalization for the gene-level count data. 3.

medRxiv preprint doi: https://doi.org/10.1101/2023.06.28.23291966; this version posted June 29, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Filtering low expression transcripts. 4. Regressing the log2(TMM+1) on the technical covariates, and the resultant residuals were used to perform association analysis. We fitted a linear mixed effects model for the residuals of the TMM normalized log2 transformed counts data and the lipid phenotypes adjusting for predicted complete blood count (CBC), constructed surrogate variables (SVs), sex, age, and family structure as variance-covariance matrix.

Genome build

All genome coordinates are given in the NCBI GRCh38/UCSC hg38 version of the human genome.

Acknowledgements

Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). G.M.P. is supported by NIH grants R01HL142711 and R01HL127564. 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). X.Lin is supported by grants R35-CA197449, U19-CA203654, R01-HL113338, and U01-HG009088. We like to acknowledge all the grants that supported this study, R01 HL121007, U01 HL072515, R01 AG18728, X01HL134588, HL 046389, HL113338, and 1R35HL135818, K01 HL135405, R03 HL154284, U01HL072507, R01HL087263, R01HL090682, P01HL045522, R01MH078143, R01MH078111, R01MH083824, U01DK085524, R01HL113323, R01HL093093, R01HL140570, R01HL142711, R01HL127564, R01HL148050, R01HL148565, HL105756, and Leducq TNE-18CVD04. The views expressed

medRxiv preprint doi: https://doi.org/10.1101/2023.06.28.23291966; this version posted June 29, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

in this manuscript are those of the authors and do not necessarily represent the views of the National Heart, Lung, and Blood Institute; the National Institutes of Health; or the U.S. Department of Health and Human Services. We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed and UK Biobank. The full study specific acknowledgements and NHLBI TOPMed Fellowship acknowledgement are detailed in **Supplementary Text**.

Author contributions

Y.W., P.N., and G.M.P. designed the study. Y.W. carried out all the primary analysis with critical inputs from P.N. and G.M.P. M.S.S carried out the replication analysis. Y.W. and J.A.H carried out the secondary analysis. Y.W., M.S.S., X.Li, Z.L., A.K.D, J.C.B., J.B., E.B., D.W.B., B.E.C., J.C.C., A.P.C., Y.C., J.E.C., P.S.D., S.K.D., P.T.E., J.S.F., M.F., B.I.F., S. Gabriel, S.Germer, R.A.G., X.G., J.H., N.H., B.H., L.H., M.R.I., R.J., R.C.K., S.LR.K., T.N.K., R.K., C.K., B.G.K., D.Levy, C.Li, C.Liu, D.Lloyd-Jone, R.JF.L., M.C.M., L.W.M., R.A.M., R.L.M., B.D.M., M.E.M., A.C.M., J.M.M., T.N., J.R.O., N.D.P., M.H.P., B.M.P., L.M.E., D.C.R., S.R., A.P.R., S.S.R., M.R., W.H-H.S., J.A.S., A.S., H.K.T., M.Y.T., K.A.V., Z.W., L.R.Y., W.Z., J.I.R., X.Lin., P.N., and G.M.P. acquired, analyzed or interpreted data. G.M.P. and P.N. and NHLBI TOPMed Lipids Working Group provided administrative, technical, or material support. Y.W. and G.M.P. wrote the first draft of the manuscript and revised it according to suggestions by the coauthors. All authors critically reviewed the manuscript, suggested revisions as needed and approved the final version.

Declaration of interests

medRxiv preprint doi: https://doi.org/10.1101/2023.06.28.23291966; this version posted June 29, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

P.N. reports investigator-initiated grant support from Amgen, Apple, AstraZeneca, and Boston Scientific, personal fees from Apple, AstraZeneca, Blackstone Life Sciences, Foresite Labs, Genentech, TenSixteen Bio, and Novartis, scientific advisory board membership of geneXwell and TenSixteen Bio, and spousal employment at Vertex, all unrelated to the present work. B.M.P. serves on the Steering Committee of the Yale Open Data Access Project funded by Johnson & Johnson. L.M.R is a consultant for the TOPMed Administrative Coordinating Center (through Westat). M.E.M. receives funding from Regeneron Pharmaceutical Inc. unrelated to this work. X. Lin is a consultant of AbbVie Pharmaceuticals and Verily Life Sciences. The remaining authors declare no competing interests.

Data availability

Individual whole-genome sequence data for TOPMed and harmonized lipids at individual sample level are available through restricted access via the TOPMed dbGaP Exchange area. Summary level genotype data from TOPMed are available through the BRAVO browser (https://bravo.sph.umich.edu/). The UK Biobank (UKB) whole-genome sequence data can be accessed through UKB Research Analysis Platform (RAP), through the UKB approval system (https://www.ukbiobank.ac.uk). The dbGaP accessions for TOPMed cohorts are as follows: Old Order Amish (Amish) *phs000956 and phs00039;*Atherosclerosis Risk in Communities study (ARIC) *phs001211 and phs000280;* Mt Sinai BioMe Biobank (BioMe) *phs001644 and phs000925;* Coronary Artery Risk Development in Young Adults (CARDIA) phs001612 and phs000285; Cleveland Family Study (CFS) *phs000954 and phs000284;* Cardiovascular Health Study (CHS) *phs001368 and phs000287;* Diabetes Heart Study (DHS) *phs001412 and phs001012;* Framingham Heart Study (FHS) *phs000974 and phs000007;* Genetic Studies of Atherosclerosis Risk (GeneSTAR) *phs001218 and phs000375;* Genetic Epidemiology Network

medRxiv preprint doi: https://doi.org/10.1101/2023.06.28.23291966; this version posted June 29, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

of Arteriopathy (GENOA) *phs001345 and phs001238;* Genetic Epidemiology Network of Salt Sensitivity (GenSalt) *phs001217 and phs000784;* Genetics of Lipid-Lowering Drugs and Diet Network (GOLDN) *phs001359 and phs000741;* Hispanic Community Health Study - Study of Latinos (HCHS_SOL) *phs001395 and phs000810;* Hypertension Genetic Epidemiology Network and Genetic Epidemiology Network of Arteriopathy (HyperGEN) *phs001293 and phs001293;* Jackson Heart Study (JHS) *phs000964 and phs000286;* Multi-Ethnic Study of Atherosclerosis (MESA) *phs001416 and phs000209;* Massachusetts General Hospital Atrial Fibrillation Study (MGH_AF) *phs001062 and phs001001;* San Antonio Family Study (SAFS) *phs001215 and phs000462;* Samoan Adiposity Study (SAS) *phs000972 and phs000914;* Taiwan Study of Hypertension using Rare Variants (THRV) *phs001387 and phs001387;* Women's Health Initiative (WHI) *phs001237 and phs000200*.

Code availability

R code for implementing the analysis is available at the public GitHub Repository https://github.com/kyleyxw/lncRNA-paper. STAAR is implemented as an open-source R package available at https://github.com/xihaoli/STAAR. STAARpipeline is implemented as an open-source R package available at https://github.com/xihaoli/STAARpipeline.

medRxiv preprint doi: https://doi.org/10.1101/2023.06.28.23291966; this version posted June 29, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

641

642

643

644

medRxiv preprint doi: https://doi.org/10.1101/2023.06.28.23291966; this version posted June 29, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

References

1. Saxena, R. *et al.* Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science (1979)* **316**, 1331–1336 (2007).
2. Kathiresan, S. *et al.* A genome-wide association study for blood lipid phenotypes in the Framingham Heart Study. *BMC Med Genet* **8**, 1–10 (2007).
3. Kathiresan, S. *et al.* Polymorphisms Associated with Cholesterol and Risk of Cardiovascular Events. *New England Journal of Medicine* **358**, 1240–1249 (2008).
4. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature 2010 466:7307* **466**, 707–713 (2010).
5. Asselbergs, F. W. *et al.* Large-scale gene-centric meta-analysis across 32 studies identifies multiple lipid loci. *Am J Hum Genet* **91**, 823–838 (2012).
6. Albrechtsen, A. *et al.* Exome sequencing-driven discovery of coding polymorphisms associated with common metabolic phenotypes. *Diabetologia* **56**, 298–310 (2013).
7. Tachmazidou, I. *et al.* A rare functional cardioprotective APOC3 variant has risen in frequency in distinct population isolates. *Nature Communications 2013 4:1* **4**, 1–6 (2013).
8. Willer, C. J. *et al.* Discovery and refinement of loci associated with lipid levels. *Nature Genetics 2013 45:11* **45**, 1274–1283 (2013).
9. Holmen, O. L. *et al.* Systematic evaluation of coding variation identifies a candidate causal variant in TM6SF2 influencing total cholesterol and myocardial infarction risk. *Nature Genetics 2014 46:4* **46**, 345–351 (2014).
10. Peloso, G. M. *et al.* Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. *Am J Hum Genet* **94**, 223–232 (2014).
11. Surakka, I. *et al.* The impact of low-frequency and rare variants on lipid levels. *Nature Genetics 2015 47:6* **47**, 589–597 (2015).
12. Tang, C. S. *et al.* Exome-wide association analysis reveals novel coding sequence variants associated with lipid traits in Chinese. *Nature Communications 2015 6:1* **6**, 1–9 (2015).
13. Liu, D. J. *et al.* Exome-wide association study of plasma lipids in >300,000 individuals. *Nature Genetics 2017 49:12* **49**, 1758–1766 (2017).
14. Lu, X. *et al.* Exome chip meta-analysis identifies novel loci and East Asian–specific coding variants that contribute to lipid levels and coronary artery disease. *Nature Genetics 2017 49:12* **49**, 1722–1730 (2017).
15. Hoffmann, T. J. *et al.* A large electronic-health-record-based genome-wide study of serum lipids. *Nature Genetics 2018 50:3* **50**, 401–413 (2018).
16. Klarin, D. *et al.* Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. *Nature Genetics 2018 50:11* **50**, 1514–1523 (2018).
17. Spracklen, C. N. *et al.* Association analyses of East Asian individuals and trans-ancestry analyses with European individuals reveal new loci associated with cholesterol and triglyceride levels. *Hum Mol Genet* **27**, 1122–1122 (2018).
18. Graham, S. E. *et al.* The power of genetic diversity in genome-wide association studies of lipids. *Nature* **600**, 675–679 (2021).
19. Kanoni, S. *et al.* Implicating genes, pleiotropy, and sexual dimorphism at blood lipid loci through multi-ancestry meta-analysis. *Genome Biol* **23**, 268 (2022).

medRxiv preprint doi: https://doi.org/10.1101/2023.06.28.23291966; this version posted June 29, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

20. 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: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Circulation* **139**, E1082–E1143 (2019).
21. Musunuru, K. *et al.* Exome Sequencing, ANGPTL3 Mutations, and Familial Combined Hypolipidemia. *New England Journal of Medicine* **363**, 2220–2227 (2010).
22. Cohen, J. C., Boerwinkle, E., Mosley, T. H. Jr. & Hobbs, H. H. Sequence Variations in PCSK9, Low LDL, and Protection against Coronary Heart Disease. *https://doi.org/10.1056/NEJMoa054013* **354**, 1264–1272 (2006).
23. Kathiresan, S. A PCSK9 Missense Variant Associated with a Reduced Risk of Early-Onset Myocardial Infarction. *https://doi.org/10.1056/NEJMc0707445* **358**, 2299–2300 (2008).
24. Uszczynska-Ratajczak, B., Lagarde, J., Frankish, A., Guigó, R. & Johnson, R. Towards a complete map of the human long non-coding RNA transcriptome. *Nature Reviews Genetics* vol. 19 535–548 Preprint at https://doi.org/10.1038/s41576-018-0017-y (2018).
25. Van Solingen, C., Scacalossi, K. R. & Moore, K. J. Long noncoding RNAs in lipid metabolism. *Curr Opin Lipidol* **29**, 224 (2018).
26. Muret, K. *et al.* Long noncoding RNAs in lipid metabolism: literature review and conservation analysis across species. *BMC Genomics 2019 20:1* **20**, 1–18 (2019).
27. Statello, L., Guo, C. J., Chen, L. L. & Huarte, M. Gene regulation by long non-coding RNAs and its biological functions. *Nature Reviews Molecular Cell Biology 2020 22:2* **22**, 96–118 (2020).
28. Bocher, O. & Génin, E. Rare variant association testing in the non-coding genome. *Human Genetics* vol. 139 1345–1362 Preprint at https://doi.org/10.1007/s00439-020-02190-y (2020).
29. Harrow, J. *et al.* GENCODE: The reference human genome annotation for the ENCODE project. *Genome Res* **22**, 1760–1774 (2012).
30. Frankish, A. *et al.* GENCODE 2021. *Nucleic Acids Res* **49**, D916–D923 (2021).
31. Hon, C. C. *et al.* An atlas of human long non-coding RNAs with accurate 5′ ends. *Nature* **543**, 199–204 (2017).
32. Zhao, L. *et al.* NONCODEV6: An updated database dedicated to long non-coding RNA annotation in both animals and plants. *Nucleic Acids Res* **49**, D165–D171 (2021).
33. Seifuddin, F. *et al.* lncRNAKB, a knowledgebase of tissue-specific functional annotation and trait association of long noncoding RNA. *Sci Data* **7**, (2020).
34. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
35. Natarajan, P. *et al.* Deep-coverage whole genome sequences and blood lipids among 16,324 individuals. *Nat Commun* **9**, (2018).
36. Selvaraj, M. S. *et al.* Whole genome sequence analysis of blood lipid levels in >66,000 individuals. *Nature Communications 2022 13:1* **13**, 1–18 (2022).
37. Li, X. *et al.* Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nat Genet* **52**, 969–983 (2020).
38. Li, Z. *et al.* A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies. *Nat Methods* **19**, 1599–1611 (2022).

medRxiv preprint doi: https://doi.org/10.1101/2023.06.28.23291966; this version posted June 29, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

39. Zhou, H. *et al.* FAVOR: functional annotation of variants online resource and annotator for variation across the human genome. *Nucleic Acids Res* (2022) doi:10.1093/nar/gkac966.
40. Halldorsson, B. V. *et al.* The sequences of 150,119 genomes in the UK Biobank. *Nature* **607**, 732–740 (2022).
41. Splansky, G. L. *et al.* The Third Generation Cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination. *Am J Epidemiol* **165**, 1328–1335 (2007).
42. Gao, X., Starmer, J. & Martin, E. R. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genet Epidemiol* **32**, 361–369 (2008).
43. Liu, C. *et al.* Whole genome DNA and RNA sequencing of whole blood elucidates the genetic architecture of gene expression underlying a wide range of diseases. *Sci Rep* **12**, 20167 (2022).
44. Rossignoli, A. *et al.* Poliovirus Receptor-Related 2: A Cholesterol-Responsive Gene Affecting Atherosclerosis Development by Modulating Leukocyte Migration. *Arterioscler Thromb Vasc Biol* **37**, 534–542 (2017).
45. Ponting, C. P. & Haerty, W. Genome-Wide Analysis of Human Long Noncoding RNAs: A Provocative Review. (2022) doi:10.1146/annurev-genom-112921.
46. Tontonoz, P. *et al.* Long noncoding RNA facilitated gene therapy reduces atherosclerosis in a murine model of familial hypercholesterolemia. *Circulation* vol. 136 776–778 Preprint at https://doi.org/10.1161/CIRCULATIONAHA.117.029002 (2017).
47. Sallam, T. *et al.* Feedback modulation of cholesterol metabolism by the lipid-responsive non-coding RNA LeXis. *Nature* **534**, 124–128 (2016).
48. Soria, L. F. *et al.* Association between a specific apolipoprotein B mutation and familial defective apolipoprotein B-100. *Proc Natl Acad Sci U S A* **86**, 587–591 (1989).
49. 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 (2010).
50. Pollin, T. I. *et al.* A Null Mutation in Human APOC3 Confers a Favorable Plasma Lipid Profile and Apparent Cardioprotection * NIH Public Access. *Science (1979)* **322**, 1702–1705 (2008).
51. Casper, J. *et al.* The UCSC Genome Browser database: 2018 update. *Nucleic Acids Res* **46**, D762–D769 (2018).
52. Lumley, T., Brody, J., Peloso, G., Morrison, A. & Rice, K. FastSKAT: Sequence kernel association tests for very large sets of markers. *Genet Epidemiol* **42**, 516 (2018).
53. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310–315 (2014).
54. Huang, Y. F., Gulko, B. & Siepel, A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat Genet* **49**, 618–624 (2017).
55. Rogers, M. F. *et al.* FATHMM-XF: Accurate prediction of pathogenic point mutations via extended features. *Bioinformatics* **34**, 511–513 (2018).
56. Ramdas, S. *et al.* A multi-layer functional genomic analysis to understand noncoding genetic variation in lipids. *Am J Hum Genet* **109**, 1366–1387 (2022).