

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

3 Yuxuan Wang¹, Margaret Sunitha Selvaraj^{2,3,4}, Xihao Li⁵, Zilin Li^{5,6,7}, Jacob A. Holdcraft¹,
4 Donna K. Arnett^{8,9}, Joshua C. Bis¹⁰, John Blangero¹¹, Eric Boerwinkle¹², Donald W. Bowden¹³,
5 Brian E. Cade^{14,15}, Jenna C. Carlson^{16,17}, April P. Carson¹⁸, Yii-Der Ida Chen¹⁹, Joanne E.
6 Curran¹¹, Paul S. de Vries¹², Susan K. Dutcher²⁰, Patrick T. Ellinor^{21,22}, James S. Floyd^{10,23},
7 Myriam Fornage²⁴, Barry I. Freedman²⁵, Stacey Gabriel²⁶, Soren Germer²⁷, Richard A. Gibbs²⁸,
8 Xiuqing Guo¹⁹, Jiang He^{29,30}, Nancy Heard-Costa^{31,32}, Bertha Hildalgo³³, Lifang Hou³⁴,
9 Marguerite R. Irvin³³, Roby Joehanes³⁵, Robert C. Kaplan^{36,37}, Sharon LR. Kardia³⁸, Tanika N.
10 Kelly³⁹, Ryan Kim⁴⁰, Charles Kooperberg³⁷, Brian G. Kral⁴¹, Daniel Levy^{31,35}, Changwei Li^{30,29},
11 Chunyu Liu^{1,31}, Don Lloyd-Jone³⁴, Ruth JF. Loos^{42,43}, Michael C. Mahaney¹¹, Lisa W. Martin⁴⁴,
12 Rasika A. Mathias⁴¹, Ryan L. Minster⁴⁵, Braxton D. Mitchell⁴⁶, May E. Montasser⁴⁶, Alanna C.
13 Morrison¹², Joanne M. Murabito^{31,47}, Take Naseri⁴⁸, Jeffrey R. O'Connell⁴⁶, Nicholette D.
14 Palmer¹³, Michael H. Preuss⁴², Bruce M. Psaty^{10,23,49}, Laura M. Raffield⁵⁰, Dabeeru C. Rao⁵¹,
15 Susan Redline⁵², Alexander P. Reiner²³, Stephen S. Rich⁵³, Muagututi'a Sefuiva Ruepena⁵⁴,
16 Wayne H-H. Sheu⁵⁵, Jennifer A. Smith³⁸, Albert Smith⁵⁶, Hemant K. Tiwari⁵⁷, Michael Y. Tsai⁵⁸,
17 Karine A. Viaud-Martinez⁵⁹, Zhe Wang⁴², Lisa R. Yanek⁴¹, Wei Zhao³⁸, NHLBI Trans-Omics
18 for Precision Medicine (TOPMed) Consortium, Jerome I. Rotter¹⁹, Xihong Lin^{3,5,60}, Pradeep
19 Natarajan^{2,3,4}, Gina M. Peloso^{1,*}

20
21 ¹Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA,

22 ²Cardiovascular Research Center and Center for Genomic Medicine, Massachusetts General

23 Hospital, Boston, MA, USA, ³Program in Medical and Population Genetics, Broad Institute of

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

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

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

86

87

88

89

90

91

92

93 *Corresponding author:
94 Gina M. Peloso, PhD
95 Department of Biostatistics
96 Boston University School of Public Health
97 Boston, MA 02118
98 gpeloso@bu.edu
99 Abstract word counts: 240
100 Text word counts: 4787
101 No. of references: 56
102 No. of figures: 3
103 No. of tables: 1

104 **Abstract**

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

123 **Introduction**

124 Blood lipid levels, including low-density lipoprotein cholesterol (LDL-C), total cholesterol (TC),
125 triglyceride (TG), high-density lipoprotein cholesterol (HDL-C), are quantitative clinically
126 important traits with well-described monogenic and polygenic bases¹⁻¹⁹. Abnormal blood lipid
127 levels contribute to risk of coronary heart disease (CHD) and, in clinical practice, several
128 treatments, including statins, PCSK9 and ANGPTL3 inhibitors²⁰⁻²², are available to reduce the
129 risk of developing CHD. Each of these therapeutics has supporting evidence of their efficacy
130 from human genetic analysis of blood lipid levels²¹⁻²³.

131
132 Long non-coding RNAs (lncRNAs) are broadly defined as transcripts greater than 200
133 nucleotides in length that biochemically resemble mRNAs but do not code for proteins²⁴.
134 lncRNAs are known to perform important regulatory functions in lipid metabolism²⁵⁻²⁷. Rare
135 variants (RVs) in lncRNAs have not been systematically explored for their impact on blood lipid
136 levels as they are not comprehensively genotyped or imputed on non-WGS platforms. In
137 addition, there are difficulties in defining testing units and selecting qualifying variants²⁸.
138 Rapidly growing knowledge about the regulatory elements of the non-coding genome²⁹⁻³³, large-
139 scale WGS studies³⁴⁻³⁶, and new statistical methods³⁷⁻³⁹ for variant set tests provide the
140 possibility to assess the associations between plasma lipid traits and the genome-wide impact of
141 lncRNAs.

142
143 We examined the associations of rare variants in lncRNA genes from high-coverage WGS of
144 66,329 participants from diverse ancestry who have blood lipid traits (LDL-C, HDL-C, TC and
145 TG) in the National Heart, Lung, and Blood Institute (NHLBI) Trans-omics for Precision

146 Medicine (TOPMed) program freeze 8 data³⁴. We show that the rare noncoding variants in
147 lncRNA genes located near known Mendelian dyslipidemia genes contribute to phenotypic
148 variation in lipid levels among unselected individuals from population-based cohorts biobanks
149 independently of common variants associated with blood lipid levels.

150

151 **Results**

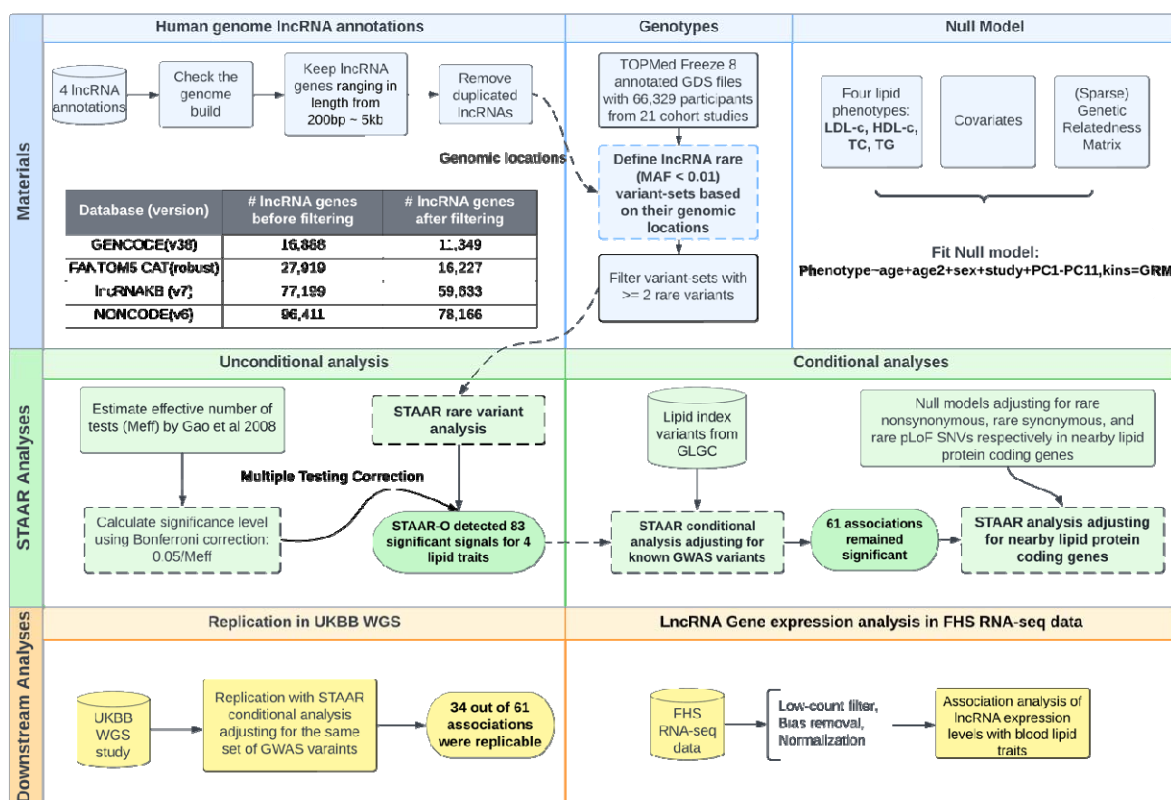
152 *Overview*

153 We performed a comprehensive evaluation of the association between quantitative blood lipid
154 traits and rare variants in lncRNA genes across the genome (**Figure 1**). We systematically
155 curated more than 165k lncRNA genes from the union of four human genome lncRNA
156 annotations, including GENCODE^{29,30}, FANTOM5 CAT³¹, NONCODE³² and lncRNAKB³³.
157 We utilized the TOPMed Freeze 8 dataset of 66,329 participants from 21 studies with WGS and
158 measured blood lipid levels and performed the rare variant (MAF <1%) association tests of
159 curated lncRNA genes with four blood lipid phenotypes: LDL-C, HDL-C, TC, and TG. We
160 further conducted the conditional analysis adjusting for known genome-wide association study
161 (GWAS) variants from the Global Lipids Genetics Consortium (GLGC)¹⁸. Associations between
162 lncRNA genes and lipids that were conditionally independent from the GWAS variants
163 (conditional P value < 6.0e-04) were then tested using STAAR procedure for conditional
164 analysis adjusting for rare nonsynonymous variants (MAF < 1%) within the closest protein
165 coding gene and the nearby known lipid monogenic genes in the region. We performed
166 replication in ~140K genomes from UK Biobank⁴⁰. We intersected our results with the gene
167 expression signatures of lipid traits in 1,505 participants from the Framingham Heart Study

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

170

171 **Figure 1. A schematic illustration of the study.**



172

173

174 *Characteristics of TOPMed participants*

175 We included 66,329 diverse participants from 21 cohort studies in the NHLBI TOPMed

176 consortium with blood lipid levels. The discovery cohorts consisted of 29,502 (44.5%) self-

177 reported White, 16,983 (25.6%) self-reported Black, 13,943 (21.0%) self-reported Hispanic,

178 4,719 (7.1%) self-reported Asian, and 1,182 (1.8%) self-reported Samoan participants

179 (Supplementary Table 1, Supplementary Text). Among the 66,329 participants, 41,182 (62%)

180 were female. The mean age of the 66,329 participants was 53 years (SD = 15). The mean ages at
181 lipid measurement varied across 21 cohorts from 25 years (SD = 3.56) for the Coronary Artery
182 Risk Development in Young Adults (CARDIA) to 73 years (SD = 5.38) for the Cardiovascular
183 Health Study (CHS). We observed that the Amish cohort had a higher concentration of LDL-C
184 (140 [SD = 43] mg/dL) and HDL-C (56 [SD = 16] mg/dL) as well as lower TG (median 63 [IQR
185 = 50] mg/dL) consistent with the known founder mutations in *APOB* and *APOC3*³⁵.

186

187 *Identification of rare lncRNA variants associated with blood lipid traits*

188 We defined lncRNA testing units using the available genomic positions in four genome
189 annotation projects described in the **Methods**. There were 11,349 lncRNA genes obtained from
190 GENCODE^{29,30}, 16,227 from FANTOM5 CAT³¹, 78,166 from NONCODE³² and 59,633 from
191 lncRNAKB³³. In total, we tested 165,375 lncRNA genes, among which, the average number of
192 rare variants in each lncRNA was 483 (SD = 572). The minimum and the maximum number of
193 rare variants among the lncRNAs being tested are 2 and 2947, respectively.

194

195 Our aggregation of lncRNAs across four lncRNA resources led to an overlap in the lncRNA
196 units, leading to non-independent tests of association of the lncRNAs with blood lipid levels. We
197 estimated the effective number of tests (M_{eff}) using a principal component analysis (PCA) based
198 approach⁴² since the traditional Bonferroni correction would be too conservative and reduce
199 power to detect association with blood lipid levels²⁸. M_{eff} was estimated as 111,550, providing a
200 significance threshold of $\alpha = 0.05/111,550 = 4.5 \times 10^{-7}$.

201

202

203 **Table 1. Summary of significant lncRNA associations for unconditional analysis,**
 204 **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

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

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

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

208

209 We applied STAAR (variant-Set Test for Association using Annotation infoRmation)

210 framework^{37,38} to identify the lncRNA rare variant (RV) sets that associated with quantitative

211 lipid traits (LDL-C, HDL-C, TC and TG) using TOPMed WGS data. STAAR-O identified 83

212 genome-wide significant associations (28 with LDL-C, 20 with TC, 19 with HDL-C, and 16 with

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

219

220 *Conditional analyses of trait-associated lncRNAs adjusting for known GWAS*

221 *variants and nonsynonymous variants within the nearby lipid monogenic genes*

222 After conditioning on known lipid-associated variants in a ± 500 kb window beyond a variant
223 set¹⁸, 61 out of 83 associations (73%) remained significant (20 with LDL-C, 14 with TC, 15 with
224 HDL-C, and 12 with TG) at the Bonferroni corrected level of $0.05/83 = 6.0 \times 10^{-4}$, indicating
225 that the associations between the lncRNA genes and lipid levels are distinct from the known
226 GWAS variants. The most significant association for LDL-C and TC was the lncRNA
227 NONHSAG026007.2 (chr19:44,892,420-44,903,056) near the *APOE-APOC1* region.

228 NONHSAG026007.2 remained significantly associated with LDL-C (P value = 2.44×10^{-15}) and
229 TC (P value = 2.17×10^{-27}) after adjusting for nearby known lipid-associated variants (**Figure 2**).

230 The most significant associations for HDL-C and TG were NONHSAG063125.1

231 (chr11:116,790,241-116,805,983) and NONHSAG09700.3 (chr11: 116,773,068-116,779,841),

232 respectively, both near *APOA5-APOC3-APOA1* region. NONHSAG063125.1 remained similarly
233 associated after conditioning on known lipid GWAS variants, while NONHSAG09700.3 became

234 even more significant (**Figure 2**). We then conditioned the GWAS-distinct associations on the

235 rare nonsynonymous variants within the closest protein coding gene and nearby lipid monogenic

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

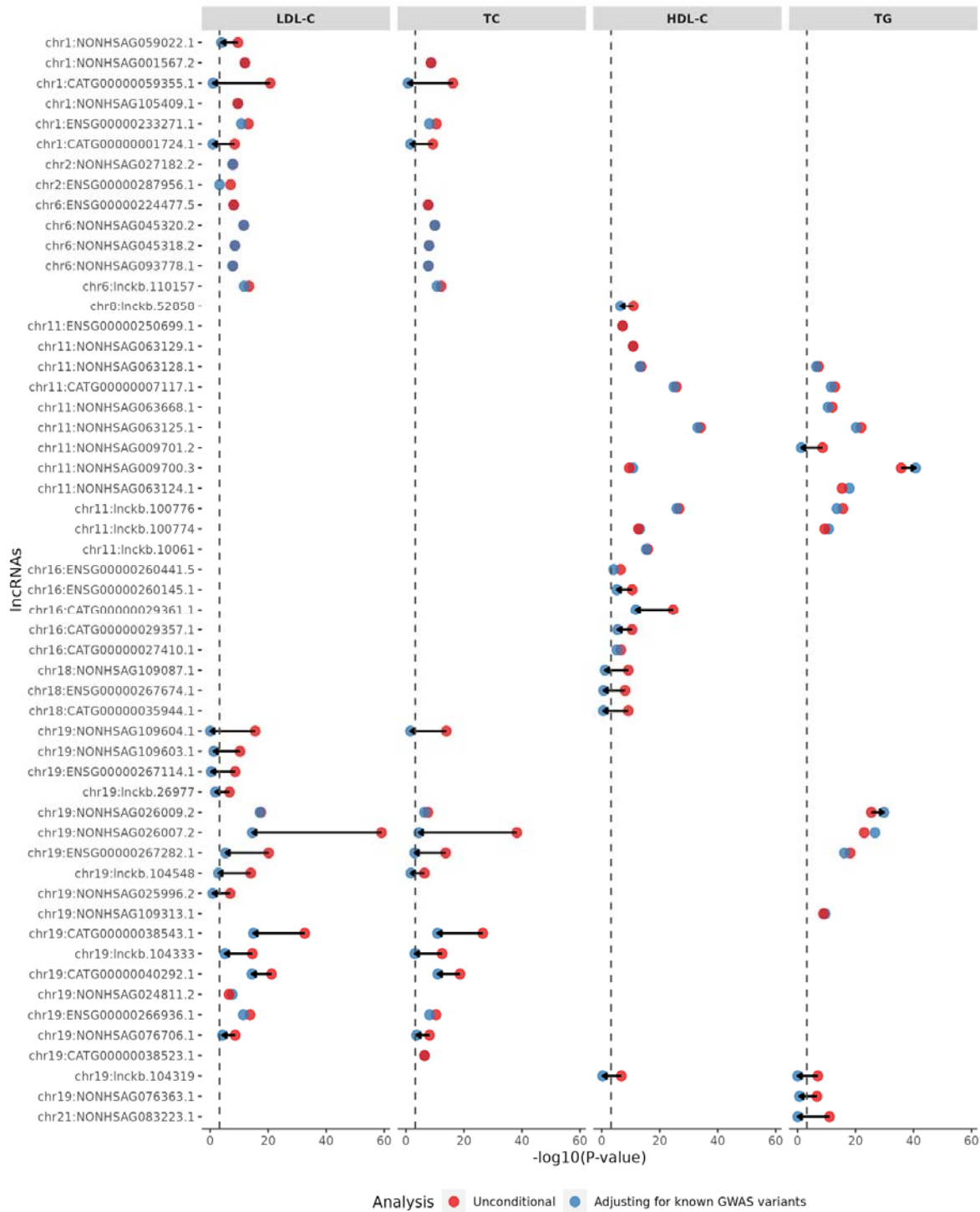
241

242 *Replication of significant lncRNA-blood lipid trait associations*

243 Replication of 61 lncRNAs associated with blood lipid levels was evaluated in 139,849 UK
244 Biobank individuals with WGS and blood lipid levels (**Supplementary Table 3**). We replicated
245 34 out of 61 (56%) lncRNA associations with blood lipid levels at a Bonferroni-corrected
246 threshold of $0.05/61 = 8.2e-04$ (**Supplementary Table 2**). The most significant associations in
247 the UK Biobank replication were NONHSAG025996.2 (chr19: 44,694,720-44,696,054) near
248 *APOE-APOC1* region for LDL-C, NONHSAG109604.1 near *APOE-APOC1* region for TC,
249 NONHSAG009700.3 near *APOA5-APOC3-APOA1* region for both HDL-C and TG
250 (**Supplementary Table 2**), which were consistent with the results from TOPMed.

251

252 **Figure 2. Significantly associated lncRNAs with four blood lipid traits (STAAR-O *P* value**
253 **< 4.5e-07).** The lncRNA genes are ordered by chromosome, followed by genomic positions.
254 Dots in red and blue represent the $-\log_{10}(\text{STAAR-O } P \text{ value})$ of the STAAR unconditional and
255 conditional analysis adjusting for known lipid-associated GWAS variants, respectively. The
256 black dashed line is the Bonferroni correction level of $0.05/83 = 6.0e-04$. Arrows indicate at least
257 10^4 fold change of STAAR-O *P* values comparing the unconditional analysis and conditional
258 analysis adjusting for known lipid-associated GWAS variants.



259

260

261

262 *lncRNA gene expression analysis in FHS RNA-seq data*

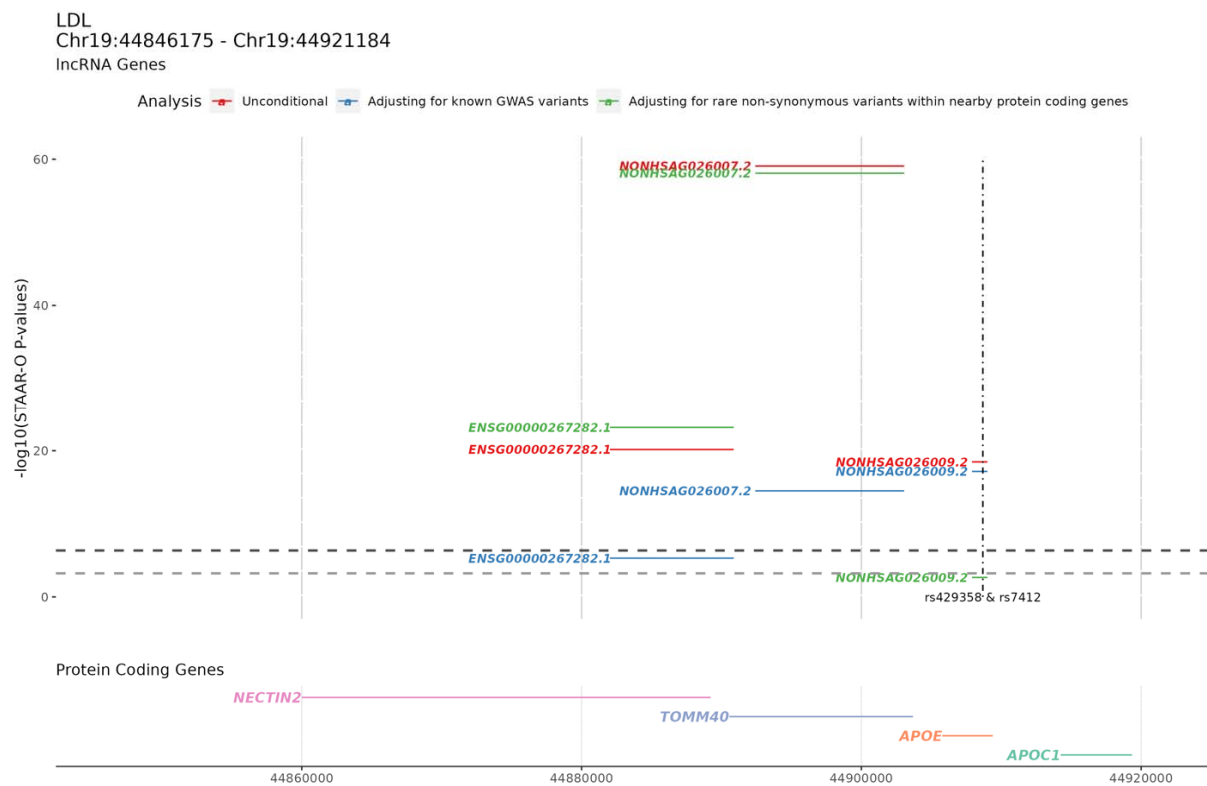
263 We overlapped the significant lipid-associated lncRNA genes with the lncRNA genes available
264 in the Framingham Heart Study (FHS) RNA-seq data generated by TOPMed⁴³. Since the gene-
265 level expression data in FHS is annotated by GENCODE v30, we limited the lncRNA genes to
266 those presented in GENCODE. Among the 54 unique lncRNA genes that are significantly
267 associated with either one of the lipid traits using TOPMed WGS data, 10 lncRNA genes are
268 annotated by GENCODE, and 8 out of 10 can be found in the FHS data. We performed
269 association analyses of expression levels of those 8 significant lipid-associated lncRNA genes
270 with blood lipid levels (LDL-C, TC, HDL-C, TG) (**Supplementary Text, Supplementary**
271 **Table 4**). In total, we tested 12 associations of lncRNA gene expression with blood lipid level
272 (**Supplementary Table 4**). The small proportion of overlapping was partially due to lncRNA
273 genes' generally lower expression. The lowly expressed genes were filtered out when processing
274 the gene expression data.

275
276 Four associations achieved Bonferroni-adjusted significance, including the gene expression level
277 of ENSG00000267282.1 (chr19:44,881,088-44,890,922) associated with LDL-C, TC, and TG,
278 and the gene expression level of ENSG00000266936.1 (chr19:11,010,917-11,016,011)
279 associated with TC. ENSG00000267282.1 is an antisense of *NECTIN2* (also known as *PVRL2*)
280 (**Figure 3**). The nectin cell adhesion molecule 2 (*NECTIN2*) protein is a cell adhesion molecule
281 involved in lipid metabolism⁴⁴. Additionally, ENSG00000267282.1 was one of the lncRNA
282 associations that we replicated in the independent UK Biobank (**Supplementary Table 2**). We
283 also queried whether the RVs in this lipid-associated lncRNA led to an alteration of the
284 corresponding lncRNA levels in the blood. However, due to the small number of overlapping

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

291

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



300

301 **Discussion**

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

312
313 Genetic variation for blood lipids levels has been observed across the allelic spectrum with
314 common, rare coding, and rare non-coding variants being associated with blood lipids levels³⁶.
315 Blood lipids have been associated with non-coding regulatory variants and coding variation in
316 genes, and now also associated with lncRNAs. We show that all the trait-associated lncRNAs are
317 in genomic regions previously associated with blood lipid traits, leading to the plausibility of
318 these results. About 75% of the associations are conditionally distinct from common regulatory
319 variation and rare protein coding variation at the same loci previously identified through GWAS
320 and whole exome sequencing studies. This indicates that the regulatory variants through
321 lncRNAs additionally contribute to the variation of blood lipid levels.

322

323

324 Despite numerous reports indicating the potential regulatory role of long non-coding RNAs
325 (lncRNAs), only a small proportion of them have substantial evidence to support such
326 claims^{25,26,45}. The fraction of lncRNAs that are functional remains unknown. Through a
327 comprehensive study of over 165,000 lncRNAs, we found that the majority of lncRNAs are not
328 associated with a lipid trait, which aligns with the argument made previously that only a few
329 human lncRNAs contribute centrally to human physiology⁴⁵. However, there are still some
330 lncRNAs that harbor variants that predispose individuals to phenotypic differences in blood lipid
331 levels. Our results suggest that investigators should first prioritize individual lncRNAs near the
332 known trait-associated loci for analysis, which is more likely to yield robust experimental
333 observations.

334
335 We further investigated one lncRNA, liver-expressed liver X receptor-induced sequence (*LeXis*),
336 which is a mediator of the complex effects of liver X receptor (LXR) signaling on hepatic lipid
337 metabolism to maintain hepatic sterol content and serum cholesterol levels^{46,47}. A potential
338 orthologue of *LeXis* in humans, TCONS_00016452 (chr9:104,990,086-104,991,780), is found in
339 a region adjacent to the human *ABCA1* gene. It didn't stand out as a significant signal for any
340 lipid trait in our study, which might suggest that it was not a functional orthologue of *LeXis*.
341 However, the rapid evolutionary turnover of lncRNAs still hinders the functional identification
342 between species^{45,47}.

343
344 Several limitations of our study should be noted. First, our RNA-seq analyses were restricted to
345 GENCODE annotation. The small proportion of overlapping RNA-seq data and WGS data limits
346 the ability to test rare lncRNA variants with their gene expression. Second, we did not correct for

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

351
352 In summary, our results from a large ancestrally diverse participants add further evidence that
353 lncRNA is an additional genomic element in known lipid gene regions that is distinct from the
354 known genes. We comprehensively evaluated 165K lncRNAs for their association with variation
355 in lipid traits and replicated most of the signals in an independent UKB WGS cohort.

356

357

358

359

360

361

362

363

364

365

366

367

368

369

370 **Methods**

371 *Discovery and replication cohorts*

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

392 **Replication cohorts.** We sought to replicate the findings using the UK Biobank WGS data for
393 139,849 genomes with blood lipid traits⁴⁰. The UK Biobank is a large, population-based
394 prospective cohort of half a million United Kingdom residents aged 40–69 years. The replication
395 cohorts consisted of 116,335 White, and 23,335 others (**Supplementary Table 3**).

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

399

400 *TOPMed WGS Freeze 8 data*

401 **Phenotype data.** We included four conventionally measured blood lipids in this study: low-
402 density lipoprotein cholesterol (LDL-C), total cholesterol (TC), triglyceride (TG), high-density
403 lipoprotein cholesterol (HDL-C). Detailed phenotype calculation and harmonization were
404 described elsewhere³⁶. Briefly, LDL-C was either directly measured or calculated by the
405 Friedewald equation when triglycerides were <400 mg/dL. We adjusted the total cholesterol by
406 dividing by 0.8 and LDL-C by dividing by 0.7 when statins were present^{10,35}. For triglycerides,
407 we additionally performed the natural log transformation for analysis, since triglycerides were
408 skewed. We then fitted a linear regression model for each phenotype to get the residuals after
409 adjusting for age, age², sex, race/ethnicity, study and the first 11 ancestral PCs (as recommended
410 by the TOPMed DCC). For Amish participants, we additionally adjusted for APOB
411 p.Arg3527Gln in LDL-C and TC, and adjusted for APOC3 p.Arg19Ter in HDL-C and TG^{48–50}.
412 The residuals were inverse rank normalized and rescaled by the standard deviation of the original
413 phenotype within each group³⁶.

414 **Genotype data.** Whole genome sequencing data were accessed from the TOPMed Freeze 8
415 release. DNA samples were sequenced at the >30× target coverage at seven centers (Broad
416 Institute of MIT and Harvard, Northwest Genomics Center, New York Genome Center, Illumina
417 Genomic Services, PSOMAGEN [formerly Macrogen], Baylor College of Medicine Human
418 Genome Sequencing Center, and McDonnell Genome Institute [MGI] at Washington
419 University)³⁴. The reads were aligned to human genome build GRCh38 using the BWA-MEM
420 algorithm. The genotype calling was performed using the TOPMed variant calling pipeline
421 (https://github.com/statgen/topmed_variant_calling). The resulting BCF files were converted to
422 SeqArray GDS format and annotated were annotated internally by curating data from multiple
423 database sources using Functional Annotation of Variant–Online Resource (FAVOR
424 (<http://favor.genohub.org>)^{37,39}). The resulting annotated GDS (aGDS) files were used in this
425 study. We computed the genetic relationship matrix (GRM) using R package *PC-relate* and
426 subtracted GRM of those samples with lipid phenotypes using R package *GENESIS*.

427

428 *Human reference genome annotations for long non-coding RNA genes*

429 Multiple lncRNA annotations are available. We obtained four long non-coding RNAs
430 (lncRNAs) annotation resources with different qualities and sizes and merged them to improve
431 comprehensiveness. They included GENCODE^{29,30}, FANTOM5 CAT³¹, NONCODE³² and
432 lncRNAKB³³.

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

437 manual, which ensures the quality and consistency of the data. We downloaded the GENCODE
438 v38 (December 2020) human release from
439 [https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_38/gencode.v38.long_non-](https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_38/gencode.v38.long_noncoding_RNAs.gtf.gz)
440 [oding_RNAs.gtf.gz](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
441 location information.

442 **FANTOM CAT.** The Functional Annotation of the Mammalian genome (FANTOM) CAGE-
443 associated transcriptome (CAT) meta-assembly combines both published sources and in-house
444 short-read assemblies. It utilized CAGE tags, which mark transcription start sites (TSSs), to
445 identify human lncRNA genes with high-confidence 5' ends. We acquired the FANTOM CAT
446 (lv3 robust) lncRNAs assembly from
447 [https://fantom.gsc.riken.jp/5/suppl/Hon_et_al_2016/data/assembly/lv3_robust/FANTOM_CAT.l](https://fantom.gsc.riken.jp/5/suppl/Hon_et_al_2016/data/assembly/lv3_robust/FANTOM_CAT.lv3_robust.only_lncRNA.gtf.gz)
448 [v3_robust.only_lncRNA.gtf.gz](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
449 (GRCh37), we lifted over to genome version hg38 (GRCh38) using the UCSC liftOver tool⁵¹.

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

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

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 complexity⁵². 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 consortium³⁴.

481 **STAAR unconditional analysis.** We applied the STAAR (variant-set test for association using
482 annotation information) framework to identify rare variants in the lncRNA variant sets that are
483 associated with four quantitative lipid traits (LDL-C, HDL-C, TG and TC). STAAR is a scalable
484 and powerful variant-set test that uses an omnibus multi-dimensional weighting scheme to
485 incorporate both qualitative functional categories and multiple in silico variant annotation scores
486 for genetic variants. STAAR accounts for population structure and relatedness and is scalable for
487 analyzing large WGS studies of continuous and dichotomous traits by fitting linear and logistic
488 mixed models^{37,38}. To perform the STAAR unconditional analysis, we first fitted a STAAR null
489 model using *fit_null_glmkin()* function to account for sample relatedness with phenotypic data,
490 covariates and (sparse) genetic relatedness matrix as input. For each of the four lipid phenotypes,
491 we adjusted for age, age2, sex, study and PC1-PC11. We calculated the *P* value for each lncRNA
492 rare variant set using STAAR-O, an omnibus test in the STAAR framework that combines *P*
493 values from multiple annotation-weighted burden tests, SKAT and ACAT-V using the ACAT
494 method. A total of 13 aggregated variant functional annotations were incorporated in STAAR-O,
495 including three integrative scores (CADD⁵³, LINSIGHT⁵⁴ and FATHMM-XF⁵⁵) and 10
496 annotation principal components (aPCs) (**Supplementary Table 5**)³⁸. All analyses were
497 performed using R packages *STAAR* (version 0.9.6) and *STAARpipeline* (version 0.9.6).

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

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

522 **Effective number of independent tests.** Although we removed redundant lncRNAs, the
523 remaining lncRNAs can still have overlapping regions across different genome annotations.
524 Therefore, we adopted a principal component analysis (PCA) based approach, the simpleM
525 method to calculate the effective number of independent tests⁴². For each chromosome, suppose
526 we had tested K lncRNA rare variant set ($lncRNA_1, lncRNA_2, \dots, lncRNA_K$) for N individuals

527 (1, 2, ..., N), we first found the minor allele counts of rare variants (MAF < 1%) carried by each
528 individual within each lncRNA rare variant set that were tested by STAAR and constructed a
529 $N \times K$ matrix. We then derived the pairwise lncRNA correlation matrix $R_{K \times K}$ that reflected the
530 correlation structure among the tests from the constructed $N \times K$ matrix. We calculated the
531 eigenvalues, $\{\lambda_i: \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_K\}$, from the pairwise lncRNA correlation matrix $R_{K \times K}$. The
532 effective number of tests (M_{eff}) for each chromosome was estimated as

533 $M_{eff} = \min(x) \text{ s.t. } \frac{\sum_{i=1}^x \lambda_i}{\sum_{i=1}^K \lambda_i} > c$, where c was a pre-defined parameter which was set to 0.95. We

534 added up the effective number of tests (M_{eff}) by each chromosome assuming independence
535 between chromosomes. The Bonferroni correction formula was then used to calculate the
536 adjusted significance level as $0.05/M_{eff}$ as used for unconditional analysis.

537

538 *LncRNA gene expression analysis*

539 **Framingham Heart Study (FHS) RNA-seq data.** We utilized FHS RNA sequencing data to
540 perform the association analyses of lncRNA expression levels with blood lipid traits. This study
541 included 1505 participants from the FHS Third Generation cohort⁴¹. Blood samples for RNA seq
542 were collected from Third Generation participants who attended the second examination cycle
543 (2008–2011). Protocols for participant examinations and collection of genetic materials were
544 approved by the Institutional Review Board at Boston Medical Center. All participants provided
545 written, informed consent for genetic studies. All research was performed in accordance with
546 relevant guidelines/regulations. The technical details for the blood draw and RNA sequencing
547 can be found elsewhere⁴³. For the association analyses (**Supplementary Text**), we first
548 processed the RNASeq Data with following steps: 1. Sample QC by removing misidentified
549 samples and sentinel control samples. 2. TMM normalization for the gene-level count data. 3.

550 Filtering low expression transcripts. 4. Regressing the $\log_2(\text{TMM}+1)$ on the technical covariates,
551 and the resultant residuals were used to perform association analysis. We fitted a linear mixed
552 effects model for the residuals of the TMM normalized \log_2 transformed counts data and the
553 lipid phenotypes adjusting for predicted complete blood count (CBC), constructed surrogate
554 variables (SVs), sex, age, and family structure as variance-covariance matrix.

555 **Genome build**

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

558

559 **Acknowledgements**

560 Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed)
561 program was supported by the National Heart, Lung and Blood Institute (NHLBI). G.M.P. is
562 supported by NIH grants R01HL142711 and R01HL127564. P.N. is supported by grants from
563 the National Heart, Lung, and Blood Institute (R01HL142711, R01HL148050, R01HL151283,
564 R01HL148565, R01HL135242, R01HL151152), Fondation Leducq (TNE-18CVD04), and
565 Massachusetts General Hospital (Paul and Phyllis Fireman Endowed Chair in Vascular
566 Medicine). X.Lin is supported by grants R35-CA197449, U19-CA203654, R01-HL113338, and
567 U01-HG009088. We like to acknowledge all the grants that supported this study, R01
568 HL121007, U01 HL072515, R01 AG18728, X01HL134588, HL 046389, HL113338, and
569 1R35HL135818, K01 HL135405, R03 HL154284, U01HL072507, R01HL087263,
570 R01HL090682, P01HL045522, R01MH078143, R01MH078111, R01MH083824,
571 U01DK085524, R01HL113323, R01HL093093, R01HL140570, R01HL142711, R01HL127564,
572 R01HL148050, R01HL148565, HL105756, and Leducq TNE-18CVD04. The views expressed

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

579

580 **Author contributions**

581 Y.W., P.N., and G.M.P. designed the study. Y.W. carried out all the primary analysis with
582 critical inputs from P.N. and G.M.P. M.S.S carried out the replication analysis. Y.W. and J.A.H
583 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.,
584 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,
585 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.,
586 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.,
587 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.,
588 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.,
589 J.I.R., X.Lin., P.N., and G.M.P. acquired, analyzed or interpreted data. G.M.P. and P.N. and
590 NHLBI TOPMed Lipids Working Group provided administrative, technical, or material support.
591 Y.W. and G.M.P. wrote the first draft of the manuscript and revised it according to suggestions
592 by the coauthors. All authors critically reviewed the manuscript, suggested revisions as needed
593 and approved the final version.

594

595 **Declaration of interests**

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

605 **Data availability**

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

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

630

631 **Code availability**

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

636

637

638

639

640

641

642

643

644

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* **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).

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. Uszczyńska-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).

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).