

# Exome wide association study for blood lipids in 1,158,017 individuals from diverse populations

Satoshi Koyama<sup>1,2,3,4,5</sup>, Zhi Yu<sup>1,2,3,4,5</sup>, Seung Hoan Choi<sup>2,3,6</sup>, Sean J. Jurgens<sup>3,5,7</sup>, Margaret Sunitha Selvaraj<sup>2,3,4,5</sup>, Derek Klarin<sup>8,9</sup>, Jennifer E. Huffman<sup>1</sup>, Shoa L. Clarke<sup>8,10</sup>, Michael N. Trinh<sup>2,3,4,5</sup>, Akshaya Ravi<sup>2,3,4,5</sup>, Jacqueline S. Dron<sup>2,3,4,5</sup>, Catherine Spinks<sup>2,3,4,5</sup>, Ida Surakka<sup>2,3,4,5,10</sup>, Aarushi Bhatnagar<sup>2,3,4,5</sup>, Kim Lannery<sup>2,3,4,5</sup>, Whitney Hornsby<sup>2,3,4,5</sup>, Scott M. Damrauer<sup>12,13</sup>, Kyong-Mi Chang<sup>12,13</sup>, Julie A Lynch<sup>14,15</sup>, Themistocles L. Assimes<sup>8,10</sup>, Philip S. Tsao<sup>8,10</sup>, Daniel J. Rader<sup>13</sup>, Kelly Cho<sup>1,16,17</sup>, Gina M. Peloso<sup>1,6</sup>, Patrick T. Ellinor<sup>2,3,4,5</sup>, Yan V. Sun<sup>18,19,20</sup>, Peter WF. Wilson<sup>18,20</sup>, Million Veteran Program, and Pradeep Natarajan<sup>1,2,3,4,5,17\*</sup>

## Affiliations:

1. VA Boston Healthcare System, Boston, MA
2. Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA
3. Cardiovascular Disease Initiative, Broad Institute of MIT and Harvard, Cambridge, MA
4. Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA
5. Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA
6. Department of Biostatistics, Boston University School of Public Health, Boston, MA
7. Department of Experimental Cardiology, Heart Center, Heart Failure and Arrhythmias, Amsterdam UMC location University of Amsterdam, Amsterdam, Netherlands
8. VA Palo Alto Healthcare System, Palo Alto, CA
9. Department of Surgery, Stanford University School of Medicine, Stanford, CA
10. Department of Medicine, Stanford University School of Medicine, Stanford, CA
11. Department of Internal Medicine, Division of Cardiology, University of Michigan, Ann Arbor, MI
12. Corporal Michael J. Crescenz VA Medical Center, Philadelphia, PA
13. University of Pennsylvania, Philadelphia, PA
14. VA Salt Lake City Health Care System, Salt Lake City, UT
15. College of Nursing and Health Sciences, University of Massachusetts, Boston, MA
16. Massachusetts General Brigham, Boston, MA
17. Department of Medicine, Harvard Medical School, Boston, MA
18. VA Atlanta Healthcare System, Decatur, GA
19. Department of Epidemiology and Global Health, Emory University Rollins School of Public Health, Atlanta, GA
20. Emory University School of Medicine, Atlanta, GA

\*Correspondence: PNATARAJAN@mgh.harvard.edu

## 33 **Abstract**

34           Rare coding alleles play crucial roles in the molecular diagnosis of genetic  
35 diseases. However, the systemic identification of these alleles has been challenging  
36 due to their scarcity in the general population. Here, we discovered and characterized  
37 rare coding alleles contributing to genetic dyslipidemia, a principal risk for coronary  
38 artery disease, among over a million individuals combining three large contemporary  
39 genetic datasets (the Million Veteran Program, n = 634,535, UK Biobank, n = 431,178,  
40 and the All of Us Research Program, n = 92,304) totaling 1,158,017 multi-ancestral  
41 individuals. Unlike previous rare variant studies in lipids, this study included 238,243  
42 individuals (20.6%) from non-European-like populations.

43           Testing 2,997,401 rare coding variants from diverse backgrounds, we identified  
44 800 exome-wide significant associations across 209 genes including 176 predicted loss  
45 of function and 624 missense variants. Among these exome-wide associations, 130  
46 associations were driven by non-European-like populations. Associated alleles are  
47 highly enriched in functional variant classes, showed significant additive and recessive  
48 associations, exhibited similar effects across populations, and resolved pathogenicity for  
49 variants enriched in African or South-Asian populations. Furthermore, we identified 5  
50 lipid-related genes associated with coronary artery disease (*RORC*, *CFAP65*, *GTF2E2*,  
51 *PLCB3*, and *ZNF117*). Among them, *RORC* is a potentially novel therapeutic target  
52 through the down regulation of LDLC by its silencing.

53           This study provides resources and insights for understanding causal  
54 mechanisms, quantifying the expressivity of rare coding alleles, and identifying novel  
55 drug targets across diverse populations.

56 Family-based discovery and characterization of rare coding alleles causative of  
57 familial hypercholesterolemia (FH) have yielded important insights for coronary artery  
58 disease (CAD), the leading cause of premature mortality among adults<sup>1,2</sup>. While FH is  
59 associated with a heightened risk for early-onset CAD, early intervention using lipid-  
60 lowering medications can considerably mitigate this risk, suppressing cumulative  
61 exposure to continuously high levels of low-density lipoprotein cholesterol (LDLC)<sup>3,4</sup>.  
62 However, FH remains substantially underdiagnosed and undertreated<sup>4-7</sup>. This highlights  
63 the need for increased efforts to identify and characterize pathogenic variants  
64 associated with FH.

65 Additionally, like other Mendelian conditions, population-based genetic analyses  
66 have often shown that expressivity (continuous effects on lipid levels) and penetrance  
67 (likelihood of CAD) may not be sufficiently high for some previously implicated  
68 pathogenic variants relative to initial descriptions in family-based studies<sup>8-13</sup>. As rare  
69 Mendelian alleles are increasingly returned to asymptomatic individuals through  
70 screening or secondary reporting<sup>14-16</sup>, allele-specific prognosis is increasingly important.

71 Furthermore, clinically curated variants are enriched among individuals  
72 genetically similar to European reference populations, reflecting biases in accumulated  
73 knowledge and data. In contrast, variants associated with non-European reference  
74 populations are more likely to be reclassified<sup>17</sup>, susceptible to population-related biased  
75 filters,<sup>18</sup> underdiagnosed due to limited data availability<sup>19</sup>.

76 To address these challenges, we assembled a large-scale, finely  
77 imputed/sequenced dataset encompassing lipid measures from over a million  
78 individuals combining the Million Veteran Program (MVP)<sup>20</sup>, UK Biobank (UKB)<sup>21</sup>, and  
79 the All of Us Research Program (AOU) cohorts, which included more than 230,000 that  
80 are genetically similar to non-European reference populations - historically  
81 underrepresented in genomic research. This diverse dataset allowed us to identify and  
82 characterize rare coding variant associations with blood lipids and validate the  
83 generalizability across populations. The summary of the estimated effects will provide a  
84 resource for further functional assessment and clinical utility.

## 85 **Study population**

86 We generated a large-scale clinical genetic dataset by imputing MVP (634,535  
87 individuals) to TOPMed imputation reference panel version r2<sup>22</sup>, which includes  
88 308,107,085 variants from 97,256 individuals representing diverse populations.  
89 Combined with whole exome sequence (WES) data in UKB (431,178 individuals)<sup>23</sup> and  
90 whole genome sequence (WGS) data in AOU (92,304 individuals)<sup>24</sup>, we generated a  
91 cohort of 1,158,017 individuals, including 238,243 (20.57%) from non-European  
92 populations (Fig. 1a, Supplementary Table 1). Large scale imputation reference panel  
93 including diverse populations allowed us to impute rare variants with high accuracy  
94 comparable to sequenced data (Extended Data Figures 1a-1c, Supplementary Notes I).

## 95 **Variant identification**

96 We curated the variants with minor allele count (MAC)  $\geq 5$  detected in  $\pm 50$  base  
97 pairs of exome target region used in UKB-WES (Methods). Annotation using 19,603  
98 protein coding transcripts identified 214,000 predicted Loss of Function (pLoF, stop gain,  
99 frameshift insertion/deletion, and canonical splice site), and 2,766,489 missense  
100 variants [missense single nucleotide variant (SNV), and in-frame insertion/deletion,  
101 Supplementary Table 2]. These variants covered 1.72% of all possible pLoF SNVs and  
102 3.72% of all possible missense SNVs (Extended Data Figure 1d, Supplementary Table  
103 3, Supplementary Notes II).

104 In addition, using the Splice-AI algorithm<sup>25</sup>, we detected 23,523 putatively cryptic  
105 splice variants [variants associated with donor/acceptor-gain, donor/acceptor-loss in  
106 distant position from canonical splice site with Delta Score  $> 0.8$ . Supplementary Notes  
107 III]. We re-classified these cryptic splice variants as pLoF and included them in  
108 association analyses. In total we identified 237,523 pLoF variants and 2,759,878,  
109 missense variants in this study (Supplementary Table 4). The MVP and AOU study  
110 populations, including diverse populations, effectively increases the variety of variants  
111 included in this study (Extended Data Figures 1b and 1c).

112 We identified at least one testable pLoF in 89.20% (17,486/19,603) of assessed  
113 transcripts and a missense variant in 95.30% (18,682/19,603). Among them, 71.29%  
114 (12,465/17,486) and 99.37% (18,666/18,682) of transcript had at least one pLoF or

115 missense variant with 80% statistical power to detect effect size of one standard  
116 deviation (SD) of phenotypes per allele (Extended Data Figures 1e and 1f,  
117 Supplementary Table 5, Supplementary Notes IV).

## 118 **Association analysis**

119 We tested linear associations of the imputed/sequenced genotypes of rare ( $5 \leq$   
120  $MAC$  and  $MAF_{POP_{MAX}} < 1\%$ ) pLoF variants or missense variants with blood lipids [total  
121 cholesterol (TC), LDLC, high-density lipoprotein cholesterol (HDL) triglycerides (TG)]  
122 using an additive model stratified by population groups (4 from MVP, 5 from UKB, and 5  
123 from AOU, Fig. 1a, Supplementary Table 1, Methods) followed by fixed effects meta-  
124 analysis including 14 population groups. In total, we tested 11,226,703 variant-  
125 phenotype combinations in the additive model. The highest Lambda GC in four tested  
126 traits was 1.025 for HDLC indicating suitable calibration (Extended Data Figure 2a). In  
127 addition, we conducted recessive model analysis for 233,971 variant-phenotype  
128 combinations with  $5 \leq$  minor homozygote counts and minor homozygote frequency  $<$   
129 1%. Exome-wide significance (EWS) was defined as  $P < 4.4 \times 10^{-9}$  [ $0.05/(11,226,703 +$   
130  $233,971)$ ].

131 We identified 800 additive EWS associations in 184 loci (202 associations in 45  
132 loci for TC; 235 in 48 for LDLC; 222 in 47 for HDLC; and 141 in 44 for TG, Extended  
133 Data Figure 2b, Supplementary Table 6), and 110 recessive EWS associations  
134 (Supplementary Table 7). The additive signals included 176 pLoF associations across  
135 40 genes and 623 missense associations across 193 genes (Figs. 1b and 1c) often with  
136 multiple associations per gene (Fig. 1d).

137 We observed significant enrichment of EWS variants in pLoF or missense  
138 variants compared to synonymous/non-coding variants [odds ratio (OR) $_{EWS/Non-EWS} =$   
139 6.33, 95% Confidence Interval (CI) = 5.02 – 7.91,  $P = 6.0 \times 10^{-41}$  for pLoF variants, and  
140 2.27 (1.98 – 2.60),  $P = 8.9 \times 10^{-32}$  for missense variants]. One of the strongest signals  
141 was the *APOB* pLoF variant (p.M3438X), which altered LDLC by  $-3.14$  SD per allele (or  
142  $-103.53$  mg/dL per allele, mean LDLC was 57.9mg/dL for 5 carriers, and 145mg/dL for  
143 409,041 non-carriers, Extended Data Figure 2c).

144 To assess the replicability of the results, we compared the effect sizes with a  
145 previous independent microarray-based rare-variant study for blood lipids (Lu et al., *Nat*  
146 *Genet* 2017, N = 358,251)<sup>26</sup>. In the replication dataset, we identified 48.4% (387/800) of  
147 EWS associations. 99.7% (386/387) of these variants showed directional concordance  
148 and 41.9% (162/387) showed significant association in the replication dataset ( $P <$   
149  $0.05/387$ ). For the eleven variants found in ten novel loci identified in this study, we  
150 found 81.8% (9/11) associations in the replication dataset. All 9 of these associations  
151 showed concordant effect direction and 5 of these showed nominal association ( $P <$   
152  $0.05$ , Extended Data Figure 2d) in the replication dataset.

### 153 Variant function predicts phenotype expressivity

154 To gain further insights into genetic associations and variant functions, we  
155 employed existing *in silico* methods for predicting variant functionality. For pLoF variants,  
156 we utilized the LOFTEE<sup>27</sup> plugin in VEP<sup>28</sup> and identified 163,643 ‘high-confidence’ pLoF  
157 variants (87.7% of pLoF variants). For missense variants, we applied 29 *in silico*  
158 deleterious prediction algorithms<sup>29</sup>, from which we derived an ensembled Missense  
159 Score (MiS, Methods) and grouped them into bins ([0, 0.5], (0.5, 0.7], (0.7, 0.9], and (0.9,  
160 1], where deleteriousness increases with increasing value. Supplementary Tables 8 and  
161 9). We observed strong linear relationships across variant deleteriousness, lower allele  
162 frequencies, and phenotype association (Fig. 2a, Supplementary Table 10). Notably,  
163 high-confidence pLoF, deleterious missense variants with a MiS (0.9, 1.0], and (0.7, 0.9]  
164 exhibited similarly constrained low MAF (median MAF 0.0023%, 0.0021%, and 0.0024%,  
165 respectively) and were more likely to be EWS [ $OR_{EWS/Non-EWS} = 7.24$  (95% CI 5.66 –  
166 91.8) and  $P = 5.2 \times 10^{-40}$  for pLoF;  $OR = 11.61$  (7.02 – 18.15) and  $P = 6.2 \times 10^{-15}$  for  
167 MiS (0.9, 1.0];  $OR = 5.02$  (3.87 – 6.43) and  $P = 1.9 \times 10^{-26}$  for MiS (0.7, 0.9]].  
168 Furthermore, the cryptic splice variants exhibited a similar level of constraint (median  
169 MAF 0.0028%) and were equally enriched for EWS [ $OR_{EWS/Non-EWS} = 5.96$  (95% CI 2.71  
170 – 11.42),  $P = 3.1 \times 10^{-5}$ ]. As an example, among the 53 pLoF variants observed in  
171 *APOB*, 19 were EWS, but these associated variants were depleted in the last exon (Fig.  
172 2b) and predicted as “low-confident” pLoF.

### 173 Distinguishing hypomorphic and hypermorphic missense variants



174 Multiple EWS pLoF associations allowed us to assess the effect directions of  
175 gene silencing in 23 gene-phenotype pairs (Fig. 1d). These included 128 pLoF variant-  
176 phenotype pairs, and all exhibited consistent effect directions except for a cryptic splice  
177 variant in *CETP* and HDLC. 87% (239/275) of missense variants showed concordant  
178 effect directions with pLoF variants in the same genes (hypomorphic variants). However,  
179 36 associations in 10 genes were found to have opposite effect direction to pLoF  
180 variants (hypermorphic variants, Supplementary Table 6). Some previously discovered  
181 hypermorphic variants included *PCSK9* [p.R469W<sup>30</sup>, p.R496W<sup>31</sup>] and *APOB*  
182 [p.R3527Q] but most are newly discovered hypermorphic variants. One such example is  
183 *LDLR* p.S849L which showed strong negative association with LDLC [ $\beta = -1.07$  (SE  
184 0.087),  $P = 3.6 \times 10^{-34}$ ] indicating gain-of-function. Another example is *APOB* p.G4395S  
185 which is of higher MAF in the African-like population [ $\beta_{MVP-AFR} = 0.433$  (SE 0.080),  $\beta_{UKB-}$   
186  $_{AFR} = 0.775$  (0.253)]. While MiS was an important factor in predicting hypomorphic  
187 associations, it did not predict hypermorphic associations (Fig. 2c).

### 188 **Cryptic splicing variants as novel candidates for loss-of-function**

189 For all identified variants in this study, we predicted the variant's potential for  
190 splice site disruption/creation using SpliceAI<sup>25</sup> and derived a Delta Score (DS) – a  
191 numeric score ranging 0 – 1 (Supplementary Notes III). The score distribution was  
192 sparse and only 0.598% (58,402/9,399,797) of variants had high DS (> 0.8). 43.5% of  
193 variants with high DS were not located in the canonical splice sites (Extended Data  
194 Figure 3a). We observed a strong enrichment of cryptic splice variants disrupting the  
195 donor structure (Donor Loss) in the splice donor 5<sup>th</sup> base (Extended Data Figure 3b),  
196 which are not typically considered as pLoF in the current practice. One representative  
197 example was rs200831171 – a splice donor 5<sup>th</sup> base variant of *APOA5* and associated  
198 with higher TG concentrations. This intronic variant has high donor loss potential  
199 ( $DS_{Donor Loss} = 0.97$ ) and was associated with increased TG levels with the largest effect  
200 size [ $\beta = 1.10$  (SE 0.079),  $P = 7.0 \times 10^{-44}$ ] among 6 EWS coding variants in *APOA5*  
201 associated with TG (Extended Data Figure 3c). Including this variant, we identified 15  
202 cryptic splice variants with EWS (Supplementary Table 6). Overall, cryptic splicing  
203 variants showed equivalent effect sizes with pLoF variants [median  $\beta_{Cryptic Splice} = 1.092$   
204 (IQR 0.601 – 1.118), normalized to pLoF as 1,  $P = 0.71$  by Wilcoxon Rank Sum test,

205 Extended Data Figure 3d], and larger effect sizes than missense variants [median  
206  $\beta_{\text{Missense}} = 0.408$  (0.136 – 0.701),  $P = 7.0 \times 10^{-4}$ ].

### 207 **Novel rare variant association outside of established lipid loci**

208 We identified associations for several variants residing outside established lipid  
209 loci (Extended Data Figures 4, Supplementary Table 6). One example is a rare  
210 missense variant GYS2 p.Y636H (MAF = 0.0431%), which showed significant  
211 associations with decreased TC, LDLC, and HDLC [ $\beta_{\text{TC}} = -0.24$ ,  $P_{\text{TC}} = 2.3 \times 10^{-15}$ ;  
212  $\beta_{\text{LDLC}} = -0.19$ ,  $P_{\text{LDLC}} = 4.0 \times 10^{-10}$ ; and  $\beta_{\text{HDL}} = -0.22$ ,  $P_{\text{HDL}} = 7.1 \times 10^{-14}$ ]. GYS2  
213 encodes glycogen synthetase 2, is expressed in the liver<sup>32</sup>, and is a causal gene for  
214 glycogen storage diseases<sup>33</sup>. Another example is the STS gene on chromosome X. A  
215 rare missense variant (p.H439R) in this gene was associated with decreased HDLC.  
216 STS encodes steroid sulfatase which is directly involved in steroid metabolism<sup>34</sup>. Other  
217 novel loci identified by this study include *SH3TC1* (TC), *ETV6* (TC), *PCSK6* (TC),  
218 *PCSK9* (HDL), *POR* (HDL), and *PTPRB* (TG).

### 219 **New insights into causal genes within established lipid loci**

220 Lead variants in genome-wide association studies (GWAS) are typically common  
221 and non-coding, and the causal gene is therefore unclear. Rare variant association  
222 study more directly interrogates gene product perturbation providing greater confidence  
223 in causal gene inference. One such example is 1q21.1, an established HDL GWAS  
224 locus comprising 21 genes (Extended Data Figure 5a). A rare pLoF variant in only  
225 *PDZK1* at 1q21.1 was associated with increased HDL levels (MAF 0.016%,  $\beta = 0.30$ ,  
226  $P = 1.1 \times 10^{-10}$ ), strongly implicating *PDZK1* as the causal gene at this locus. The gene  
227 product of *PDZK1* is known to interact with the known HDL-related gene *SCARB1*<sup>35</sup>.  
228 Another example is *SREBF1*, which is a master regulator for lipogenesis<sup>36</sup>.  
229 rs114001633 is a rare missense variant in *SREBF1* and associated with higher TC  
230 levels (MAF = 0.74%,  $\beta = 0.0442$ ,  $P = 2.8 \times 10^{-9}$ ) and 372kb from the index GWAS non-  
231 coding variant (Extended Data Figure 5b). Other examples included a missense  
232 association on the androgen receptor (*AR*) p.Q799E in chromosome X with HDL  
233 (Extended Data Figure 5c) and *CREB3L1* in chromosome 11 with TG (Extended Data  
234 Figure 5d).



235 While 58% (87/150) of these putatively effector genes harboring coding variants  
236 with EWS were the nearest genes of GWAS lead variants in the loci, the rest (42.0%)  
237 were not (Supplementary Table 11). By systemic conditioning analysis and introducing  
238 rare-coding alleles as covariates, we confirmed independence of rare-coding  
239 associations and common genetic associations (Extended Data Figures 6a and 6b,  
240 Supplementary Notes V). Reflecting functional relevance, we observed stronger  
241 enrichment of genes harboring rare coding variants with EWS than the nearest genes to  
242 the common variant GWAS signals (Extended Data Figures 7a, 7b, and 7c,  
243 Supplementary Table 12, Supplementary Notes VI).

#### 244 **Population enriched coding associations and shared effect sizes**

245 Inclusion of the diverse populations enabled testing for associations with  
246 ancestry enriched alleles. By intra-population meta-analysis, we identified 655 signals in  
247 European-like (EUR) populations, 124 in African-like (AFR) populations, and 45 in  
248 Admix-American-like (AMR) populations (Fig. 3a). Most of these signals are population-  
249 specific (631/655 associations were specific for EUR, 105/124 for AFR, and 18/45 for  
250 AMR), and overall we identified 130 lipid associated alleles that were only significant in  
251 non-EUR populations. These alleles are exclusively or dominantly found in AFR/AMR  
252 populations (Fig. 3b).

253 While we observed significant differences in variant frequencies, we found highly  
254 similar effect sizes between genetically dissimilar groups ( $R^2 \sim 0.9$ , Fig. 3c) for EWS  
255 variants. One example is a stop gain variant in *PCSK9* (Fig. 3d), which is dominant in  
256 AFR ( $MAF_{AOU-AFR} = 0.951\%$ ,  $MAF_{AOU-AMR} < 0.211\%$ ,  $MAF_{MVP-AFR} = 0.828\%$ ,  $MAF_{MVP-EUR}$   
257  $= 0.009\%$ ,  $MAF_{MVP-HIS} = 0.036\%$ ,  $MAF_{UKB-AFR} = 0.978$ ), but included consistently large  
258 positive effects on HDLC levels [median  $\beta = -1.036$  (range  $-1.140 - -0.874$ )] across  
259 populations.

260 As demonstrated with the polygenic risk score, estimates from large population  
261 studies are expected to be valuable resources for assessing individual risk. To explore  
262 the feasibility of a rare variant-based risk score, we estimated the carrier frequencies of  
263 these alleles in the study populations. The prevalence of lipid-related variants was  
264 67.5% in MVP and 74.0% in UKB overall, but there were significant differences among

265 genetically similar groups, with the highest in EUR (MVP 75.7%, UKB 75.2%) and the  
266 lowest in ASN/EAS (MVP 15.4%, UKB 5.5%), likely due to differences in the size of the  
267 discovery analysis (Extended Data Figure 8, Supplementary Table 13).

## 268 **Contribution of rare coding variants in trait variance**

269 Recent studies suggest additional contribution of the rare variants to the trait  
270 variance is not explained by common variants. Using LD-independent rare coding  
271 variants with EWS association, we estimated phenotype variance explained (PVE) for  
272 each trait. Collectively, rare coding variations contributed to additional 2.03 – 3.75 % of  
273 PVE in blood lipids corresponding 15.8 – 22.1% of PVE by common variants (Extended  
274 Data Figure 9a, Supplementary Table 14). While per-variant PVEs are slightly lower in  
275 rare variants [median (IQR) 0.00645% (0.00392% – 0.0119%)] than common variants  
276 [0.00667% (0.00446% – 0.0129%),  $P = 0.002$  by Wilcoxon Rank Sum test], sum of the  
277 PVE of rare variants showed substantially larger per-variant PVE. The largest PVE by  
278 rare coding variants in a single gene was observed in *PCSK9* for LDLC and TC (1.17%  
279 and 0.86%, respectively), followed by *APOB* for LDLC and TC (0.97% and 0.65%),  
280 *APOC3* for TG and HDLC (0.942% and 0.422%), *LDLR* for LDLC (0.31%). Notably, in  
281 20.6% (39/189) of lead variant - gene pairs, the sum of PVE by rare coding variants  
282 exceeded PVE by GWAS leading variant (Extended Data Figures 9c and 9d).

## 283 **Insights from recessive modeling**

284 We identified 110 variant-trait pairs with significant associations in the recessive  
285 model ( $P < 4.4 \times 10^{-9}$ , Fig. 4, Supplementary Table 7). Among these associations, we  
286 observed several examples of recessive effect sizes substantially larger than expected  
287 from a purely additive model. One example is *ANGPTL4* p.E40K on TG with larger  
288 effect sizes in the recessive model ( $\beta_{\text{Recessive}} = -0.845$ ) compared to the additive  
289 expectation ( $2 \times \beta_{\text{Additive}} = -0.544$ ). Another example is *TM6SF2* p.L156P which showed  
290 > 3 times higher effect size on LDLC in the recessive model [ $\beta_{\text{Recessive}} = -0.942$ ,  $P = 1.1$   
291  $\times 10^{-32}$ ] compared to the additive expectation ( $2 \times \beta_{\text{Additive}} = -0.307$ ). Heterozygosity for  
292 this variant has been linked to hepatic triglyceride accumulation and impaired VLDL (a  
293 hepatic- precursor of LDL) intracellular trafficking<sup>37</sup>. Another example was observed in  
294 *HBB* p. E7V (rs3334) – the causal variant for sickle cell anemia<sup>38,39</sup> and TC or LDLC.

295 While the additive associations were weak for these traits ( $P_{TC} = 0.0005$  and  $P_{LDLC} =$   
296  $0.017$ ), the recessive associations showed the largest effect sizes ( $\beta_{TC-Recessive} = -1.26$ ,  
297  $P_{TC-Recessive} = 2.9 \times 10^{-19}$  and  $\beta_{LDLC-Recessive} = -1.12$ ,  $P_{LDLC-Recessive} = 8.3 \times 10^{-12}$ ) among  
298 recessive associations. Strong recessive associations were also observed in *ABHD15*  
299 pLoF and lower TG ( $\beta_{TG-Recessive} = -0.586$ ,  $P_{TG-Recessive} = 5.8 \times 10^{-11}$ ). In the heterozygote  
300 state, the association was not observed ( $P = 0.12$ ). *ABHD15* is known to interact with  
301 *PDE3B* and associated with insulin signaling<sup>40</sup>. While recessive inheritance has been  
302 emphasized in the context of FH, we did not detect strong recessive associations in the  
303 previously suggested recessive genes.

### 304 Pathogenicity reassessment of FH variants

305 Curated pathogenic variants play a crucial role in the molecular diagnosis of  
306 familial hypercholesterolemia (FH). To contribute to this essential resource, we re-  
307 evaluated curated variants using our population-scale genomic dataset. By intersecting  
308 6,520 FH-related variants reported in ClinVar database<sup>41</sup> with 1,601 tested variants in  
309 this study, we identified 86 pathogenic/likely pathogenic (P/LP) variants, 268  
310 benign/likely benign (B/LB) variants, and 704 variants of uncertain significance (VUS) in  
311 *PCSK9/APOB/LDLR*. The B/LB variants showed a higher allele frequency compared to  
312 other classes (Fig. 5a). More than half of the P/LP variants (45 out of 83) are associated  
313 with higher LDLC levels ( $P < 0.05/1,601$ , Fig. 5b) with median  $\beta = 1.58 \text{ SD}_{LDLC}$  per allele  
314 (range 0.51 - 2.61). Importantly, despite fixed clinical categories of pathogenicity,  
315 expressivity varied and was overlapping (Fig. 5c).

316 We identified eight variants across the B/LB/VUS categories with equivalent  
317 effect sizes [median  $\beta = 1.66$  (range 1.43 - 1.86), Supplementary Table 15] to P/LP  
318 variants, including two missense variants in *PCSK9* (p.E40K, p.E197K), one in *APOB*  
319 (p.K3524T), and four in *LDLR* (p.H327Y, p.R440G, p.L456P, p.A705P). Among these,  
320 *LDLR* p.H327Y is enriched in SAS [MAF = 0.048%,  $\beta = 1.75$  (SE 0.32),  $P = 5.4 \times 10^{-8}$ ]  
321 but the pathogenicity of this variant was inconclusive in ClinVar. However, its highly  
322 significant association with a large effect size on LDLC apart from the median effect size  
323 of established P/LP variants supports a pathogenic role of this variant in FH. Another  
324 variant, *LDLR* p.L456P, was enriched in AFR [MAF = 0.0066%,  $\beta = 1.66$  (SE 0.25),  $P =$

325  $6.4 \times 10^{-11}$ ]. We also identified a previously considered pathogenic variant with a  
326 negative effect size. A missense variant in *APOB* (p.R490W) showed a strong negative  
327 association [ $\beta = -2.78$  (SE 0.32)] with LDLC levels. This variant is predicted to be a  
328 cryptic splice variant with a high DS ( $DS_{\text{Donor Gain}} = 0.98$ ), suggesting it introduces a loss-  
329 of-function change in the *APOB* gene and decrease blood LDLC level.

### 330 **Clinical outcomes of lipid associated alleles**

331 To connect the lipid related alleles and clinical outcomes, we tested for 800 lipid  
332 associated alleles identified in this study with prevalent/incident CAD. We used logistic  
333 regression framework to test for significant associations between the lipids associated  
334 variants and the occurrence of CAD (Methods). We observed positive associations of  
335 TC, LDLC, TG with CAD risk (Fig. 6, Supplementary Table 16) including several strong  
336 associations for known FH pathogenic variants in LDLC (p.C197Y, p.C184Y, Splicing  
337 variant). On the other hand, HDLC levels were not uniformly associated with CAD risk,  
338 as exemplified by the known association between higher HDLC level and increase CAD  
339 risk by *SCARB1*. Several established lipid related genes (*PCSK9*, *APOB*, *NPC1L1*,  
340 *ANGPTL3/4*, *APOC3*, and *LDLR*) were associated with lower LDLC/TG and decreased  
341 CAD risk with nominal significance ( $P_{\text{CAD}} < 0.05$ ). Overall, we identified five genes  
342 significantly associated with CAD ( $FDR_{\text{CAD}} < 0.05$ ) including *RORC*, *CFAP65*, *GTF2E2*,  
343 *PLCB3*, and *ZNF117*. Among these genes, *RORC* has high potential as a new  
344 therapeutic target to prevent CAD. In vitro and in vivo studies suggested beneficial  
345 effect of silencing *RORC* in the development of atherosclerotic disease<sup>42,43</sup> consistent  
346 with protective effect of pLoF in *RORC* for CAD observed in this study.

347 In this study, we conducted the largest rare variant association study for blood  
348 lipids to-date. The substantial sample size enabled the analysis of single rare variants  
349 as opposed to more conventional aggregation of rare variants into a statistical unit for  
350 burden testing. This analysis not only advances novel mechanistic insights but also  
351 improves the clinical interpretation of Mendelian dyslipidemia genotypes beyond the  
352 current clinical classification schema. Overall, this study demonstrated the capability of  
353 population-based analyses of to identify rare coding alleles with both mechanistic and  
354 clinical implications.

355 Importantly, our study expands allelic diversity by including large cohorts from  
356 non-European-like populations resulting in the discovery of 130 alleles that are  
357 exclusively or dominantly observed in the non-European-like populations. alleles, We  
358 typically observed consistent and highly similar effect sizes across populations despite  
359 differences in allele frequencies. The transferability of associated rare coding alleles  
360 may reflect the causality of these alleles and consistent with our observations from the  
361 systematic evaluation of rare variant burden testing across various traits<sup>44</sup>.

362 In addition to insights specific to blood lipids, our study provides several  
363 observations that may be generalizable. Specifically, our expansive rare variant  
364 association study in highly heritable phenotypes allowed qualitative/quantitative  
365 assessment of variant characteristics behind significant associations. In our study,  
366 associated variants are significantly enriched in functional variant classes (high  
367 confidence pLoF or deleterious missense variants) highlighting the importance of further  
368 effort for precise classification of variant functionality. We further implemented machine  
369 learning-based splice site prediction<sup>25</sup>, and successfully re-classified previously  
370 underestimated variant class. The cryptic splice variants showed constraint pattern and  
371 enrichment by associated variants equivalent to canonical pLoF variants.

372 The broad range of allele-specific analyses also allowed us to infer the variant  
373 effects on gene function. First, we observed highly concordant effect direction of pLoF  
374 alleles (> 99%) on the same gene proxied by phenotypic expression confirming the  
375 findings from previous study<sup>45</sup>. Also, most (87%) missense variants showed concordant  
376 effect directions with pLoFs, however, the remainder (13%) had opposite effects,

377 indicative of hypermorphic characteristics. *In silico* deleterious prediction is not effective  
378 to capture these hypermorphic alleles and these variants might be missed by variant  
379 filters for gene-based testing despite their empiric functional significance. Increasingly  
380 available large-scale genetic analysis across diverse phenotypes and populations,  
381 focusing on rare coding variations, may expand the list of hypermorphic alleles and  
382 inform models to better detect this phenomenon across genes and domains.

383 We also conducted recessive modeling and identified multiple strong  
384 associations. Intriguingly, the recessive associations are robustly shared across studies  
385 and populations. Aligned with previous studies focusing on binary traits<sup>46</sup>, some rare  
386 alleles have a prominent recessive effect not captured by standard additive modeling,  
387 suggesting a contribution to the missing heritability.

388 Additionally, using estimated effect size and statistical significance driven by  
389 population scale association analysis, we re-assessed a curated database considered a  
390 gold standard for clinical genetic diagnosis. We confirmed the accuracy of most variant  
391 annotations aligned with previous study<sup>47</sup> and further provided evidence toward  
392 potential re-classification of the pathogenicity of other variants. Especially, we found two  
393 non-European enriched candidate variants, aligned with previous studies. including  
394 ours,<sup>19,48,49</sup> that reported under diagnosis of genetic disease in the non-European  
395 population. Importantly, we observed a range of expressivity for pathogenic alleles that  
396 was associated with clinical outcomes.

397 In conclusion, we conducted a rare variant focused genetic study for blood lipids  
398 involving over a million individuals, yielding hundreds of rare alleles associated with  
399 blood lipids and improved mechanistic understanding of rare variant associations. Our  
400 study suggests that population-scale rare variant analysis is now adequately powered  
401 for heritable phenotypes, allowing for the classification of rare pathogenic alleles and  
402 providing new insights into variant expressivity/penetrance, toward improved diagnosis  
403 and more quantitative prognosis.



## 404 **References**

- 405 1. Wiegman, A., *et al.* Familial hypercholesterolaemia in children and adolescents:  
406 gaining decades of life by optimizing detection and treatment. *European Heart*  
407 *Journal* **36**, 2425-2437 (2015).
- 408 2. Gidding, S.S., *et al.* The Agenda for Familial Hypercholesterolemia: A Scientific  
409 Statement From the American Heart Association. *Circulation* **132**, 2167-2192  
410 (2015).
- 411 3. Verschuren, J., *et al.* Efficacy of statins in familial hypercholesterolaemia: a long  
412 term cohort study. *BMJ* **337**, a2423 (2008).
- 413 4. Neil, A., *et al.* Reductions in all-cause, cancer, and coronary mortality in statin-  
414 treated patients with heterozygous familial hypercholesterolaemia: a prospective  
415 registry study. *European Heart Journal* **29**, 2625-2633 (2008).
- 416 5. Pijlman, A.H., *et al.* Evaluation of cholesterol lowering treatment of patients with  
417 familial hypercholesterolemia: a large cross-sectional study in The Netherlands.  
418 *Atherosclerosis* **209**, 189-194 (2010).
- 419 6. Nordestgaard, B.G., *et al.* Familial hypercholesterolaemia is underdiagnosed and  
420 undertreated in the general population: guidance for clinicians to prevent  
421 coronary heart disease: Consensus Statement of the European Atherosclerosis  
422 Society. *European Heart Journal* **34**, 3478-3490 (2013).
- 423 7. Sturm, A.C., *et al.* Clinical Genetic Testing for Familial Hypercholesterolemia:  
424 JACC Scientific Expert Panel. *J Am Coll Cardiol* **72**, 662-680 (2018).
- 425 8. Benn, M., Watts, G.F., Tybjaerg-Hansen, A. & Nordestgaard, B.G. Mutations  
426 causative of familial hypercholesterolaemia: screening of 98 098 individuals from  
427 the Copenhagen General Population Study estimated a prevalence of 1 in 217.  
428 *Eur Heart J* **37**, 1384-1394 (2016).
- 429 9. Natarajan, P., *et al.* Aggregate penetrance of genomic variants for actionable  
430 disorders in European and African Americans. *Science Translational Medicine* **8**,  
431 364ra151-364ra361 (2016).
- 432 10. Sun, Y.V., *et al.* Effects of Genetic Variants Associated with Familial  
433 Hypercholesterolemia on Low-Density Lipoprotein-Cholesterol Levels and  
434 Cardiovascular Outcomes in the Million Veteran Program. *Circulation: Genomic*  
435 *and Precision Medicine* **11**(2018).
- 436 11. Forrest, I.S., *et al.* Population-Based Penetrance of Deleterious Clinical Variants.  
437 *JAMA* **327**, 350 (2022).
- 438 12. Clarke, S.L., *et al.* Coronary Artery Disease Risk of Familial  
439 Hypercholesterolemia Genetic Variants Independent of Clinically Observed  
440 Longitudinal Cholesterol Exposure. *Circ Genom Precis Med* **15**, e003501 (2022).
- 441 13. Dewey, F.E., *et al.* Distribution and clinical impact of functional variants in 50,726  
442 whole-exome sequences from the DiscovEHR study. *Science* **354**, aaf6814  
443 (2016).

- 444 14. Green, R.C., *et al.* ACMG recommendations for reporting of incidental findings in  
445 clinical exome and genome sequencing. *Genetics in Medicine* **15**, 565-574  
446 (2013).
- 447 15. Richards, S., *et al.* Standards and guidelines for the interpretation of sequence  
448 variants: a joint consensus recommendation of the American College of Medical  
449 Genetics and Genomics and the Association for Molecular Pathology. *Genetics*  
450 *in Medicine* **17**, 405-424 (2015).
- 451 16. Blout Zawatsky, C.L., *et al.* Returning actionable genomic results in a research  
452 biobank: Analytic validity, clinical implementation, and resource utilization. *The*  
453 *American Journal of Human Genetics* **108**, 2224-2237 (2021).
- 454 17. Sharo, A.G., Zou, Y., Adhikari, A.N. & Brenner, S.E. ClinVar and HGMD genomic  
455 variant classification accuracy has improved over time, as measured by implied  
456 disease burden. *Genome Medicine* **15**(2023).
- 457 18. Kessler, M.D., *et al.* Challenges and disparities in the application of personalized  
458 genomic medicine to populations with African ancestry. *Nature Communications*  
459 **7**, 12521 (2016).
- 460 19. Manrai, A.K., *et al.* Genetic Misdiagnoses and the Potential for Health Disparities.  
461 *New England Journal of Medicine* **375**, 655-665 (2016).
- 462 20. Hunter-Zinck, H., *et al.* Genotyping Array Design and Data Quality Control in the  
463 Million Veteran Program. *Am J Hum Genet* **106**, 535-548 (2020).
- 464 21. Bycroft, C., *et al.* The UK Biobank resource with deep phenotyping and genomic  
465 data. *Nature* **562**, 203-209 (2018).
- 466 22. Taliun, D., *et al.* Sequencing of 53,831 diverse genomes from the NHLBI  
467 TOPMed Program. *Nature* **590**, 290-299 (2021).
- 468 23. Backman, J.D., *et al.* Exome sequencing and analysis of 454,787 UK Biobank  
469 participants. *Nature* **599**, 628-634 (2021).
- 470 24. Bick, A.G., *et al.* Genomic data in the All of Us Research Program. *Nature* **627**,  
471 340-346 (2024).
- 472 25. Jaganathan, K., *et al.* Predicting Splicing from Primary Sequence with Deep  
473 Learning. *Cell* **176**, 535-548 e524 (2019).
- 474 26. Lu, X., *et al.* Exome chip meta-analysis identifies novel loci and East Asian–  
475 specific coding variants that contribute to lipid levels and coronary artery disease.  
476 *Nature Genetics* **49**, 1722-1730 (2017).
- 477 27. Karczewski, K.J., *et al.* The mutational constraint spectrum quantified from  
478 variation in 141,456 humans. *Nature* **581**, 434-443 (2020).
- 479 28. McLaren, W., *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122  
480 (2016).
- 481 29. Liu, X., Li, C., Mou, C., Dong, Y. & Tu, Y. dbNSFP v4: a comprehensive  
482 database of transcript-specific functional predictions and annotations for human  
483 nonsynonymous and splice-site SNVs. *Genome Med* **12**, 103 (2020).

- 484 30. Allard, D., *et al.* Novel mutations of the PCSK9 gene cause variable phenotype of  
485 autosomal dominant hypercholesterolemia. *Human Mutation* **26**, 497-497 (2005).
- 486 31. Hopkins, P.N., *et al.* Characterization of Autosomal Dominant  
487 Hypercholesterolemia Caused by PCSK9 Gain of Function Mutations and Its  
488 Specific Treatment With Alirocumab, a PCSK9 Monoclonal Antibody. *Circulation:  
489 Cardiovascular Genetics* **8**, 823-831 (2015).
- 490 32. Aguet, F., *et al.* The GTEx Consortium atlas of genetic regulatory effects across  
491 human tissues. *Science* **369**, 1318-1330 (2020).
- 492 33. Orho, M., *et al.* Mutations in the liver glycogen synthase gene in children with  
493 hypoglycemia due to glycogen storage disease type 0. *Journal of Clinical  
494 Investigation* **102**, 507-515 (1998).
- 495 34. Reed, M.J., Purohit, A., Woo, L.W.L., Newman, S.P. & Potter, B.V.L. Steroid  
496 Sulfatase: Molecular Biology, Regulation, and Inhibition. *Endocrine Reviews* **26**,  
497 171-202 (2005).
- 498 35. Assanasen, C., *et al.* Cholesterol binding, efflux, and a PDZ-interacting domain of  
499 scavenger receptor–BI mediate HDL-initiated signaling. *Journal of Clinical  
500 Investigation* **115**, 969-977 (2005).
- 501 36. Goldstein, J.L., Debose-Boyd, R.A. & Brown, M.S. Protein Sensors for  
502 Membrane Sterols. *Cell* **124**, 35-46 (2006).
- 503 37. Ehrhardt, N., *et al.* Hepatic Tm6sf2 overexpression affects cellular ApoB-  
504 trafficking, plasma lipid levels, hepatic steatosis and atherosclerosis. *Human  
505 Molecular Genetics* **26**, 2719-2731 (2017).
- 506 38. Pauling, L., Itano, H.A., Singer, S.J. & Wells, I.C. Sickle Cell Anemia, a Molecular  
507 Disease. *Science* **110**, 543-548 (1949).
- 508 39. Ingram, V.M. A Specific Chemical Difference Between the Globins of Normal  
509 Human and Sickle-Cell Anæmia Hæmoglobin. *Nature* **178**, 792-794 (1956).
- 510 40. Xia, W., *et al.* Loss of ABHD15 Impairs the Anti-lipolytic Action of Insulin by  
511 Altering PDE3B Stability and Contributes to Insulin Resistance. *Cell Reports* **23**,  
512 1948-1961 (2018).
- 513 41. Landrum, M.J., *et al.* ClinVar: improving access to variant interpretations and  
514 supporting evidence. *Nucleic Acids Res* **46**, D1062-D1067 (2018).
- 515 42. Zou, H., Yang, N., Zhang, X. & Chen, H.-W. ROR $\gamma$  is a context-specific master  
516 regulator of cholesterol biosynthesis and an emerging therapeutic target in  
517 cancer and autoimmune diseases. *Biochemical Pharmacology* **196**, 114725  
518 (2022).
- 519 43. Cai, D., *et al.* ROR $\gamma$  is a targetable master regulator of cholesterol biosynthesis in  
520 a cancer subtype. *Nature Communications* **10**(2019).
- 521 44. Jurgens, S.J., *et al.* Rare coding variant analysis for human diseases across  
522 biobanks and ancestries. *Nature Genetics* **56**, 1811-1820 (2024).

- 523 45. Barton, A.R., Sherman, M.A., Mukamel, R.E. & Loh, P.-R. Whole-exome  
524 imputation within UK Biobank powers rare coding variant association and fine-  
525 mapping analyses. *Nature Genetics* **53**, 1260-1269 (2021).
- 526 46. Heyne, H.O., *et al.* Mono- and biallelic variant effects on disease at biobank scale.  
527 *Nature* **613**, 519-525 (2023).
- 528 47. Halford, J.L., *et al.* Endophenotype effect sizes support variant pathogenicity in  
529 monogenic disease susceptibility genes. *Nature Communications* **13**(2022).
- 530 48. Sun, K.Y., *et al.* A deep catalog of protein-coding variation in 985,830 individuals.  
531 *bioRxiv*, 2023.2005.2009.539329 (2023).
- 532 49. Koyama, S., *et al.* Decoding Genetics, Ancestry, and Geospatial Context for  
533 Precision Health. *medRxiv*, <https://doi.org/10.1101/2023.1110.1124.23297096>  
534 (2023).
- 535 50. Graham, S.E., *et al.* The power of genetic diversity in genome-wide association  
536 studies of lipids. *Nature* **600**, 675-679 (2021).
- 537 51. Verma, A., *et al.* Diversity and Scale: Genetic Architecture of 2,068 Traits in the  
538 VA Million Veteran Program. (Cold Spring Harbor Laboratory, 2023).
- 539 52. Klarin, D., *et al.* Genetics of blood lipids among ~300,000 multi-ethnic  
540 participants of the Million Veteran Program. *Nature genetics* **50**, 1514-1523  
541 (2018).
- 542 53. Clarke, S.L., *et al.* Race and Ethnicity Stratification for Polygenic Risk Score  
543 Analyses May Mask Disparities in Hispanics. *Circulation* **146**, 265-267 (2022).
- 544 54. Chang, C.C., *et al.* Second-generation PLINK: rising to the challenge of larger  
545 and richer datasets. *Gigascience* **4**, 7 (2015).
- 546 55. Poterba, T., *et al.* The Scalable Variant Call Representation: Enabling Genetic  
547 Analysis Beyond One Million Genomes. *bioRxiv*,  
548 <https://doi.org/10.1101/2024.1101.1109.574205> (2024).
- 549 56. Li, H. Toward better understanding of artifacts in variant calling from high-  
550 coverage samples. *Bioinformatics* **30**, 2843-2851 (2014).
- 551 57. Poplin, R., *et al.* A universal SNP and small-indel variant caller using deep neural  
552 networks. *Nature Biotechnology* **36**, 983-987 (2018).
- 553 58. Fang, H., *et al.* Harmonizing Genetic Ancestry and Self-identified Race/Ethnicity  
554 in Genome-wide Association Studies. *The American Journal of Human Genetics*  
555 **105**, 763-772 (2019).
- 556 59. Ross, P.B., Song, J., Tsao, P.S. & Pan, C. Trellis for efficient data and task  
557 management in the VA Million Veteran Program. *Scientific Reports* **11**(2021).
- 558 60. Mbatchou, J., *et al.* Computationally efficient whole-genome regression for  
559 quantitative and binary traits. *Nat Genet* **53**, 1097-1103 (2021).
- 560 61. Mägi, R. & Morris, A.P. GWAMA: software for genome-wide association meta-  
561 analysis. *BMC Bioinformatics* **11**, 288 (2010).

- 562 62. Frankish, A., *et al.* GENCODE reference annotation for the human and mouse  
563 genomes. *Nucleic Acids Res* **47**, D766-D773 (2019).
- 564 63. Xie, Z., *et al.* Gene Set Knowledge Discovery with Enrichr. *Curr Protoc* **1**, e90  
565 (2021).
- 566 64. Liu, D.J., *et al.* Exome-wide association study of plasma lipids in >300,000  
567 individuals. *Nature Genetics* **49**, 1758-1766 (2017).
- 568 65. McKenna, A., *et al.* The Genome Analysis Toolkit: a MapReduce framework for  
569 analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-  
570 1303 (2010).



## 571 **Online Methods**

### 572 **Ethics oversight**

573 This study received ethics approval from the Veterans Affairs Central Institutional  
574 Review Board under Institutional Review Board protocol number 16-06. The study  
575 protocols were approved under protocol numbers 2016P002395 and 2021P002228 by  
576 the Mass General Brigham Institutional Review Board. The analysis for UKB was  
577 performed under application number 7089.

### 578 **Blood Lipids Phenotyping**

579 The UK Biobank (UKB) is a volunteer cohort of approximately 500,000 residents aged  
580 40 to 69 years of age living in the United Kingdom recruited 2006 - 2010<sup>21</sup>. In the UKB,  
581 blood lipids were measured using blood samples collected at the enrollment. We  
582 adjusted total cholesterol (TC) and low density lipoprotein cholesterol (LDLC) levels by  
583 dividing 0.7 for individuals prescribed lipid-lowering medication at the enrolment as  
584 previously described<sup>50</sup>.

585 The Million Veteran Program (MVP) is a national hospital-based cohort initiated in 2011  
586 by the United States Department of Veterans Affairs (VA). Recruitment was conducted  
587 in the VA affiliated hospitals across the United States<sup>20,51</sup>. In the MVP, lipid phenotypes  
588 were derived from longitudinal lipid measurements over the time. For TC, LDLC and  
589 triglycerides (TG), we utilized the highest value recorded, while for high density  
590 lipoprotein cholesterol (HDL), we selected the lowest value as previously described<sup>52</sup>.

591 The All of Us Research Program is a U.S.-based population cohort that began  
592 enrollment in 2018 under the National Institutes of Health (NIH). Participants were  
593 enrolled through a network of more than 340 recruitment sites. In AOU, lipid phenotypes  
594 were derived similarly to those in MVP. To minimize potential overlap between AOU and  
595 MVP participants, we excluded AOU participants who answered the baseline survey,  
596 indicating they are receiving healthcare from the Veterans Affairs (n = 13,400), from the  
597 analysis.

### 598 **CAD Phenotyping**

599 In the UKB and AOU, we ascertained CAD cases based on at least one of the following  
600 criteria: a) any ICD code in the in-hospital record or death registry (I21 – I25 in ICD10;  
601 410 - 414 in ICD9), or b) any procedure code for coronary revascularization (K40 – K45,  
602 K49, K50, and K75 in OPCS4, 33510 – 33523, 33533 – 33536, 92920 – 92950 in  
603 CPT4). In the MVP, we used a previously established CAD definition<sup>53</sup>. ICD9, ICD10,  
604 and CPT codes along with self-report were used to determine CAD cases and controls.  
605 Qualifying codes were those pertaining to acute myocardial infarction (inpatient only),  
606 stable ischemic heart disease (inpatient or outpatient), and coronary revascularization  
607 (inpatient and outpatient). Cases were individuals who had at least 2 qualifying codes



608 on different dates within a 12-month period. Controls were individuals who carried no  
609 codes and who did not self-report a history of coronary artery disease.

### 610 **Quality control for microarray genotyping in UKB**

611 We conducted sample quality control as follows. Among 488,175 individuals, we  
612 removed samples with aneuploidy (N = 651), sex-gender mismatch implying phenotypic  
613 quality issues (N = 378), higher heterozygosity or missing outlier (N = 739) leading to a  
614 total of 1811 (0.4%) samples removed. 486,364 quality control passed individuals  
615 remained (9,454 AFR, 2,413 AMR, 2,582 EAS, 461,352 EUR, 10,563 SAS).

### 616 **Population ascertainment in UKB**

617 Using reference population data from 1000 genomes project and microarray genotypes  
618 in UKB using the Affymetrix UK BiLEVE Axiom and UK Biobank Axiom arrays, we  
619 determined genetically determined population ascertainment. First, we extracted quality-  
620 controlled variants from 1000 genomes data [non-palindromic single nucleotide variant  
621 (SNV), minor allele frequency (MAF) > 1%, a population specific Hardy Weinberg  
622 equilibrium  $P$ -value >  $1 \times 10^{-6}$ ]. Next, we extracted the intersection of the quality  
623 controlled 1000 genomes data and the study population. Using intersected variants, we  
624 pruned variants on the 1000 genomes data using PLINK2<sup>54</sup> software (Jun 3, 2022,  
625 release) with --indep-pairwise option (window size 50, sliding window size 10,  $R^2 < 0.2$ ).  
626 Which yields 224,993 variants. Using pruned variants, we calculated the SNV weights  
627 for genetic principal components. Then we projected study participants to the principal  
628 components space. Using the 1000 genomes reference population annotation, we  
629 trained the k-nearest neighbor model using class R package (version 7.3). Then we split  
630 study cohort into the five genetic populations (African-like, AFR; Admixed-American-like,  
631 AMR; East-Asian-like, EAS; European-like, EUR; and South-Asian-like, SAS) and  
632 conducted association analyses separately to minimize potential effects of  
633 heterogeneity.

### 634 **Quality control for WES in UKB**

635 For genotype level quality control, first, we utilized Hail's<sup>55</sup> `split\_multi\_hts` function to  
636 divide multiallelic sites. We then filtered out low-quality genotypes based on the  
637 following criteria: i) Genotyping quality less than or equal to 20. ii) Genotype depth (DP)  
638 either less than or equal to 10 or greater than 200. iii) For heterozygous genotypes:  
639  $(DP_{\text{Reference}} + DP_{\text{Alternate}})/(DP_{\text{Total}}) > 0.9$  and  $DP_{\text{Alternate}}/DP_{\text{Total}} > 0.2$ . iv) For alternate  
640 homozygous genotypes:  $DP_{\text{Alternate}}/DP_{\text{Total}} > 0.9$ .

641 These processes retained 26,645,535 variants in the 454,756 sequenced samples. We  
642 excluded i) 6,131,710 variants due to high missingness (missing rate > 10%). ii) 47,441  
643 variants that deviated from the Hardy-Weinberg equilibrium ( $P_{\text{Hardy-Weinberg equilibrium}} < 1 \times$   
644  $10^{-15}$ ). iii) 364,207 variants located within low-complexity regions<sup>56</sup>. Cumulatively, we  
645 excluded 6,289,813 variants, resulting in 20,355,722 retained variants.

646 Among 454,756 individuals whole exome sequenced (450K UKB release using Deep  
647 Variant<sup>57</sup> for variant detection), we identified 452,929 individuals overlapping array  
648 genotyped data. For these individuals, we conducted sample-level quality control. First,  
649 we calculated array-exome genotype discordance rate and F-statistics in non-pseudo  
650 autosomal region X-chromosome variants to detect potential sample or phenotypic  
651 swapping in exome data. For this analysis, we used pruned and stringent variant quality  
652 control criteria (missingness < 1%, MAF > 0.1%). We identified 27 potential sex-  
653 swapping (27 Females with F- statistics > 0.6, 0 Males with F- statistics < 0.6) and 0  
654 discordant genotypes between exome and array data (non-reference homozygote  
655 concordance rate < 0.8). We calculated array-exome discordance using pruned, non-  
656 palindromic, high-quality exome data and the corresponding array data. Next, we  
657 removed samples with a high missing rate (> 10%, N = 12). Then, we filtered samples  
658 using the following autosomal quality control outlier metrics [outside mean  $\pm$  8 standard  
659 deviation (SD)]: heterozygous/homozygous rate (N = 761), transition/transversion rate  
660 (N = 0), SNV/Insertion-Deletion ratio (N = 2), number of singletons (N = 283). In total,  
661 1,052 (0.2%) samples were removed, resulting in 451,877 samples in the final dataset.  
662 After removing these samples, we also removed 226,083 monomorphic variants in the  
663 dataset retaining 20,129,639 variants among the 451,877 samples in the final dataset.

#### 664 **Quality control for microarray genotyping, population ascertainment, and** 665 **imputation in MVP**

666 Genotyping was performed using the custom Axiom array (MVP1.0), and variant and  
667 sample quality control was described in detail previously<sup>20</sup>. We used the latest release  
668 (release 4) data for this analysis. Release 4 data included array-genotypes and genetic  
669 dosage imputed to the TOPMed imputation r2 reference panel<sup>22</sup>. The quality-controlled  
670 sample size was 657,242. We grouped participants into four population groups [AFR,  
671 Asian-like (ASN), EUR, and Hispanic-like, HIS] following the harmonized ancestry and  
672 race/ethnicity (HARE) algorithm previously established in MVP<sup>58</sup>.

#### 673 **Quality control for WGS in MVP**

674 To confirm the accuracy, sensitivity, and specificity of the imputation, we have utilized  
675 the initial release of whole genome sequencing data in the MVP study. This data was  
676 collected and sequenced with a focus on elucidating the pathophysiology of COVID-19  
677 infection from their genomes. The sequencing was performed using Illumina's  
678 Sequencing by Synthesis technology to a targeted depth of 30x. Individual variant  
679 calling from 10,413 samples was performed on the cloud-based data and task  
680 management framework Trellis<sup>59</sup>. In summary, reads were aligned with BWA-MEM  
681 (version 0.7.15) on the GRCh38 reference genome, and variant calling was performed  
682 in GATK 4.1.0.0 using the haplotypeCaller function. Genotypes of all samples were  
683 aggregated into a matrix table using gVCF Combiner implemented in Hail<sup>55</sup> for  
684 additional quality-control steps. In summary, we retained high-quality genotypes by

685 applying the following steps: I. Variants in low complexity regions and ENCODE  
686 blacklist regions were removed. II. Variants within regions of atypical sequencing depth  
687 ( $DP < 10$  or  $DP > 400$ ) were discarded. For haploid genotypes on sex chromosomes, a  
688 minimum  $DP > 5$  was required. III. Genotypes were retained if sites were: a.  
689 Homozygous reference with Genotype Quality  $> 20$ , or b. Alternate homozygotes with  
690 Phred-scaled likelihood of the genotype for reference homozygotes ( $PL[0]$ )  $> 20$ , and  
691 the ratio of depth for alternate alleles ( $DP_{ALT}$ ) to total depth at the site ( $DP_{ALT}/DP_{SITE}$ )  $>$   
692  $0.9$ , or, c. Heterozygous with  $PL[0] > 20$ , and the ratio of the sum of  $DP_{ALT}$  and depth for  
693 reference alleles ( $DP_{REF}$ ) to  $DP_{SITE}$  [ $(DP_{ALT} + DP_{REF})/DP_{SITE}$ ]  $> 0.9$ , and  $DP_{ALT}/DP_{SITE} >$   
694  $0.2$ . III. Variants with high missing rate ( $> 0.8$ ) and population wide  $P_{\text{Hardy-Weinberg equilibrium}}$   
695  $\leq 1 \times 10^{-5}$  for variants with minor allele frequency (MAF)  $\geq 1\%$ , and  $P_{\text{Hardy-Weinberg equilibrium}}$   
696  $\leq 1 \times 10^{-6}$  for variants with MAF  $< 1\%$  were discarded. IV. Samples with low call rate ( $\leq$   
697  $0.97$ ) or low overall sequencing coverage (mean depth  $\leq 18$ ) were excluded. This  
698 processing resulted in 187,790,701 variants in 10,390 individuals.

### 699 **Quality control for WGS in AOU**

700 We curated genotypes from the jointly called WGS call set (version 7) provided AOU<sup>24</sup>.  
701 We split multiallelic site to biallelic variants using hail's split\_multi\_hts function, then, low  
702 quality genotypes flagged as FAIL in FT field was set as missing. The genotypes were  
703 export as bgen files and converted to pgen files for quality control procedure and  
704 downstream analysis. We filtered variants i) flagged in the FILTER column in original  
705 VDS, ii) located in the low complexity region, iii) low call rate ( $< 90\%$ ), iv) monomorphic,  
706 v) population specific Hardy Weinberg equilibrium P-value  $< 1 \times 10^{-15}$ . Finally, we  
707 excluded flagged individuals ( $n = 549$ ) and genotype missing rate more than 1% ( $n =$   
708  $396$ ).

### 709 **Exome-wide association analysis**

710 We utilized the association analysis framework implemented in Regenie software  
711 (version 3.1.3)<sup>60</sup>. We used array-based autosomal genotypes for step 1 excluding  
712 variants with MAF  $< 1\%$ ,  $P_{\text{Hardy-Weinberg equilibrium}} < 1 \times 10^{-15}$ , call rate  $< 98\%$ , and located  
713 in the Major Histocompatibility Complex region (Chromosome 6 23 - 37 megabases).  
714 We pruned variants using PLINK2<sup>54</sup> software (Jun 3, 2022, release) with --indep-  
715 pairwise option (window size 1000, sliding window size 100,  $R^2 < 0.9$ ) by genetic  
716 population in each cohort. The association model was adjusted by age, age<sup>2</sup>, sex, and  
717 the first ten genetic principal components and blood lipids measurements were inverse  
718 rank normalized. For CAD analysis, we used firth logistic regression implemented in  
719 Regenie software. We tested all quality-controlled genotypes in exome sequence data  
720 in UKB. For the MVP whole genome imputed dataset and AOU WGS dataset, we  
721 restricted the analysis to the exome sequence targeting file used for UKB exome  
722 sequencing (<https://biobank.ndph.ox.ac.uk/ukb/refer.cgi?id=3801>) with 50bp flanking in  
723 both sides of the target region. After generating summary statistics for each cohort

724 (MVP, AOU, and UKB) and each population (AFR, HIS, ASN, EUR in MVP and AFR,  
725 AMR, EAS, EUR, SAS in UKB and AOU), we meta-analyzed the results using the  
726 GWAMA<sup>61</sup> software (version 2.2.2).

727 The obtained summary statistics were meta-analyzed using GWAMA<sup>61</sup> fixed effect  
728 model. We computed effect size and *P*-values all the variants in exome region  
729 irrespective to the variant annotation. While we used summary statistics for  
730 synonymous/non-coding variants as reference to contextualize coding associations, the  
731 statistical significance of these variants was not considered throughout study. In this  
732 study, we primarily applied additive model for the association analysis. In addition to the  
733 additive model, we performed association analyses modeling recessive effects. For the  
734 additive model, we restricted the analysis for the variants with  $5 \leq \text{MAC}$  and  $\text{MAF}_{\text{POP}_{\text{MAX}}} < 1\%$   
735 before meta-analysis. For recessive model, we also restricted the analysis for the  
736 variants with  $5 \leq$  estimated minor homozygote counts (number of participants  $\times$   $\text{MAF}^2$ )  
737 and estimated minor homozygote frequency ( $\text{MAF}^2$ )  $< 1\%$ . We reported unadjusted *P*-  
738 values without correction for multiple testing throughout the manuscript.

739 Per the reporting guidelines by the MVP, we have masked the MAF of variants with a  
740 MAC of less than equal to 12 in the summary statistics. This measure is implemented to  
741 prevent the potential identification of individuals participating in the study. The directions  
742 of the effects in the table and summary statistics indicate the variant effect in the  
743 alphabetical order:  $\text{AOU}_{\text{AFR}}$ ,  $\text{AOU}_{\text{AMR}}$ ,  $\text{AOU}_{\text{EAS}}$ ,  $\text{AOU}_{\text{EUR}}$ ,  $\text{AOU}_{\text{SAS}}$ ,  $\text{MVP}_{\text{AFR}}$ ,  $\text{MVP}_{\text{ASN}}$ ,  
744  $\text{MVP}_{\text{EUR}}$ ,  $\text{MVP}_{\text{HIS}}$ ,  $\text{UKB}_{\text{AFR}}$ ,  $\text{UKB}_{\text{AMR}}$ ,  $\text{UKB}_{\text{EAS}}$ ,  $\text{UKB}_{\text{EUR}}$ , and  $\text{UKB}_{\text{SAS}}$ .

#### 745 **Variant annotation**

746 We utilized a single transcript for each gene based on Gencode v41<sup>62</sup> canonical and  
747 coding transcript (coding transcript set,  $n = 19,603$ , Supplementary Dataset) for all  
748 annotations. First, we annotated tested variants (variants within  $\pm 50$  bases from target  
749 region in the UKB exome) with the VEP<sup>28</sup> software (version 107, aligned with Gencode  
750 v41) and selected annotations on the coding transcript set. If we found variants  
751 overlapping in more than two transcripts in the coding transcripts set, we selected  
752 higher functional consequence. If the consequences were equivalent, we selected the  
753 annotation on the longest transcript. We selected predicted loss of function (pLoF,  
754 IMPACT HIGH) or missense (IMPACT MODERATE) variants as coding variants. Next,  
755 we ran Splice AI<sup>25</sup> for all the variant tested for the coding transcript set with default  
756 parameters. Splice AI returns DS which represent potential for cryptic splicing  
757 (Supplementary Notes III). We treated non-pLoF variants with  $\text{DS} > 0.8$  as cryptic splice  
758 variants and reclassified them as pLoF.

#### 759 **Missense Score**

760 For further classification of missense variants, we applied ensemble prediction using 29  
761 in-silico prediction models to assess the deleteriousness of missense single nucleotide

762 variants. Using pre-computed in-silico predictions in the dbNSFP<sup>29</sup> database (version  
763 4.2), we annotated all the missense variants using the dbNSFP plugin for VEP. We  
764 binarized the predictions into 'Deleterious' or 'Tolerant' using an algorithm-specific  
765 threshold. We computed the Missense Score, ranging from 0 to 1, is calculated by  
766 dividing the number of Deleterious predictions by the total number of available  
767 algorithms for the corresponding variant.

### 768 **Pathway enrichment analysis**

769 We compared pathway enrichments between genes identified by rare coding variant  
770 associations in this study (1) and those identified by common variant association studies  
771 (2). The analysis was restricted to genes included in the coding transcript set (n =  
772 19,603). For genes supported by rare variants (1), we chose those with the smallest P-  
773 value within their respective loci. For genes supported by common variants (2), we  
774 selected the closest gene to the lead variant identified in a recent large-scale genome  
775 wide association study (GWAS) by Graham et al., as reported in *Nature* in 2021<sup>50</sup>. For  
776 the selected gene sets, enrichment was tested using the enrichR (version 3.0) R  
777 package<sup>63</sup>, considering the following pathway sets: Reactome\_2022;  
778 KEGG\_2021\_Human; GO\_Biological\_Process\_2023; GO\_Cellular\_Component\_2023;  
779 GO\_Molecular\_Function\_2023; ChEA\_2022; ENCODE\_TF\_ChIP-seq\_2015; ENCODE  
780 and ChEA Consensus TFs from ChIP-X; and Enrichr Submissions TF-Gene  
781 Cooccurrence. Enrichment was considered significant if the P-value was lower than the  
782 threshold adjusted by the Holm method.

### 783 **Replication analysis**

784 For replication, we utilized data from a previous large-scale exome array study by Lu et  
785 al., as reported in *Nat Genet* in 2017<sup>64</sup>. All variants were updated to the hg38 reference  
786 genome using the LifterVcf function in GATK<sup>65</sup>. We then combined the updated  
787 summary statistics from the previous study with those from our current study. In total,  
788 we identified 387 combinations of variants and phenotypes that matched between both  
789 studies. The concordance of effect sizes and statistical significance was assessed. A  
790 directional concordance was noted if the effect direction was the same in both the  
791 replication dataset and our study. Statistical significance was defined by a P-value in the  
792 replication dataset that was smaller than the Bonferroni-adjusted threshold ( $P <$   
793  $0.05/387$ ).

### 794 **Variance Explained**

795 Per variant explained variance (*Var*) was computed by the following formula<sup>50</sup>:

$$Var = 2f(1 - f)\beta^2$$

796 We calculated the variance explained by common variants using the index variant from  
797 the latest GWAS for common variants (511 variants for TC, 442 variants for LDLC, 562



798 variants for HDLC, 480 variants for TG)<sup>50</sup>. To eliminate linked variants in exome wide  
799 significant (EWS) rare variants, we employed the *clump* function in PLINK1.9<sup>54</sup>. With the  
800 MVP imputed genotypes and UKB WES, we clumped EWS variants using an  $R^2$   
801 threshold of 0.01. By this process, variants in linkage disequilibrium ( $R^2 \geq 0.01$ ) with any  
802 variants that had smaller P-values were excluded. Additionally, EWS variants do not  
803 present in the MVP or UKB were omitted from the analysis. As a result, 172, 195, 182,  
804 and 121 variants from MVP and 179, 197, 185, and 128 variants from UKB were  
805 retained for TC, LDLC, HDLC, and TG, respectively.

## 806 **Conditioning analysis**

807 To evaluate the independence of genetic signals derived from rare coding variants and  
808 common variants, we executed a conditional analysis where rare coding variants were  
809 incorporated as covariates. This analysis was performed in addition to using the  
810 standard covariates applied in our primary analyses, which included sex, age, the age<sup>2</sup>,  
811 and the first ten genetic PCs. For conditioning purpose, we utilized genotype data for  
812 rare coding variants with EWS as covariates. In the MVP, we introduced 185, 207, 203,  
813 and 131 rare coding variants as covariates for TC, LDLC, HDLC, and TG, respectively.  
814 Similarly, in the UKB, 197, 224, 209, and 140 variants were introduced to the model for  
815 TC, LDLC, HDLC, and TG, respectively. By comparing the  $\beta$  and *P*-values obtained  
816 from the analyses conducted with and without these genotype covariates, we aimed to  
817 ascertain the extent to which signals from rare variants are dependent on or  
818 independent from those associated with common variants. This approach leveraged the  
819 *condition* function available in the Regenie software package. Conditioning was done in  
820 both Step 1 and Step 2.

## 821 **Pathogenic variant reclassification**

822 We curated pathogenic alleles for Familial Hypercholesterolemia, a well-known  
823 monogenic condition linked to severe hypercholesterolemia and premature coronary  
824 artery disease, from the ClinVar<sup>41</sup> database, downloading bulk data on August 16, 2022.  
825 We first extracted genetic regions corresponding to the *PCSK9*, *APOB*, and *LDLR*  
826 genes from the VCF file. Using the same pipeline as for the tested variants, we  
827 annotated these variants and excluded pLoF variants for *PCSK9* and *APOB* due to their  
828 known reduction of LDLC levels. We then classified variants as necessary with  
829 conflicting interpretations by majority vote. We calculated the difference in evidence  
830 [Number of Pathogenic + Likely Pathogenic – (Benign + Likely Benign + Uncertain  
831 Significance)]; if the score was greater than 0, the variants were considered  
832 Pathogenic/Likely pathogenic; otherwise, they were considered Benign. This process  
833 resulted in a categorized list of variants in three classes, i) Pathogenic/Likely  
834 Pathogenic (P/LP), ii) Uncertain Significance (VUS), and iii) Benign/Likely Benign (B/LB).  
835 We intersected these variants with those in *APOB/LDLR/PCSK9* tested in this study and  
836 defined the pathogenic effect size by taking the median of positive effect sizes from



837 known pathogenic variants. To identify a subset of VUS to be reclassified as P/LP, we  
838 ranked the variants by their effect sizes and grouped them, accordingly, ensuring that  
839 the median effect size of this group was larger than that of the known P/LP variants.

#### 840 **External data**

841 For replication, we obtained summary statistics from previous exome array-based study  
842 (Lu et. al. *Nat Genet* 2017)<sup>26</sup>. We lifted summary statistics from hg19 coordinate to hg38  
843 using LiftOverVcf function in picard software. We removed insertions/deletions due to  
844 ambiguousness of alleles (n = 24) and failed in lifting (n = 71). In total, we successfully  
845 lifted > 99.96% (292,322/292,417) of variants in the data. For common variant  
846 integration analysis, we obtained summary statistics from the latest GWAS (Graham et.  
847 al. *Nature* 2021)<sup>50</sup>. We utilized summary statistics from trans population meta-analysis  
848 (with\_BF\_meta-analysis\_AFR\_EAS\_EUR\_HIS\_SAS\_\*\_INV\_ALL\_with\_N\_1.gz, for  
849 autosomes and meta-  
850 analysis\_chrX\_AFR\_EAS\_EUR\_HIS\_SAS\_\*\_INV\_ALL\_with\_N\_1.gz for X  
851 chromosome). All summary statistics were downloaded from the Global Lipids Genetics  
852 Consortium website (<http://www.lipidgenetics.org>). For those summary statistics, we  
853 successfully lifted more than 99.82% of variants.

#### 854 **Data availability**

855 Full summary statistics will be publically available after the acceptance of the  
856 manuscript through dbGAP. The individual data for AOU, MVP, and UKB is available  
857 upon application to the respective organizations. The analysis codes and supplemental  
858 data are available at Zenodo (<https://zenodo.org/doi/10.5281/zenodo.11092802>). The  
859 docker/singularity images used in the analysis are publically available through docker  
860 hub (<https://hub.docker.com/u/skoyamamd>).

## 861 **Acknowledgments**

862 This research is based on data from the Million Veteran Program, Office of Research  
863 and Development, Veterans Health Administration, and was supported by award  
864 #BX004821. This publication does not represent the views of the Department of Veteran  
865 Affairs or the United States Government. The analysis of UK Biobank was performed  
866 under the application number 7089. S.K. is supported by Japan Society for the  
867 Promotion of Science (202160643), Uehara Memorial Foundation, and National Heart  
868 Lung and Blood Institute (NHLBI, K99HL169733). Z.Y. is supported by National Human  
869 Genome Research Institute (K99HG012956). S.H.C is supported by NHLBI  
870 (R01HL127564). S.J.J. is supported by the Dutch Heart Foundation (grant no. 03-007-  
871 2022-0035). M.S.S is supported by TOPMed (2022-6842.02), D.K. is supported by the  
872 Department of Veterans Affairs (VA, IK2BX005759-01), the American Heart Association  
873 (DOI: <https://doi.org/10.58275/AHA.23SCEFIA1153369.pc.gr.173943>), and the  
874 Baszucki Research Initiative provided to Stanford Vascular Surgery. This work was  
875 supported in part via funding from VA Merit Award I01 BX003362 (K.M.C., P.S.T.) from  
876 the VA Office of R&D. P.N. and G.M.P are supported by NHLBI (R01HL142711,  
877 R01HL127564).

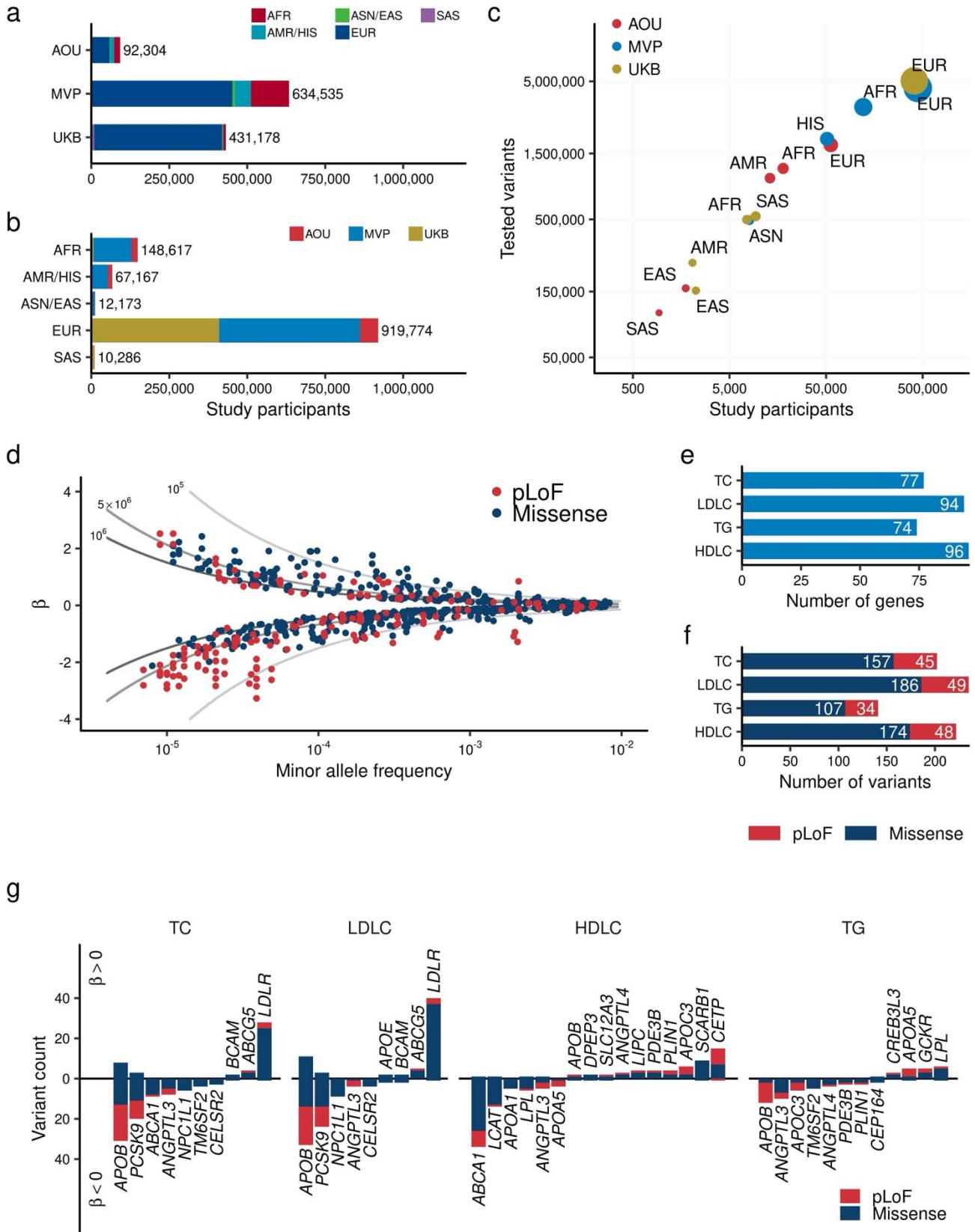
## 878 **Author contributions**

879 S.K., P.T.E., Y.V.S., P.W.W, P.N. conceptualized this project. S.K., Z.Y., D.K., J.E.H.,  
880 K.C. curated phenotype data. S.K., S.H.C, S.J.J., M.S.S., D.K., J.E.H., J.S.D., P.S.T.  
881 curated genotype data. S.K., Z.Y., S.H.C, S.J.J., M.S.S., M.N.T., A.R. analyzed data.  
882 S.K., J.E.H., M.N.T., A.R., J.S.D., C.S., I.S., S.M.D., K.M.C., T.L.A., D.J.R., G.M.P.,  
883 P.T.E., Y.V.S., P.W.W, P.N. interpreted data. S.K., Y.V.S., P.W.W, P.N. prepared the  
884 initial draft S.K., Z.Y., S.J.J., M.S.S., D.K., J.S.D., C.S., I.S., S.M.D., K.M.C., T.L.A.,  
885 D.J.R., G.M.P., P.T.E., Y.V.S., P.W.W, P.N. provided critical review and edits for the  
886 manuscript. W.H., P.S.T., K.C., P.T.E., Y.V.S., P.W.W, P.N. supervised the project. A.B.,  
887 K.L., W.H., P.S.T., K.C., P.T.E., Y.V.S., P.W.W managed the project administration.  
888 K.C., P.T.E., Y.V.S., P.W.W, P.N. obtained funding for the project.

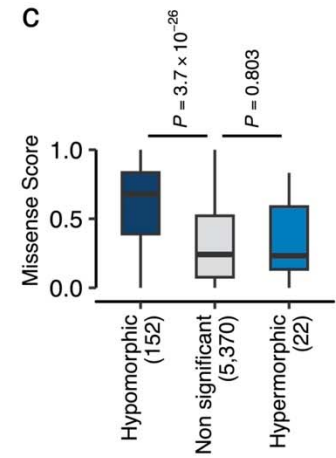
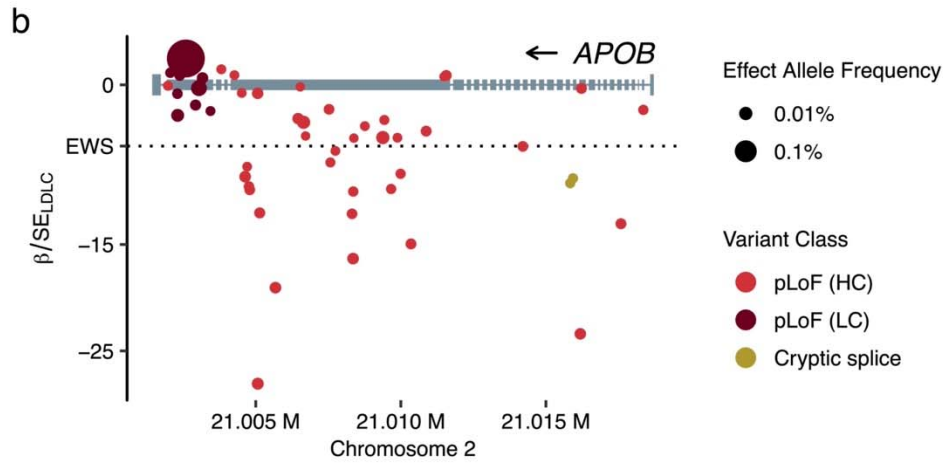
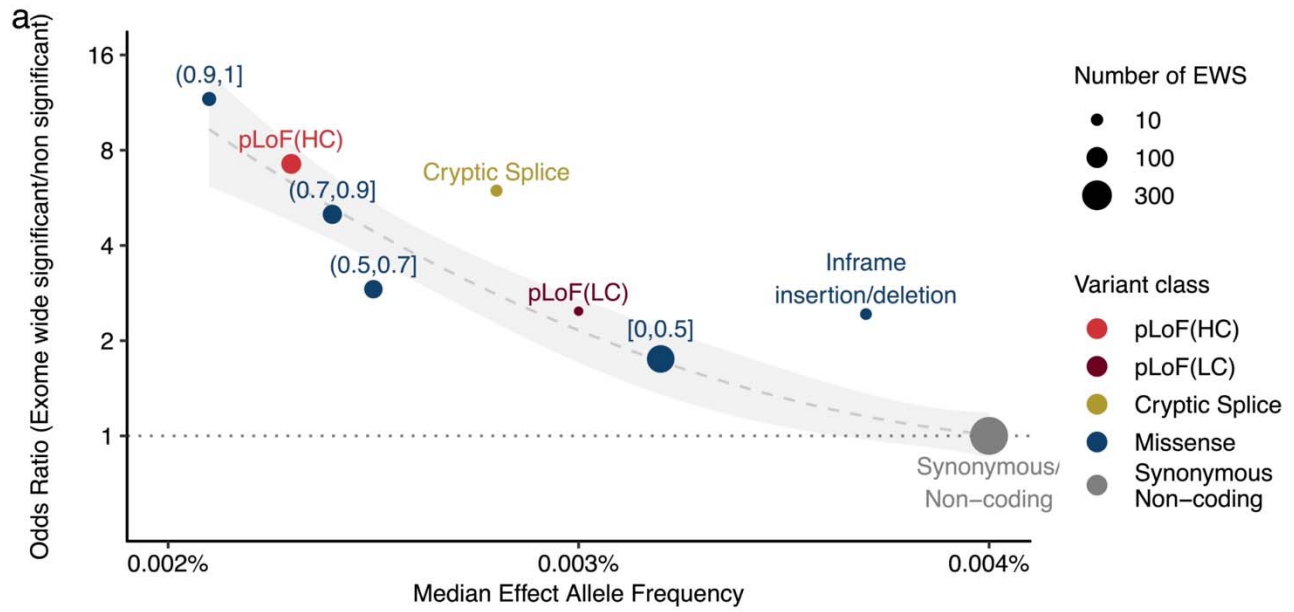
## 889 **Competing interest declaration**

890 D.K. is a scientific advisor and reports consulting fees from Bitterroot Bio, Inc unrelated  
891 to the present work. P.N. reports research grants from Allelica, Amgen, Apple, Boston  
892 Scientific, Genentech / Roche, and Novartis, personal fees from Allelica, Apple,  
893 AstraZeneca, Blackstone Life Sciences, Creative Education Concepts, CRISPR  
894 Therapeutics, Eli Lilly & Co, Foresite Labs, Genentech / Roche, GV, HeartFlow, Magnet  
895 Biomedicine, Merck, and Novartis, scientific advisory board membership of Esperion  
896 Therapeutics, Preciseli, and TenSixteen Bio, scientific co-founder of TenSixteen Bio,  
897 equity in MyOme, Preciseli, and TenSixteen Bio, and spousal employment at Vertex  
898 Pharmaceuticals, all unrelated to the present work.

899 **Figures**



901 **Fig. 1 | Exome wide association study for blood lipids over one million individuals a.** and  
902 **b.** Overview of the study. The number of individuals included in the analysis by study (a) and  
903 by population (b). **c.** Correlation between the number of individuals and identified variants in  
904 the target region. The horizontal axis shows the number of individuals in each population by  
905 study. The vertical axis shows the number of variants identified in the corresponding  
906 population. The size of point is proportional to the number of individuals. **d.** Distribution of  
907 effect sizes for exome-wide significant associations is shown. Each dot represents a variant-  
908 trait pair with significant association in this study (Methods). All four blood lipids are plotted.  
909 The horizontal axis indicates the minor allele frequency, while the vertical axis displays the  
910 effect size for each allele from the regression model ( $\beta$ ), with the unit of effect size normalized  
911 to the standard deviations of blood lipids. The lines represent the statistical power of 80% at  
912 sample sizes of one million (dark gray), 500,000 (medium gray), and 100,000 (light gray)  
913 individuals. **c.** Minor allele frequency of associated variants by variant impact. The rectangles  
914 illustrate the interquartile range of the minor allele frequencies, with the bottom and top edges  
915 representing the first and third quartiles, respectively. The line inside the rectangle denotes the  
916 median and the whiskers extend from the quartiles to the smallest and largest observed values,  
917 within a distance no greater than 1.5 times the interquartile range. **d.** Direction of the effects for  
918 associated variants. Variants positively associated with the blood lipids are displayed on the  
919 positive side of the vertical axis. The height of each bar represents the number of variants in  
920 that category. Bar colors indicate variant classes, with blue for missense variants and red for  
921 pLoF variants. AFR, African-like population; AMR, Admixed-American-like population; ASN,  
922 Asian-like population; EAS, East-Asian-like population; EUR, European-like population; HIS,  
923 Hispanic-like population; SAS, South-Asian-like population; TC, Total Cholesterol; LDLC, Low  
924 Density Lipoprotein Cholesterol; HDLC, High Density Lipoprotein Cholesterol; TG,  
925 Triglycerides; pLoF, predicted Loss of Function; EWS, Exome Wide Significance.

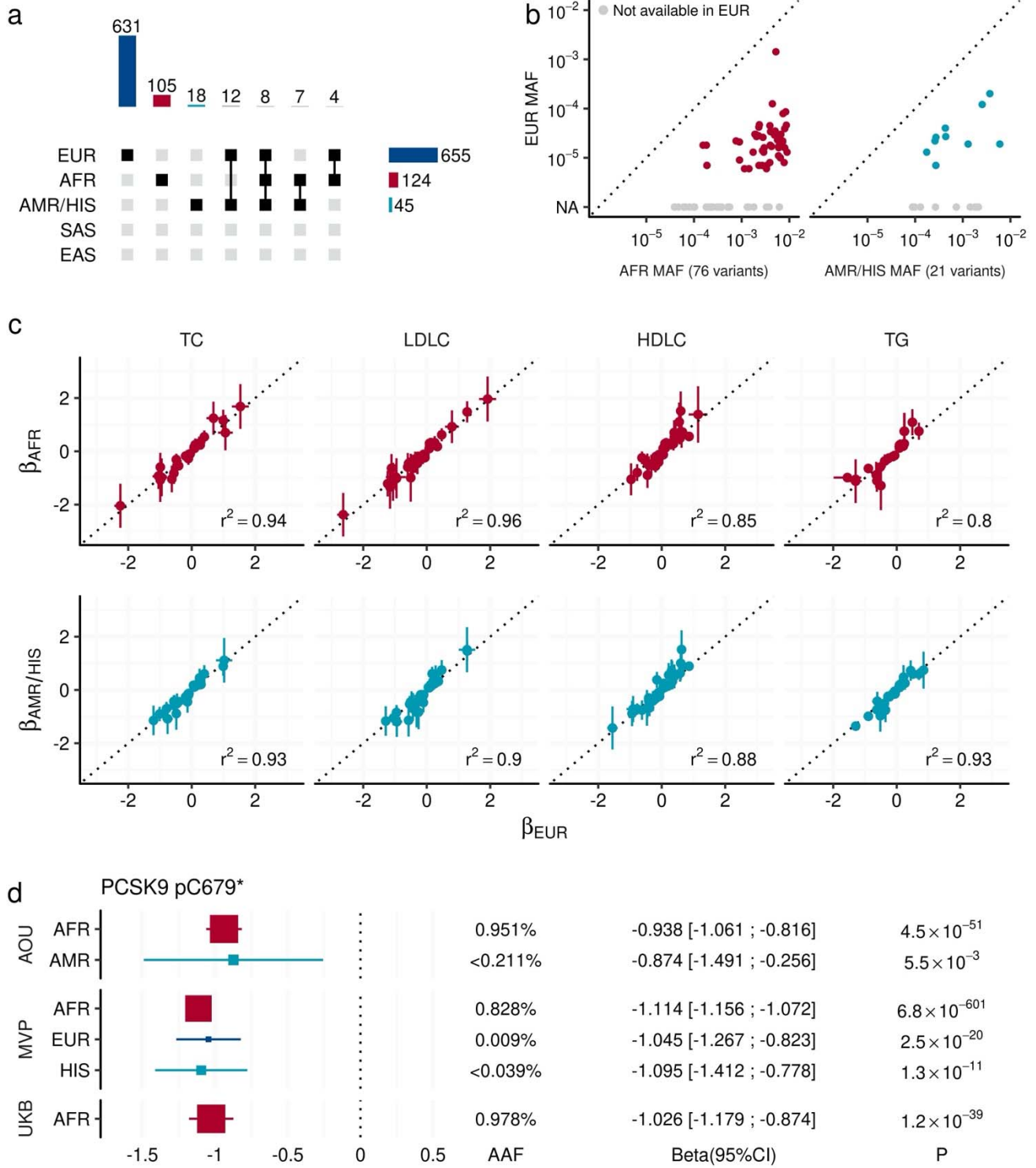


926

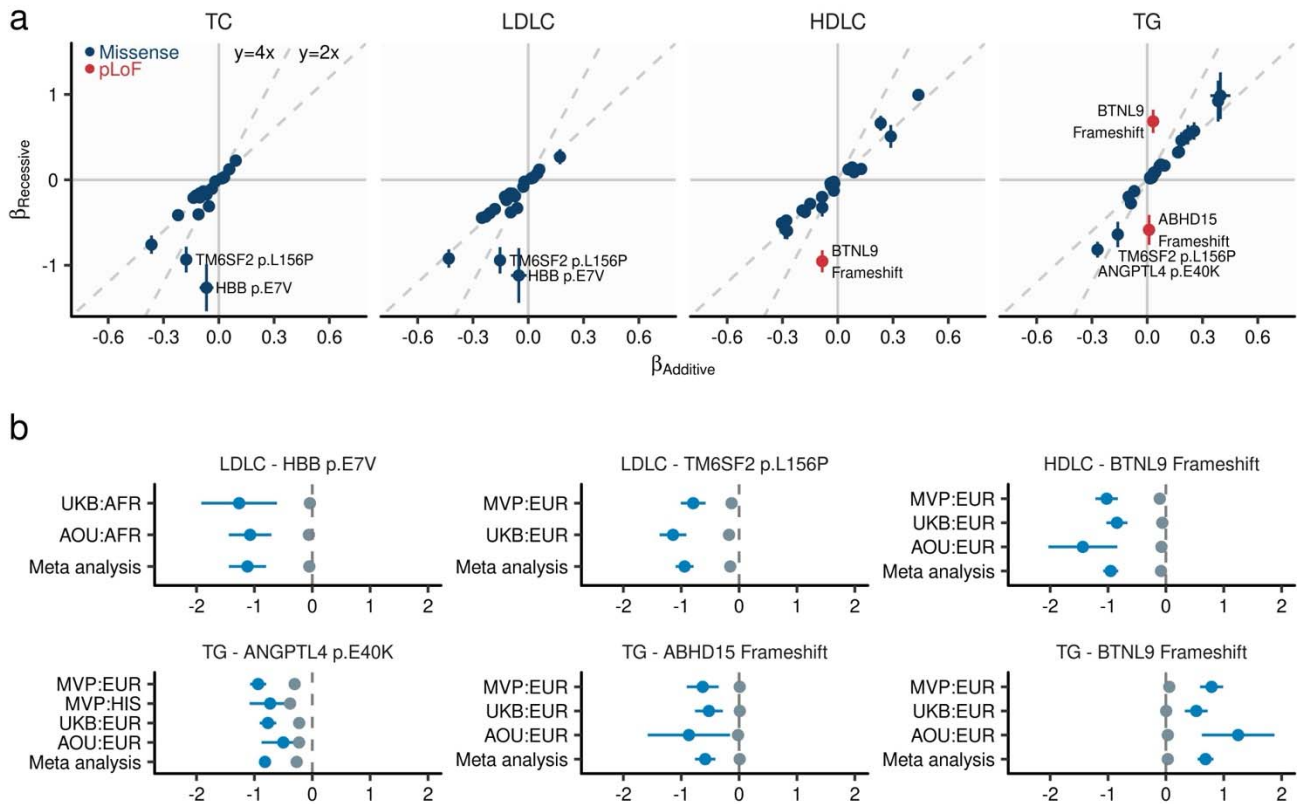


927 **Fig. 2 | Different expressivity of rare coding variants by variant classes a.** Variant  
928 deleteriousness, constraints, and statistical associations. The panel represents variant classes  
929 as pLoF (red), Missense (blue), and Synonymous/Non-coding (gray, used as reference). The  
930 ranges associated with the blue points depict the Missense Score for missense variants. We  
931 computed the Missense Score for missense single nucleotide variants by using 29 in-silico  
932 deleteriousness prediction algorithms. The score was calculated as the number of deleterious  
933 predictions divided by the number of available algorithms for each variant, with values ranging  
934 from 0 to 1 (Methods). Based on the Missense Scores, missense variants were grouped into  
935 bins. pLoF variants were grouped by LOFTEE predictions. The horizontal axis indicates the  
936 median minor allele frequency for each variant class, while the vertical axis shows the odds  
937 ratios of EWS to non-EWS variants in reference to Synonymous/Non-coding variants. Odds  
938 ratios were estimated by Fisher's Exact test. Circle size corresponds to the number of variants  
939 achieving EWS in each variant class. The dashed curve is the estimated line, and the shaded  
940 area is its 95% confidence interval. **b.** Penetrance of pLoF variants in the *APOB*. Gray  
941 rectangles represent the *APOB* gene model. Circles correspond to genetic variants examined  
942 in this study, with circle size denoting effect allele frequency, and color signifying variant class.  
943 The horizontal axis outlines genomic coordinates (hg38), whereas the vertical axis indicates Z-  
944 values (Beta/Standard Error) for LDLC association calculated by liner mixed model (Methods).  
945 **c.** Different distributions of Missense Scores (See above) observed in hypermorphic and  
946 hypomorphic variants. The box plot displays the distribution of Missense Scores for Missense  
947 variants within genes that have at least one EWS association by pLoF. A hypomorphic variant  
948 is defined as having the same directional association with EWS pLoF association. The *P*-  
949 values were calculated by two-sided Wilcoxon's rank-sum test. The *P*-values were not  
950 adjusted for multiple testing correction. Conversely, a hypermorphic variant is defined as  
951 having an opposite directional association to EWS. pLoF, predicted Loss of Function; HC, High  
952 Confidence; LC, Low Confidence; EWS, Exome Wide Significance; LDLC, Low Density  
953 Lipoprotein Cholesterol.

It is made available under a [CC-BY-NC 4.0 International license](https://creativecommons.org/licenses/by-nc/4.0/).

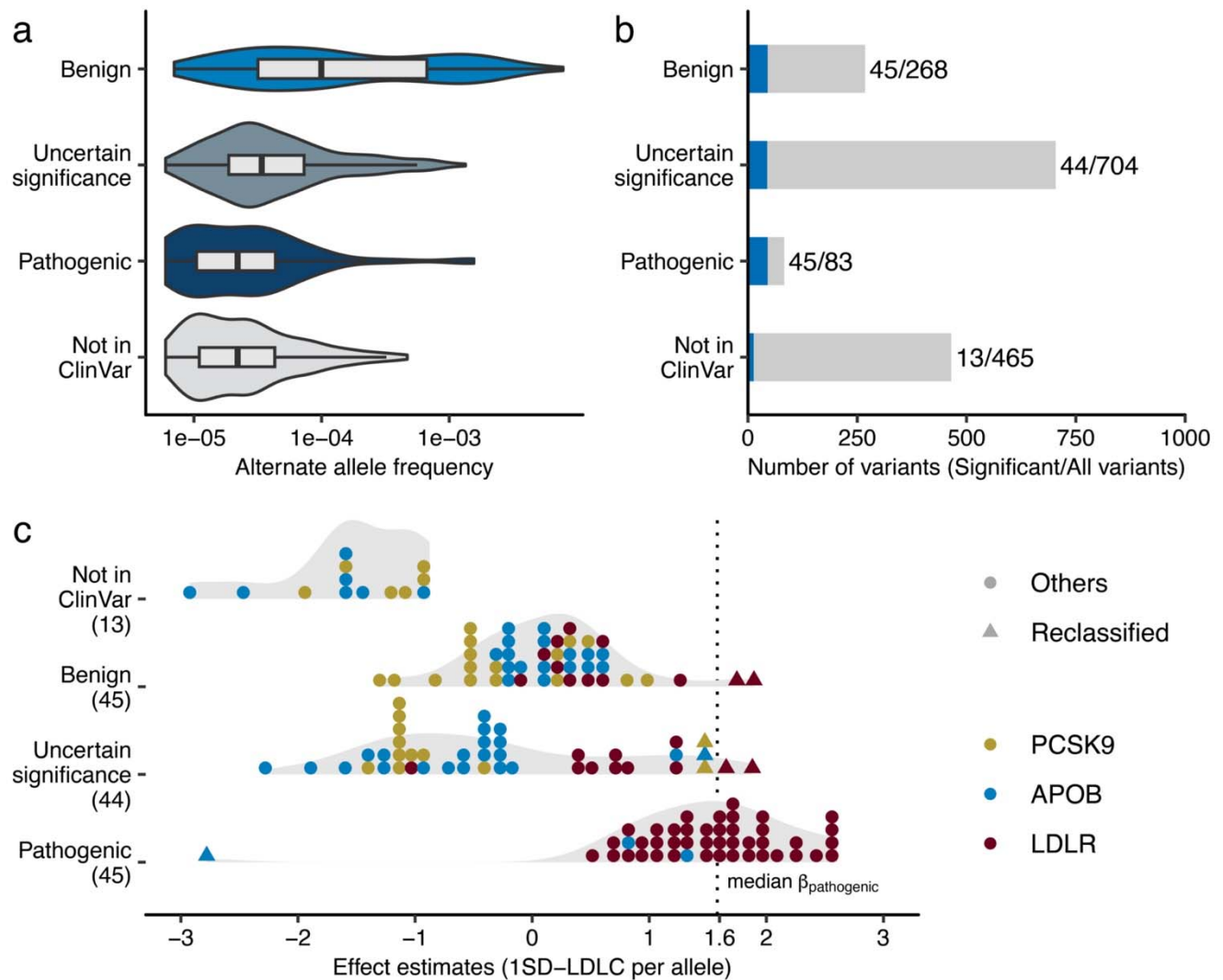


955 **Fig. 3 | Shared allelic effects across diverse populations a.** The upset plot describes the  
956 combinations of populations that observed EWS signals through intra-population meta-analysis.  
957 The bar chart at the top quantifies the number of EWS associations across various  
958 combinations of populations. Each bar represents the total number of associations observed  
959 for specific combinations of populations, as indicated by the connected points in the central  
960 matrix. The central matrix shows the population combinations involved in each set of  
961 associations, where filled squares indicate the populations included in a particular combination.  
962 In the right panel, the horizontal bar chart shows the number of associations observed within  
963 each population individually. **b.** Allele frequency comparison for non-EUR specific signals.  
964 Each point represents an EWS association that is significant only in non-EUR groups (AFR in  
965 the left panel and AMR in the right panel). The vertical axes show the minor allele frequency in  
966 EUR, while the horizontal axes show the minor allele frequency in AFR or AMR. Gray points  
967 indicate variants that were not tested in the EUR group due to low allele frequencies. **c.**  
968 Observed effect sizes across studies and populations. Each point indicates variant-trait pair  
969 with EWS. The horizontal axis shows the effect sizes in the EUR population. The vertical axes  
970 show the effect sizes in AFR and AMR/HIS populations. The error bars represent the 95%  
971 confidence interval.  $R^2$  indicates the squared Pearson's correlation coefficients of effect sizes.  
972 **d.** Consistent effect size of PCSK9 p.C679\* (stop gain) variant across multiple populations.  
973 The rectangles indicate effect sizes of PCSK9 p.C679\* on blood LDLC level in the studied  
974 population. The error bars show its 95% confidence interval. The size of rectangles is  
975 proportional to AAF. The *P*-values were calculated by linear mixed model with two-sided test.  
976 The *P*-values were not adjusted for multiple testing correction. AAF, Alternate Allele  
977 Frequency; EWS, Exome Wide Significance; AFR, African-like population; ASN, Asian-like  
978 population; AMR, Admixed-American-like population; EAS, East-Asian-like population; EUR,  
979 European-like population; HIS, Hispanic-like population; SAS, South-Asian-like population; TC,  
980 Total Cholesterol; HDLC, High density lipoprotein cholesterol; LDLC, Low Density Lipoprotein  
981 Cholesterol; TG, Triglycerides. MVP, Million Veteran Program; UKB, UK Biobank; AOU, All of  
982 Us Research Program.



983

984 **Fig. 4 | Recessive alleles associated with blood lipids** **a.** Comparison of effect sizes  
 985 between additive and recessive models. The horizontal axis displays the effect size as  
 986 estimated by linear mixed model under additive assumption, while the vertical axis shows the  
 987 effect size estimated under recessive assumption (Methods). Each dot indicates a genetic  
 988 variant, with the error bar representing the 95% confidence interval. Dashed lines represent  
 989 the predictions of recessive effect sizes based on the additive model estimates ( $y = 2x$ ) and  
 990 estimates that are twice as large ( $y = 4x$ ) as those from the additive model. **b.** Effect size from  
 991 population-wise or meta-analysis estimates for variants with the largest deviations in recessive  
 992 estimates from the predicted effect sizes based on additive model estimates. Gray dots  
 993 represent additive effect sizes, while dark blue dots correspond to recessive effect sizes  
 994 calculated by linear mixed model. Error bars indicate 95% confidence intervals. TC, Total  
 995 Cholesterol; LDLC, Low Density Lipoprotein Cholesterol; HDLC, High Density Lipoprotein  
 996 Cholesterol; TG, Triglycerides; MVP, Million Veteran Program; UKB, UK Biobank; AFR,  
 997 African-like population; AMR, Admixed-American-like population; ASN, Asian-like population;  
 998 EAS, East-Asian-like population; EUR, European-like population; HIS, Hispanic-like  
 999 population; SAS, South-Asian-like population.



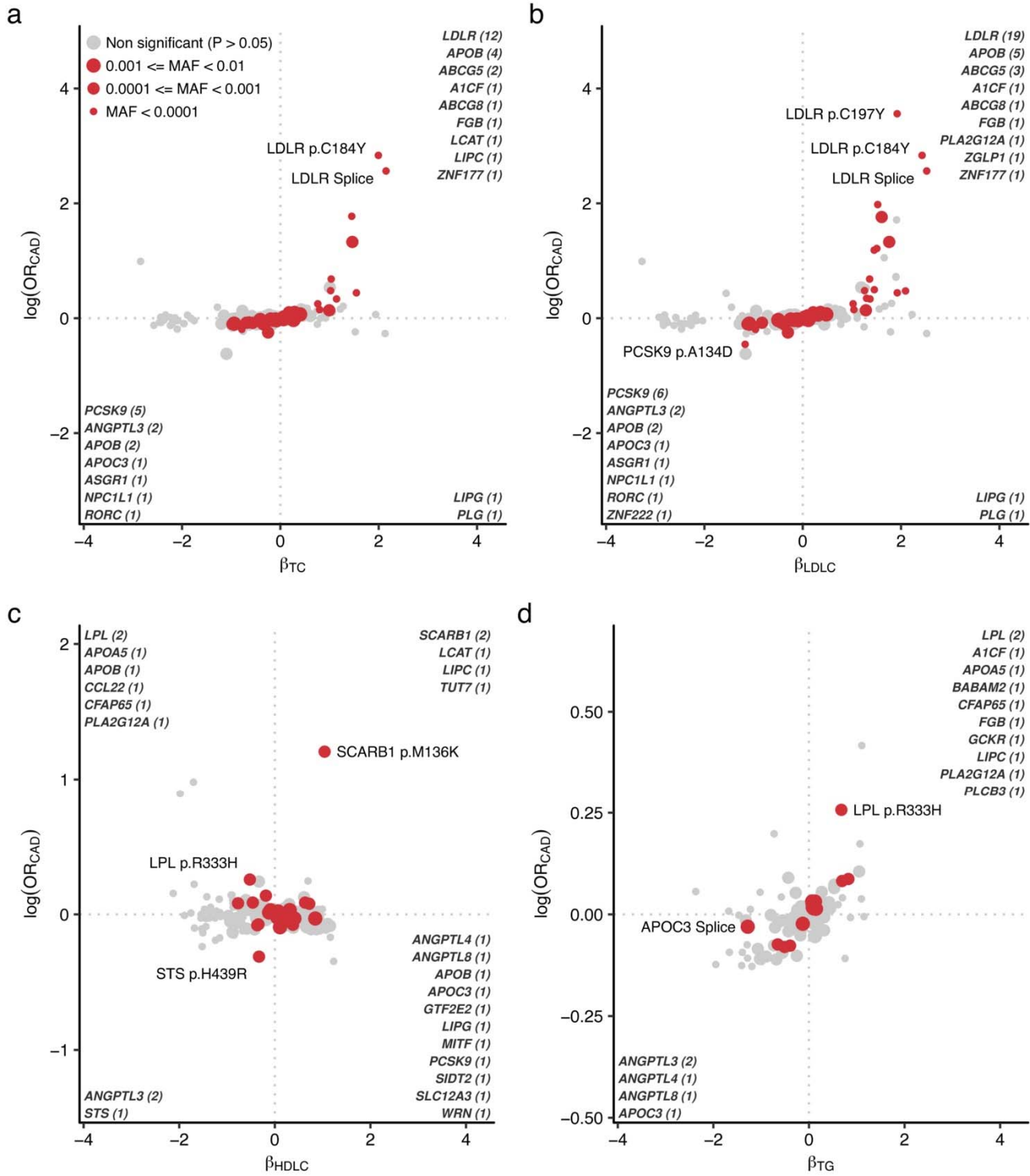
1000

1001 **Fig. 5 | Re-evaluation of clinically curated pathogenic variants for FH**

1002 **a.** Variant allele frequencies of FH-related ClinVar variants observed in the study. The rectangles illustrate the interquartile range of the minor allele frequencies, with the bottom and top edges representing the first and third quartiles, respectively. The line inside the rectangle denotes the median and the whiskers extend from the quartiles to the smallest and largest observed values, within a distance no greater than 1.5 times the interquartile range. **b.** Phenotype associations of FH-related ClinVar variants. The height of the bar indicates total number of variants in the category, and the blue color indicates the proportion of the variants significantly associated with clinical LDLC levels in this study. Statistical significance determined using Bonferroni adjustment. **c.** Distribution of the effect sizes for ClinVar FH associated variants determined in this study. Each dot represents a variant in *PCSK9*, *APOB*, or *LDLR*. The color of each dot indicates the associated gene. The dashed, vertical line indicates median effect size for established pathogenic variants. Triangles indicate variants of uncertain significance with large effect sizes, as well as pathogenic variants with a negative effect size on clinical LDLC levels. SD, Standard Deviation; LDLC, Low Density Lipoprotein Cholesterol. FH, Familial Hypercholesterolemia.



It is made available under a [CC-BY-NC 4.0 International license](https://creativecommons.org/licenses/by-nc/4.0/).



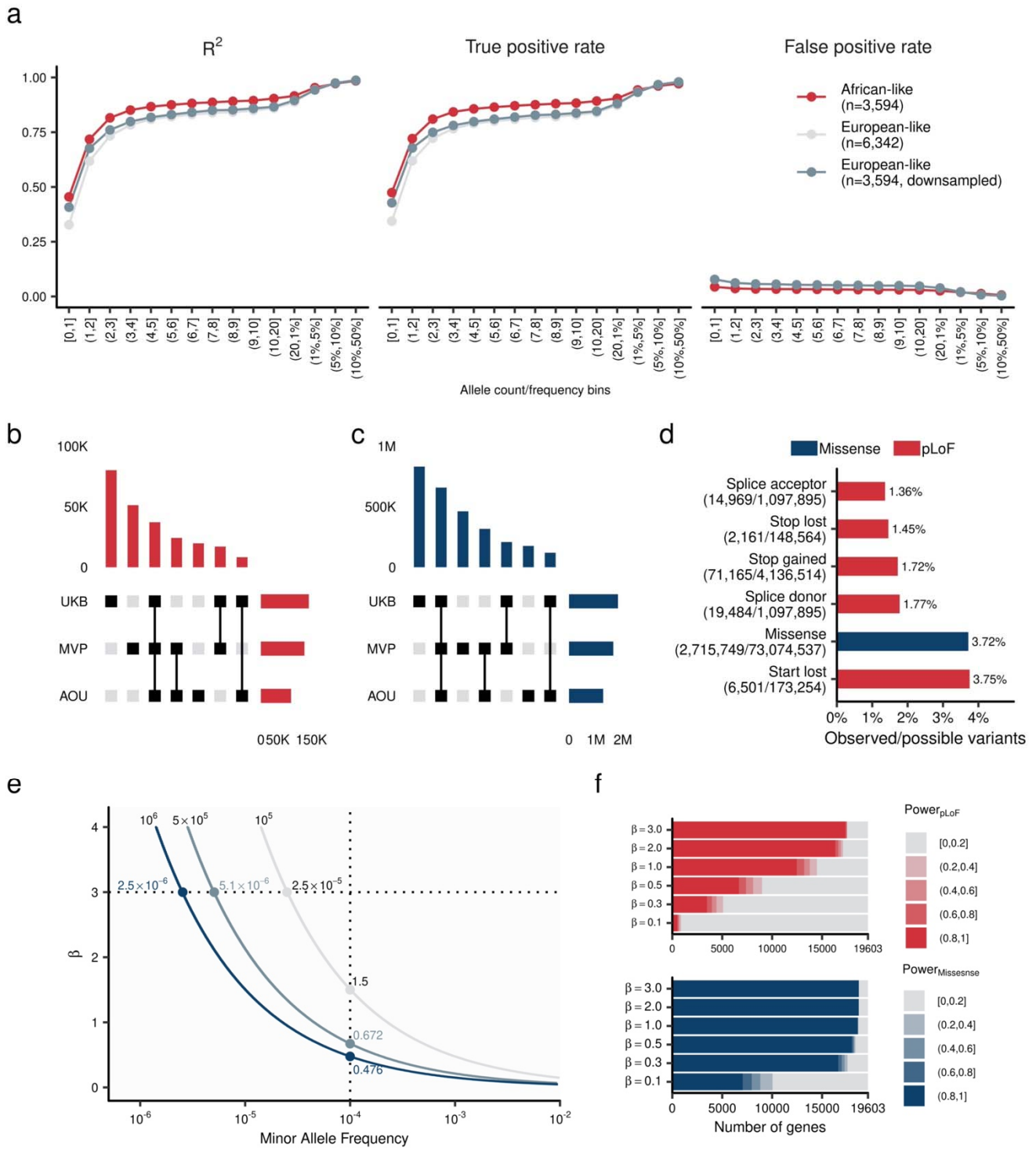
1016



1017 **Fig. 6 | CAD risks in blood lipid associated alleles** Scatter plots indicate effect size in lipids  
1018 on the horizontal axes and log odds ratio for CAD on the vertical axes. Nominally associated  
1019 ( $P < 0.05$ ) variants with CAD were highlighted in red and the sizes of the points indicating  
1020 minor allele frequency. The associated gene names are highlighted in the corner of quadrant  
1021 and the number of associations were indicated. CAD, coronary artery disease; OR, odds ratio;  
1022 MAF, minor allele frequency; TC, total cholesterol; LDLC, low density lipoprotein cholesterol;  
1023 HDLC, high density lipoprotein cholesterol; TG, triglycerides.

1024

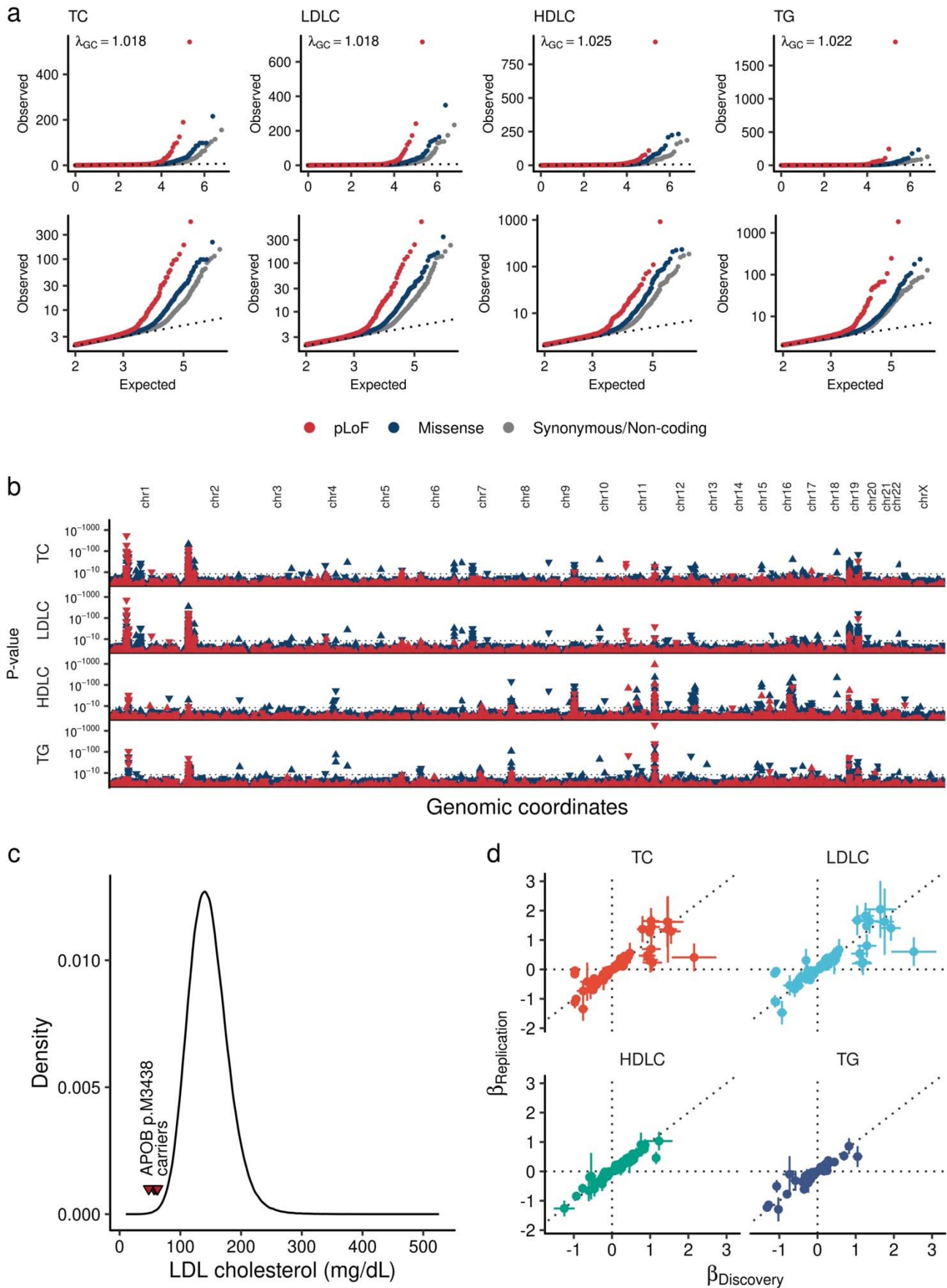
## Extended Data Figures



1025

1026 **Extended Data Figure 1 | The imputation quality, allelic diversity, coverage, and power**  
1027 **in the study a.** Imputation accuracy in MVP whole genome imputation data by TOPMed  
1028 imputation reference panel. Each dot indicates mean  $R^2$  (Squared Pearson's correlation  
1029 coefficient), TPR, and FPR by population and MAC/MAF bins. TPR and FPR were computed  
1030 by comparing dichotomized hard-called dosage (imputed data) and dichotomized sequenced  
1031 genotype (WGS data, Supplementary Notes I). **b** and **c.** Shared and unique variants across  
1032 MVP, UKB, and AOU for pLoF (b) and missense (c) variants. The central matrices define the  
1033 variant sharing status between MVP, UKB, and AOU. The top panel quantifies the variants  
1034 within the groups defined in the central matrices. The right panel summarizes the count of  
1035 variants in each study. **d.** Variant coverage. The relative proportions of SNVs identified in this  
1036 study is shown as a fraction of all possible SNVs within the target transcripts. **f.** Simulated  
1037 power curves for different sample sizes. The horizontal axis indicates minor allele frequency,  
1038 and the vertical axis indicates effect size. The dark blue line indicates 80% power curve at 1  
1039 million sample size, the intermediate curve indicates 500K sample size, and the gray curve  
1040 indicates 100K sample size, respectively. **e.** Power curve for tested genes in this study. The  
1041 curves indicate most powered pLoF/missense variants in each gene estimated by simulated  
1042 effect size ( $\beta$ ) and observed allele frequency. The color intensity corresponds with  $\beta$ . **f.** Gene  
1043 based power estimation. The color of the bar charts indicates the highest power of the coding  
1044 variant in the gene. The top panel shows pLoF variants and the bottom panel shows missense  
1045 variants.  $\beta$  indicate simulated effect size. TPR, True Positive Rate; FPR, False Positive Rate;  
1046 MAC, Minor Allele Count; MAF, Minor Allele Frequency; WGS, Whole Genome Sequence;  
1047 MVP, Million Veteran Program; UKB, UK Biobank; TOPMed, Trans-Omics for Precision  
1048 Medicine; AFR, African-like population; AMR, Admixed-American-like population; ASN, Asian-  
1049 like population; EAS, East-Asian-like population; EUR, European-like population; HIS,  
1050 Hispanic-like population; SAS, South-Asian-like population. pLoF, predicted Loss of Function.

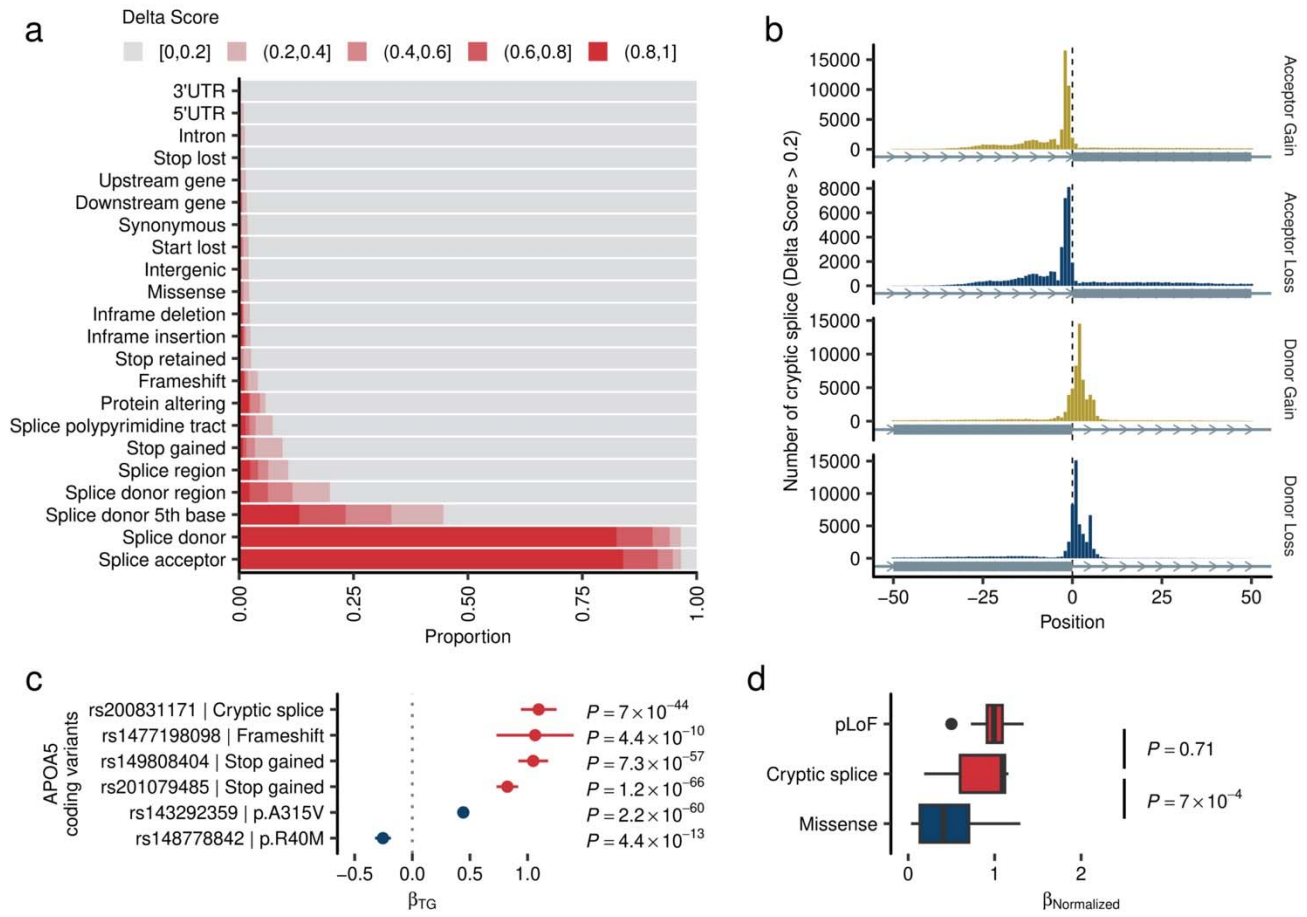
It is made available under a [CC-BY-NC 4.0 International license](https://creativecommons.org/licenses/by-nc/4.0/).



1052 **Extended Data Figure 2 | Exome wide association analysis over a million individuals a.**  
1053 Quantile-quantile plot. Upper panels are quantile-quantile plots for four tested lipid traits. Each  
1054 dot indicates a tested variant. Colors indicate variant annotation. Dotted lines show expected  
1055 distribution. Lower panel focused variants with expected  $P$ -value  $< 0.01$ . **b.** 184 exome wide  
1056 significant loci. The horizontal axis shows genomic coordinates, and the vertical axis shows  $P$ -  
1057 values. The red triangles indicate pLoF variant and blue indicate missense variant. The upward  
1058 triangles indicate trait increasing associations, and downward triangles indicate trait  
1059 decreasing associations. The  $P$ -values were calculated by linear mixed model with two-sided  
1060 test. The  $P$ -values were not adjusted for multiple testing correction. **c.** Penetrant association of  
1061 APOB p.M3438X. The curve indicates LDLC distribution of the European-like population in the  
1062 UKB (N = 409,046). The red triangles indicate LDLC level of the carriers of APOB p.M3438X. **d.**  
1063 Replication evidence in the independent study for associated variants. Each point represents  
1064 rare-coding genetic variants that are significantly associated with blood lipids in this study. The  
1065 horizontal axes display the effect sizes from this study (Discovery,  $N_{MAX} = 1,057,837$ ), while the  
1066 vertical axes present the effect sizes from the previous exome-array study (Replication,  $N_{MAX} =$   
1067  $358,251$ , Lu et al., *Nat Genet* 2017). The error bars represent the 95% confidence intervals in  
1068 each study. TC, Total Cholesterol; LDLC, Low Density Lipoprotein Cholesterol; HDLC, High  
1069 density lipoprotein cholesterol; TG, Triglycerides; GC; Genomic Control; pLoF, predicted Loss  
1070 of Function; Chr, Chromosome.



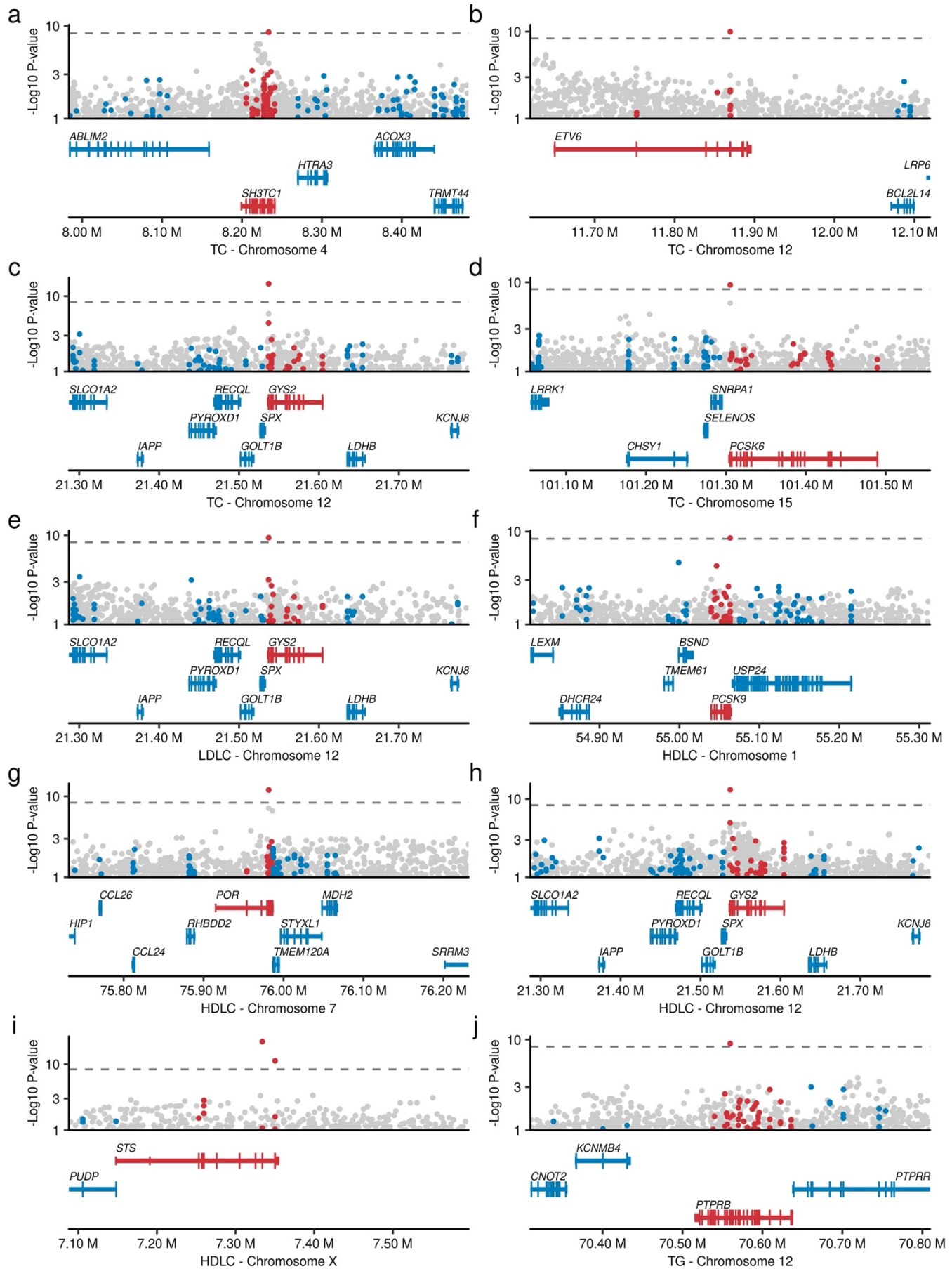
It is made available under a [CC-BY-NC 4.0 International license](https://creativecommons.org/licenses/by-nc/4.0/).



1071

1072 **Extended Data Figure 3 | Cryptic splice variants affect human blood lipids** **a.** Distribution  
1073 of cryptic splice variants across canonical variant classes. The bar graphs illustrate the  
1074 proportion of cryptic splice variants within the canonical annotations, with the colors of the bars  
1075 indicating the Delta Score (DS). **b.** Distribution of cryptic splice variants around exon-intron  
1076 boundary. The histogram shows the positions of cryptic splicing variants (DS > 0.8) in relation  
1077 to the exon-intron boundary. Exons are represented by blue rectangles. **c.** Strong expressivity  
1078 of *APOA5* cryptic splice variant. Each dot indicates effect size of variant calculated by linear  
1079 mixed model. The unit of effect size is a standard error of blood triglycerides. The error bar  
1080 indicates 95% confidence interval of effect size. Red dots indicate pLoF variants and blue dots  
1081 indicate missense variants. The *P*-values were calculated by linear mixed model with two-  
1082 sided test. The *P*-values were not adjusted for multiple testing correction. **d.** Strong  
1083 expressivity of Cryptic splice variants. The horizontal axis shows the normalized effect sizes for  
1084 pLoF, pLoF (cryptic splice) and missense variants. The analysis was restricted to the genes  
1085 both harboring pLoF, cryptic splice, and missense variants. Boxplot shows the median value  
1086 as the centerline; box boundaries show the first and third quartiles and whiskers extending 1.5  
1087 times the interquartile range. The *P*-values were calculated by the Wilcoxon rank-sum test.  
1088 The *P*-values were not adjusted for multiple testing correction. UTR, Untranslated Region;  
1089 pLoF, predicted Loss of Function; TG, Triglycerides.

It is made available under a [CC-BY-NC 4.0 International license](https://creativecommons.org/licenses/by-nc/4.0/).



1091 **Extended Data Figure 4 | Novel loci driven by rare coding variants** The horizontal axes  
1092 represent genomic coordinates, while the vertical axes denote the negative log<sub>10</sub> P-values.  
1093 Red dots illustrate the association of rare coding variants in genes with significant variants. In  
1094 contrast, blue dots show the association of rare coding variants in genes without significant  
1095 variants. Gray dots represent common variant associations from a previous study (Graham et  
1096 al., *Nature* 2021). The dashed line in the upper panel indicates the exome-wide significance  
1097 threshold ( $P < 4.4 \times 10^{-9}$ ). The lower panel illustrates the coding genes within the locus; genes  
1098 harboring significant variants are highlighted in red, and others are in blue. The *P*-values were  
1099 calculated by linear mixed model with two-sided test. The *P*-values were not adjusted for  
1100 multiple testing correction. TC, Total Cholesterol; LDLC, Low Density lipoprotein cholesterol;  
1101 HDLC, High Density Lipoprotein Cholesterol; TG, Triglycerides.

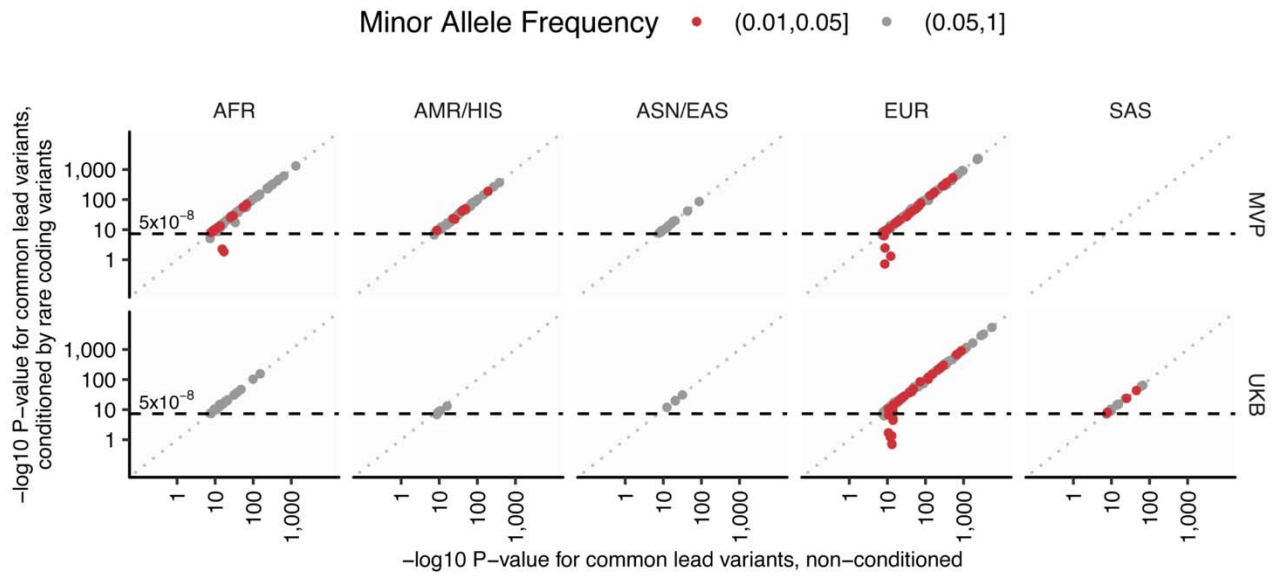


1103 **Extended Data Figure 5 | Implicated causal genes in the established lipid associated loci**

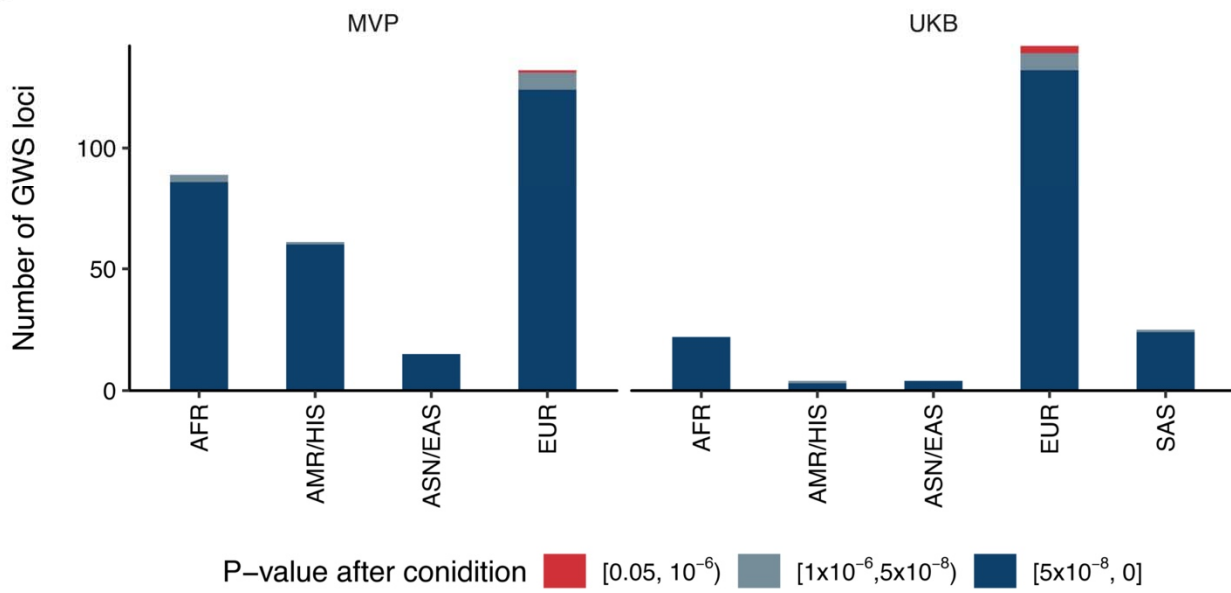
1104 The horizontal axes represent genomic coordinates, while the vertical axes denote the  
1105 negative log<sub>10</sub> P-values for *PDZK1* (a), *SREBF1* (b), *AR* (c), and *CREB3L1* (d). Red dots  
1106 illustrate the association of rare coding variants in genes with significant variants. In contrast,  
1107 blue dots show the association of rare coding variants in genes without significant variants.  
1108 Gray dots represent common variant associations from a previous study (Graham et al.,  
1109 *Nature* 2021). The dashed line in the upper panel indicates the exome-wide significance  
1110 threshold ( $P < 4.4 \times 10^{-9}$ ). The lower panel illustrates the coding genes within the locus; genes  
1111 harboring significant variants are highlighted in red, and others are in blue. The *P*-values were  
1112 calculated by linear mixed model with two-sided test. The *P*-values were not adjusted for  
1113 multiple testing correction. The LDLC, low-density lipoprotein cholesterol; HDLC, high-density  
1114 lipoprotein cholesterol; MVP, Million Veteran Program; UKB, UK Biobank; AFR, African-like  
1115 population; EUR, European-like population; HIS, Hispanic-like population.



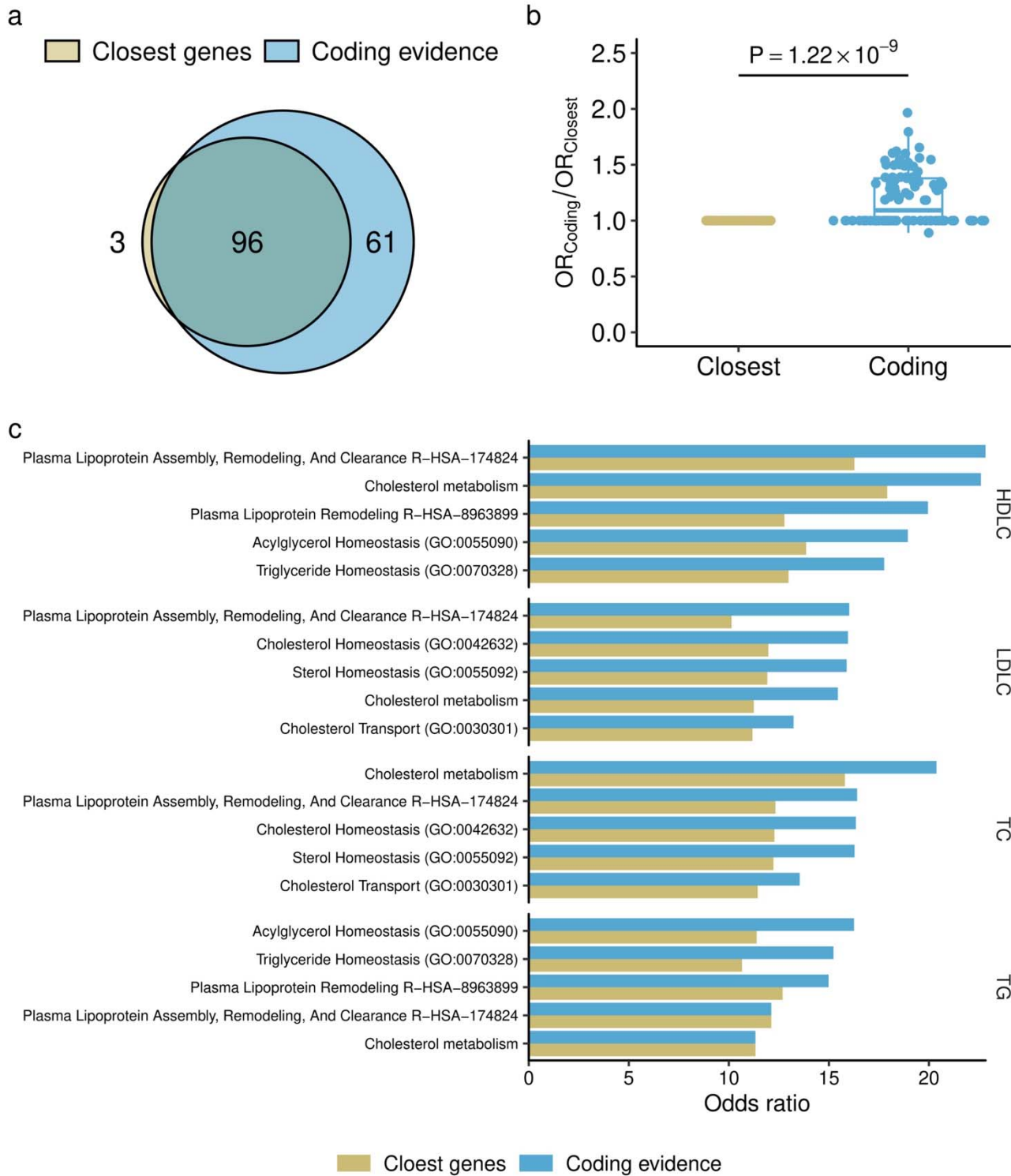
**a**



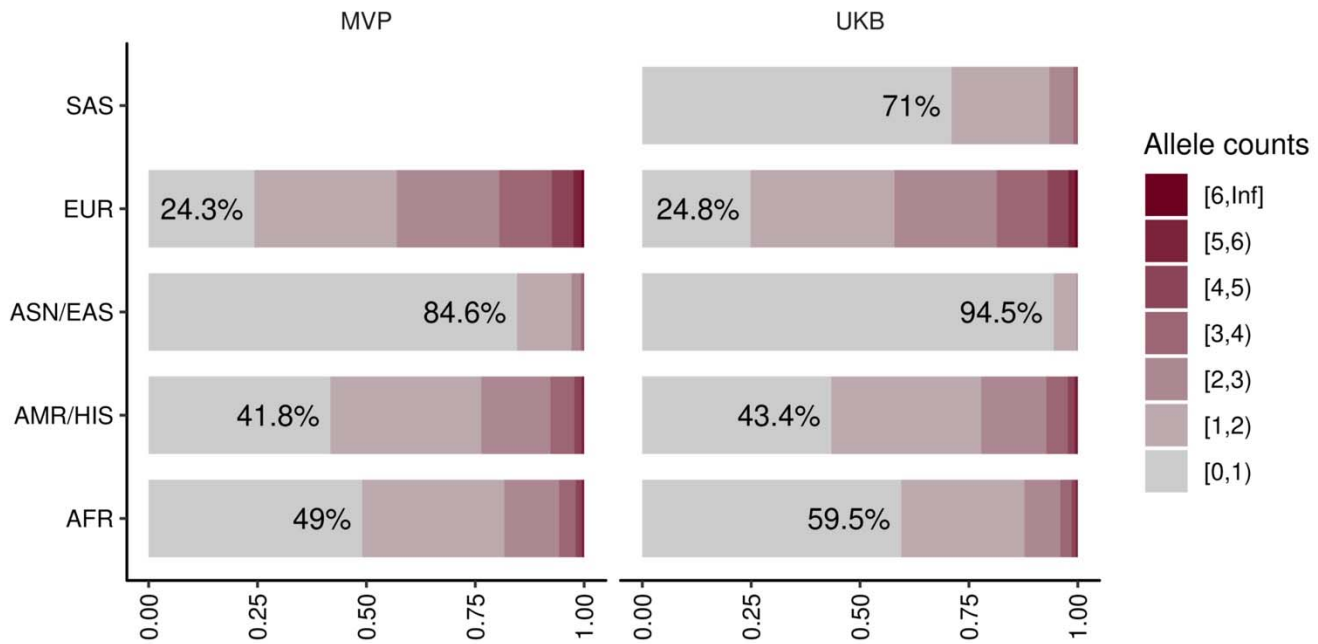
**b**



1117 **Extended Data Figure 6 | Independence of common genetic signals and rare genetic**  
1118 **signals a.** Each dot indicates common genetic variant (MAF  $\geq$  1%) associated with blood lipids  
1119 within the loci identified by rare genetic associations in this study. We compare non-  
1120 conditioned and conditioned statistics in this figure to assess the independence of common  
1121 genetic signals and rare genetic signals. In conditioned analysis, we introduced all the  
1122 associated rare variant genotypes as covariates in the linear regression model (Methods and  
1123 Supplementary Notes V). The horizontal axes show  $-\log_{10}$  P-values without conditioning and  
1124 the vertical axes show them with conditioning by rare variant genotypes with EWS. The P-  
1125 values were calculated by linear regression model with two-sided test. The P-values were not  
1126 adjusted for multiple testing correction. **b.** The number of common genetic signals affected by  
1127 rare genetic signals were summarized in the bar chart. The bar chart indicates number of  
1128 common genetic signals, and the color classifies the signals based on the P-values of common  
1129 genetic signals after conditioning by rare genetic signals. MVP, Million Veteran Program; UKB,  
1130 UK Biobank; AFR, African-like population; AMR, Admixed-American-like population; ASN,  
1131 Asian-like population; EAS, East-Asian-like population; EUR, European-like population; HIS,  
1132 Hispanic-like population; SAS, South-Asian-like population; TC, Total Cholesterol; LDLC, Low  
1133 Density Lipoprotein Cholesterol; HDLC, High Density Lipoprotein Cholesterol; TG,  
1134 Triglycerides.

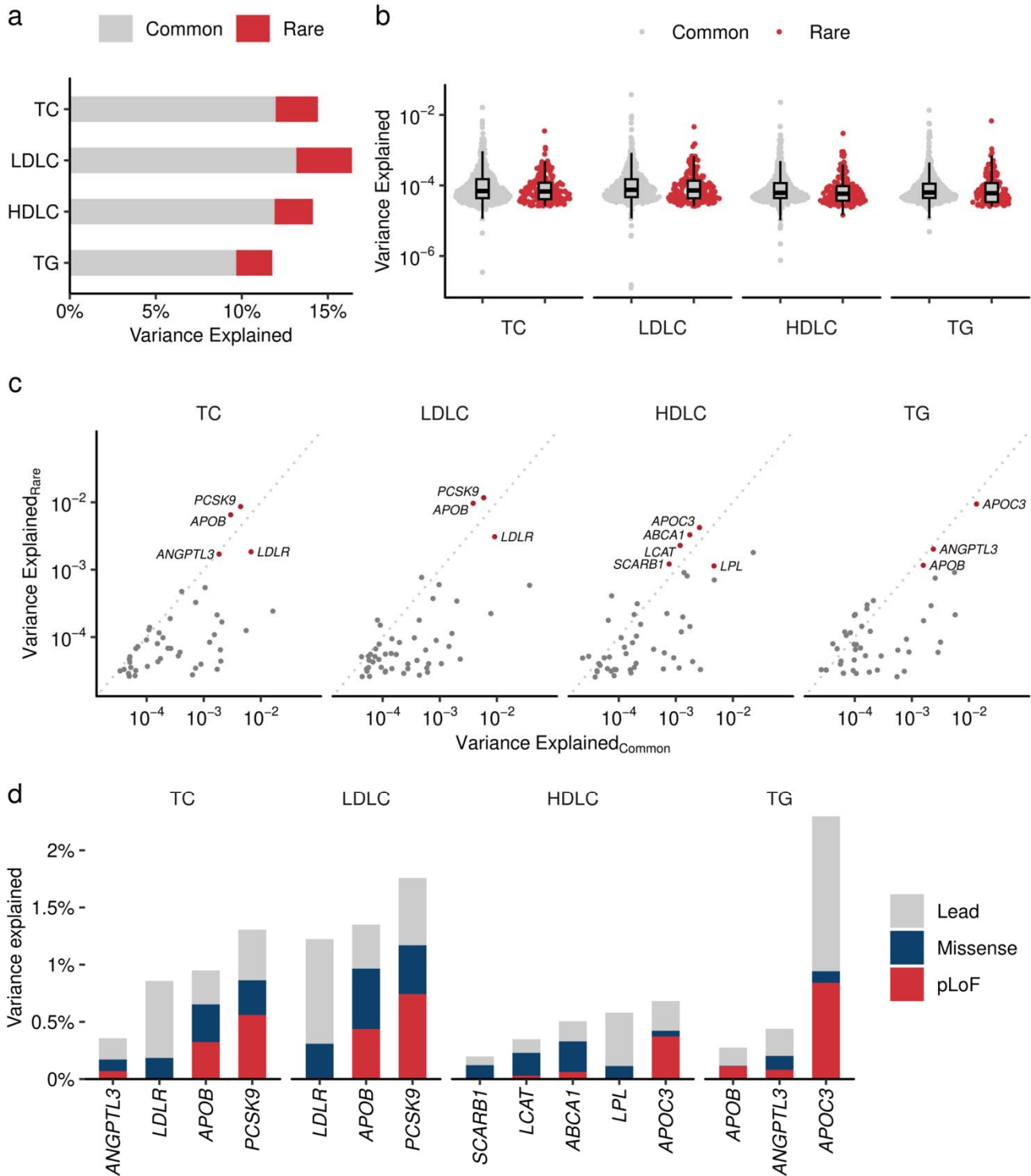


1136 **Extended Data Figure 7 | Enhanced enrichment of associated genes in the causal**  
1137 **pathway a.** Pathway enrichment by common and rare genetic signals. Venn diagram showing  
1138 significantly enriched pathways for gene sets based on common and rare variant associations.  
1139 The gene set for common variants was defined by the nearest genes to the lead common  
1140 variant (Graham et al., *Nature* 2021), while the gene set for rare variants was defined by the  
1141 genes harboring exome-wide significant (EWS) associations in this study. **b.** Pairwise  
1142 comparison of odds ratios for gene sets ( $n = 96$ ) associated with both common and rare  
1143 variants. The vertical axis shows the relative odds ratio ( $OR_{Common}/OR_{Rare}$ ). The  $P$ -value was  
1144 computed by paired Wilcoxon's rank-sum test. Boxplot shows the median value as the  
1145 centerline; box boundaries show the first and third quartiles and whiskers extending 1.5 times  
1146 the interquartile range. **c.** Pathway enrichment analysis was performed on genes harboring  
1147 rare coding variants associated with lipids and on genes closest to common variant  
1148 associations with blood lipids. The top five enriched pathways for each trait are displayed. The  
1149 horizontal axis denotes the odds ratio, with red bars indicating the odds ratios for the gene set  
1150 with rare variants and blue bars for the gene set with common variants. GO, Gene Ontology;  
1151 TC, Total Cholesterol; HDLC, High Density Lipoprotein Cholesterol; LDLC, Low Density  
1152 Lipoprotein Cholesterol; TG, Triglycerides.



1153

1154 **Extended Data Figure 8 | Limited discovery in non-European lipid associated alleles** This  
1155 figure shows the proportion of the individuals with lipid associated alleles identified in this study.  
1156 The colors of bar charts indicate allele counts of lipid associated alleles possessed by  
1157 individuals. The percentages in the bars are showing the proportion of the individuals without  
1158 lipid associated alleles in the population. MVP, Million Veteran Program; UKB, UK-Biobank;  
1159 AFR, African-like population, AMR, Admixed-American-like population, HIS, Hispanic-like  
1160 population, ASN, Asian-like population; EAS, East-Asian-like population; EUR, European-like  
1161 population; SAS, South-Asian-like population.



1162

1163 **Extended Data Figure 9 | Contribution of rare coding variants to trait variance**

1164 **a.** Phenotypic variance explained (PVE) by common and rare variants. The height of the bar  
 1165 chart indicates the PVE by GWAS lead variant (yellow) and the sum of rare coding variants in  
 1166 the locus (dark blue). PVE is computed by the formula  $2f(1-f)\beta^2$ , where  $f$  is the allele frequency  
 1167 and  $\beta$  is the effect size. **b.** PVE by individual variants. Grey dots indicate common (Graham et  
 1168 al. Nature 2021) and red dots indicate rare (current study) variants. Boxplot shows the median



1169 value as the centerline; box boundaries show the first and third quartiles and whiskers  
1170 extending 1.5 times the interquartile range. **c.** Trait variance by rare coding variant and  
1171 common genetic signals. The horizontal axis indicates PVE by lead variant in the GWAS loci.  
1172 The vertical axis indicates the sum of PVEs by rare coding variants in the locus. **d.** The  
1173 cumulative contribution of lead and rare coding variants for trait variance. PVE by each rare  
1174 variant in representative genes. Lead variant in the locus in gray, the sum of PVEs by pLoF in  
1175 red and missense in dark blue. PVE, Phenotypic Variance Explained; GWAS, Genome Wide  
1176 Association Study. TC, Total Cholesterol; High Density Lipoprotein Cholesterol; LDLC, Low  
1177 Density Lipoprotein Cholesterol; TG, Triglycerides; pLoF, predicted Loss of Function.