

# 1 Cross-ancestry genetic architecture and prediction for cholesterol traits

2

3 Md. Moksedul Momin<sup>1,2,3,4\*</sup>, Xuan Zhou<sup>1,2,4</sup>, Elina Hyppönen<sup>1,4,5</sup>, Beben Benyamin<sup>1,2,4</sup>, and  
4 S. Hong Lee<sup>1,2,4\*</sup>

5

6 <sup>1</sup>Australian Centre for Precision Health, University of South Australia, Adelaide, SA, 5000,  
7 Australia

8 <sup>2</sup>UniSA Allied Health and Human Performance, University of South Australia, Adelaide, SA,  
9 5000, Australia

10 <sup>3</sup>Department of Genetics and Animal Breeding, Faculty of Veterinary Medicine, Chattogram  
11 Veterinary and Animal Sciences University (CVASU), Khulshi, Chattogram, 4225,  
12 Bangladesh

13 <sup>4</sup>South Australian Health and Medical Research Institute (SAHMRI), University of South  
14 Australia, Adelaide, SA, 5000, Australia

15 <sup>5</sup>UniSA Clinical and Health Sciences, University of South Australia, Adelaide, SA, Australia

16

17 Correspondance : [Cvasu.Momin@gmail.com](mailto:Cvasu.Momin@gmail.com) and [Hong.Lee@unisa.edu.au](mailto:Hong.Lee@unisa.edu.au)

18

## 19 Abstract

20 While cholesterol is essential for human life, a high level of cholesterol is closely linked with  
21 the risk of cardiovascular diseases. Genome-wide association studies (GWASs) have been  
22 successful to identify genetic variants associated with cholesterol, which have been conducted  
23 mostly in white European populations. Consequently, it remains mostly unknown how genetic  
24 effects on cholesterol vary across ancestries. Here, we estimate cross-ancestry genetic  
25 correlation to address questions on how genetic effects are shared across ancestries for  
26 cholesterol. We find significant genetic heterogeneity between ancestries for total- and LDL-  
27 cholesterol. Furthermore, we show that single nucleotide polymorphisms (SNPs), which have  
28 concordant effects across ancestries for cholesterol, are more frequently found in the regulatory  
29 region, compared to the other genomic regions. Indeed, the positive genetic covariance between  
30 ancestries is mostly driven by the effects of the concordant SNPs, whereas the genetic  
31 heterogeneity is attributed to the discordant SNPs. We also show that the predictive ability of  
32 the concordant SNPs is significantly higher than the discordant SNPs in the cross-ancestry  
33 polygenic prediction. The list of concordant SNPs for cholesterol is available in GWAS  
34 Catalog (<https://www.ebi.ac.uk/gwas/>; details are in web resources section). These findings  
35 have relevance for the understanding of shared genetic architecture across ancestries,  
36 contributing to the development of clinical strategies for polygenic prediction of cholesterol in  
37 cross-ancestral settings

38

## 39 Introduction

40 Cholesterol is a type of lipid that is essential for human life, forming an essential structural  
41 component of the cell membrane<sup>1-3</sup>. While cholesterol is necessary for human body to function,  
42 too much cholesterol can harm the body. High cholesterol is linked with a high risk of  
43 cardiovascular diseases (CVDs), such as coronary heart disease, stroke, and peripheral vascular

44 **NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.**  
disease, which are the leading cause of death worldwide<sup>4</sup>, accounting for 32% of all deaths in

45 2019<sup>5</sup>. Specifically, elevated low-density lipoprotein (LDL) and decreased high-density  
46 lipoprotein (HDL) cholesterol are associated with increased CVD risk<sup>6-9</sup>. These cholesterol  
47 traits are heritable and known to be polygenic<sup>6, 10, 11</sup>. Reported heritability estimates for total-,  
48 LDL- and HDL-cholesterols are typically in the range of 20 to 60%<sup>12</sup>.

49

50 Over the last two decades, genome-wide association studies (GWASs) have successfully  
51 identified several genome-wide significant single nucleotide polymorphisms (SNPs) associated  
52 with cholesterol traits<sup>4, 13-15</sup>. While these findings have provided important insights into the  
53 genetics of cholesterol, most GWAS for cholesterol to date have been conducted in populations  
54 of white European ancestry<sup>16-18</sup>. Although the number of GWASs representing non-European  
55 populations are gradually increasing, they still remain greatly underrepresented in the efforts  
56 of gene discovery<sup>16, 19</sup>. Consequently, how genetic effects on cholesterol vary across ancestries  
57 remain mostly unknown<sup>20, 21</sup>. It is also not clear to what extent the associated genetic variants  
58 discovered in European populations are relevant for other ancestries (e.g., South Asian and  
59 African ancestries), and if the polygenic risk prediction of cholesterol can be applied across  
60 ancestries<sup>22-25</sup>.

61

62 The genetic effects on most complex traits are likely to vary at least to some extent across  
63 different ancestry groups<sup>26, 27</sup>. Cross-ancestry genetic correlation analyses can dissect the  
64 shared genetic architecture between diverse ancestries, also allowing to leverage power from  
65 diverse sources of information<sup>28</sup>. While common causal variants for cholesterol are likely to be  
66 shared across ancestries, their per-allele effect sizes may depend on allele frequencies that can  
67 differ across ancestries due to different evolutionary force such as selection and genetic drift<sup>29</sup>.  
68 Moreover, each ancestry has a unique genetic background that may affect the magnitude and  
69 direction of per-allele effect sizes for complex traits such as cholesterol<sup>30</sup>. It has been reported  
70 that the relationship between allele frequency and per-allele effect size varies across different  
71 ancestries, which should be properly accounted for. otherwise, the estimation of cross-ancestry  
72 genetic correlation can be biased<sup>31, 32</sup>.

73

74 Cross-ancestry genetic prediction can reduce the potential health disparity for non-European  
75 populations that are still underrepresented in public genomic databases including GWAS and  
76 polygenic risk scores (PRS)<sup>33</sup>. It is crucial to understand the source of genetic heterogeneity  
77 across ancestries in the genetic prediction. In general, it is not likely that SNP effects estimated  
78 from a single ancestry group are always applicable to other ancestries, which has a practical

79 relevance. For example, several studies have reported that the predictive ability of complex  
80 traits including cholesterol was poor for Africans, East-Asians, South-Asians and Latinos,  
81 when using SNP effects estimated in Europeans<sup>19, 34, 35</sup>. To obtain more reliable cross-ancestry  
82 genetic prediction, it may be important to restrict to functionally homogenous genes or  
83 common causal variants across ancestries<sup>28, 36</sup>. We hypothesize that SNPs in strong linkage  
84 disequilibrium (LD) with the functionally homogenous genes have concordant effects, i.e., the  
85 same direction of SNP effects, across ancestries.

86

87 In this study, we estimate cross-ancestry genetic correlation to address the question about how  
88 genetic effects are shared across ancestries for cholesterol traits, accounting for the relationship  
89 between allele frequency and per-allele effect size<sup>31</sup>. In the estimation of cross-ancestry genetic  
90 correlation, we also investigate the role of concordant SNPs that are derived from comparing  
91 SNP effects between two independent GWAS datasets of UK Biobank and Biobank Japan  
92 (BBJ). We evaluate the transferability of genetic prediction across different ancestry groups  
93 and suggest a list of SNPs that are suitable for the use in polygenic risk prediction in cross-  
94 ancestry analyses.

95

96

## 97 **Results**

### 98 **Overview of methods**

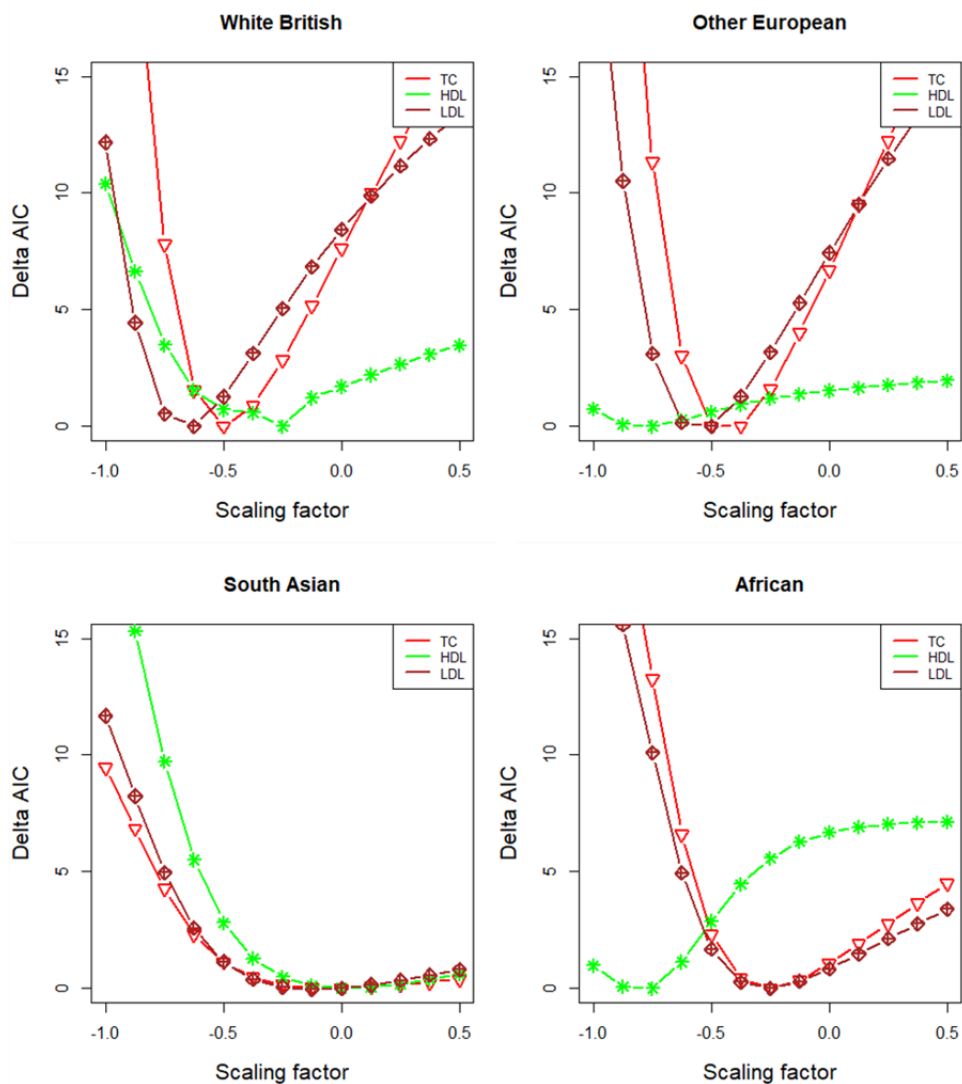
99 The total numbers of individuals and SNPs for each ancestry after stringent quality control (QC)  
100 (Methods) are shown in **Supplementary Table 1**. From the quality-controlled data of 288,837  
101 white British people, we randomly selected 30,000 individuals to be used in the analyses of  
102 cross-ancestry genetic correlations. The remaining 258,792 individuals were used as the  
103 discovery dataset in the cross-ancestry genetic risk prediction and in the classification of  
104 concordant SNPs (referred to as UKBB discovery). In the cross-ancestry genetic analysis of  
105 total-, HDL- and LDL-cholesterol, four ancestry groups were included, i.e. the 30,000 white  
106 British ancestry group, 26,457 other European, 6,199 south Asian and 6,179 African ancestry  
107 groups (**Supplementary Table 1**). We accounted for the relationship between per-allele effect  
108 size and allele frequency<sup>31, 37</sup> by using trait-specific and ancestry-specific  $\alpha$  that was explicitly  
109 estimated for each trait and each ancestry, using Akaike information criterion (AIC)<sup>31, 32</sup>. We  
110 used the common SNPs for each pair of ancestries to estimate the cross-ancestry genetic  
111 correlation, using the bivariate GREML approach<sup>38</sup>, accounting for the relationship between  
112 allele frequency and per-allele effect size<sup>31</sup>. We further investigated if the set of concordant

113 SNPs, which were derived by comparing UKBB and Biobank Japan discovery GWAS  
114 summary statistics for cholesterol, is enriched in the regulatory region, compared to the other  
115 genomic regions. The list of concordant SNPs for total-, HDL- and LDL-cholesterol are now  
116 available in GWAS catalogue. Cross-ancestry genetic covariance was partitioned, based on the  
117 sets of concordant and discordant SNPs, to see how the genetic heterogeneity is attributed to  
118 those SNP sets (see **Supplementary Table 4**). Finally, cross-ancestry polygenic prediction was  
119 performed based on the sets of concordant and discordant SNPs.

120

#### 121 **Determining trait-specific and ancestry-specific scale factor ( $\alpha$ ) for each ancestry**

122 The scale factor ( $\alpha$ ) can account for the relationship between allele frequency and per-allele  
123 effect size, that is, per-allele effect sizes vary, proportional to  $[p(1-p)]^\alpha$ , where  $p$  is the allele  
124 frequency<sup>32, 39, 40</sup>. It is also reported that the scale factor is not uniformly distributed across  
125 ancestries, and there may be an optimal  $\alpha$  value for each specific ancestry group<sup>31</sup>. Following  
126 the previous approach<sup>31, 32</sup>, we investigated various  $\alpha$  values ranging between -1 and 0.5 to  
127 determine the ancestry specific  $\alpha$  value of each ancestry group for total-, LDL- and HDL-  
128 cholesterol. To determine optimal  $\alpha$ , we compared the Akaike Information Criteria (AIC)  
129 values across different heritability models with various  $\alpha$  values for each trait and each ancestry  
130 (**Figure 1**). Detailed values of log-likelihood and AIC are provided in **Supplementary Table**  
131 **6-9**. As expected, optimal  $\alpha$  values are not uniformly distributed across traits and across  
132 ancestries (Figure 1). These identified  $\alpha$  values are subsequently used in the estimation of cross-  
133 ancestry genetic correlations to dissect the shared genetic architecture and investigate genetic  
134 heterogeneity across ancestries for the cholesterol related traits.



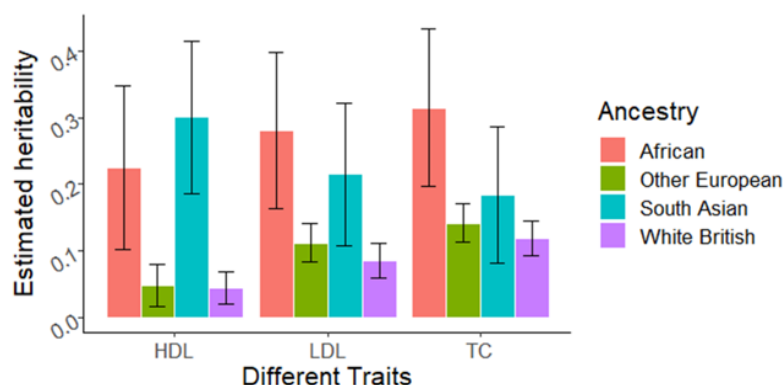
135  
 136 **Figure 1: Determining the optimal ancestry-specific scaling factors ( $\alpha$ ) for each trait.** The  
 137  $\alpha$  value reflects the relationship between allele frequency and per-allele effect size and can vary  
 138 across ancestries and traits.  $\Delta AIC$  values are plotted against scaling factors,  $\alpha$ , for each ancestry  
 139 group. The lowest AIC (i.e.,  $\Delta AIC=0$ ) indicates the best model. The sample sizes are 30,000,  
 140 26,457, 6,199, and 6,179 for white British, other European, South Asian, and African ancestry  
 141 groups, respectively. TC: total-cholesterol, HDL: high-density lipoprotein cholesterol, LDL:  
 142 low-density lipoprotein cholesterol.  
 143

#### 144 Heritability ( $h^2$ ) estimates across ancestries

145 The estimated SNP-based heritabilities of total-, LDL- and HDL-cholesterol are presented in  
 146 Figure 1. The estimates are significantly different particularly between European and African  
 147 ancestries. For total-cholesterol, there is a significant difference in SNP-based heritability  
 148 estimates between African vs. European ( $p$ -value=4.26e-03), and African vs white British ( $p$ -  
 149 value=1.14e-03). Similarly, the estimate of LDL-cholesterol is significantly lower in white  
 150 British ( $p$ -value= 1.11e-03) and other European ( $p$ -value= 5.19e-03) than African ancestry,  
 151 which agrees with the previous findings based on twin studies<sup>41</sup>. We also observed significant

152 heterogeneity of SNP-based heritability for the HDL-cholesterol between South Asian and  
153 other Europeans, between South Asian and white British.

154



155

156 **Figure 2: Estimated SNP-based heritability across ancestries for cholesterol traits.** The  
157 main bars indicate SNP-based heritability estimates, and the error bars indicate 95% confidence  
158 intervals. TC= Total-cholesterol, HDL= high-density lipoprotein cholesterol, LDL= low-  
159 density lipoprotein cholesterol.

160

### 161 **Estimated cross-ancestry genetic correlations**

162 The estimated cross-ancestry genetic correlations ( $r_g$ ) for cholesterol traits are presented in

163 **Figure 3.** For total-cholesterol, we observed a genetic heterogeneity between South Asian vs.

164 white British ( $r_g = 0.399$ ; SE= 0.143;  $p$ -value= 2.65e-05), South Asian vs. other European ( $r_g =$

165 0.353; SE=0.133;  $p$ -value= 1.14e-06) and South Asian vs. African ancestry ( $r_g = 0.188$ ;

166 SE=0.197;  $p$ -value= 3.76e-05). There is also a genetic heterogeneity between African vs. white

167 British ( $r_g = 0.473$ ; SE=0.127;  $p$ -value= 3.33e-05) and African vs. other European ancestry ( $r_g =$

168 0.315; SE=0.122;  $p$ -value=1.96e-08). In contrast, white British and other European are

169 genetically homogenous ( $r_g = 0.954$ ; SE=0.087;  $p$ -value= 5.96e-01) (**Figure 3 and**

170 **Supplementary Table 10**). For LDL-cholesterol, results are similar to total-cholesterol. There

171 is a significant genetic heterogeneity between South Asian vs. white British ( $r_g = 0.296$ ;

172 SE=0.155;  $p$ -value=5.57e-06), South Asian vs. other European ( $r_g = 0.177$ ; SE=0.138;  $p$ -

173 value=2.46e-09), South Asian vs. African ( $r_g = 0.110$ ; SE=0.190;  $p$ -value=2.81e-06), and

174 African vs. other European ancestry ( $r_g = 0.409$ ; SE=0.147;  $p$ -value=2.81e-06) (**Figure 3 and**

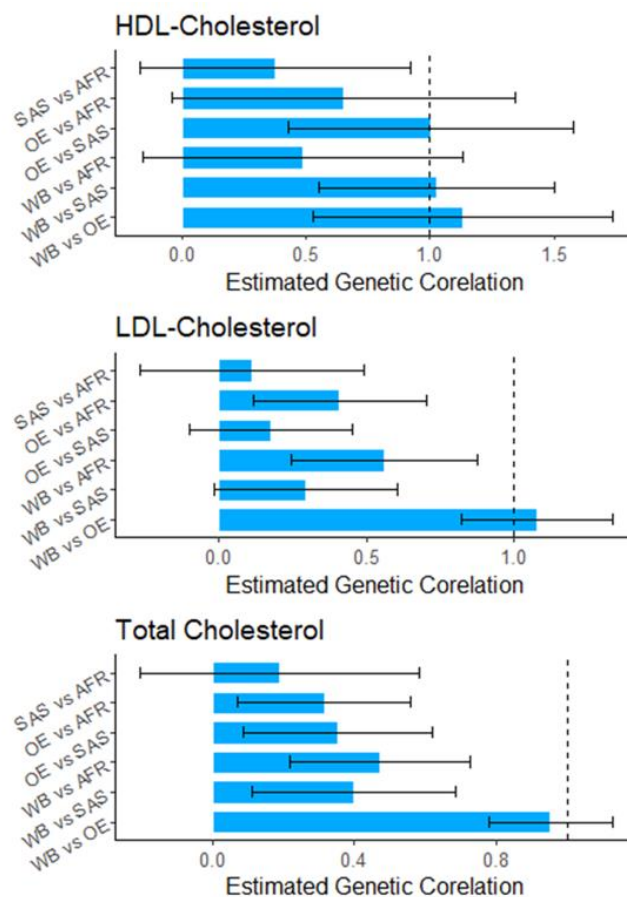
175 **Supplementary Table 11**). As expected, the cross-ancestry genetic correlation between other

176 European and white British was close to 1 ( $r_g = 1.084$ ; SE=0.128;  $p$ -value=5.12e-01). We did

177 not observe genetic heterogeneity among the pairs of ancestry groups for HDL-cholesterol

178 (**Figure 3 and Supplementary Table 12**).

179



180

181 **Figure 3: Estimated cross-ancestry genetic correlations.** The main bars indicate estimated  
182 cross-ancestry genetic correlations, and the error bars indicate 95% confidence intervals of the  
183 estimates. WB = White British, OE = Other European, SAS = South Asian, AFR = African.

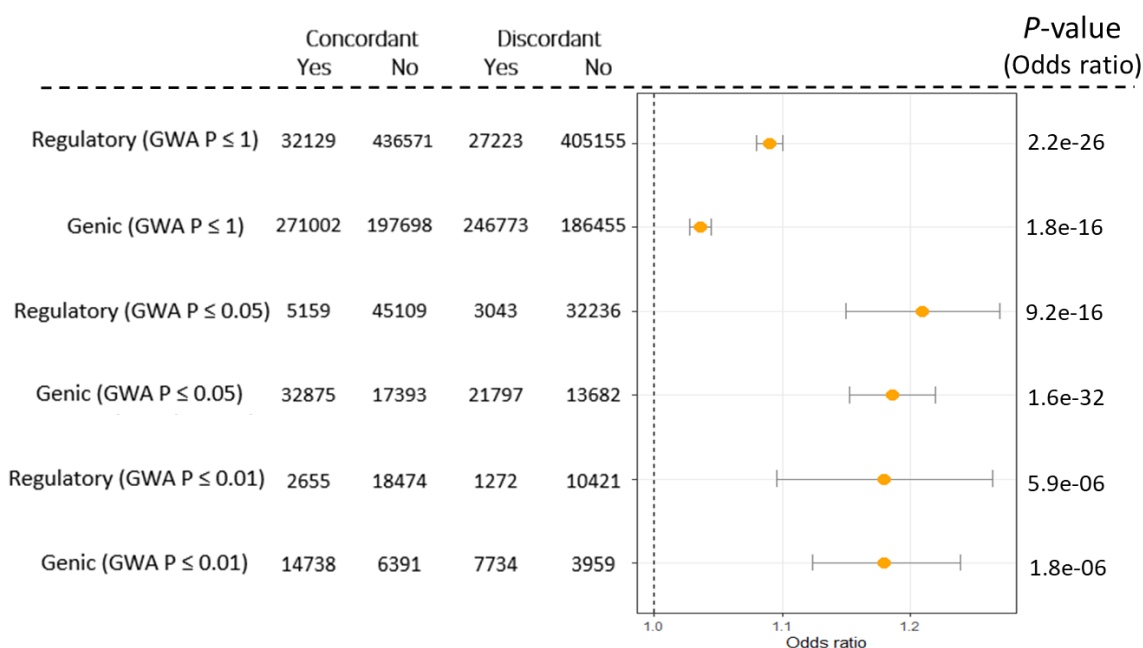
184

### 185 **Genomic partitioning of cross-ancestry genetic covariance using concordant and** 186 **discordant SNPs between two diverse ancestries**

187 Some genes are functionally homogeneous across ancestries while the other genes may not be<sup>36,</sup>  
188 <sup>42, 43</sup>. It can be hypothesised that the functionally homogenous genes are enriched in the  
189 regulatory regions, and they contribute more to phenotypic variation within and between  
190 ancestries, compared to the other genes. We obtained a set of concordant SNPs (a proxy of  
191 functionally homogenous genes) for total-, HDL-, and LDL-cholesterols, by comparing the  
192 direction of SNP effects between two diverse ancestries, using the GWAS summary statistics  
193 of UK Biobank and Biobank Japan. For the UK Biobank GWAS, we used 258,792 white  
194 British individuals who are not overlapping with anyone in the 4 ancestry groups used in our  
195 study (white British, other European, South Asian, and African). For the Biobank Japan, we  
196 used GWAS summary statistics that are publicly available. In this concordance/discordance  
197 analysis, we considered the same HapMap3 SNPs used in the genetic correlation analyses

198 above. The numbers of concordant and discordant SNPs for each pair of ancestries are  
 199 presented in **Supplementary Table 4**.

200  
 201 First, we quantified if the concordance SNPs are more frequently found in the regulatory or  
 202 genic region, compared to the other genomic regions for total-cholesterol. **Figure 4** shows that  
 203 the number of concordant SNPs in the regulatory region is significantly higher than the non-  
 204 regulatory region (OR= 1.09,  $p$ -value=2.2e-26 for  $p$ -value  $\leq 1$ ; OR= 1.21,  $p$ -value=9.2e-16 for  
 205  $p$ -value  $\leq 0.05$ ; OR= 1.18,  $p$ -value=1.6e-06 for  $p$ -value  $\leq 0.01$ ). When selecting SNPs with a  
 206 genome-wide association (GWA)  $p$ -value  $> 0.05$  or 0.01, the odds ratio increases (**Figure 4**).  
 207 Similarly, the number of concordant SNPs in the genic region is significantly higher than the  
 208 non-genic region (OR= 1.03,  $p$ -value=1.8e-16 for  $p$ -value  $\leq 1$ ; OR= 1.17,  $p$ -value=1.6e-32 for  
 209  $p$ -value  $\leq 0.05$ ; OR= 1.18,  $p$ -value=1.8e-06 for  $p$ -value  $\leq 0.01$ ) (**Figure 4**). Similar results were  
 210 observed when using the HDL- and LDL-cholesterol traits (**Supplementary Figure 1 and 2**).



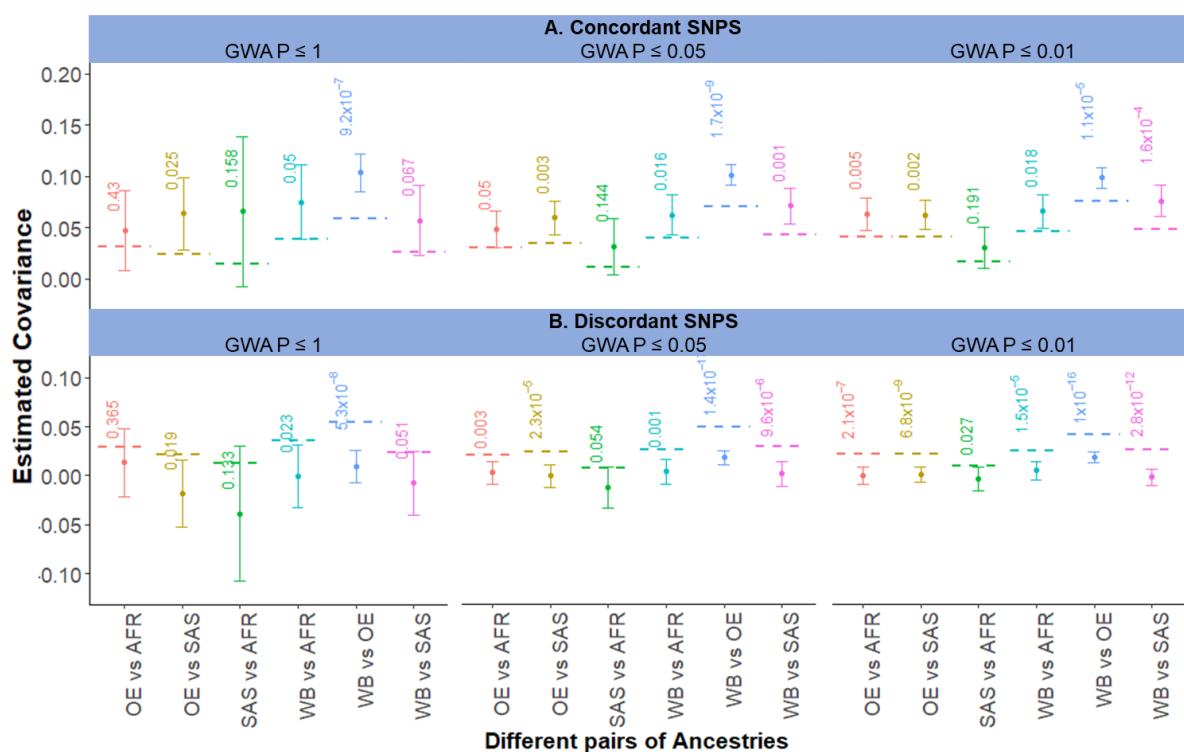
211  
 212 **Figure 4: A forest plot with odds ratios indicating that concordant SNPs are more**  
 213 **frequently found in the regulatory or genic region.** This analysis is for total-cholesterol  
 214 phenotypes. Error bar represents 95% confidence intervals. The  $p$ -value of odds ratio indicates  
 215 that the odds ratio is significantly different from 1. For regulatory or genic region, a genome-  
 216 wide association (GWA)  $p$ -value threshold  $\leq 1$ , 0.05 or 0.01 was used to select a set of  
 217 concordant and discordant SNPs using UK Biobank GWAS summary statistics for total-  
 218 cholesterol.  
 219

220 Subsequently, we partitioned genetic covariance components attributed to the two sets of  
 221 genomic regions (concordant vs. discordant SNPs). We estimated two genomic relationship  
 222 matrixes (GRM), using the sets of concordant and discordant SNPs, which were simultaneously



223 fitted in a bivariate multiple random-effects model. When considering the set of concordant  
 224 SNPs, the estimated genetic covariances between other European (OE) vs. south Asian (SAS),  
 225 white British (WB) vs. African (AFR) and WB vs. OE were significantly higher than the  
 226 expectation (the proportion of the concordant SNPs) for total-cholesterol (**Figure 5**). On the  
 227 other hand, the estimated genetic covariances for these pairs of ancestries were significantly  
 228 lower than the expectation when using discordant SNPs (**Figure 5**). For HDL-cholesterol  
 229 (**Supplementary Figure 3**) and LDL-cholesterol (**Supplementary Figure 4**), a similar result  
 230 was observed that the estimated genetic covariances between OE vs. SAS, WB vs. OE and WB  
 231 vs. SAS were significantly deviated from the expectation. When using SNPs with genome-  
 232 wide association  $p$ -values  $< 0.05$  or  $0.01$  (**Supplementary Table 4**), the estimated genetic  
 233 covariances due to concordant and discordant SNPs were more significantly deviated from the  
 234 expectation in general (**Figure 5, Supplementary Figure 3 and 4**). It is also noted that the  
 235 estimated genetic covariances for the set of discordant SNPs were not higher than zero (Figure  
 236 5), implying that the genetic heterogeneity of cholesterol traits across ancestry might be mostly  
 237 due to the set of discordant SNPs. This also shows that the set of concordant SNPs may be  
 238 useful in cross-ancestry polygenic risk predictions. The results are similar when genome-wide  
 239 association  $p$ -values from BBJ are used (**Supplementary Figure 5**).

240  
 241



242  
 243 **Figure 5: Estimated genetic covariances for concordant and discordant SNPs for total-**  
 244 **cholesterol.** Concordant and discordant SNPs were derived from the comparison of SNP

245 effects between two independent GWAS datasets of UK Biobank and BBJ. In this concordant  
246 or discordant analysis, a set of SNPs with genome-wide association (GWA)  $p$ -values  $< 1, 0.05$   
247 or  $0.01$  was used, where the GWA  $p$ -values were from UK Biobank GWAS for total-  
248 cholesterol. The main bars represent estimated cross-ancestry genetic covariance using the set  
249 of genome-wide SNPs, and the error bars indicate 95% confidence intervals. The horizontal  
250 dashed line indicates the expected genetic covariance, assuming all SNPs contribute equally to  
251 the genetic covariance, i.e., the expected genetic covariance = the estimated total genetic  
252 covariance  $\times$  the proportion of number of concordant SNPs, where the estimated total genetic  
253 covariance is based on all the SNPs including both concordant and discordant SNPs. The value  
254 with each bar indicates a  $p$ -value testing the null hypothesis that the estimated genetic  
255 covariance is not significantly different from the expectation. WB = White British, OE = Other  
256 European, SAS = South Asian, AFR = African.  
257

258 We further investigated the impact of concordant SNPs in a cross-ancestry polygenic risk  
259 prediction. We used the UKBB discovery dataset, which is independent from the four target  
260 datasets including white British, other European, South Asian, and African ancestries, to  
261 estimate SNP effects and obtain GWAS summary statistics for cholesterol traits. Using the  
262 GWAS summary statistics, we constructed polygenic risk scores for the individuals in the  
263 target datasets. The predictive ability ( $R^2$ ) of polygenic risk scores for total-cholesterol is  
264 significantly higher when using the set of concordant SNPs than when using the set of  
265 discordant SNPs for both within- and cross-ancestry predictions (**Figure 6, Supplementary**  
266 **Figure 4**) ( $p$ -values for the difference between concordant and discordant PRS SNPs are  $3.8e-$   
267  $33, 2.2e-25, 1.3e-04$  and  $5.3e-04$  for white British, other European, South Asian and African,  
268 respectively). Although not significant,  $R^2$  is slightly higher when using the set of concordant  
269 SNPs, compared to when using the total set of SNPs (**Figure 6**), suggesting that including  
270 discordant SNPs may have adverse effects on the cross-ancestry risk predictions. When  
271 accounting for the proportion of concordant SNPs, a similar result was observed in that  
272 concordant SNPs performed better than discordant SNPs in within- and cross-ancestry risk  
273 predictions (**Supplementary Figure 5**). A similar finding was observed when using BBJ  
274 discovery GWAS summary statistics, i.e., the cross-ancestry prediction accuracy of the  
275 concordant SNPs significantly higher than the discordant SNPs (**Supplementary Figure 4**).  
276 Interestingly, the concordant SNPs performs notably better than the total set of SNPs when  
277 predicting white British, other European and south Asian ancestries (**Supplementary Figure**  
278 **4**). Results are invariant when considering LDL- and HDL-cholesterol (Supplementary Figures  
279 6-7).  
280

281 For HDL cholesterol, it is notable that the accuracy of cross-ancestry prediction can be higher  
 282 than within-ancestry prediction (e.g., South Asian vs. White British in Supplementary Figure  
 283 6). We further confirmed this result with a clump-and-threshold (C + T) based PRS method  
 284 (PRSice)<sup>44</sup> and compared the significance of difference (Supplementary Figure 8). It shows  
 285 that PGS generated from White British GWAS provides a significantly higher predictive  
 286 accuracy for South Asian (p-value = 6.67e-16) and African ancestry groups (p-value = 7.35e-  
 287 04), compared to White British. This may have an important implication in genomic medicine  
 288 for underrepresented non-European populations.  
 289  
 290



291  
 292 **Figure 6: The predictive ability ( $R^2$ ) of polygenic risk scores for total-cholesterol when**  
 293 **using the set of concordant, discordant, or total SNPs for cross-ancestry risk predictions.**  
 294 UK Biobank GWAS was used as the discovery dataset (n= 258,792), while target datasets were  
 295 other European (n=26,457), south Asian (n=6,199) and African ancestry (n=6,179).

296 **Left panels:** The main bars represent  $R^2$  values and error bars correspond to 95% confidence  
297 interval.

298 **Right panels:** Dot points represent the differences between  $R^2$  values, error bars correspond  
299 to 95% confidence intervals of the differences, and  $p$ -values indicate that the differences of  $R^2$   
300 are significantly different from zero (null hypothesis).  $P$ -values was estimated using an R-  
301 package (r2redux)<sup>45</sup> based on Wald's test statistics.  
302

### 303 **Discussion**

304 Cholesterol is an essential structural component of the cell membrane, which is necessary for  
305 the body to function<sup>1,2</sup>. However, the risk of CVD is associated with a high level of cholesterol  
306 that can be determined by genetic risk factors<sup>4,46,47</sup>. Although the genetic study of cholesterol  
307 has been conducted, it is not clear how genetic effects on cholesterol vary across different  
308 ancestries. In this study, we explicitly estimated cross-ancestry genetic correlations to  
309 investigate the shared genetic architecture across ancestries for cholesterol. Importantly, we  
310 appropriately accounted for the relationship between allele frequency and per-allele effect size  
311 by modelling the ancestry-specific scale factor for cholesterol, which can provide more reliable  
312 estimates<sup>31</sup>.

313  
314 The reliable estimation of cross-ancestry genetic correlation allows us to understand the shared  
315 genetic architecture across ancestries, providing crucial information when for various  
316 downstream analyses of complex traits such as cross-ancestry GWAS and cross-ancestry  
317 polygenic risk score prediction. Moreover, this may inform best practices for cross-ancestry  
318 meta-analysis, multi-ancestry disease mapping, and the transferability of epidemiological  
319 findings. Our analysis shows that in general, total- and LDL-cholesterol are both genetically  
320 heterogeneous across ancestries, whereas HDL-cholesterol is not<sup>48</sup>. This finding has important  
321 implications for the power of cross-ancestry GWASs and cross-ancestry polygenic risk score  
322 prediction, which for HDL-cholesterol may be much higher than that for total- and LDL-  
323 cholesterols (**Supplementary Figure 6**).

324  
325 To identify genetic variants that contribute to the genetic heterogeneity, we investigated  
326 concordant and discordant SNP sets that were identified by comparing the direction of SNP  
327 effects between UK Biobank and Biobank Japan GWAS summary statistics, noting that the  
328 two datasets are independent from the four target ancestry groups used in this study. The  
329 concordant SNPs may be associated with genes that are functionally homogeneous across  
330 ancestries<sup>49</sup>, and we show in this study that the concordant SNPs are more often located in the

331 regulatory or genic regions, compared to other genomic regions. We also show that such strong  
332 genetic heterogeneity across ancestries for cholesterol can be attributed to the discordant SNPs,  
333 but not to the concordant SNPs. We provide evidence that the set of concordant SNPs can be  
334 useful in the cross-ancestry polygenic risk predictions, which may improve the transferability  
335 of polygenic risk scores to clinical practice<sup>16, 50, 51</sup>.

336

337 There are a number of limitations in this study. For determining optimal  $\alpha$ , we did not consider  
338 the relationship between LD and per-allele effect sizes, i.e., as in LDK-thin model<sup>32</sup> that  
339 requires a substantial reduction of the number of SNPs. We also acknowledge that the  
340 conclusions from cross-ancestry analyses (cross-ancestry correlation and genomic prediction)  
341 in this study are restricted to common variants ( $MAF \geq 0.01$ ) and HapMap3 SNPs only; as  
342 these are robust and reliable for dissecting cross-ancestry genetic architecture<sup>52, 53</sup>. A moderate  
343 sample size (limited power of the data) was used to estimate optimal scale factors ( $\alpha$ ) for south  
344 Asian and African populations. Therefore, the genetic heterogeneity needs to be explored with  
345 larger sample size. The concordant SNPs were identified by comparing the direction of SNP  
346 effects between white British (UKBB) and East Asian (BBJ) populations, because adequate  
347 data was not available from other ancestries. When public genomic databases have sufficient  
348 resources across ancestries, we can have a finer set of concordant SNPs by comparing SNP  
349 effects across various ancestries.

350

351 In conclusion, there is a significant genetic heterogeneity between ancestries for total- and  
352 LDL-cholesterol, which is mostly driven by the set of discordant SNPs. Interestingly, the  
353 concordant SNPs are more frequently found in the regulatory region as annotated by an  
354 independent study<sup>54</sup>, and restricting to concordant SNPs can provide better accuracy for cross-  
355 ancestry polygenic prediction for cholesterol. Our findings contribute to knowledge about the  
356 genetic architecture of cholesterol that is shared across ancestries. The proposed cross-ancestry  
357 polygenic prediction can be potentially useful in clinical practice. Our analysis protocol can be  
358 extended to a wide range of other complex traits and diseases.

359

360

## 361 **Methods**

### 362 **Ethical statement**

363 We used publicly available from the UK Biobank (<https://www.ukbiobank.ac.uk/>). Science  
364 protocol and operational procedures for the UK Biobank have been reviewed and approved by

365 the North-West Multi-Centre Research Ethics Committee (MREC), National Information  
366 Governance Board for Health & Social Care (NIGB), and Community Health Index Advisory  
367 Group (CHIAG). The UK Biobank has obtained consent from all participants. The access of  
368 the UK Biobank data was approved under the reference number 14575 (“Whole genome  
369 approaches for dissecting (shared) genetic architecture and individual risk prediction of  
370 complex traits in human populations”). Publicly available GWAS summary statistics of  
371 Biobank Japan (BBJ) were used, following BBJ’s guidelines (<http://jenger.riken.jp/en/>). The  
372 research ethics approval of this study has been obtained from the University of South Australia  
373 Human Research Ethics Committee.

374

375

### 376 **Participants and stratification of ancestries**

377 Data from the UK Biobank contains 501,748 participants recruited between 2006 and 2010<sup>55</sup>.  
378 The participants were recruited from 22 assessment centres in England, Wales, and Scotland,  
379 ranging in age from 37 to 73 years old<sup>56</sup>. All the phenotypic data for cholesterol traits under  
380 this study are derived from baseline survey. Principal component analysis was applied to the  
381 UK Biobank individuals to stratify participants<sup>57</sup> into four different ancestries following  
382 previous approach<sup>31</sup>.

383

384

### 385 **Genotypic data and quality control**

386 We used the second release of the UK biobank (<https://www.ukbiobank.ac.uk/>) genotype data  
387 comprising 488,377 individuals and 92,693,895 imputed autosomal SNPs. The individuals  
388 were genotyped by Affymetrix UK BiLEVE Axiom array and Affymetrix UK Biobank  
389 Axiom® array. Combination of UK10K and Haplotype Reference Consortium (HRC) data  
390 were considered as the reference dataset for the imputation of the UK Biobank genotypic  
391 dataset<sup>58</sup>. In this analysis, to dissect the genetic architecture of disease and complex traits, we  
392 retained only HapMap3 SNPs in this analysis<sup>52</sup>, which are also considered robust and reliable  
393 for estimating heritability, genetic correlation<sup>52, 59</sup>. Stringent quality control (QC) procedure  
394 was applied to each ancestry to select high quality individuals and high-quality SNPs. SNPs  
395 QC criteria include, SNPs excluded with an INFO score (used to indicate the quality of  
396 genotype imputation) <0.6<sup>60-62</sup>, call rate <0.95, a MAF <0.01 and a Hardy–Weinberg  
397 equilibrium  $p$ -value <10<sup>-4</sup>. We also exclude population outliers (individuals outside  $\pm 6SD$ ) and  
398 related individuals (--rel-cutoff 0.05) using PLINK<sup>63</sup>.

399

400 Individual level QC criteria include samples with genotype missing rate  $>0.05$ , gender  
401 mismatch (reported gender does not fit with the genetically assigned sex determined from gene  
402 data), poor genotype quality or a sex chromosome aneuploidy was excluded from the main  
403 analyses. For the ease of computation, we reduced the number individuals in white British  
404 ancestry. The total number of individuals and total number of SNPs after QC shown in  
405 **Supplementary Table 1**. The number of common SNPs across different pairs of ancestries  
406 presented in **Supplementary Table 2** and the number of common SNP for each genomic  
407 region (genomic partitioning) between ancestries presented in **Supplementary Table 3**.

408

#### 409 **Functional annotation of the genome**

410 The common SNPs between populations were partitioned into genomic region using genomic  
411 annotation reported by Gusev et. al.<sup>54</sup>, where they partitioned the genome into coding, UTR,  
412 promoter, intron, DHS and intergenic regions. For the genomic partitioning analysis, we  
413 include promoter, coding, UTR, and DHS regions as regulatory regions<sup>64-66</sup>, and introns (an  
414 integral part of a gene)<sup>67,68</sup> and the intergenic regions as non-genic regions. We also partitioned  
415 the whole genome into two predefined functional categories as genic (includes SNPs from  
416 promoter, coding, untranslated, intron and DHS region) and non-genic regions (intergenic  
417 region).

418

#### 419 **Concordant and discordant SNP annotation**

420 To identify concordant and discordant SNPs we compared SNP effects between two  
421 independent GWAS datasets of white British from UK Biobank and Biobank Japan (BBJ). The  
422 BBJ summary statistics data are publicly available (<http://jenger.riken.jp/en/result>). We  
423 excluded SNPs that were ambiguous or had a strand issue. After excluding these SNPs, there  
424 were 4,113,630 SNPs that are common between UKBB and BBJ. To determine concordant and  
425 discordant SNPs, we compared the direction of SNP effects between white British from UKBB  
426 and BBJ. We used only HapMap3 SNPs from 4,113,630 SNPs for concordant and discordant  
427 analysis across different ancestry pairs (**Supplementary Table 4**).

428

429 There were four possible combinations of direction of SNP effects (beta):

430 (+beta, +beta) if the SNP effects are positive in both GWAS.

431 (+beta, -beta) if the SNP effects are positive and negative in the UKBB and BBJ GWAS.

432 (-beta, +beta) if the SNP effects are negative and positive in the UKBB and BBJ GWAS.

433 (-beta, -beta) if the SNP effects are negative in both GWAS.  
434 Each SNP should be in one of four possible combinations and belongs to either concordance  
435 or discordance. SNPs belonged to  $((+\text{beta}, +\text{beta}) \cup (-\text{beta}, -\text{beta}))$  were considered concordant,  
436 otherwise discordant, i.e.  $((+\text{beta}, -\text{beta}) \cup (-\text{beta}, +\text{beta}))$ .

437

438

## 439 **Data analysis**

### 440 **Phenotypic adjustment of main traits**

441 Prior to model fitting, all cholesterol traits were adjusted for demographic variables, the UK  
442 biobank assessment centre (as factor), genotype measurement batch (as factor) and population  
443 structure measured by the first 10 principal components (PCs)<sup>64, 69</sup> using linear models in *R-*  
444 *software* (4.0.3). Demographic variable includes sex, birth year, education, and Townsend  
445 deprivation index (**Supplementary Table 5**). Information of educational qualifications  
446 converted to education levels (years) for all the UK Biobank individuals<sup>70</sup>.

447

### 448 **Determining scale factors for GCTA- $\alpha$ model**

449 GCTA model assumes all the SNPs has equal contribution to the genetic variance (has no LD  
450 weights), whereas LDAK-thin model<sup>32</sup> explicitly considers LD among SNPs. The previously  
451 recommended and widely used  $\alpha$  are -0.50 and -0.125 for GCTA model<sup>71</sup> and LDAK-thin  
452 model<sup>32</sup>, respectively. Here we have used 13 different values of  $\alpha$  (between -1 and 0.5)  
453 following GCTA model (termed as GCTA-  $\alpha$  model)<sup>31</sup>. In order to perform a cross-ancestry  
454 genetic correlation analysis of cholesterol traits (total cholesterol, HDL cholesterol, and LDL  
455 cholesterol), we determined and used optimal  $\alpha$  based on GCTA models for each trait and  
456 ancestry. We did not consider another widely used LDAK-thin model as it will reduce number  
457 of common SNPs between ancestry due to LD-pruning.

458

## 459 **Statistical models**

### 460 **Univariate Linear Mixed Model**

461 The univariate Linear Mixed Model (LMM) for can be written as,

$$462 \mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{g} + \mathbf{e} \quad (1)$$

463 Where  $\mathbf{y}$  is the vector of phenotypic observation,  $\mathbf{b}$  is the vector of fixed effects,  $\mathbf{g}$  is the vector  
464 of additive genetic value and  $\mathbf{e}$  is the vector of the residuals. The random effects ( $\mathbf{g}$  and  $\mathbf{e}$  are  
465 presumed to be distributed normally with mean zero where  $\mathbf{X}$  and  $\mathbf{Z}$  are incidence matrices



466

467 Heritability was estimated using the genetic and residual variances obtained from the univariate  
468 LMM, which can be expressed as

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2} \quad (2)$$

469 Here,  $\sigma_g^2$  is the genetic variance and  $\sigma_e^2$  is residual variance. Estimation assumed  
470 environmental homogeneity

471

### 472 **Bivariate Linear Mixed Model**

473 The bivariate Linear Mixed Model (LMM) was used to estimate heritability and cross-ancestry  
474 genetic correlation using individual level genetic data written as,

$$475 \mathbf{y}_1 = \mathbf{X}_1 \mathbf{b}_1 + \mathbf{Z}_1 \mathbf{g}_1 + \mathbf{e}_1 \quad \text{for ancestry 1 (3)}$$

$$476 \mathbf{y}_2 = \mathbf{X}_2 \mathbf{b}_2 + \mathbf{Z}_2 \mathbf{g}_2 + \mathbf{e}_2 \quad \text{for ancestry 2 (4)}$$

477 Where  $\mathbf{y}_1$  and  $\mathbf{y}_2$  are vector of phenotypic observation,  $\mathbf{b}_1$  and  $\mathbf{b}_2$  are the vector of fixed  
478 effects,  $\mathbf{g}_1$  and  $\mathbf{g}_2$  are vector of additive genetic value and  $\mathbf{e}_1$  and  $\mathbf{e}_2$  are the vector of  
479 residuals. The random effects ( $\mathbf{g}_1, \mathbf{g}_2$  and  $\mathbf{e}_1, \mathbf{e}_2$ ) are presumed to be distributed normally with  
480 mean zero where  $\mathbf{X}$  and  $\mathbf{Z}$  are incidence matrices i.e. i.e.  $\mathbf{g}_i \sim N(0, \mathbf{A}\sigma_{gi}^2)$  and  $\mathbf{e}_i \sim N(0, \mathbf{I}\sigma_{ei}^2)$ .

481

482 The variance covariance matrix of observed phenotypes can be written as

$$483 \mathbf{V} = \begin{bmatrix} \mathbf{Z}_1 \mathbf{A} \sigma_{g_1}^2 \mathbf{Z}_1' + \mathbf{I} \sigma_{e_1}^2 & \mathbf{Z}_1 \mathbf{A} \sigma_{g_{12}}^2 \mathbf{Z}_2' \\ \mathbf{Z}_2 \mathbf{A} \sigma_{g_{21}}^2 \mathbf{Z}_1' & \mathbf{Z}_2 \mathbf{A} \sigma_{g_2}^2 \mathbf{Z}_2' + \mathbf{I} \sigma_{e_2}^2 \end{bmatrix} \quad (5)$$

484 where,  $\mathbf{A}$  is the genomic relationship matrix (GRM)<sup>72-74</sup>, which can be estimated based on the  
485 genome-wide SNP information, and  $\mathbf{I}$  is an identity matrix which implicitly assumes across  
486 individuals of environmental effects and measurement error. The terms,  $\sigma_{g_1}^2$  ( $\sigma_{g_2}^2$ ) and  
487  $\sigma_{e_1}^2$  ( $\sigma_{e_2}^2$ ) indicate the genetic and residual variance of the trait for the two-ancestry group, and  
488  $\sigma_{g_{12}}^2$  ( $\sigma_{g_{21}}^2$ ) is the genetic covariances between the two ancestry groups. It is noted that there is  
489 no parameter to model residual correlation in  $\mathbf{V}$  because there are no multiple phenotypic  
490 measures for any individual, i.e., the phenotypes of the first (second) trait are available only  
491 for the first (second) ancestry group.

492

493 Cross-ancestry genetic correlation between two random genetic effects can be computed either  
494 directly as genetic covariance standardized by the square root of the product of the genetic

495 variances of the two random genetic effects (equation 6) or indirectly by the correlation  
496 coefficient of SNP effect sizes<sup>38, 75</sup>.

$$r_{\mathbf{g}_i \mathbf{g}_j} = \frac{\sigma_{\mathbf{g}_i \mathbf{g}_j}}{\sqrt{\sigma_{\mathbf{g}_i}^2 \cdot \sigma_{\mathbf{g}_j}^2}} \quad (6)$$

497

#### 498 **GREML analysis to estimate heritability and cross-ancestry genetic correlations**

499 Bivariate GREML is the cornerstone method to estimate SNP heritability and cross-ancestry  
500 genetic correlation using common SNPs across ancestries. The SNPs frequency, heritability  
501 model (relationship between heritability and MAF), and the scale factor ( $\alpha$ ) varied across  
502 ancestries<sup>31</sup>. We used a recently proposed approach of estimating GRM<sup>31</sup> in combined  
503 population, that accounts ancestry specific  $\alpha$  and ancestry specific allele frequencies for  
504 estimating heritability and cross-ancestry genetic correlation. Both estimation of GRM and  
505 GREML analysis was implemented in *mtg2*<sup>76</sup>.

506

#### 507 **Genomic prediction**

508 The polygenic score (PGS) is obtained from by aggregating and quantifying single nucleotide  
509 polymorphism (SNP) effects. PGS of an individual ( $k$ ) can be defined as cumulative effect of  
510 SNP counts with a standard equation as:

$$PGS = \sum_{j=1}^m \beta_j x_{jk}$$

512 Here,  $\beta_j$  is the SNP effect from discovery GWAS,  $m$  is the total number of SNPs included in  
513 the predictor,  $x_{jk}$  is the number of copies (0,1, or 2) of trait associated SNP  $j$  in the genotype  
514 of individual  $k$ .

515

516

#### 517 **Web resources and code availability**

518 The genotype and phenotype data of the UK Biobank can be accessed through procedures  
519 described on its webpage (<https://www.ukbiobank.ac.uk/>) and summary statistics of BMI and  
520 total-, LDL- and HDL-cholesterol from Biobank Japan (BBJ) can be obtained from its website  
521 (<http://jenger.riken.jp/en/result>)

522 MTG2, <https://sites.google.com/site/honglee0707/mtg2>

523 PLINK2 version can be downloaded from <https://www.cog-genomics.org/plink/>

524 *r2redux* R-package (<https://github.com/mommy003/r2redux> from GitHub or from CRAN)

525 The GWAS summary statistics dataset that is generated in this current study and supports the  
526 findings have been deposited in the NHGRI-EBI GWAS catalogue with the accession codes  
527 GCST90244051, GCST90244052, GCST90244053, GCST90244054, GCST90244055 and  
528 GCST90244056; (<https://www.ebi.ac.uk/gwas/>). GWAS for all SNPs and concordant SNPs for  
529 total-, HDL- and LDL-cholesterol can be accessed in following links

530 GWAS of total cholesterol (all SNP)  
531 ([https://ftp.ebi.ac.uk/pub/databases/gwas/summary\\_statistics/GCST90244001-  
532 GCST90245000/GCST90244051/](https://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST90244001-GCST90245000/GCST90244051/))

533 GWAS of total cholesterol (concordant SNP)  
534 ([https://ftp.ebi.ac.uk/pub/databases/gwas/summary\\_statistics/GCST90244001-  
535 GCST90245000/GCST90244052/](https://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST90244001-GCST90245000/GCST90244052/))

536 GWAS of HDL-cholesterol (all SNP)  
537 ([https://ftp.ebi.ac.uk/pub/databases/gwas/summary\\_statistics/GCST90244001-  
538 GCST90245000/GCST90244053/](https://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST90244001-GCST90245000/GCST90244053/))

539 GWAS for HDL-cholesterol (concordant SNP)  
540 ([https://ftp.ebi.ac.uk/pub/databases/gwas/summary\\_statistics/GCST90244001-  
541 GCST90245000/GCST90244054/](https://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST90244001-GCST90245000/GCST90244054/))

542 GWAS for LDL-cholesterol (all SNP)  
543 ([https://ftp.ebi.ac.uk/pub/databases/gwas/summary\\_statistics/GCST90244001-  
544 GCST90245000/GCST90244055/](https://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST90244001-GCST90245000/GCST90244055/))

545 GWAS for LDL-cholesterol (concordant SNP)  
546 ([https://ftp.ebi.ac.uk/pub/databases/gwas/summary\\_statistics/GCST90244001-  
547 GCST90245000/GCST90244056/](https://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST90244001-GCST90245000/GCST90244056/))

548

549

## 550 **Acknowledgements**

551 This research is supported by the Australian Research Council (DP190100766). We thank the  
552 staff and participants of the UK Biobank and Biobank Japan for their important contributions.  
553 Our reference number approved by UK Biobank is 14575. The analyses were performed using  
554 computational resources provided by the Australian Government through Gadi under the  
555 National Computational Merit Allocation Scheme (NCMAS), and HPCs (Statgen server)  
556 managed by UniSA IT.

557

## 558 **Declaration of interest**

559 The authors declare that they do not have any competing interests.

560

561

562

563

## 564 **References**

- 565 1. Ding, X., et al., *The role of cholesterol metabolism in cancer*. American journal of cancer  
566 research, 2019. **9**(2): p. 219.
- 567 2. Yan, S., et al., *Bufalin enhances TRAIL-induced apoptosis by redistributing death receptors in*  
568 *lipid rafts in breast cancer cells*. Anti-cancer drugs, 2014. **25**(6): p. 683-689.
- 569 3. Craig, M., S.N.S. Yarrarapu, and M. Dimri, *Biochemistry, cholesterol*. 2018.
- 570 4. Musunuru, K. and S. Kathiresan, *Genetics of common, complex coronary artery disease*. Cell,  
571 2019. **177**(1): p. 132-145.
- 572 5. WHO, *World Health Organization; Cardiovascular diseases (CVDs)*. 2021.
- 573 6. Trinder, M., G.A. Francis, and L.R. Brunham, *Association of monogenic vs polygenic*  
574 *hypercholesterolemia with risk of atherosclerotic cardiovascular disease*. JAMA cardiology  
575 2020. **5**(4): p. 390-399.
- 576 7. Verbeek, R., et al., *Cardiovascular disease risk associated with elevated lipoprotein (a)*  
577 *attenuates at low low-density lipoprotein cholesterol levels in a primary prevention setting*.  
578 European heart journal, 2018. **39**(27): p. 2589-2596.
- 579 8. Go, A.S., et al., *Heart disease and stroke statistics—2013 update: a report from the American*  
580 *Heart Association*. Circulation, 2013. **127**(1): p. e6-e245.
- 581 9. Andaleon, A., L.S. Mogil, and H.E. Wheeler, *Gene-based association study for lipid traits in*  
582 *diverse cohorts implicates BACE1 and SIDT2 regulation in triglyceride levels*. PearJ 2018. **6**: p.  
583 e4314.
- 584 10. Trinder, M., et al., *Polygenic contribution to low-density lipoprotein cholesterol levels and*  
585 *cardiovascular risk in monogenic familial hypercholesterolemia*. Circulation: Genomic  
586 Precision Medicine, 2020. **13**(5): p. 515-523.
- 587 11. Motazacker, M.M., et al., *Evidence of a polygenic origin of extreme high-density lipoprotein*  
588 *cholesterol levels*. Arteriosclerosis, Thrombosis and Vascular biology, 2013. **33**(7): p. 1521-  
589 1528.
- 590 12. Weiss, L.A., et al., *The sex-specific genetic architecture of quantitative traits in humans*. Nature  
591 genetics, 2006. **38**(2): p. 218-222.
- 592 13. Ma, L., et al., *Genome-wide association analysis of total cholesterol and high-density*  
593 *lipoprotein cholesterol levels using the Framingham heart study data*. BMJ medical genetics,  
594 2010. **11**(1): p. 1-11.
- 595 14. Klarin, D., et al., *Genetics of blood lipids among ~ 300,000 multi-ethnic participants of the*  
596 *Million Veteran Program*. Nature Genetics, 2018. **50**(11): p. 1514-1523.
- 597 15. Liu, D.J., et al., *Exome-wide association study of plasma lipids in > 300,000 individuals*. Nature  
598 genetics, 2017. **49**(12): p. 1758-1766.
- 599 16. Martin, A.R., et al., *Clinical use of current polygenic risk scores may exacerbate health*  
600 *disparities*. Nature Genetics, 2019. **51**(4): p. 584-591.
- 601 17. Bustamante, C.D. and E.G. Burchard, *De la Vega FM. Genomics for the world*. Nature, 2011.  
602 **475**(7355): p. 163-5.
- 603 18. Oh, S.S., et al., *Making precision medicine socially precise. Take a deep breath*. American  
604 Journal of Respiratory and Critical Care Medicine, 2016.
- 605 19. Duncan, L., et al., *Analysis of polygenic risk score usage and performance in diverse human*  
606 *populations*. Nature Communications, 2019. **10**(1): p. 1-9.
- 607 20. Morris, A.P., *Transethnic meta-analysis of genomewide association studies*. Genetic  
608 Epidemiology, 2011. **35**(8): p. 809-822.
- 609 21. Okada, Y., et al., *Genetics of rheumatoid arthritis contributes to biology and drug discovery*.  
610 Nature, 2014. **506**(7488): p. 376-381.
- 611 22. Brown, B.C., et al., *Transethnic genetic-correlation estimates from summary statistics*. The  
612 American Journal of Human Genetics, 2016. **99**(1): p. 76-88.
- 613 23. Galinsky, K.J., et al., *Estimating cross-population genetic correlations of causal effect sizes*.  
614 Genetic Epidemiology, 2019. **43**(2): p. 180-188.

- 615 24. Veturi, Y., et al., *Modeling heterogeneity in the genetic architecture of ethnically diverse*  
616 *groups using random effect interaction models*. Genetics, 2019. **211**(4): p. 1395-1407.
- 617 25. Takeuchi, F., et al., *Interethnic analyses of blood pressure loci in populations of East Asian and*  
618 *European descent*. Nature Communications, 2018. **9**(1): p. 1-16.
- 619 26. Rosenberg, N.A., et al., *Genome-wide association studies in diverse populations*. Nature  
620 Reviews Genetics, 2010. **11**(5): p. 356-366.
- 621 27. McClellan, J. and M.-C. King, *Genetic heterogeneity in human disease*. Cell, 2010. **141**(2): p.  
622 210-217.
- 623 28. Benyamin, B., et al., *Cross-ethnic meta-analysis identifies association of the GPX3-TNIP1 locus*  
624 *with amyotrophic lateral sclerosis*. Nature Communications, 2017. **8**(1): p. 1-7.
- 625 29. Ding, K. and I.J. Kullo, *Evolutionary genetics of coronary heart disease*. Circulation, 2009.  
626 **119**(3): p. 459-467.
- 627 30. Shi, H., et al., *Population-specific causal disease effect sizes in functionally important regions*  
628 *impacted by selection*. Nature communications, 2021. **12**(1): p. 1-15.
- 629 31. Momin, M.M., et al., *A novel method for an unbiased estimate of cross-ancestry genetic*  
630 *correlation using individual-level data*. bioRxiv, 2021.
- 631 32. Speed, D., et al., *Reevaluation of SNP heritability in complex human traits*. Nature Genetics,  
632 2017. **49**(7): p. 986-992.
- 633 33. Martin, A.R., et al., *Human demographic history impacts genetic risk prediction across diverse*  
634 *populations*. The American Journal of Human Genetics, 2017. **100**(4): p. 635-649.
- 635 34. Márquez-Luna, C., et al., *Multiethnic polygenic risk scores improve risk prediction in diverse*  
636 *populations*. Genetic Epidemiology, 2017. **41**(8): p. 811-823.
- 637 35. Vilhjálmsson, B.J., et al., *Modeling linkage disequilibrium increases accuracy of polygenic risk*  
638 *scores*. The American Journal of Human Genetics, 2015. **97**(4): p. 576-592.
- 639 36. Lam, M., et al., *Comparative genetic architectures of schizophrenia in East Asian and European*  
640 *populations*. Nature Genetics, 2019. **51**(12): p. 1670-1678.
- 641 37. Neshat, M., et al., *A novel hyper-parameter can increase the prediction accuracy in a single-*  
642 *step genetic evaluation*. BioRxiv, 2022: p. 2022.07.03.498620.
- 643 38. Lee, S.H., et al., *Estimation of pleiotropy between complex diseases using single-nucleotide*  
644 *polymorphism-derived genomic relationships and restricted maximum likelihood*.  
645 Bioinformatics, 2012. **28**(19): p. 2540-2542.
- 646 39. Speed, D., et al., *Improved heritability estimation from genome-wide SNPs*. The American  
647 Journal of Human Genetics, 2012. **91**(6): p. 1011-1021.
- 648 40. Speed, D., J. Holmes, and D.J. Balding, *Evaluating and improving heritability models using*  
649 *summary statistics*. Nature Genetics, 2020. **52**(4): p. 458-462.
- 650 41. Iliadou, A., et al., *Heritabilities of lipids in young European American and African American*  
651 *twins*. Twin Research and Human Genetics, 2005. **8**(5): p. 492-498.
- 652 42. Peterson, R.E., et al., *Genome-wide association studies in ancestrally diverse populations:*  
653 *opportunities, methods, pitfalls, and recommendations*. Cell, 2019. **179**(3): p. 589-603.
- 654 43. Marigorta, U.M. and A. Navarro, *High trans-ethnic replicability of GWAS results implies*  
655 *common causal variants*. PLoS Genet, 2013. **9**(6): p. e1003566.
- 656 44. Euesden, J., C.M. Lewis, and P.F. O'Reilly, *PRSice: polygenic risk score software*. Bioinformatics,  
657 2015. **31**(9): p. 1466-1468.
- 658 45. Momin, M.M., et al., *Significance tests for R<sup>2</sup> of out-of-sample prediction using polygenic*  
659 *scores*. The American Journal of Human Genetics, 2023.
- 660 46. Nelson, R.H., *Hyperlipidemia as a risk factor for cardiovascular disease*. Primary Care: Clinics  
661 in Office Practice, 2013. **40**(1): p. 195-211.
- 662 47. Tall, A.R., et al., *Addressing dyslipidemic risk beyond LDL-cholesterol*. The Journal of Clinical  
663 Investigation, 2022. **132**(1).
- 664 48. Kuchenbaecker, K., et al., *The transferability of lipid loci across African, Asian and European*  
665 *cohorts*. Nature Communications, 2019. **10**(1): p. 1-10.

- 666 49. Cao, C., *Analysis of Concordance and Discordance in Genetic Association Studies via Forward-*  
667 *Backward Scoring Scheme*, Masters Thesis. 2020, The Ohio State University.
- 668 50. Huang, Q.Q., et al., *Transferability of genetic loci and polygenic scores for cardiometabolic*  
669 *traits in British Pakistani and Bangladeshi individuals*. Nature communications, 2022. **13**(1): p.  
670 1-11.
- 671 51. Lewis, C.M. and E. Vassos, *Polygenic risk scores: from research tools to clinical instruments*.  
672 Genomic Medicine, 2020. **12**: p. 1-11.
- 673 52. Tropf, F.C., et al., *Hidden heritability due to heterogeneity across seven populations*. Nature  
674 Human Behaviour, 2017. **1**(10): p. 757-765.
- 675 53. Bulik-Sullivan, B., et al., *ReproGen Consortium Psychiatric Genomics Consortium Genetic*  
676 *Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3 An atlas of*  
677 *genetic correlations across human diseases and traits*. Nat Genet, 2015. **47**(11): p. 1236-1241.
- 678 54. Gusev, A., et al., *Partitioning heritability of regulatory and cell-type-specific variants across 11*  
679 *common diseases*. The American Journal of Human Genetics, 2014. **95**(5): p. 535-552.
- 680 55. Fry, A., et al., *Comparison of sociodemographic and health-related characteristics of UK*  
681 *Biobank participants with those of the general population*. American Journal of Epidemiology,  
682 2017. **186**(9): p. 1026-1034.
- 683 56. Ollier, W., T. Sprosen, and T. Peakman, *UK Biobank: from concept to reality*. Future Medicine,  
684 2005.
- 685 57. Novembre, J. and M. Stephens, *Interpreting principal component analyses of spatial*  
686 *population genetic variation*. Nature Genetics, 2008. **40**(5): p. 646-649.
- 687 58. Loh, P.-R., et al., *Reference-based phasing using the Haplotype Reference Consortium panel*.  
688 Nature Genetics, 2016. **48**(11): p. 1443.
- 689 59. Bulik-Sullivan, B., et al., *ReproGen Consortium Psychiatric Genomics Consortium Genetic*  
690 *Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3 An atlas of*  
691 *genetic correlations across human diseases and traits*. Nature Genetics, 2015. **47**(11): p. 1236-  
692 1241.
- 693 60. Border, R., et al., *Imputation of behavioral candidate gene repeat variants in 486,551 publicly-*  
694 *available UK Biobank individuals*. European Journal of Human Genetics, 2019. **27**(6): p. 963-  
695 969.
- 696 61. Lee, S.H., W.S.P. Weerasinghe, and J.H. Van Der Werf, *Genotype-environment interaction on*  
697 *human cognitive function conditioned on the status of breastfeeding and maternal smoking*  
698 *around birth*. Scientific Reports, 2017. **7**(1): p. 1-12.
- 699 62. Peyrot, W.J., et al., *Does childhood trauma moderate polygenic risk for depression? A meta-*  
700 *analysis of 5765 subjects from the psychiatric genomics consortium*. Biological Psychiatry,  
701 2018. **84**(2): p. 138-147.
- 702 63. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage*  
703 *analyses*. The American Journal of Human Genetics, 2007. **81**(3): p. 559-575.
- 704 64. Zhou, X., H.K. Im, and S.H. Lee, *CORE GREML for estimating covariance between random*  
705 *effects in linear mixed models for complex trait analyses*. Nature Communications, 2020.  
706 **11**(1): p. 1-11.
- 707 65. Ni, G., et al., *Estimation of genetic correlation via linkage disequilibrium score regression and*  
708 *genomic restricted maximum likelihood*. The American Journal of Human Genetics, 2018.  
709 **102**(6): p. 1185-1194.
- 710 66. Meuleman, W., et al., *Index and biological spectrum of human DNase I hypersensitive sites*.  
711 Nature 2020. **584**(7820): p. 244-251.
- 712 67. Yadav, M.L. and B. Mohapatra, *Intergenic regions, also known as spacer DNA*. 2018.
- 713 68. Gilbert, W., *Genes-in-pieces revisited*. Science, 1985. **228**: p. 823-825.
- 714 69. Jin, J., et al. *Principal components ancestry adjustment for Genetic Analysis Workshop 17 data*.  
715 *in BMC Proceedings*. 2011. BioMed Central.

- 716 70. Okbay, A., et al., *Genome-wide association study identifies 74 loci associated with educational*  
717 *attainment*. Nature, 2016. **533**(7604): p. 539-542.
- 718 71. Yang, J., et al., *GCTA: a tool for genome-wide complex trait analysis*. The American Journal of  
719 Human Genetics, 2011. **88**(1): p. 76-82.
- 720 72. VanRaden, P.M., *Efficient methods to compute genomic predictions*. Journal of Dairy Science,  
721 2008. **91**(11): p. 4414-4423.
- 722 73. Yang, J., et al., *Common SNPs explain a large proportion of the heritability for human height*.  
723 Nature Genetics, 2010. **42**(7): p. 565-569.
- 724 74. Amin, N., C.M. Van Duijn, and Y.S. Aulchenko, *A genomic background based method for*  
725 *association analysis in related individuals*. PloS One, 2007. **2**(12): p. e1274.
- 726 75. Bulik-Sullivan, B.K., et al., *LD Score regression distinguishes confounding from polygenicity in*  
727 *genome-wide association studies*. Nature genetics, 2015. **47**(3): p. 291.
- 728 76. Lee, S.H. and J.H. Van der Werf, *MTG2: an efficient algorithm for multivariate linear mixed*  
729 *model analysis based on genomic information*. Bioinformatics, 2016. **32**(9): p. 1420-1422.
- 730