

# 1      **StocSum: stochastic summary statistics for whole**

## 2                   **genome sequencing studies**

### 3

4      Nannan Wang<sup>1</sup>, Bing Yu<sup>1</sup>, Goo Jun<sup>1</sup>, Qibin Qi<sup>2</sup>, Ramon A. Durazo-Arvizu<sup>3,4</sup>, Sara  
5      Lindstrom<sup>5,6</sup>, Alanna C. Morrison<sup>1</sup>, Robert C. Kaplan<sup>2</sup>, Eric Boerwinkle<sup>1,7</sup>, Han Chen<sup>1,8</sup>

### 6

7      <sup>1</sup>Human Genetics Center, Department of Epidemiology, Human Genetics and  
8      Environmental Sciences, School of Public Health, The University of Texas Health Science  
9      Center at Houston, Houston, TX, USA.

10  
11      <sup>2</sup>Department of Epidemiology & Population Health, Albert Einstein College of Medicine,  
12      Bronx, NY, USA.

13  
14      <sup>3</sup>The Saban Research Institute, Children's Hospital Los Angeles, Los Angeles, California.

15  
16      <sup>4</sup>Department of Pediatrics, Keck School of Medicine, University of Southern California,  
17      Los Angeles, CA, USA

18  
19      <sup>5</sup>Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA,  
20      USA.

21  
22      <sup>6</sup>Department of Epidemiology, School of Public Health, University of Washington, 3980  
23      15th Ave NE, Seattle, WA, USA.

24  
25      <sup>7</sup>Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA.

26  
27      <sup>8</sup>Center for Precision Health, School of Biomedical Informatics, The University of Texas  
28      Health Science Center at Houston, Houston, TX, USA.

29  
30      Correspondence: Han Chen ([Han.Chen.2@uth.tmc.edu](mailto:Han.Chen.2@uth.tmc.edu))

31 **Abstract**

32 Genomic summary statistics, usually defined as single-variant test results from genome-  
33 wide association studies, have been widely used to advance the genetics field in a wide  
34 range of applications. Applications that involve multiple genetic variants also require their  
35 correlations or linkage disequilibrium (LD) information, often obtained from an external  
36 reference panel. In practice, it is usually difficult to find suitable external reference panels  
37 that represent the LD structure for underrepresented and admixed populations, or rare  
38 genetic variants from whole genome sequencing (WGS) studies, limiting the scope of  
39 applications for genomic summary statistics. Here we introduce StocSum, a novel  
40 reference-panel-free statistical framework for generating, managing, and analyzing  
41 stochastic summary statistics using random vectors. We develop various downstream  
42 applications using StocSum including single-variant tests, conditional association tests,  
43 gene-environment interaction tests, variant set tests, as well as meta-analysis and LD score  
44 regression tools. We demonstrate the accuracy and computational efficiency of StocSum  
45 using two cohorts from the Trans-Omics for Precision Medicine Program. StocSum will  
46 facilitate sharing and utilization of genomic summary statistics from WGS studies,  
47 especially for underrepresented and admixed populations.

48

49

50 Key words: genomic summary statistics, whole genome sequencing, rare variants, LD  
51 score regression, underrepresented populations

52

53 **Main**

54 International consortia for genomic epidemiology research on complex diseases and  
55 quantitative traits have generated a great abundance of genomic summary statistics<sup>1–11</sup>.  
56 These summary statistics are often in the form of regression coefficients and their standard  
57 errors (and/or z scores) from single-variant tests for common genetic variants, typically  
58 defined as those with a minor allele frequency (MAF) of greater than 5% or 1%, in genome-  
59 wide association studies (GWAS). Genomic summary statistics contain important  
60 information for researchers without direct access to individual-level genotype data and  
61 sharing genomic summary results is now commonly mandated by scientific journals and  
62 funding agencies. Genomic summary statistics also play a crucial role for cross-  
63 institutional (both national and international) collaborations where individual-level data are  
64 difficult to share due to ethical and legal restrictions.

65

66 Genomic summary statistics have been used to address different scientific questions in  
67 genetic and genomic research, such as meta-analysis<sup>12,13</sup>, heritability estimation<sup>14–16</sup>,  
68 conditional analysis<sup>17</sup>, variant set<sup>18–21</sup> and gene-based tests<sup>22,23</sup>, multiple phenotype  
69 analysis<sup>24–26</sup>, genetic correlation or co-heritability estimation<sup>27,28</sup>, and others<sup>29,30</sup>. Many of  
70 these methods also require information on the linkage disequilibrium (LD) or correlation  
71 structure between genetic variants, which is commonly derived from external reference  
72 panels<sup>14–17,23</sup>. While these methods usually have good performance for common variants  
73 in populations of European ancestry, it has been challenging to extend the scope of  
74 summary statistic-based applications to other ancestry groups and admixed populations<sup>14</sup>  
75 as well as rare variants<sup>15</sup>, defined as those with MAF < 5% or 1%, since the LD patterns in  
76 an external reference panel often do not match with those in the study sample.

77

78 Current large-scale whole genome sequencing (WGS) projects, such as the National Heart,  
79 Lung, and Blood Institute's (NHLBI's) Trans-Omics for Precision Medicine (TOPMed)  
80 program, the National Human Genome Research Institute's (NHGRI's) Centers for  
81 Common Disease Genomics (CCDG) initiative, and the National Institute on Aging's  
82 (NIA's) Alzheimer's Disease Sequencing Project (ADSP), have unveiled hundreds of  
83 millions of rare variants from diverse populations. Making efficient and flexible use of

84 these WGS resources and derived genomic summary results is paramount to facilitate  
85 international collaborations and scientific discoveries. However, managing and  
86 coordinating large-scale consortium efforts on rare variant meta-analyses has been quite  
87 challenging, since many existing meta-analysis software programs such as seqMeta<sup>31</sup>,  
88 MetaSKAT<sup>18</sup>, RVTESTS<sup>32</sup>, RAREMETAL<sup>33</sup> and SMMAT<sup>21</sup>, require the correlation (or  
89 LD) matrices for rare variants to be computed internally in the study samples. In rare  
90 variant tests<sup>21,34-40</sup>, variant set definitions often need to be pre-specified (e.g., by genomic  
91 motifs such as genes or physical windows). Therefore, researchers have to recreate the LD  
92 matrices every time they want to redefine a variant set (e.g., by including more variants in  
93 a test region or combining two testing windows). This requires additional computational  
94 resources, making it difficult for researchers to efficiently leverage the richness of the data.  
95 On the other hand, sharing terabytes or even petabytes of individual-level WGS and  
96 phenotype data across research groups is a daunting task, and the risk of privacy breaches  
97 generally increases as more copies of individual-level data are being shared. Although  
98 individual-level WGS data can now be accessed through cloud-based computing platforms  
99 such as the Analysis Commons<sup>41</sup>, BioData Catalyst and AnVIL, and recently developed  
100 analysis tools such as STAARpipeline<sup>42</sup> have greatly improved rare variant analyses  
101 especially for the noncoding genome, research groups are still largely constrained by the  
102 computational costs they can afford in running WGS data analysis using individual-level  
103 data directly.

104

105 Ideally, computing genomic summary statistics only once and then recycling them for  
106 different variant set definitions and weighting schemes is a more efficient strategy for WGS  
107 analysis on rare variants. Downstream analyses using summary statistics would not depend  
108 on the sample size  $N$  and therefore could be easily performed on a desktop computer.  
109 However, there are critical barriers in scaling existing statistical methods based on GWAS  
110 summary statistics up to allow for summary statistics based on WGS studies. First,  
111 calculating traditional pairwise LD measures from individual-level genomic data is  
112 computationally intensive. In general, a covariance matrix of size  $M \times M$  is desired  
113 (**Fig.1a**), where  $M$  is the total number of variants, which has already exceeded 700 million  
114 in TOPMed. In practice, genotype data are usually saved by chromosome, but  $M$  is still on

115 the scale of millions even for the shortest chromosome, making pairwise LD calculations  
116 on the whole genome (or one chromosome) computationally infeasible. Second, although  
117 restricting LD calculations to only genetic variants in close proximity (e.g., the sliding  
118 window strategy<sup>43</sup> and the banded sparse LD matrices in 500kb windows<sup>44</sup>) is more  
119 computationally efficient than calculating the full  $M \times M$  covariance matrix, it does not  
120 allow for the flexibility of testing distant genetic variants jointly. As there is growing  
121 evidence that the three-dimensional organization of chromosomes profoundly affects gene  
122 regulation<sup>29,45-52</sup>, LD matrices generated through sliding windows cannot be used if the  
123 variant set of interest contains genetic variants that are located far away from each other.  
124 In addition, LD statistics used in rare variant tests can greatly depend on the phenotype of  
125 interest (e.g., the phenotype distributions in minor allele carriers vs. non-carriers for each  
126 variant), and generally cannot be pre-computed using WGS data without the phenotype  
127 information.

128

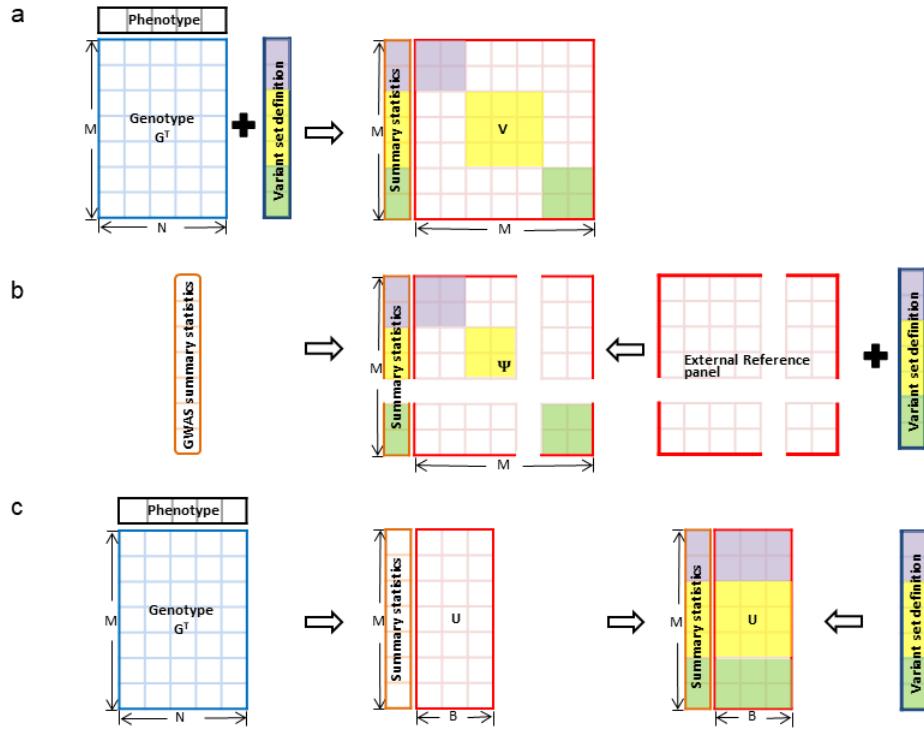
129 In addition, many existing methods using genomic summary statistics based on common  
130 variants rely on LD information from external reference panels<sup>14-17,23</sup> (**Fig. 1b**). These  
131 methods have been widely applied to common variants in primary populations of European  
132 ancestry. Extension of these methods to underrepresented and admixed populations,  
133 however, has been noted as a challenge<sup>14,27</sup> due to lack of appropriate reference panels that  
134 accurately represent the LD structure.

135

136 In this study, we propose the StocSum framework as illustrated in **Fig. 1c** to extend the  
137 scope of summary statistic-based applications. For methods that require between-variant  
138 correlation or LD matrices, we use a stochastic summary statistic matrix  $\mathbf{U}$  to replace the  
139 traditional pairwise LD matrix  $\mathbf{V}$ . Specifically, by using  $B$  independent and identically  
140 distributed random vectors to represent the parametric distribution of any model-based  
141 residuals from a complex statistical model that accounts for potential sample correlations,  
142 matrix  $\mathbf{U}$  can be quickly computed by matrix multiplication of the  $N \times M$  genotype matrix  
143  $\mathbf{G}$  and these  $B$  random vectors. The size of  $\mathbf{U}$  scales linearly with  $M$  and  $B$  (i.e.,  $O(MB)$ ),  
144 compared to quadratically in the form of a traditional pairwise LD matrix  $\mathbf{V}$ . The stochastic  
145 summary statistic matrix  $\mathbf{U}$  can always be computed in linear time with the sample size  $N$

146 (*i.e.*,  $O(NMB)$ ), regardless of any complex sample correlation structures, compared to  
147  $O(NM^2)$  for the traditional pairwise LD matrix  $\mathbf{V}$  in classical linear and logistic regression  
148 models for unrelated individuals, or mixed effect models to account for sample correlations  
149 in the presence of a sparse and block-diagonal relatedness matrix with bounded block sizes  
150 (e.g., a population-based family study, with known pedigrees). The complexity for  
151 computing  $\mathbf{V}$  could further increase to  $O(N^2M + NM^2)$  if the relatedness matrix used in  
152 the mixed effect model is not block-diagonal (e.g., the genetic relationship matrix, or  
153 GRM). We also develop downstream applications using StocSum, including single-variant,  
154 conditional association, gene-environment interaction, variant set tests, as well as meta-  
155 analysis and LD score regression tools. This framework can flexibly accommodate changes  
156 of variant set definitions in analysis plans. For example, in variant set tests for rare variants,  
157 we can efficiently calculate the LD matrix for any variant sets by simply looking up  $\frac{\mathbf{U}\mathbf{U}^T}{B}$   
158 rather than rerunning the analysis with individual-level genotype data to update LD  
159 matrices for new variant sets. Compared with using external reference panels which might  
160 not well represent the LD structure in study samples from underrepresented and admixed  
161 populations, StocSum can be used to better calibrate the LD information in a wide range  
162 of genomic summary statistic-based applications.

163



164

165 **Figure 1: The StocSum framework.** **a)** Traditional methods calculate the correlation or  
166 LD matrix  $V$  from individual-level genotype data. To reduce the computational burden, the  
167 full  $M \times M$  matrix is usually not computed in practice, but rather replaced by a block-  
168 diagonal or banded sparse matrix based on pre-defined variant sets, at the cost of losing the  
169 flexibility in testing distant genetic variants jointly. **b)** The approximate LD matrix  $\Psi$  is  
170 obtained from external reference panels when individual-level genotypes are not available,  
171 in many genomic summary statistic-based applications. However, variants may be  
172 excluded if they do not exist in the reference panel. **c)** StocSum generates stochastic  
173 summary statistics  $U$  from random vectors, which can be used to efficiently look up the  
174 covariance among arbitrary variant sets that are not pre-defined.  $M$ , the number of variants.  
175  $N$ , the sample size.  $B$ , the number of random vectors used to construct stochastic summary  
176 statistics  $U$ .

## 177 Results

### 178 Overview of the method

179 We describe StocSum under the generalized linear mixed model (GLMM) framework. It  
180 can also be applied to simpler statistical models such as generalized linear models<sup>53</sup> and  
181 extended to more complex models such as generalized additive mixed models<sup>54</sup>. The  
182 GLMM can be written as:

183

$$\text{g}(\mu_i) = \mathbf{X}_i \boldsymbol{\alpha} + \tilde{\mathbf{G}}_i \boldsymbol{\beta} + b_i \quad (1)$$

184 where  $\text{g}(\cdot)$  is a monotonic link function of  $\mu_i$ , and  $\mu_i = E(y_i | \mathbf{X}_i, \tilde{\mathbf{G}}_i, b_i)$  is the conditional  
185 mean of the phenotype  $y_i$  given  $p$  covariates  $\mathbf{X}_i$ ,  $q$  genotypes  $\tilde{\mathbf{G}}_i$  and random effects  $b_i$ , for  
186 individual  $i$  of  $N$  samples. The phenotype  $y_i$  follows a distribution in the exponential  
187 family, such as a normal distribution for continuous phenotypes, or a Bernoulli distribution  
188 for binary phenotypes. Here  $\boldsymbol{\alpha}$  is a length  $p$  column vector of fixed covariate effects  
189 including an intercept term. The genotype matrix  $\tilde{\mathbf{G}} = (\tilde{\mathbf{G}}_1^T \tilde{\mathbf{G}}_2^T \cdots \tilde{\mathbf{G}}_N^T)^T$  is an  $N \times q$  matrix  
190 for  $q$  ( $q \geq 1$ ) genetic variants and  $\boldsymbol{\beta}$  is a length  $q$  genotype effect vector. We assume that  
191  $\mathbf{b} = (b_1 \ b_2 \ \cdots \ b_N)^T$  is a length  $N$  column vector of random effects and  $\mathbf{b} \sim \sum_{k=1}^K \tau_k \boldsymbol{\Phi}_k$ ,  
192 where  $\tau_k$  are the variance component parameters and  $\boldsymbol{\Phi}_k$  are known  $N \times N$  dense or  
193 sparse relatedness matrices which account for multiple layers of correlation structures, such  
194 as genetic relatedness, hierarchical designs, shared environmental effects and repeated  
195 measures from longitudinal studies.

196

197 For both single-variant ( $q = 1$ ) and variant set ( $q > 1$ ) tests, we only need to fit the null  
198 model  $\text{g}(\mu_{0i}) = \mathbf{X}_i \boldsymbol{\alpha} + b_i$  without fixed genetic effects one time, then each test can be  
199 constructed using single-variant scores  $\mathbf{S}$  and  $q \times q$  covariance matrices  $\tilde{\mathbf{V}} = \tilde{\mathbf{G}}^T \mathbf{P} \tilde{\mathbf{G}}$ ,  
200 where  $\mathbf{P}$  is the projection matrix from this model<sup>21,55</sup>. Denote  $M$  as the total number of  
201 genetic variants on the whole genome (or one chromosome). To avoid computing the full  
202  $M \times M$  matrix  $\mathbf{V}$  or its block-diagonal version for every  $q$  variants  $\tilde{\mathbf{V}}$  directly from  
203 individual-level data or an external reference panel, StocSum leverages a length  $N$  random  
204 vector  $\mathbf{R}_b$  from a multivariate normal distribution with mean  $\mathbf{0}$  and covariance matrix  $\mathbf{P}$ .

205 Then it repeats this simulation process  $B$  times and combines these random vectors into an  
206  $N \times B$  random matrix  $\mathbf{R} = (\mathbf{R}_1 \ \mathbf{R}_2 \ \dots \ \mathbf{R}_B)$ . Denoting  $\mathbf{U} = \mathbf{G}^T \mathbf{R}$  as the stochastic  
207 summary statistics for  $M$  genetic variants on the whole genome (or one chromosome), for  
208 arbitrary  $q$  variants ( $q < B$ ), we can extract the corresponding rows from the  $M \times B$   
209 stochastic summary statistics matrix  $\mathbf{U}$  as  $\tilde{\mathbf{U}}$  and use  $\frac{\tilde{\mathbf{U}}\tilde{\mathbf{U}}^T}{B}$  to estimate the covariance matrix  
210  $\tilde{\mathbf{V}}$ .

211  
212 To implement StocSum and various downstream genetic analysis applications, our  
213 framework comprises four major steps: (1) fitting a generalized linear mixed model under  
214 the null hypothesis, e.g.,  $\mathbf{g}(\mu_{0i}) = \mathbf{X}_i \boldsymbol{\alpha} + b_i$ , estimating variance component parameters,  
215 residuals and the projection matrix  $\mathbf{P}$ ; (2) generating an  $N \times B$  random matrix  $\mathbf{R}$ , with each  
216 column of  $\mathbf{R}$  simulated from a multivariate normal distribution with mean  $\mathbf{0}$  and  
217 covariance matrix  $\mathbf{P}$ ; (3) using individual-level genotypes  $\mathbf{G}$  to compute score statistics  
218 from residuals, and the stochastic summary statistics matrix  $\mathbf{U} = \mathbf{G}^T \mathbf{R}$ ; and (4) computing  
219  $P$  values in each downstream application (see **Methods**). The first three steps could be  
220 shared by multiple genetic analysis applications including single-variant, conditional  
221 association, gene-environment interaction, and variant set tests. We could also estimate LD  
222 scores efficiently in the stochastic summary statistics framework, thus extending its  
223 application to underrepresented and admixed populations (see **Methods**).

## 224 **Single-variant tests**

225 To evaluate the performance of StocSum in single-variant tests, we used TOPMed WGS  
226 freeze 8 data from the Hispanic Community Health Study/Study of Latinos (HCHS/SOL).  
227 After quality control we had data for 120M variants in 7,297 individuals (**Methods**). We  
228 first compared  $P$  values calculated by StocSum with different numbers of random vector  
229 replicates  $B$  and GMMAT<sup>55</sup> using individual-level genotypes in a genome-wide single-  
230 variant analysis of blood low-density lipoprotein (LDL) cholesterol levels (**Fig. 2a-d**). The  
231  $P$  values calculated from StocSum were compared with those from GMMAT using  
232 individual-level data. No systematic genomic inflation was observed from the quantile-  
233 quantile (Q-Q) plots (**Fig. S1**). StocSum  $P$  values were close to GMMAT when the number

234 of random vector replicates  $B$  ranged from 100 to 10,000 (Fig.2b-2d). We did observe that  
235 a small  $B$  ( $B=10$ ) led to inaccurate  $P$  values (Fig. 2a).

236

237 To demonstrate the computational efficiency of StocSum, we ran GMMAT and StocSum  
238 ( $B=1,000$ ) on the same computing platform where 64 cores were used in parallel computing  
239 for both programs. Both runtime and memory usage of StocSum were much lower  
240 compared to GMMAT. For example, it took about 50.2 CPU hours to run chromosome 1  
241 with 9.7M variants using StocSum, which was 4.6-fold faster than GMMAT. Meanwhile,  
242 StocSum only had 29.3% of the memory footprint compared to GMMAT. Across all 22  
243 autosomes, StocSum was 4.4-fold faster than GMMAT, with about 25.1% of the memory  
244 footprint compared to GMMAT (Fig. 2e-f). As expected, both the run time and memory  
245 footprint increased with a larger  $B$ . However, the run time and memory footprint of  
246 StocSum when  $B=10,000$  were still only 29.3% and 50.6% compared to GMMAT,  
247 respectively.

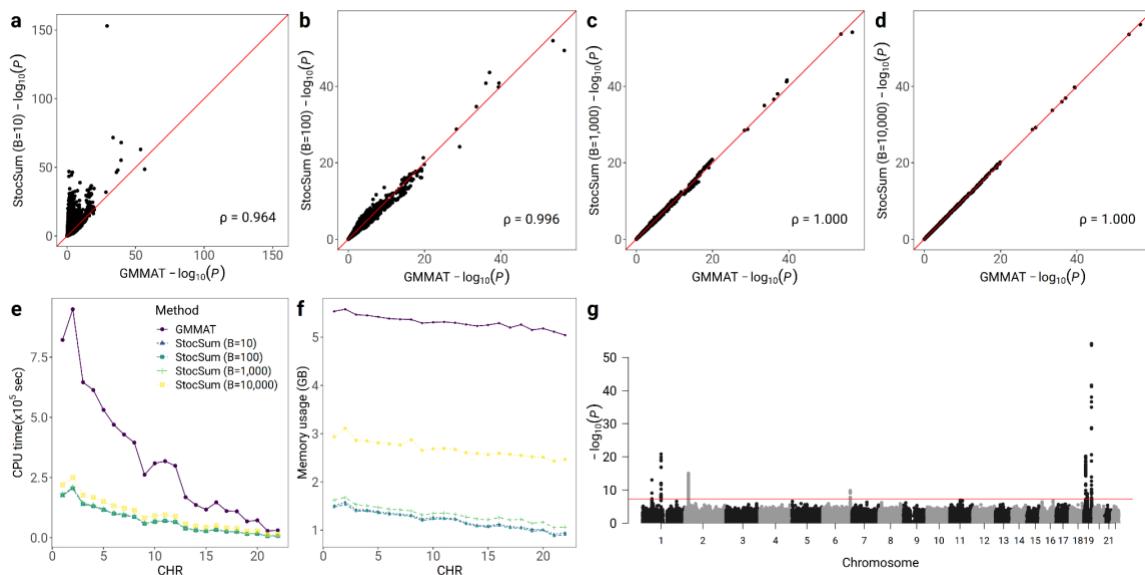
248

249 Using StocSum, we ran a WGS study of LDL cholesterol in HCHS/SOL and identified  
250 seven genome-wide significant ( $P$  values  $< 5 \times 10^{-8}$ ) regions mapped to genes *PCSK9* and  
251 *CELSR2* on chromosome 1, *APOB* on chromosome 2, *LPA* on chromosome 6, *LDLR*,  
252 *SUGP1*, and *APOE* on chromosome 19 (Fig. 2g, Table S1), all of which had been  
253 previously reported to be associated with LDL<sup>4,56-59</sup>.

254

255 We also compared StocSum with fastGWA<sup>60</sup>, another widely used single-variant test tool  
256 (Figs. S2-3). To make a fair comparison on the same statistical model, we only included  
257 one random effect term for genetic relatedness, without allowing for heteroscedasticity in  
258 the null model for GMMAT and StocSum. Both fastGWA and GMMAT results were very  
259 similar (Figs.S2-3). In this different null model, StocSum  $P$  values were still consistent  
260 with GMMAT when  $B$  ranged from 100 to 10,000. The CPU time used by fastGWA was  
261 generally stable for different chromosomes (Fig.S4a). The total CPU time for the whole  
262 genome analysis was similar for StocSum ( $B=1,000$ ) and fastGWA. The memory usage of  
263 fastGWA was slightly larger (about 1.7-fold) compared to StocSum with  $B=1,000$   
264 (Fig.S4b).

265



266

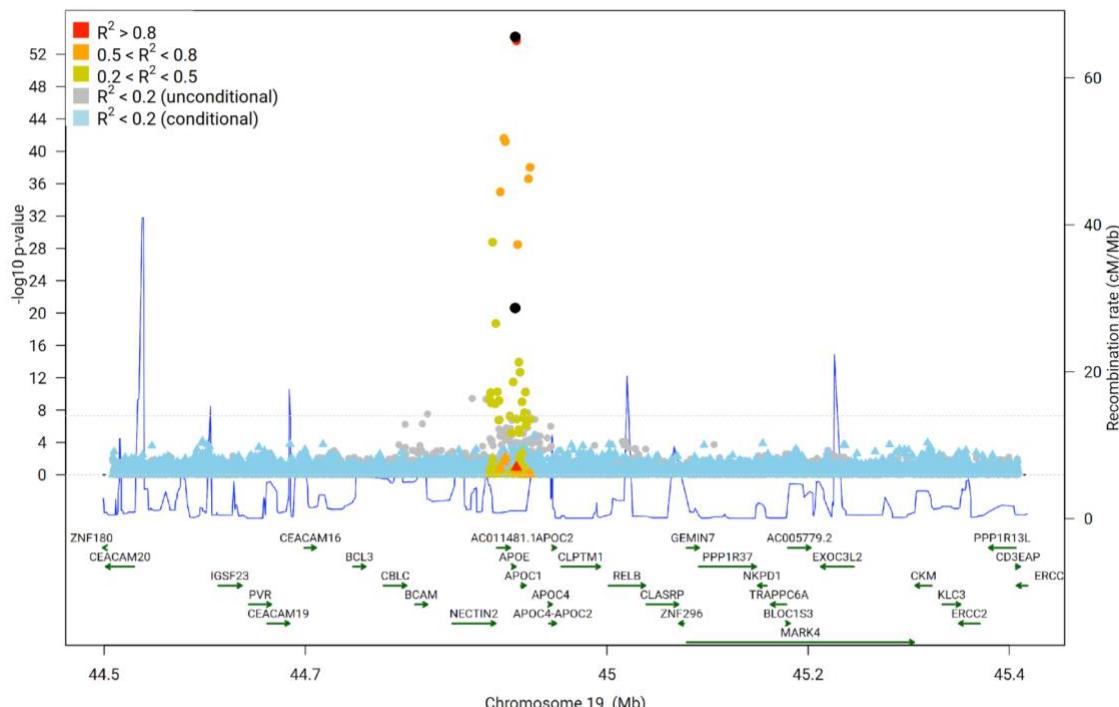
267 **Figure 2: StocSum in single-variant tests.** a-d) comparison of  $P$  values from GMMAT  
 268 and StocSum with the number of random vector replicates  $B$  being equal to 10 (a), 100 (b),  
 269 1,000 (c) and 10,000 (d). The x axis and the y axis represent  $-\log_{10}(P)$  from single-variant  
 270 tests using GMMAT and StocSum, respectively. The red line denotes the reference line of  
 271 equality. Spearman's rank correlation coefficients are shown at the bottom right. e)  
 272 comparison of CPU time between GMMAT and StocSum. The x axis represents the  
 273 chromosome numbers, and the y axis represents the CPU time in  $10^5$  seconds. For  
 274 GMMAT, the CPU time consists of fitting the null model and conducting the association  
 275 test. For StocSum, the CPU time is the sum of four steps: fitting the null model, generating  
 276 the random vectors, computing the single-variant score statistics and the stochastic  
 277 summary statistics, and computing the  $P$  values. f) comparison of memory usage by  
 278 GMMAT and StocSum. The x axis represents the chromosome numbers and the y axis  
 279 represents the memory footprint per core in GB. The data used in this test consisted of  
 280 120M variants from 7,297 individuals in HCHS/SOL. All tests were performed on a high-  
 281 performance computing server, with 64 cores used for parallel computing. g) the  
 282 Manhattan plot of single-variant test on LDL in the HCHS/SOL study using StocSum. The  
 283 x-axis represents the physical chromosome and position of each variant and the y-axis  
 284 represents  $-\log_{10}(P)$  from the StocSum single-variant test. Only variants with  $MAF > 0.5\%$

285 were included in the Manhattan plot. The red line indicates the genome-wide significance  
286 level on the log scale,  $-\log_{10}(5 \times 10^{-8})$ .

287 **Conditional association tests**

288 We implemented StocSum for conditional association tests and applied it to the seven  
289 genome-wide significant regions identified in **Fig. 2g**. The sentinel variant in the *APOE*  
290 gene region is chr19: 44908822 (rs7412) with  $P = 7.1 \times 10^{-55}$ . There are 26 common variants  
291 with MAF > 0.5% close to this sentinel variant in this region, with a  $P$  value less than  $5 \times 10^{-8}$   
292 (**Fig. 3**). After conditioning on the sentinel variant, we identified a secondary association  
293 variant chr19: 44908684 (rs429358) with conditional  $P = 8.2 \times 10^{-15}$ . After conditioning on  
294 both rs7412 and rs429358, all other variants in the region had  $P$  values  $> 1.8 \times 10^{-3}$ ,  
295 indicating that no additional independent associations exist in this region. We also observed  
296 similar patterns in the other six regions (**Table S1**), after conditioning on either one or two  
297 top associated variants in each region (**Fig. S5 a-f**).

298



299

300 **Figure 3: A regional plot of StocSum conditional association test results in the APOE**  
301 **region.** Variants with MAF > 0.5% in a 1Mb window near association variants rs7412 and  
302 rs429358 (highlighted in black dots). Original single variant test  $P$  values are shown in dots

303 and conditional  $P$  values are shown in triangles. Variants in four LD categories are shown  
304 in different colors based on the maximum squared correlation to the sentinel variant rs7412  
305 and the secondary association variant rs429358 calculated in HCHS/SOL. The horizontal  
306 dashed line indicates the genome-wide significance level on the log scale,  $-\log_{10}(5 \times 10^{-8})$ . The blue curve shows recombination rates from all populations in the 1000 Genome  
307 Project.  
308

### 309 **Gene-environment interaction tests**

310 We next developed and implemented a one-degree-of-freedom gene-environment  
311 interaction test and a two-degree-of-freedom joint test of the genetic main effects and the  
312 gene-environment interactions in the StocSum framework. We benchmarked our tests with  
313 MAGEE using individual-level data<sup>61</sup>. No systematic genomic inflation was observed from  
314 the quantile-quantile (Q-Q) plots (Fig. S6). Fig. S7 shows  $P$  values from a gene-sex  
315 interaction analysis on waist-hip ratio (WHR) in HCHS/SOL. MAGEE and StocSum  $P$   
316 values were highly consistent, with Spearman's correlations of 1.000, 0.998, 0.999,  
317 respectively, for the marginal genetic effect test, the gene-environment interaction test and  
318 the joint test. We identified four potential loci from marginal genetic effect tests, three  
319 with significant gene-sex interactions, and four from the joint tests, at the suggestive  
320 significance level of  $5 \times 10^{-7}$ , including six previously reported genome-wide significant  
321 loci in gene regions *COBLL1*, *IGF2R*, *AOAH*, *IQSEC3*, *TEKT5*, and *MAPT*<sup>62-68</sup> (Table S2).

### 322 **Variant set tests**

323 We also used TOPMed WGS freeze 8 data and LDL cholesterol levels from the  
324 HCHS/SOL study to illustrate variant set tests in the StocSum framework. We compared  
325  $P$  values calculated by StocSum with different numbers of random vector replicates  $B$  and  
326 SMMAT<sup>21</sup> using individual-level genotypes in a genome-wide 20 kb non-overlapping  
327 sliding window analysis on all genetic variants, using a beta density weight on the MAF  
328 with parameters 1 and 25. We noted that 20 kb was probably wider than what was  
329 commonly used in WGS sliding window analyses<sup>43</sup>, but we chose this window size to  
330 evaluate the performance of StocSum variant set tests in an extreme scenario not in favor  
331 of StocSum, because there could be many windows with the number of variants  $q > B$ . In  
332 this case,  $\frac{\tilde{U}\tilde{U}^T}{B}$  from StocSum would not be an appropriate estimate for the  $q \times q$

333 covariance matrix  $\tilde{\mathbf{V}}$  computed directly from individual-level data, since only  $B$  singular  
334 values could be computed from the  $q \times B$  matrix  $\tilde{\mathbf{U}}$ .

335

336 **Figs. 4a-d** shows comparisons of  $P$  values from SMMAT using individual-level genotypes  
337 and StocSum with  $B$  ranging from 10 to 10,000. When  $B=1,000$  or 10,000,  $P$  values from  
338 the two methods were highly consistent (**Figs. 4c-d**). For windows with small SMMAT  $P$   
339 values, StocSum tended to overestimate these  $P$  values when  $B=10$  or 100 (**Fig. 4a-b**),  
340 possibly because only 10 or 100 singular values from  $\tilde{\mathbf{U}}$  was insufficient to approximate  
341 the eigenvalues from the  $q \times q$  covariance matrix  $\tilde{\mathbf{V}}$  from SMMAT.

342

343 StocSum variant set tests are computationally efficient (**Figs. 4e-f**). It only took StocSum  
344 ( $B=1,000$ ) 2.7 CPU hours to finish variant set tests on chromosome 1 using 20 kb sliding  
345 windows, which was 9.7-fold faster than SMMAT using individual-level data. Across the  
346 autosomes, there were a total of 134,739 non-overlapping 20 kb windows containing at  
347 least one variant. On average, the StocSum ( $B=1,000$ ) CPU time was about 14.3% of the  
348 SMMAT CPU time. Meanwhile, StocSum ( $B=1,000$ ) only required about 68.1% of the  
349 memory compared to SMMAT. StocSum with  $B=10,000$  utilized more CPU time than  
350 SMMAT since  $B$  was larger than the sample size ( $N=7,297$ ), making the  $M \times B$  stochastic  
351 summary statistics matrix  $\mathbf{U}$  even larger in size compared to the  $N \times M$  genotype matrix  
352  $\mathbf{G}$ . In this 20 kb sliding window analysis using StocSum variant set tests, we identified four  
353 regions associated with LDL levels in HCHS/SOL<sup>4,56-59</sup>, at the significance level of  
354  $0.05/134,739=3.7 \times 10^{-7}$  (**Fig. 4g, Table S3**).

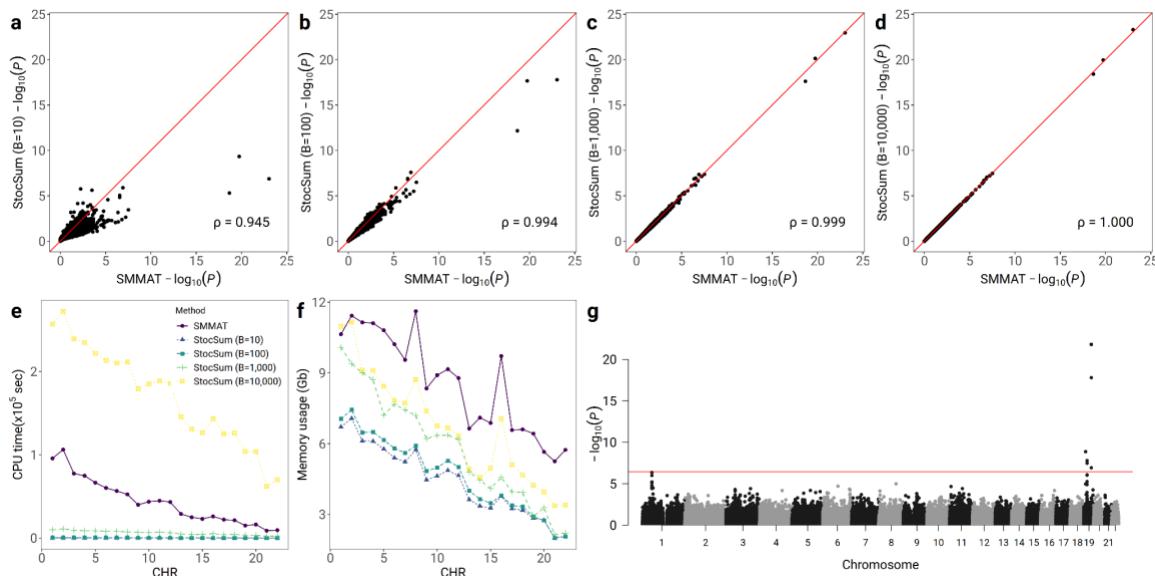
355

356 We next compared StocSum with fastBAT for variant set tests. fastBAT utilizes single-  
357 variant summary statistics from fastGWA and LD information from a reference panel such  
358 as the 1000 Genomes Project<sup>69</sup>. To make a fair comparison on the same statistical model  
359 and same weights used in variant set tests, we only included one random effect term for  
360 genetic relatedness, without allowing for heteroscedasticity in the null model for SMMAT  
361 and StocSum, and a beta density weight on the MAF with parameters 0.5 and 0.5 (which  
362 is equivalent to rescaling each variant with a unit variance as implemented in fastBAT).  
363 For fastBAT, we compared five different reference panels, including an internal reference

364 panel using individual-level genotypes from the original study sample (called fastBAT  
365 (Sample)), as well as four external reference panels from the 1000 Genomes Project<sup>69</sup>:  
366 European populations (fastBAT (Eu)), European and African populations (fastBAT  
367 (EuAf)), European and American populations (fastBAT (EuAm)), and European, African  
368 and American populations (fastBAT (EuAfAm)). Variant set test  $P$  values from SMMAT,  
369 StocSum ( $B=1,000$ ), and fastBAT (Sample) were highly concordant (**Figs. S9-10**), with  
370 pairwise Spearman correlation coefficients being greater than 0.99. However, fastBAT  
371 with external reference panels, i.e., fastBAT (Eu), fastBAT (EuAf), fastBAT (EuAm),  
372 fastBAT (EuAfAm), gave inaccurate variant set test  $P$  values compared to SMMAT using  
373 individual-level genotypes. The correlation coefficients of  $\log_{10}(P)$  between SMMAT and  
374 fastBAT with Eu, EuAf, EuAm, EuAfAm reference panels were 0.59, 0.77, 0.66, and 0.78,  
375 respectively (**Fig. S10**). Since Hispanic/Latino adults are three-way admixed populations  
376 with European, African and Amerindian ancestries, it is not surprising that an external  
377 reference panel from only European populations could not represent the LD structure in  
378 HCHS/SOL samples accurately. Interestingly, although including African and American  
379 populations in the external reference panel did improve the concordance of fastBAT  $P$   
380 values compared to SMMAT, fastBAT using the internal reference panel clearly  
381 outperformed all external reference panels that we investigated. In addition, when an  
382 external reference panel was used, variants not included in the panel would have to be  
383 excluded, leading to loss of unique variants in the study samples. This highlights the  
384 importance of choosing an accurate reference panel for fastBAT, and the best reference  
385 panel for study samples from underrepresented, admixed or isolated populations are the  
386 study samples themselves. StocSum represents the LD structure in any variant sets through  
387 a stochastic summary statistic matrix  $\mathbf{U}$  directly derived from study samples rather than  
388 external reference panels, thus providing accurate variant set test results. Meanwhile,  
389 StocSum with  $B=1,000$  was slightly faster (1.7-fold) than fastBAT (Sample) on the whole  
390 genome (**Fig. S11a**), with a dramatically reduced memory footprint (3.6%) compared to  
391 fastBAT (Sample) (**Fig. S11b**).  
392  
393 To illustrate StocSum variant set tests beyond sliding windows, we compared StocSum  
394 ( $B=1,000$ ) with SMMAT when the variant sets composed of different regions that were

395 physically farther away. These variant sets were defined by merging chromatin loops of  
396 H3K27ac HiChIP interaction in the GM12878 cell line<sup>70–72</sup>. As the definition of variant  
397 sets changed, SMMAT required rerunning the analysis using individual-level genotypes,  
398 while StocSum variant set tests could directly extract information about these new variant  
399 sets from the same pre-computed stochastic summary statistic matrix  $U$ , which yielded  
400 highly accurate  $P$  values (Fig. S12a), while using much less CPU time and memory (Figs.  
401 S12b-c).

402



403

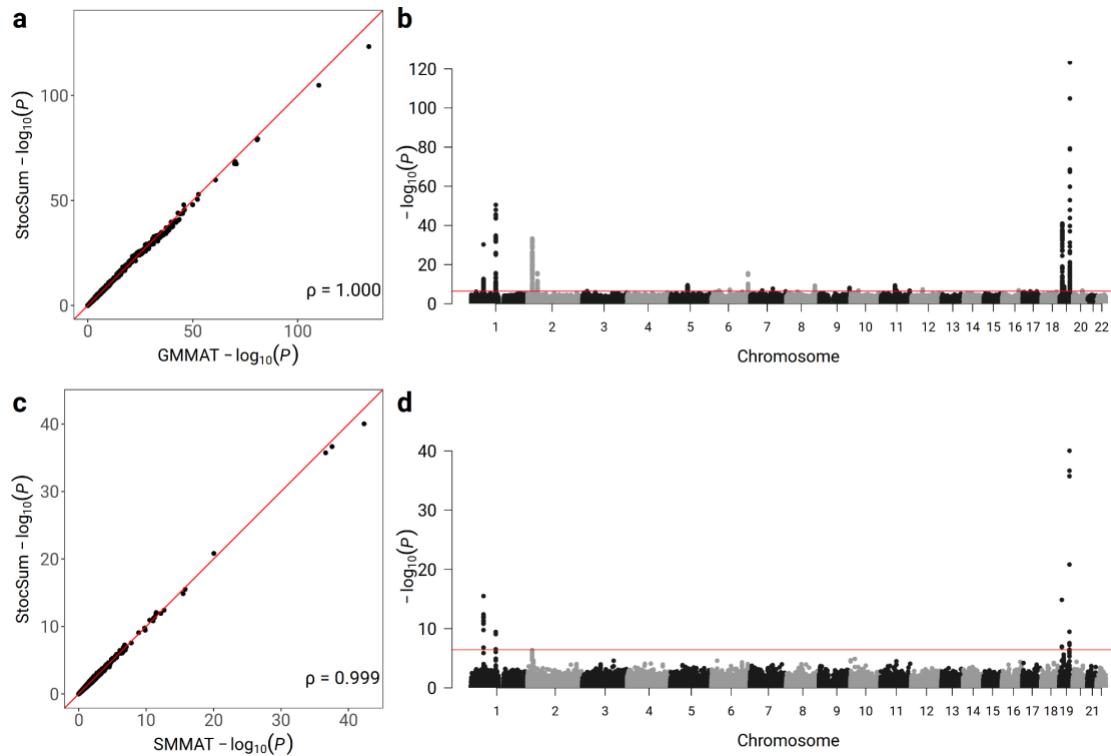
404 **Figure 4: StocSum in variant set tests.** Comparison of  $P$  values from SMMAT and  
405 StocSum with the number of random vector replicates  $B$  being equal to 10 (a), 100 (b),  
406 1,000 (c) and 10,000 (d) in a 20 kb sliding window analysis on the whole genome. The x  
407 axis and the y axis represent  $-\log_{10}(P)$  from a whole genome 20 kb sliding window analysis,  
408 using variant set tests from SMMAT and StocSum, respectively, with a beta density weight  
409 on the MAF with parameters 1 and 25. The red line denotes the reference line of equality.  
410 Spearman's rank correlation coefficients are shown at the bottom right. e, comparison of  
411 CPU time between SMMAT and StocSum. The x axis represents the chromosome numbers  
412 and the y axis represents the CPU time in  $10^5$  seconds. For SMMAT, the CPU time did not  
413 include fitting the null model or reading the variant set definitions. For StocSum, the CPU  
414 time did not include computing stochastic summary statistics from individual-level data. f,  
415 comparison of memory usage by SMMAT and StocSum. The x axis represents the

416 chromosome number and the y axis represents the memory footprint in GB. The data used  
417 in this test consisted of 120M variants from 7,297 individuals in HCHS/SOL, including all  
418 variants regardless of their MAF (such as singletons and doubletons). All tests were  
419 performed on a high-performance computing server, with a single thread for each  
420 chromosome. g, the Manhattan plot of 20 kb sliding window variant set tests on LDL in  
421 the HCHS/SOL study using StocSum. The x-axis represents the start physical chromosome  
422 and position of each variant set and the y-axis represents  $-\log_{10}(P)$  from the StocSum  
423 variant set test corresponding to SMMAT. The red line indicates the genome-wide  
424 significance level on the log scale,  $-\log_{10}(3.7 \times 10^{-7})$ .

425 **Meta-analysis**

426 Meta-analysis in the StocSum framework can be performed by combining the stochastic  
427 summary statistic matrices  $\mathbf{U}$  from different studies. To illustrate how single-variant and  
428 variant set tests can be conducted in a meta-analysis, we combined the stochastic summary  
429 statistic matrices  $\mathbf{U}$  from three studies: longitudinal LDL levels as repeated measures in  
430 African-Americans (AA) from the Atherosclerosis Risk in Communities (ARIC) study  
431 (70M variants from 2,045 individuals) visits 1-6, European-Americans (EA) from ARIC  
432 (92M variants from 6,327 individuals) visits 1-6, and baseline LDL levels as cross-  
433 sectional measures in Hispanic/Latino adults from HCHS/SOL (120M variants from 7,297  
434 individuals).  $P$  values from StocSum ( $B=1,000$ ) were highly concordant with GMMAT  
435 results from longitudinal LDL level analyses, for both ARIC AA and EA subgroups (**Fig.**  
436 **S13**), which further demonstrated the robustness of StocSum in different populations.  $P$   
437 values from StocSum meta-analysis ( $B=1,000$ ) were highly concordant with those from  
438 GMMAT single-variant meta-analysis (**Fig. 5a**) and SMMAT variant set meta-analysis  
439 (**Fig. 5c**). We identified 14 LDL loci from StocSum meta-analysis ( $B=1,000$ ) single-variant  
440 tests<sup>4,56-59,73-76</sup> (**Fig. 5b, Table S4**), at the significance level of  $5 \times 10^{-8}$ . In variant set tests  
441 (**Fig. 5d, Table S5**), we identified four regions associated with LDL levels from StocSum  
442 meta-analysis ( $B=1,000$ ), at the significance level of  $3.7 \times 10^{-7}$ .

443



444

445 **Figure 5: StocSum in meta-analysis.** a, comparison of single-variant meta-analysis  $P$   
446 values from GMMAT and StocSum with the number of random vector replicates  $B$  being  
447 equal to 1,000. The x axis and the y axis represent  $-\log_{10}(P)$  from single-variant meta-  
448 analysis using GMMAT and StocSum, respectively. The red line denotes the reference line  
449 of equality. Spearman's rank correlation coefficients are shown at the bottom right. b, the  
450 Manhattan plot of single-variant tests on LDL in the meta-analysis of ARIC AA and EA,  
451 and HCHS/SOL studies using StocSum. The x-axis represents the physical chromosome  
452 and position of each variant and the y-axis represents  $-\log_{10}(P)$  from the StocSum single-  
453 variant test. Only variants with  $MAF > 0.5\%$  were included in the Manhattan plot. The red  
454 line indicates the genome-wide significance level on the log scale,  $-\log_{10}(5 \times 10^{-8})$ . c,  
455 comparison of variant set meta-analysis  $P$  values from SMMAT and StocSum with the  
456 number of random vector replicates  $B$  being equal to 1,000. The x axis and the y axis  
457 represent  $-\log_{10}(P)$  from variant set meta-analysis using SMMAT and StocSum,  
458 respectively. The red line denotes the reference line of equality. Spearman's rank  
459 correlation coefficients are shown at the bottom right. d, the Manhattan plot of variant set  
460 tests on LDL in the meta-analysis of ARIC AA and EA, and HCHS/SOL studies using  
461 StocSum. The x-axis represents the start physical chromosome and position of each variant

462 set and the y-axis represents  $-\log_{10}(P)$  from the StocSum variant set test corresponding to  
463 SMMAT. The red line indicates the genome-wide significance level on the log scale,  $-\log_{10}(3.7 \times 10^{-7})$ . All tests were performed on a high-performance computing server, with a  
464 single thread for each chromosome.  
465

#### 466 **LD score regression**

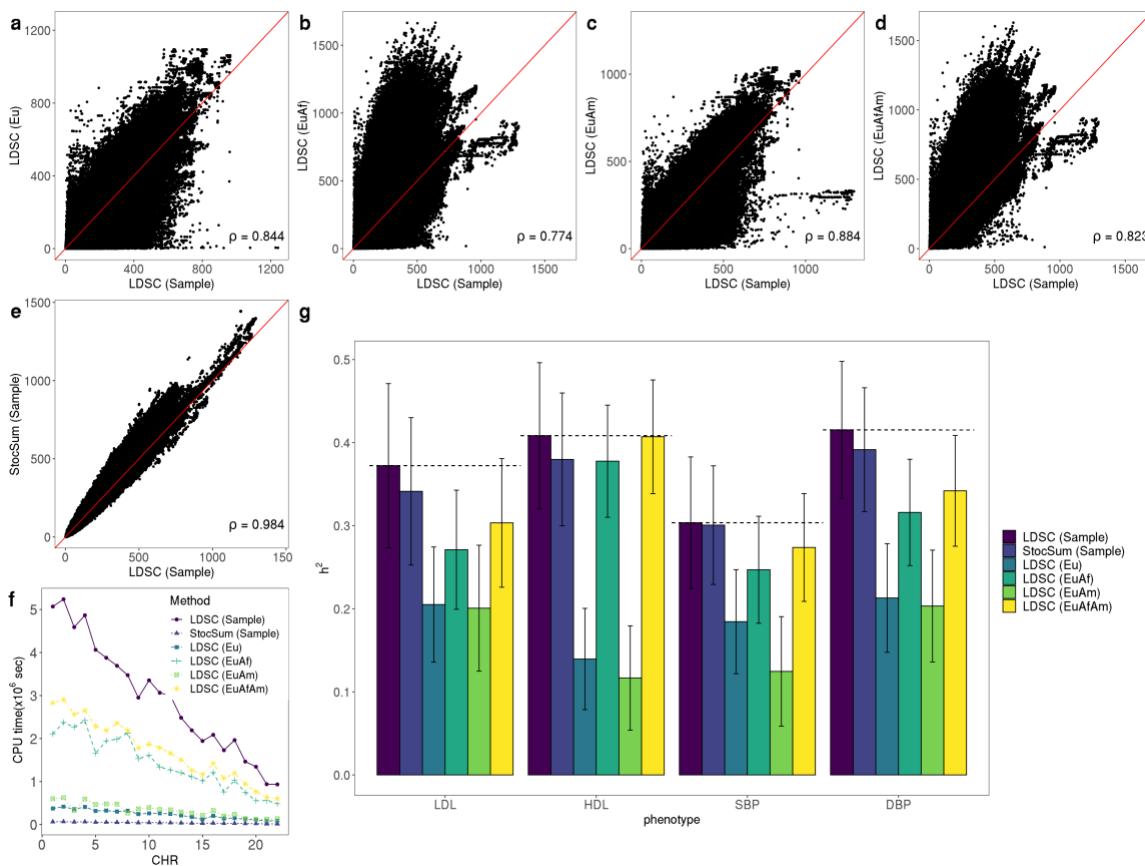
467 StocSum can also be used to extend the LD Score Regression (LDSC) framework<sup>14</sup> to  
468 underrepresented, admixed or isolated populations, without external reference panels. In  
469 this example, we compared LD scores and heritability estimates of four traits: LDL, high-  
470 density lipoprotein (HDL) cholesterol levels, systolic blood pressure (SBP), and diastolic  
471 blood pressure (DBP) from Hispanic/Latino adults in HCHS/SOL. LD scores were  
472 calculated using six different approaches: 1) StocSum (Sample): StocSum ( $B=1,000$ ) on  
473 HCHS/SOL study samples; 2) LDSC (Sample): LDSC using HCHS/SOL study samples as  
474 internal reference panels; 3) LDSC (Eu): LDSC using European populations from the 1000  
475 Genomes Project as external reference panels; 4) LDSC (EuAf): LDSC using European  
476 and African populations from the 1000 Genomes Project as external reference panels; 5)  
477 LDSC (EuAm): LDSC using European and American populations from the 1000 Genomes  
478 Project as external reference panels; and 6) LDSC (EuAfAm): LDSC using European,  
479 African and American populations from the 1000 Genomes Project as external reference  
480 panels. LD scores computed from StocSum (Sample) and LDSC using external reference  
481 panels were compared with those computed from LDSC (Sample).  
482

483 LD scores from StocSum (Sample) were much closer to those from LDSC (Sample) (**Fig. 6e**), compared to LDSC results using external reference panels (**Fig. 6a-d**). Moreover, there  
484 seems to be an upward bias for many variants in LDSC (EuAf) and LDSC (EuAfAm)  
485 results, when African populations from the 1000 Genomes Project were included in the  
486 reference panel (**Fig. 6b,d**), highlighting the challenges in selecting appropriate external  
487 reference panels for LD score estimation in underrepresented, admixed or isolated  
488 populations. StocSum (Sample) required only 1.4% of CPU time used by LDSC (Sample)  
489 (**Fig. 6f**). It was also 6-fold to 42-fold faster than LDSC using different external reference  
490 panels.  
491

492

493 Using LD scores from these six approaches, we compared heritability estimates of four  
494 traits LDL, HDL, SBP and DBP in HCHS/SOL (Fig. 6g). StocSum (Sample) results were  
495 consistently observed to be close to LDSC (Sample) heritability estimates, for all these  
496 traits. Heritability estimates from LDSC using external reference panels tended to be lower  
497 than LDSC (Sample), especially when African populations from the 1000 Genomes Project  
498 were excluded in the reference panel. For example, heritability estimates from LDSC  
499 (EuAm) were about 46.1%, 71.4%, 59.0%, and 51.1% lower compared to those from  
500 LDSC (Sample), for LDL, HDL, SBP, and DBP traits. Heritability estimates partitioned  
501 by different MAF bins also showed that StocSum (Sample) results were consistent with  
502 those from LDSC (Sample) (Fig. S14). Overall, StocSum is better suited for conducting  
503 LD score regression in Hispanic/Latino adults, while LDSC needs a reference panel that  
504 matches the LD structure in the study samples.

505



506

507 **Figure 6: StocSum in LD score regression and heritability estimation.** a-e, comparison  
508 of LD scores from LDSC (Sample) (x-axis) and different alternative methods (y-axis). a,  
509 LDSC (Eu). b, LDSC (EuAf). c, LDSC (EuAm). d, LDSC (EuAfAm). e, StocSum  
510 (Sample). Spearman's rank correlation coefficients are shown at the bottom right. f,  
511 comparison of CPU time between StocSum and LDSC in LD score calculations. The x axis  
512 represents the chromosome numbers and the y axis represents the CPU time in  $10^6$  seconds.  
513 g, heritability estimates using LD scores from LDSC and StocSum. The error bars show  
514 point estimates  $\pm$  standard errors. LD scores were estimated from LDSC (Sample) and  
515 StocSum (Sample) using HCHS/SOL study samples, or LDSC on external reference panels  
516 using European, African and/or American populations from the 1000 Genomes Project:  
517 LDSC (Eu), LDSC (EuAf), LDSC (EuAm), and LDSC (EuAfAm).

## 518 **Discussion**

519 We have developed and implemented StocSum, a novel framework for generating,  
520 managing, and using stochastic summary statistics for WGS studies. We showed that in all  
521 the example applications that use between-variant correlation or LD matrices, either from  
522 the study samples or external reference panels, such as conditional association tests, variant  
523 set tests and LD score regression, we could use a much smaller stochastic summary statistic  
524 matrix  $\mathbf{U}$  to replace the between-variant correlation or LD matrices, and flexibly extract  
525 the pairwise LD information between any variants on the same chromosome. This strategy  
526 was highly accurate and computationally efficient. The size of  $\mathbf{U}$  scales linearly with the  
527 number of genetic variants  $M$ , compared to quadratically in the form of traditional pairwise  
528 LD matrices. The computing time for the stochastic summary statistic matrix  $\mathbf{U}$  always  
529 scales linearly with both the sample size  $N$  and the number of genetic variants  $M$  (the same  
530 complexity with reading the data), regardless of any complex sample correlation structures.  
531 This matrix only needs to be computed once for each phenotype in both cross-sectional  
532 and longitudinal studies, and can be reused in single-variant tests, conditional association  
533 tests, and variant set tests with different variant set definitions.

534 StocSum leverages stochastic algorithms to reduce the computational burden in WGS  
535 studies. Similar algorithms have previously been applied to principal component

537 analysis<sup>77,78</sup>, heritability<sup>79</sup> and genetic correlation estimation<sup>80</sup>, and it is our hope that the  
538 StocSum framework can be extended to a wide range of other applications to genomic  
539 summary statistics that currently require external reference panels, thus facilitating use of  
540 genomic summary statistics from WGS studies. This is especially important for  
541 underrepresented, admixed, and/or isolated populations, for which appropriate reference  
542 panels are difficult to find. We have shown for variant set tests (Fig. S9-10) and LD score  
543 regression (Fig. 6) that external reference panels did not perform well even when all three  
544 ancestry populations for Hispanic/Latino adults were included, and the performance was  
545 even worse when a European-only reference panel was used. By using StocSum instead of  
546 external reference panels, more genetic research can be conducted in diverse populations  
547 that will equally benefit all humans.

548

549 StocSum will likely also facilitate international collaborations on genomic epidemiological  
550 research using WGS data, so that meta-analysis for rare genetic variants can be easily  
551 conducted without sharing individual-level WGS data across borders. Such collaborations  
552 have largely focused on common genetic variants in the past, by sharing genomic summary  
553 statistics. With the decreasing cost and increasing availability of WGS data, large-scale  
554 meta-analysis efforts on rare genetic variants are currently very difficult to coordinate, as  
555 variant sets determining how rare genetic variants should be grouped need to be pre-  
556 defined. In contrast, in the StocSum framework, researchers can combine the stochastic  
557 summary statistic matrices  $U$  from different studies first, and then decide how the variants  
558 should be grouped. When analysis plans change, there is no need to rerun any analyses  
559 using individual-level data, thus encouraging use of WGS data in international consortia.

560

561 WGS data are big in size and often difficult to share. Although large-scale studies such as  
562 the UK Biobank<sup>11</sup>, the TOPMed program<sup>10</sup>, and the CCDG initiative, have made plans to  
563 host their WGS data on cloud-computing platforms to facilitate access, it is still  
564 computationally expensive to directly analyzing individual-level data, making it financially  
565 difficult for small research groups to contribute to scientific discoveries using WGS data.  
566 The StocSum framework will democratize access to WGS resources, as we expect these  
567 high-level summary data will be generated by central analysis centers who are familiar

568 with and have direct access to individual-level phenotype and WGS data, and broadly  
569 shared with the scientific community. All downstream analyses using StocSum are free of  
570 the sample size  $N$  and could be performed on a laptop. It is also an eco-friendly strategy by  
571 avoiding different research teams running individual-level WGS data analyses on the same  
572 phenotypes, which are at least  $O(NM)$  operations for each team, thus saving a lot of  
573 electricity in computation.

574

575 There are also several limitations. We have demonstrated concordance of StocSum results  
576 as compared to methods that directly use individual-level data, for both common and rare  
577 variants, but it does not imply these results are statistically valid in all scenarios. For  
578 example, asymptotic  $P$  values from GMMAT may not be well-calibrated for extremely  
579 unbalanced cases:control ratios from Biobank studies<sup>81</sup>. This issue likely also exists for  
580 StocSum tests, given the concordance of StocSum and GMMAT results, and would require  
581 further adjustments or approximations. Moreover, although LD scores and heritability  
582 estimates from StocSum matched well with those from LDSC using internal reference  
583 panels (Fig. 6), these heritability estimates are likely underestimates and may not compare  
584 with estimates from other studies, due to the relatively small sample size in HCHS/SOL.  
585 Also, the choice of the number of random vector replicates  $B$  depends on the scientific  
586 questions to be investigated in downstream analyses. It does not depend on the sample size  
587  $N$ , although we note that for small studies with  $N < B$ , it might be more computationally  
588 expensive to use StocSum, compared to directly using individual-level data. In this study  
589 we have recommended using  $B=1,000$  in all applications, and it worked well in variant set  
590 tests for both regions with the number of variants  $q \leq B$  and  $q > B$  (Fig. S8). However,  
591 when it is of interest to test a very wide region with  $q$  being much greater than  $B$ , such as  
592 topologically associating domains and chromosome-wide association by class of histone  
593 markers<sup>82</sup>, the performance of StocSum is not guaranteed. Nevertheless, we expect  
594 StocSum to be a computationally efficient and eco-friendly framework for WGS studies  
595 that will facilitate genetic research in diverse populations, international collaborations, and  
596 equal access to WGS resources for the scientific community.

597 **Methods**

598 **Stochastic summary statistics**

599

600 We first define the basic null model in the StocSum framework. Under the null hypothesis  
601 of no genetic fixed effects  $H_0: \boldsymbol{\beta} = 0$ , model (Eq.(1)) (see Results) reduces to  
602

$$\mathbb{g}(\mu_{0i}) = \mathbf{X}_i \boldsymbol{\alpha} + b_i. \quad (2)$$

603 Here  $\mathbb{g}(\cdot)$  is a monotonic link function of  $\mu_{0i}$ , and  $\mu_{0i} = E(y_i | \mathbf{X}_i, b_i)$  is the conditional  
604 mean of the phenotype  $y_i$  under the null hypothesis  $H_0: \boldsymbol{\beta} = 0$ , given  $p$  covariates  $\mathbf{X}_i$   
605 (including an intercept) and random effects  $b_i$ , for individual  $i$  of  $N$  samples. Let  $\hat{\boldsymbol{\mu}}_0 =$   
606  $(\hat{\mu}_{01}, \hat{\mu}_{02}, \dots, \hat{\mu}_{0N})^T$  be a length  $N$  column vector for the estimated values of  $\mu_{0i}$ ,  $\hat{\phi}$  be an  
607 estimate of the dispersion parameter (or the residual variance for continuous traits in linear  
608 mixed models)  $\phi$ , and  $\hat{\tau}_k$  be the estimates for variance component parameters  $\tau_k$   
609 corresponding to  $N \times N$  relatedness matrices  $\boldsymbol{\Phi}_k$ , from the null model (Eq.(2)), we define  
610  $\mathbf{P} = \hat{\boldsymbol{\Sigma}}^{-1} - \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{X} (\mathbf{X}^T \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \hat{\boldsymbol{\Sigma}}^{-1}$  as the projection matrix, where  $\mathbf{X} =$   
611  $(\mathbf{X}_1^T \mathbf{X}_2^T \dots \mathbf{X}_N^T)^T$  is a  $N \times p$  covariate matrix, and  $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Omega}}^{-1} + \sum_{k=1}^K \hat{\tau}_k \boldsymbol{\Phi}_k$  with  $\hat{\boldsymbol{\Omega}}^{-1} =$   
612  $\hat{\phi} \mathbf{I}_n$  for continuous traits in linear mixed models, and  $\hat{\boldsymbol{\Omega}}^{-1} = \text{diag} \left\{ \frac{1}{\hat{\mu}_{0i}(1-\hat{\mu}_{0i})} \right\}$  for binary  
613 traits in logistic mixed models<sup>55</sup>.

614

615 StocSum leverages a length  $N$  random vector  $\mathbf{R}_b$  from a multivariate normal distribution  
616 with mean  $\mathbf{0}$  and covariance matrix  $\mathbf{P}$ , repeats this simulation process  $B$  times and  
617 combines  $\mathbf{R}_b$  ( $1 \leq b \leq B$ ) into an  $N \times B$  random matrix  $\mathbf{R} = (\mathbf{R}_1 \mathbf{R}_2 \dots \mathbf{R}_B)$ . In our  
618 implementation, we first decompose relatedness matrices  $\boldsymbol{\Phi}_k = \mathbf{Z}_k \mathbf{Z}_k^T$ , where  $\mathbf{Z}_k$  is an  
619  $N \times L_k$  matrix ( $L_k \leq N$ ). For low-rank relatedness matrices (such as those indicating  
620 observations from the same sample in longitudinal studies),  $\mathbf{Z}_k$  is often known as the  
621 random effect design matrix, with  $L_k$  being the rank of  $\boldsymbol{\Phi}_k$ . For sparse block-diagonal  
622 relatedness matrices (such as positive definite kinship matrices),  $\mathbf{Z}_k$  is the Cholesky  
623 decomposition of  $\boldsymbol{\Phi}_k$ , which is also sparse block-diagonal. We construct the  $N \times B$  random

624 matrix as  $\mathbf{R} = \sqrt{\hat{\phi}}\mathbf{r}_0 + \sum_{k=1}^K \sqrt{\hat{\tau}_k} \mathbf{Z}_k \mathbf{r}_k$ , in which  $\mathbf{r}_0$  is an  $N \times B$  random matrix and  $\mathbf{r}_k$   
625 ( $1 \leq k \leq K$ ) are  $L_k \times B$  random matrices, with all entries in  $\mathbf{r}_0$  and  $\mathbf{r}_k$  simulated from a  
626 standard normal distribution.

627

628 For an  $N \times M$  genotype matrix  $\mathbf{G}$  for  $M$  variants on the whole genome (or on one  
629 chromosome), the  $M \times B$  stochastic summary statistic matrix  $\mathbf{U}$  can be calculated as  $\mathbf{U} =$   
630  $\mathbf{G}^T \mathbf{R}$ . In the next sections, we describe how the stochastic summary statistics can be used  
631 in various downstream genetic analysis applications.

### 632 Single-variant tests

633 We are interested in conducting single-variant tests for the null hypothesis  $H_0: \beta = 0$ , using  
634 the score test. The GMMAT single-variant score is  $S = \frac{\mathbf{g}^T(\mathbf{y} - \hat{\mu}_0)}{\hat{\phi}}$ , where  $\mathbf{g} =$   
635  $(g_1 g_2 \dots g_N)^T$  is a length  $N$  column genotype vector for the variant of interest,  $\mathbf{y} =$   
636  $(y_1 y_2 \dots y_N)^T$  is a length  $N$  column vector for the phenotype (Chen et al., 2016). The  
637 variance of the score is  $Var(S|H_0) = \mathbf{g}^T \mathbf{P} \mathbf{g}$ .

638

639 Denote the  $j$ th row of the stochastic summary statistic matrix  $\mathbf{U}$  (for variant  $j$ ,  $1 \leq j \leq M$ )  
640 by a length  $B$  row vector  $\mathbf{U}_{j\cdot}$ , we can show that the variance  $Var(S|H_0)$  of single-variant  
641 score  $S$  for variant  $j$  can be estimated as  $\frac{1}{B} \mathbf{U}_{j\cdot} \mathbf{U}_{j\cdot}^T$ , without using any individual-level data.  
642 The asymptotic  $P$  value is then computed using the single-variant score  $S^{55}$  and its variance  
643 estimated from the stochastic summary statistic matrix  $\mathbf{U}$ , for each variant of interest.

### 644 Conditional association tests

645 Assume  $\dot{\mathbf{G}}$  is an  $N \times c$  genotype matrix for  $c \geq 1$  association genetic variants to be  
646 conditioned on and  $\mathbf{g}$  is a length  $N$  column genotype vector for the variant of interest in  
647 the conditional association test. The single-variant score conditional on the variant set  $\dot{\mathbf{G}}$  is

$$648 S_{\mathbf{g}|\dot{\mathbf{G}}} = S_{\mathbf{g}} - \mathbf{g}^T \mathbf{P} \dot{\mathbf{G}} (\dot{\mathbf{G}}^T \mathbf{P} \dot{\mathbf{G}})^{-1} S_{\dot{\mathbf{G}}}.$$

649 The variance of the conditional score is  $Var(S_{\mathbf{g}|\dot{\mathbf{G}}}) = \mathbf{g}^T \mathbf{P} \mathbf{g} - \mathbf{g}^T \mathbf{P} \dot{\mathbf{G}} (\dot{\mathbf{G}}^T \mathbf{P} \dot{\mathbf{G}})^{-1} \dot{\mathbf{G}}^T \mathbf{P} \mathbf{g}^{17}$ .

650

651 In the StocSum framework,  $S_g$  and  $\mathbf{U}_g$  are the single-variant score and stochastic summary  
 652 statistics corresponding to the variant of interest in the conditional association test and  $\mathbf{S}_{\hat{G}}$   
 653 (a length  $c$  vector) and  $\mathbf{U}_{\hat{G}}$  (a  $c \times B$  matrix) are the single-variant score and stochastic  
 654 summary statistics corresponding to the association variants to be conditioned on. The  
 655 conditional score can be computed as

656 
$$S_{g|\hat{G}} = S_g - \mathbf{U}_g \mathbf{U}_{\hat{G}}^T (\mathbf{U}_{\hat{G}} \mathbf{U}_{\hat{G}}^T)^{-1} \mathbf{S}_{\hat{G}},$$

657 and the conditional stochastic summary statistics can be computed as

658 
$$\mathbf{U}_{g|\hat{G}} = \mathbf{U}_g - \mathbf{U}_g \mathbf{U}_{\hat{G}}^T (\mathbf{U}_{\hat{G}} \mathbf{U}_{\hat{G}}^T)^{-1} \mathbf{U}_{\hat{G}}.$$

659

660 The variance  $Var(S_{g|\hat{G}})$  of the conditional score  $S_{g|\hat{G}}$  can be estimated as  $\frac{1}{B} \mathbf{U}_{g|\hat{G}} \mathbf{U}_{g|\hat{G}}^T$ .  
 661 The asymptotic  $P$  value is computed using the conditional score  $S_{g|\hat{G}}$  and its variance  
 662 estimated from the stochastic summary statistics  $\mathbf{U}_{g|\hat{G}}$ , for each variant of interest in the  
 663 conditional association test.

664

665 **Gene-environment interaction tests**

666 We introduce a general model for testing  $m$  gene-environment interaction (GEI) terms in  
 667 the GLMM framework. The full model including the genetic main effect and GEI effects  
 668 is

$$g(\mu_i) = \mathbf{X}_i \boldsymbol{\alpha} + g_i \beta + \mathbf{H}_i \boldsymbol{\gamma} + b_i, \quad (3)$$

669 where  $g_i$  is the genotype for the variant of interest for individual  $i$ ,  $\beta$  is a scalar of the  
 670 genetic main effect,  $\mathbf{H}_i$  is a length  $m$  row vector for the GEI terms, which include the  
 671 products of  $g_i$  and  $m$  environmental factors (a subset from  $p$  covariates in  $\mathbf{X}_i$ ), and  $\boldsymbol{\gamma}$  is a  
 672 length  $m$  column vector for GEI effects. We note that under the constraint  $\boldsymbol{\gamma} = 0$ ,  $\beta$  also  
 673 represents the marginal genetic effect. Other notations follow the null model (Eq.(2)).

674

675 The single-variant score for the marginal genetic effect is  $S_g = \frac{\mathbf{g}^T (\mathbf{y} - \hat{\mu}_0)}{\hat{\phi}}$  and its variance is  
 676  $Var(S_g) = \mathbf{g}^T \mathbf{P} \mathbf{g}$ . The single-variant score for the GEI effects is  $S_H = \frac{\mathbf{H}^T (\mathbf{y} - \hat{\mu}_0)}{\hat{\phi}}$  and its  
 677  $m \times m$  covariance matrix is  $Var(S_H) = \mathbf{H}^T \mathbf{P} \mathbf{H}$ , where  $\mathbf{H} = (\mathbf{H}_1^T \ \mathbf{H}_2^T \ \cdots \ \mathbf{H}_N^T)^T$  is a  $N \times m$

678 matrix for the GEI terms. The score for GEI effects adjusting for the marginal genetic effect  
679 can be approximated by  $S_{H|g} = S_H - \mathbf{H}^T \mathbf{P} \mathbf{g} (\mathbf{g}^T \mathbf{P} \mathbf{g})^{-1} S_g$ <sup>61</sup>, with a covariance matrix  
680  $Var(S_{H|g}) = \mathbf{H}^T \mathbf{P} \mathbf{H} - \mathbf{H}^T \mathbf{P} \mathbf{g} (\mathbf{g}^T \mathbf{P} \mathbf{g})^{-1} \mathbf{g}^T \mathbf{P} \mathbf{H}$ . The marginal genetic effect can be tested  
681 using the quadratic form  $S_g^T Var(S_g)^{-1} S_g$ , which follows a chi-square distribution with 1  
682 degree of freedom under the null hypothesis of no marginal genetic effects. The GEI effects  
683 can be tested using  $S_{H|g}^T Var(S_{H|g})^{-1} S_{H|g}$ , which follows a chi-square distribution with  
684  $m$  degrees of freedom under the null hypothesis of no gene-environment interactions. The  
685 joint test, which evaluates both marginal genetic effects and GEI effects, can be constructed  
686 by the sum of these two chi-square statistics, since  $S_H$  and  $S_{H|g}$  are asymptotically  
687 independent. The joint test statistic follows a chi-square distribution with  $1 + m$  degrees  
688 of freedom under the null hypothesis of no marginal genetic effects or gene-environment  
689 interactions.

690  
691 In the StocSum framework, we first compute stochastic summary statistics for the marginal  
692 genetic effect  $\mathbf{U}_g = \mathbf{g}^T \mathbf{R}$  and GEI effects  $\mathbf{U}_H = \mathbf{H}^T \mathbf{R}$  using individual-level data. We can  
693 use  $\frac{1}{B} \mathbf{U}_g \mathbf{U}_g^T$ ,  $\frac{1}{B} \mathbf{U}_H \mathbf{U}_H^T$ , and  $\frac{1}{B} \mathbf{U}_g \mathbf{U}_H^T$  to estimate the variance of the marginal genetic effect  
694 score  $Var(S_g)$ , the covariance matrix of the GEI effect score  $Var(S_H)$ , and the covariance  
695 of  $S_g$  and  $S_H$ , respectively. The adjusted scores can be constructed as  $S_{H|g} = S_H -$   
696  $\mathbf{U}_H \mathbf{U}_g^T (\mathbf{U}_g \mathbf{U}_g^T)^{-1} S_g$ , and its variance  $Var(S_{H|g})$  can be approximated as  $\frac{1}{B} \{ \mathbf{U}_H \mathbf{U}_H^T -$   
697  $\mathbf{U}_H \mathbf{U}_g^T (\mathbf{U}_g \mathbf{U}_g^T)^{-1} \mathbf{U}_g \mathbf{U}_H^T \}$ .  
698

## 699 Variant set tests

700 We include four variant set tests: the burden test<sup>34-37</sup>, SKAT<sup>38</sup>, SKAT-O<sup>83</sup>, and the efficient  
701 hybrid test of burden and SKAT<sup>21,39</sup>, in the StocSum framework. Here we consider a  
702 variant set including  $q$  genetic variants ( $q > 1$ ) and denote  $\tilde{\mathbf{S}}$  as a length  $q$  single-variant  
703 score vector, and  $\tilde{\mathbf{G}}$  as an  $N \times q$  genotype matrix (a subset of the  $N \times M$  genotype matrix  
704  $\mathbf{G}$  on the whole genome, or on one chromosome). We note that our examples are not a  
705 complete list of all variant set tests that are commonly used, but any other variant set tests

706 that would require  $q \times q$  covariance matrices could also be implemented using stochastic  
707 summary statistics.

708

709 The burden test statistic can be constructed as

$$710 \quad T_{Burden} = \tilde{\mathbf{S}}^T \mathbf{W} \mathbf{1}_q \mathbf{1}_q^T \mathbf{W} \tilde{\mathbf{S}},$$

711 where  $\mathbf{W} = diag\{w_j\}$  is a pre-specified  $q \times q$  diagonal weight matrix, and  $\mathbf{1}_q$  is a length  
712  $q$  vector of 1's. The weights can be a function of the MAF<sup>36,38</sup>, or functional annotation  
713 scores such as CADD<sup>84,85</sup>, FATHMM-XF<sup>86</sup>, and annotation principal components from  
714 STAAR<sup>87</sup>. Under the null hypothesis, the statistic  $T_{Burden}$  asymptotically follows  
715  $\xi_{Burden} \chi_1^2$ , where the scaling factor  $\xi_{Burden} = \mathbf{1}_q^T \mathbf{W} \tilde{\mathbf{G}}^T \mathbf{P} \tilde{\mathbf{G}} \mathbf{W} \mathbf{1}_q = \mathbf{1}_q^T \mathbf{W} \tilde{\mathbf{V}} \mathbf{W} \mathbf{1}_q$  (where  $\tilde{\mathbf{V}}$   
716 is a  $q \times q$  covariance matrix for the single-variant score vector  $\tilde{\mathbf{S}}$ ), and  $\chi_1^2$  is a chi-square  
717 distribution with 1 df. In the StocSum framework,  $\xi_{Burden}$  can be estimated as  
718  $\frac{1}{B} \mathbf{1}_q^T \mathbf{W} \tilde{\mathbf{U}} \tilde{\mathbf{U}}^T \mathbf{W} \mathbf{1}_q = \frac{1}{B} \tilde{\mathbf{u}}^T \tilde{\mathbf{u}}$ , where  $\tilde{\mathbf{U}}$  is a  $q \times B$  matrix (a subset of the  $M \times B$  stochastic  
719 summary statistic matrix  $\mathbf{U}$ ), and  $\tilde{\mathbf{u}} = \tilde{\mathbf{U}}^T \mathbf{W} \mathbf{1}_q$  is a length  $B$  vector (i.e., column sum of  
720  $\mathbf{W} \tilde{\mathbf{U}}$ ).

721

722 The SKAT statistic can be constructed as

$$723 \quad T_{SKAT} = \tilde{\mathbf{S}}^T \mathbf{W} \mathbf{W} \tilde{\mathbf{S}}.$$

724 Under the null hypothesis,  $T_{SKAT}$  asymptotically follows  $\sum_{j=1}^q \xi_{SKAT,j} \chi_{1,j}^2$ , where  $\chi_{1,j}^2$  are  
725 independent chi-square distributions with 1 df, and  $\xi_{SKAT,j}$  are the eigenvalues of  $\mathbf{E}_{SKAT} =$   
726  $\mathbf{W} \tilde{\mathbf{G}}^T \mathbf{P} \tilde{\mathbf{G}} \mathbf{W} = \mathbf{W} \tilde{\mathbf{V}} \mathbf{W}$ . In the StocSum framework,  $\xi_{SKAT,j}$  can be estimated as the square  
727 of the singular values of  $\frac{1}{\sqrt{B}} \mathbf{W} \tilde{\mathbf{U}}$  (**Supplementary Note 1**).

728

729 In SKAT-O, the variance component statistic  $T_\rho$  given a weight parameter  $\rho$  ( $0 \leq \rho \leq 1$ )  
730 is

$$731 \quad T_\rho = \rho T_{Burden} + (1 - \rho) T_{SKAT}.$$

732 If  $\rho = 1$ ,  $T_\rho$  becomes the burden test statistic  $T_{Burden}$ ; if  $\rho = 0$ ,  $T_\rho$  becomes the SKAT  
733 statistic  $T_{SKAT}$ . SKAT-O searches for an optimal  $\rho$  by minimizing the  $P$  value of  $T_\rho$ .  
734 Specifically, the  $q \times q$  weighted covariance matrix  $\mathbf{E}_{SKAT} = \mathbf{W} \tilde{\mathbf{V}} \mathbf{W}$  is decomposed into

735 two parts  $\Xi_{Burden} = \Xi_{SKAT} \mathbf{1}_q (\mathbf{1}_q^T \Xi_{SKAT} \mathbf{1}_q)^{-1} \mathbf{1}_q^T \Xi_{SKAT}$  and  $\Xi_{SKAT|Burden} = \Xi_{SKAT} -$   
 736  $\Xi_{Burden}$ , used in subsequent one-dimensional numerical integration to compute the SKAT-  
 737 O P value. In the StocSum framework,  $\Xi_{Burden}$  can be estimated as  $\frac{1}{B} \tilde{\mathbf{U}}_{Burden} \tilde{\mathbf{U}}_{Burden}^T$ ,  
 738 where  $\tilde{\mathbf{U}}_{Burden} = \mathbf{W} \tilde{\mathbf{U}} \tilde{\mathbf{u}} (\tilde{\mathbf{u}}^T \tilde{\mathbf{u}})^{-1} \tilde{\mathbf{u}}^T$ , and  $\Xi_{SKAT|Burden}$  can be estimated as  
 739  $\frac{1}{B} \tilde{\mathbf{U}}_{SKAT|Burden} \tilde{\mathbf{U}}_{SKAT|Burden}^T$ , where  $\tilde{\mathbf{U}}_{SKAT|Burden} = \mathbf{W} \tilde{\mathbf{U}} - \tilde{\mathbf{U}}_{Burden}$ .

740

741 In the efficient hybrid test to combine the burden test and SKAT, the adjusted SKAT  
 742 statistic  $T_{SKAT|Burden}$  can be approximated by

$$743 T_{SKAT|Burden} = \tilde{\mathbf{S}}^T \mathbf{W} \left\{ \mathbf{I}_q - \mathbf{1}_q (\mathbf{1}_q^T \Xi_{SKAT} \mathbf{1}_q)^{-1} \mathbf{1}_q^T \Xi_{SKAT} \right\} \left\{ \mathbf{I}_q - \Xi_{SKAT} \mathbf{1}_q (\mathbf{1}_q^T \Xi_{SKAT} \mathbf{1}_q)^{-1} \mathbf{1}_q^T \right\} \mathbf{W} \tilde{\mathbf{S}}.$$

745 Under the null hypothesis,  $T_{SKAT|Burden}$  asymptotically follows  $\sum_{j=1}^q \xi_{SKAT|Burden_j} \chi_{1,j}^2$ ,  
 746 where  $\chi_{1,j}^2$  are independent chi-square distributions with 1 df and  $\xi_{SKAT|Burden_j}$  are the  
 747 eigenvalues of  $\Xi_{SKAT|Burden}$ . In the StocSum framework, these eigenvalues can be  
 748 estimated as the square of the singular values of  $\frac{1}{\sqrt{B}} \tilde{\mathbf{U}}_{SKAT|Burden}$  (**Supplementary Note**  
 749 **2**).

750

## 751 **Meta-analysis**

752 In a traditional meta-analysis on a region with  $q$  genetic variants from  $L$  studies, we use  
 753 the single-variant scores  $\tilde{\mathbf{S}}_l$  and the covariance matrix  $\tilde{\mathbf{V}}_l$  from each study  $l$  ( $1 \leq l \leq L$ ).  
 754 The variant set meta-analysis can be performed using the summary scores  $\tilde{\mathbf{S}} = \sum_{l=1}^L \tilde{\mathbf{S}}_l$  and  
 755 the summary covariance matrix  $\tilde{\mathbf{V}} = \sum_{l=1}^L \tilde{\mathbf{V}}_l^{18,19,21,31,33}$ . The single-variant meta-analysis  
 756 only requires  $\tilde{\mathbf{S}}$  and the diagonal elements of  $\tilde{\mathbf{V}}$ . In the StocSum framework, we compute  
 757  $\tilde{\mathbf{U}} = \sum_{l=1}^L \tilde{\mathbf{U}}_l$  instead of  $\tilde{\mathbf{V}}$ . Assuming  $q < B$ , each column of  $\tilde{\mathbf{U}}_l$  follows a multivariate  
 758 normal distribution with mean  $\mathbf{0}$  and covariance matrix  $\tilde{\mathbf{V}}_l$ , and  $\tilde{\mathbf{U}}_l$  are independent across  
 759  $L$  studies assuming no sample overlaps or between-study relatedness. Therefore, each  
 760 column of  $\tilde{\mathbf{U}}$  follows a multivariate normal distribution with mean  $\mathbf{0}$  and covariance matrix  
 761  $\tilde{\mathbf{V}}$ . In our implementation, we first compute the stochastic summary statistic matrix  $\mathbf{U} =$

762  $\sum_{l=1}^L \mathbf{U}_l$  for all  $M$  genetic variants on the whole genome (or one chromosome), regardless  
763 of how variants should be grouped, and then extract  $q$  genetic variants by taking a subset  
764 of  $\mathbf{U}$  only when computing  $P$  values, for both single-variant meta-analysis and variant set  
765 meta-analysis.

766 **LD score regression**

767 LD Score Regression (LDSC) has been widely applied to GWAS summary statistics to  
768 estimate confounding bias, heritability explained by genotyped variants, heritability  
769 enrichments of functional categories, and genetic correlations<sup>14,15,88</sup>. The classical LDSC  
770 model can be written as

771 
$$E \left[ \chi^2_j | l_j \right] = \frac{Nh^2 l_j}{M} + Na + 1,$$

772 where  $\chi^2_j$  denotes the  $\chi^2$  statistic of variant  $j$  from GWAS summary statistics;  $l_j =$   
773  $\sum_k r_{jk}^2$  is the LD score of variant  $j$  with  $r_{jk}^2$  being the squared Pearson correlation  
774 coefficient of genotypes between variants  $j$  and  $k$ ,  $N$  is the sample size,  $M$  is the total  
775 number of variants,  $a$  is a measure of confounding bias, and  $h^2$  is the heritability of the  
776 phenotype. In practice, LDSC calculates  $l_j$  by summing up  $\hat{r}_{adj_{jk}}^2$  for all variants  $k$  in  
777 specific window around the index variant  $j$ . The adjusted correlation estimate  $\hat{r}_{adj_{jk}}$  can  
778 be computed from the sample correlation estimate  $\hat{r}_{jk}$  using

779 
$$\hat{r}_{adj_{jk}}^2 = \hat{r}_{jk}^2 - \frac{1 - \hat{r}_{jk}^2}{N-2}.$$

780 Sample correlation coefficients  $\hat{r}_{jk}$  can be estimated as  $\frac{w_j \mathbf{G}_{.j}^T \mathbf{L} \mathbf{G}_{.k} w_k}{N-1}$ , where  $\mathbf{G}_{.j}$  is the  $j$ th  
781 column of the genotype matrix  $\mathbf{G}$ , representing variant  $j$ ,  $\mathbf{L} = \left( \mathbf{I}_N - \mathbf{1}_N (\mathbf{1}_N^T \mathbf{1}_N)^{-1} \mathbf{1}_N^T \right)$   
782 is an  $N \times N$  idempotent projection matrix, and  $w_j = \frac{1}{\sqrt{2f_j(1-f_j)}}$  ( $f_j$  is the MAF of variant  
783  $j$ ) is a weight that standardizes  $\mathbf{G}_{.j}$  to a unit variance.

784  
785 In the StocSum framework, we construct the  $N \times B$  random matrix as  $\mathbf{R} = \mathbf{L} \mathbf{r}_0$ , where  $\mathbf{r}_0$   
786 is an  $N \times B$  random matrix with all entries simulated from a standard normal distribution.  
787 For an  $N \times M$  genotype matrix  $\mathbf{G}$  for all  $M$  genetic variants on the whole genome (or one

788 chromosome), we compute the  $M \times B$  stochastic summary statistic matrix  $\mathbf{U} = \mathbf{W}\mathbf{G}^T\mathbf{R}$ ,  
789 where  $\mathbf{W} = \text{diag}\{w_j\}$  is an  $M \times M$  diagonal weight matrix. For variant  $j$ , we subset  $M_j$   
790 variants within the flanking region (with a default window width of 1000 Kb) to get the  
791 corresponding  $M_j \times B$  subset  $\tilde{\mathbf{U}}$ . The adjusted correlation coefficient  $\tilde{r}_{adj_{jk}}$  for  $\tilde{r}_{jk}$  from  
792 StocSum is computed as (**Supplementary Note 3**)

$$793 \quad \tilde{r}_{adj_{jk}}^2 = \tilde{r}_{jk}^2 - \frac{1 - \tilde{r}_{jk}^2}{B-2} - \frac{1 - \tilde{r}_{jk}^2}{N-2}.$$

794 The LD score  $l_j$  of variant  $j$  could be estimated by summarizing stochastic summary  
795 statistics of  $M_j$  variants in flanking region,

$$796 \quad l_j = \sum_{k=1}^{M_j} \tilde{r}_{adj_{jk}}^2 = \left\{ \sum_{k=1}^{M_j} \left( 1 + \frac{1}{B-2} + \frac{1}{N-2} \right) \tilde{r}_{jk}^2 \right\} - \frac{M_j}{B-2} - \frac{M_j}{N-2} \\ 797 \quad = \left( 1 + \frac{1}{B-2} + \frac{1}{N-2} \right) \left( \frac{\tilde{\mathbf{U}} \tilde{\mathbf{U}}_{j.}^T}{B(N-1)} \circ \frac{\tilde{\mathbf{U}} \tilde{\mathbf{U}}_{j.}^T}{B(N-1)} \right)^T \mathbf{1}_{M_j} - \frac{M_j}{B-2} - \frac{M_j}{N-2}.$$

798 in which  $\circ$  denotes the Hadamard product, and  $\tilde{\mathbf{U}}_{j.}$  is the  $j$ th row of  $\tilde{\mathbf{U}}$ .

## 799 Whole genome sequence and phenotype data

800  
801 The Trans-Omics for Precision Medicine (TOPMed), sponsored by the National Heart,  
802 Lung and Blood Institute (NHLBI), generates scientific resources to enhance our  
803 understanding of fundamental biological processes that underlie heart, lung, blood and  
804 sleep disorders (HLBS)<sup>10</sup>. WGS of the TOPMed samples was performed over multiple  
805 studies, years and sequencing centers. The TOPMed freeze 8 WGS data include 138K  
806 samples from 72 studies. The sequence reads were aligned to the human genome build  
807 GRCh38 using BWA-MEM following the protocol published previously<sup>89</sup>. To perform  
808 variant quality control, a support vector machine classifier was trained on known variant  
809 sites (positive labels) and Mendelian inconsistent variants (negative labels). Further variant  
810 filtering was done for variants with excess heterozygosity and Mendelian discordance.  
811 Sample quality control measures included: concordance between annotated and inferred  
812 genetic sex, concordance between prior array genotype data and TOPMed WGS data, and  
813 pedigree checks<sup>10</sup>.

814

815 In this paper, our analysis includes genotypes and phenotypes from two TOPMed studies,  
816 Hispanic Community Health Study/Study of Latins (HCHS/SOL) and the Atherosclerosis  
817 Risk in Communities (ARIC) study.

818

819 **HCHS/SOL data.** The HCHS/SOL is a multi-center study of Hispanic/Latino populations  
820 with the goal of determining the role of acculturation in the prevalence and development  
821 of diseases, and to identify other traits that impact Hispanic/Latino health<sup>90</sup>. Participants  
822 were recruited using a multi-stage probability sample design, as described previously<sup>90,91</sup>.  
823 The HCHS/SOL is composed of six different background groups including Central  
824 Americans, Cubans, Dominicans, Mexicans, Puerto Ricans, and South Americans<sup>7</sup>. A total  
825 of 123,004,674 variants from 7,684 HCHS/SOL participants in TOPMed were available  
826 for genetic association analyses.

827

828 Low-density lipoprotein (LDL) cholesterol levels were used as an illustrating example in  
829 single-variant tests, conditional association tests, variant set tests, meta-analysis, and LD  
830 score regression. Additional phenotypes including high-density lipoprotein (HDL)  
831 cholesterol levels, systolic blood pressure (SBP), and diastolic blood pressure (DBP) were  
832 also used as examples in LD score regression. To account for the effect of lipid-lowering  
833 medication, LDL cholesterol levels for study participants who took statins were adjusted  
834 by dividing raw values by 0.7, following previous studies<sup>57,92,93</sup>. Both LDL and HDL  
835 cholesterol levels were set to missing for study participants with unknown statins use,  
836 unknown fibric/nicotinic acids use, or those who took only fibric/nicotinic acids but no  
837 statins. SBP and DBP were adjusted by adding 15 mmHg and 10 mmHg for study  
838 participants self-reporting use of any antihypertensive medication, respectively<sup>76</sup>. The  
839 waist-hip ratio (WHR) was used as an illustrating example in gene-environment interaction  
840 tests.

841

842 **ARIC data.** The cohort component of the ARIC study began in 1987, and each of the four  
843 ARIC field centers (Washington County, MD; Forsyth County, NC; Jackson, MS; and  
844 Minneapolis, MN) randomly selected and recruited a cohort sample of approximately 4,000  
845 individuals aged 45-64 from a defined population in their community. A total of 15,792

846 participants received an extensive examination, including medical, social, and  
847 demographic data<sup>94</sup>. These participants were examined with the first (baseline) exam  
848 occurring in 1987-89, the second in 1990-92, the third in 1993-95, the fourth in 1996-98,  
849 the fifth in 2011-13, and the sixth in 2016-17. The TOPMed WGS study over-sampled  
850 ARIC participants with incident venous thromboembolism (VTE). We removed  
851 samples/visits with missing phenotype (LDL) or covariates (age, sex, BMI, field center,  
852 and top five ancestry principal components), resulting in 26,668 observations from 6,327  
853 ARIC EA samples and 7,514 observations from 2,045 ARIC AA samples. After removing  
854 low-quality variants with a genotype call rate less than 90% and monomorphic markers,  
855 there were 91,715,717 and 69,958,574 variants in ARIC EA and AA samples, respectively.  
856

857 Longitudinal LDL cholesterol levels from the baseline exam until up to the 6th exam were  
858 used as an illustrating example in single-variant and variant set meta-analyses. To account  
859 for the effect of lipid-lowering medication, LDL cholesterol levels for study participants  
860 who took statins were adjusted by dividing raw values by 0.7<sup>57,92,93</sup>. LDL cholesterol levels  
861 were set to missing for study participants with unknown statins use, unknown cholesterol  
862 medication use, or inconsistent information from statins use and cholesterol medication  
863 use.  
864

865 **Reference data from 1000 Genomes.** Individual-level WGS data from the 1000 Genomes  
866 Project<sup>95</sup> were used as reference panels in fastBAT variant set tests and LD score  
867 regression. Only high-quality variants with a genotype call rate  $\geq 95\%$  and passed the  
868 quality control filters were included. Four reference panels were constructed with different  
869 combinations of super-populations: European (Eu), European and African (EuAf),  
870 European and American (EuAm), and European, African and American (EuAfAm), with  
871 23,654,568, 45,780,202, 31,334,904, and 49,350,7868 variants from 503, 894, 682, and  
872 1,073 samples, respectively.  
873

## 874 **Statistical Analyses**

875 **Single-variant tests.** We removed samples with missing values in the phenotype LDL  
876 cholesterol levels or covariates (age, sex, body mass index [BMI], field center, sampling

877 weight, Hispanic/Latino background groups, and top five ancestry principal components)  
878 and excluded variants with a genotype call rate less than 90% and monomorphic markers  
879 in single-variant test comparisons. After quality control, a total of 120,066,450 variants  
880 from 7,297 HCHS/SOL samples were available for analysis. We included age, age<sup>2</sup>, sex,  
881 age × sex, age<sup>2</sup> × sex, BMI, field center, sampling weight, Hispanic/Latino background  
882 groups and top five ancestry principal components as fixed-effects covariates. We rank-  
883 normalized residuals after regressing the phenotype LDL cholesterol levels on fixed-effects  
884 covariates, and then used them as the phenotype in downstream null model fitting and  
885 association tests<sup>96</sup>. Three random effects representing household, census block, and kinship  
886 effects were included to account for sample relatedness. We also allowed the residual  
887 variance to be different across 6 Hispanic/Latino background groups (i.e., Central  
888 American, Cuban, Dominican, Mexican, Puerto Rican, and South American), in a  
889 heteroscedastic linear mixed model<sup>7</sup> for both GMMAT and StocSum. The *P* values from  
890 StocSum were compared to those from GMMAT using individual-level data. The default  
891 value of the number of random vectors *B* in StocSum was set to 1,000. To benchmark the  
892 numerical accuracy and required computational resources, the number of random vectors  
893 *B* changed from 10 (StocSum (B=10)), 100 (StocSum (B=100)), 1,000 (StocSum  
894 (B=1,000)), to 10,000 (StocSum (B=10,000)).

895  
896 To compare with fastGWA<sup>60</sup> in single-variant analysis, we dropped household and census  
897 block random effects, and only included a kinship random effects to account for sample  
898 relatedness. We also assumed an equal residual variance across 6 Hispanic/Latino  
899 background groups in the linear mixed model for GMMAT and StocSum, to make a fair  
900 comparison with fastGWA.

901  
902 **Conditional association tests.** We performed conditional analyses for the seven regions  
903 associated with LDL at the genome-wide significance level of  $5 \times 10^{-8}$  from the single-  
904 variant analysis in HCHS/SOL (**Table S1**). For each region, we started with a sentinel  
905 variant with the smallest *P* value, and computed conditional association test *P* values for  
906 all variants in the flanking region (1 Mb) after adjusting for the sentinel variant. If there  
907 were any variants in a region with a conditional *P* <  $5 \times 10^{-8}$ , we then selected the variant

908 with the smallest conditional *P* value as the secondary association variant, and performed  
909 conditional analyses after adjusting for both association variants.

910

911 **Gene-environment interaction tests.** We compared gene-environment interaction tests in  
912 StocSum with MAGEE single-variant interaction tests using individual-level data. We  
913 focused on gene-sex interaction effects on an anthropometric phenotype waist-hip ratio  
914 (WHR) which shows strong evidence of sex dimorphism, using WGS data from  
915 HCHS/SOL. We included age, age<sup>2</sup>, sex, age × sex, age<sup>2</sup> × sex, BMI, field center, sampling  
916 weight, Hispanic/Latino background groups, top five ancestry principal components (PCs),  
917 and sex by top five ancestry PC interactions as fixed-effects covariates. After removing  
918 samples with missing values in the phenotype WHR or covariates, and variants with a  
919 genotype call rate less than 90% and monomorphic markers, a total of 122,076,760 variants  
920 from 7,636 HCHS/SOL samples were available for analysis. Similar to the single-variant  
921 analysis, we followed a two-step approach<sup>96</sup> and used rank-normalized WHR residuals as  
922 the phenotype in null model fitting and gene-sex interaction tests. We included three  
923 random effects representing household, census block and kinship effects to account for  
924 sample relatedness, and used a heteroscedastic linear mixed model by allowing the residual  
925 variance to be different across the 12 sex by Hispanic/Latino background groups. The  
926 marginal genetic effect, gene-sex interaction, and joint test *P* values from StocSum were  
927 compared to corresponding test results from MAGEE single-variant interaction tests using  
928 individual-level data.

929

930 **Variant set tests.** We compared variant set tests using StocSum versus SMMAT using  
931 individual-level data. After removing samples with missing values in the phenotype LDL  
932 cholesterol levels or covariates, and variants with a genotype call rate less than 90% and  
933 monomorphic markers, a total of 120,066,450 variants from 7,297 HCHS/SOL samples  
934 were available for analysis. We used the same null model as previously described in the  
935 single-variant tests for GMMAT and StocSum, and conducted a sliding window analysis<sup>43</sup>  
936 with 20kb non-overlapping windows. We applied a beta density function with parameters  
937 1 and 25 on the MAF as variant weights<sup>38</sup> in both SMMAT and StocSum. SMMAT requires  
938 individual-level data to conduct variant set tests. In contrast, StocSum directly uses the

939 single-variant summary statistics and stochastic summary statistics previously computed  
940 for single-variant tests.

941

942 To compare with fastBAT<sup>97</sup> in variant set tests, we used the same kinship-only null model  
943 with equal residual variance as previously described in the single-variant test comparison  
944 for fastGWA, GMMAT, and StocSum. We also changed variant weights using a beta  
945 density function with parameters 0.5 and 0.5 on the MAF (also known as the Madsen-  
946 Browning weights)<sup>98</sup>, equivalent to rescaling the genotypes to a unit variance in fastBAT.  
947 Four external reference panels from 1000 Genomes (Eu, EuAf, EuAm, EuAfAm), as well  
948 as an internal reference panel using the HCHS/SOL study samples, were used to estimate  
949 LD between variants in each set in fastBAT.

950

951 In a second example, we also applied StocSum to variant set tests using windows defined  
952 by functional genomic units. We collected Hi-C data generated from an *in situ* Hi-C  
953 protocol on human GM12878 B-lymphoblastoid cells<sup>49</sup>, in which the crosslinked DNA was  
954 pulled down followed by Illumina sequencing. The whole genome was split into non-  
955 overlapping segments with a bin size of 10kb (i.e., contact matrices were generated at base  
956 pair delimited resolutions of 10kb), and a total of 17,224 pairs of contacts were defined.  
957 Each segment pair can be considered as a long-distance DNA crosslink. We grouped  
958 variants from each contact pair as a variant set, including two 10kb windows which may  
959 not be located in close proximity on the primary structure of DNA (the linear sequence),  
960 to evaluate the performance of StocSum on variant sets that are physically farther away  
961 and not typically covered using fixed-size sliding windows.

962

963 **Meta-analysis.** We combined StocSum on LDL cholesterol levels from ARIC and  
964 HCHS/SOL in single-variant and variant set meta-analysis. For HCHS/SOL, we used  
965 single-variant summary statistics and stochastic summary statistics previously computed  
966 for single-variant tests on LDL cholesterol levels. For ARIC, we first fit two linear mixed  
967 models separately for EA and AA, treating LDL cholesterol levels from up to 6 visits as  
968 repeated measures for each participant, and then computed single-variant summary  
969 statistics and stochastic summary statistics. We included age, age<sup>2</sup>, sex, age × sex, age<sup>2</sup> ×

970 sex, BMI, field center, and top five ancestry principal components as fixed-effects  
971 covariates. We rank-normalized residuals after regressing the phenotype LDL cholesterol  
972 levels on fixed-effects covariates, and then used them as the phenotype<sup>96</sup>. In each ARIC  
973 dataset (EA and AA), variants with a genotype call rate less than 90% and monomorphic  
974 markers were excluded. After quality control, there were a total of 91,715,717 variants  
975 from 6,327 ARIC EA samples, and 69,958,574 variants from 2,045 ARIC AA samples.

976

977 We took the union of all variants and combined ARIC EA, ARIC AA, and HCHS/SOL  
978 summary statistics in a traditional single-variant meta-analysis using GMMAT, and a  
979 traditional variant set meta-analysis using SMMAT. In StocSum meta-analysis, we  
980 combined stochastic summary statistics from ARIC EA, ARIC AA and HCHS/SOL into a  
981 single file by adding together stochastic summary statistics for the same variant across  
982 three studies. We assigned 0 to both the single-variant summary statistic and stochastic  
983 summary statistic for a variant that was not observed in a study, since it did not contribute  
984 to the test statistic. In variant set meta-analysis, we applied a beta density function with  
985 parameters 1 and 25 on the MAF as variant weights, and conducted a 20kb sliding window  
986 analysis, for both SMMAT and StocSum.

987

988 **LD score regression.** In LD score regression, we only included common genetic variants  
989 with MAF  $\geq 1\%$  in HCHS/SOL. Following previous guidelines<sup>14,16,99,100</sup>, we excluded  
990 variants within the major histocompatibility complex (MHC; chromosome 6: 25-34Mb)  
991 and variants in regions with exceptionally long-range LD (**Table S6**). After quality control,  
992 11,190,311 common variants with MAF  $> 1\%$  from 7,289 HCHS/SOL study samples were  
993 used in StocSum to calculate LD score. We used single-variant summary statistics from  
994 GWAS of LDL, HDL, SBP and DBP in HCHS/SOL using GMMAT. Covariates included  
995 age, age<sup>2</sup>, sex, age  $\times$  sex, age<sup>2</sup>  $\times$  sex, BMI, field center, sampling weight, Hispanic/Latino  
996 background groups, and top five ancestry principal components. The same HCHS/SOL  
997 study samples were used as an internal reference panel in the LDSC program and StocSum  
998 to calculate LD scores, i.e., LDSC (Sample) and StocSum (Sample). With the same filters,  
999 four external reference panels from the 1000 Genomes Project were used in the LDSC  
1000 program to calculate LD scores, i.e., LDSC (Eu), LDSC (EuAf), LDSC (EuAm), LDSC

1001 (EuAfAm), including 9,092,238, 14,296,986, 9,410,628, 13,819,023 common variants  
1002 with MAF > 1% (1000 Genomes Project Consortium), from 503 Eu, 894 EuAf, 682 EuAm,  
1003 and 1,073 EuAfAm samples, respectively. With the LD scores from these internal and  
1004 external references, the LDSC program was used to estimate heritability. For both LDSC  
1005 and StocSum, we used a 1 Mb window around each index variant to calculate its LD score.  
1006  
1007 To evaluate the performance of StocSum, we also compared heritability estimates from  
1008 LDSC (Sample) and StocSum (Sample) partitioned by different MAF bins. Common  
1009 variants from HCHS/SOL and external reference panels were divided into 6 MAF bins,  
1010 i.e.,  $1\% < \text{MAF} \leq 5\%$ ,  $5\% < \text{MAF} \leq 10\%$ ,  $10\% < \text{MAF} \leq 20\%$ ,  $20\% < \text{MAF} \leq 30\%$ ,  $30\% < \text{MAF} \leq 40\%$ , and  $40\% < \text{MAF} \leq 50\%$ . Partitioned LD scores for different MAF bins  
1011 were calculated by LDSC and StocSum, i.e., LDSC (Sample), LDSC (Eu), LDSC (EuAf),  
1012 LDSC(EuAm), LDSC (EuAfAm), and StocSum (Sample). Partitioned heritability was  
1013 estimated by the LDSC program with summary statistics for the phenotype LDL and  
1014 partitioned LD scores.  
1015

## 1016 **Reference**

- 1017 1. Morris, A. P. *et al.* Large-scale association analysis provides insights into the  
1018 genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet* **44**, 981–  
1019 990 (2012).
- 1020 2. Manning, A. K. *et al.* A genome-wide approach accounting for body mass index  
1021 identifies genetic variants influencing fasting glycemic traits and insulin resistance.  
*Nat Genet* **44**, 659–669 (2012).
- 1023 3. Lambert, J.-C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new  
1024 susceptibility loci for Alzheimer's disease. *Nat Genet* **45**, 1452–1458 (2013).
- 1025 4. Global Lipids Genetics Consortium. Discovery and refinement of loci associated  
1026 with lipid levels. *Nat Genet* **45**, 1274–1283 (2013).
- 1027 5. Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for  
1028 obesity biology. *Nature* **518**, 197–206 (2015).
- 1029 6. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for  
1030 inflammatory bowel disease and highlight shared genetic risk across populations.  
*Nat Genet* **47**, 979–986 (2015).
- 1032 7. Conomos, M. P. *et al.* Genetic diversity and association studies in US  
1033 Hispanic/Latino populations: applications in the Hispanic Community Health  
1034 Study/Study of Latinos. *The American Journal of Human Genetics* **98**, 165–184  
1035 (2016).

1036 8. Graham, S. E. *et al.* The power of genetic diversity in genome-wide association  
1037 studies of lipids. *Nature* **600**, 675–679 (2021).

1038 9. Mikhaylova, A. V. *et al.* Whole-genome sequencing in diverse subjects identifies  
1039 genetic correlates of leukocyte traits: The NHLBI TOPMed program. *The  
1040 American Journal of Human Genetics* **108**, 1836–1851 (2021).

1041 10. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed  
1042 Program. *Nature* **590**, 290–299 (2021).

1043 11. Halldorsson, B. V. *et al.* The sequences of 150,119 genomes in the UK Biobank.  
1044 *Nature* **607**, 732–740 (2022).

1045 12. Lin, D. Y. & Zeng, D. Meta-analysis of genome-wide association studies: no  
1046 efficiency gain in using individual participant data. *Genetic Epidemiology: The  
1047 Official Publication of the International Genetic Epidemiology Society* **34**, 60–66  
1048 (2010).

1049 13. Han, B. & Eskin, E. Random-effects model aimed at discovering associations in  
1050 meta-analysis of genome-wide association studies. *The American Journal of  
1051 Human Genetics* **88**, 586–598 (2011).

1052 14. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from  
1053 polygenicity in genome-wide association studies. *Nat Genet* **47**, 291–295 (2015).

1054 15. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using  
1055 genome-wide association summary statistics. *Nat Genet* **47**, 1228–1235 (2015).

1056 16. Speed, D. & Balding, D. J. SumHer better estimates the SNP heritability of  
1057 complex traits from summary statistics. *Nat Genet* **51**, 277–284 (2019).

1058 17. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary  
1059 statistics identifies additional variants influencing complex traits. *Nat Genet* **44**,  
1060 369–375 (2012).

1061 18. Lee, S., Teslovich, T. M., Boehnke, M. & Lin, X. General framework for meta-  
1062 analysis of rare variants in sequencing association studies. *The American Journal  
1063 of Human Genetics* **93**, 42–53 (2013).

1064 19. Liu, D. J. *et al.* Meta-analysis of gene-level tests for rare variant association. *Nat  
1065 Genet* **46**, 200–204 (2014).

1066 20. Feng, S. *et al.* Methods for Association Analysis and Meta-Analysis of Rare  
1067 Variants in Families. *Genet Epidemiol* **39**, 227–238 (2015).

1068 21. Chen, H. *et al.* Efficient variant set mixed model association tests for continuous  
1069 and binary traits in large-scale whole-genome sequencing studies. *The American  
1070 Journal of Human Genetics* **104**, 260–274 (2019).

1071 22. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using  
1072 reference transcriptome data. *Nat Genet* **47**, 1091–1098 (2015).

1073 23. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide  
1074 association studies. *Nat Genet* **48**, 245–252 (2016).

1075 24. Zhu, X. *et al.* Meta-analysis of correlated traits via summary statistics from  
1076 GWASs with an application in hypertension. *The American Journal of Human  
1077 Genetics* **96**, 21–36 (2015).

1078 25. Liu, Z. & Lin, X. Multiple phenotype association tests using summary statistics in  
1079 genome-wide association studies. *Biometrics* **74**, 165–175 (2018).

1080 26. Turley, P. *et al.* Multi-trait analysis of genome-wide association summary statistics  
1081 using MTAG. *Nat Genet* **50**, 229–237 (2018).

1082 27. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and  
1083 traits. *Nat Genet* **47**, 1236–1241 (2015).

1084 28. Lu, Q. *et al.* A powerful approach to estimating annotation-stratified genetic  
1085 covariance via GWAS summary statistics. *The American Journal of Human  
1086 Genetics* **101**, 939–964 (2017).

1087 29. Wang, Y. *et al.* The 3D Genome Browser: a web-based browser for visualizing 3D  
1088 genome organization and long-range chromatin interactions. *Genome Biol* **19**, 1–  
1089 12 (2018).

1090 30. Zhang, Y., Qi, G., Park, J.-H. & Chatterjee, N. Estimation of complex effect-size  
1091 distributions using summary-level statistics from genome-wide association studies  
1092 across 32 complex traits. *Nat Genet* **50**, 1318–1326 (2018).

1093 31. Voorman, A., Brody, J., Chen, H., Lumley, T. & David, B. seqMeta: An R  
1094 package for meta-analyzing region-based tests of rare DNA variants. Preprint at  
1095 (2017).

1096 32. Zhan, X., Hu, Y., Li, B., Abecasis, G. R. & Liu, D. J. RVTESTS: an efficient and  
1097 comprehensive tool for rare variant association analysis using sequence data.  
1098 *Bioinformatics* **32**, 1423–1426 (2016).

1099 33. Feng, S., Liu, D., Zhan, X., Wing, M. K. & Abecasis, G. R. RAREMETAL: fast  
1100 and powerful meta-analysis for rare variants. *Bioinformatics* **30**, 2828–2829  
1101 (2014).

1102 34. Morgenthaler, S. & Thilly, W. G. A strategy to discover genes that carry multi-  
1103 allelic or mono-allelic risk for common diseases: a cohort allelic sums test  
1104 (CAST). *Mutation Research/Fundamental and Molecular Mechanisms of  
1105 Mutagenesis* **615**, 28–56 (2007).

1106 35. Li, B. & Leal, S. M. Methods for detecting associations with rare variants for  
1107 common diseases: application to analysis of sequence data. *The American Journal  
1108 of Human Genetics* **83**, 311–321 (2008).

1109 36. Madsen, B. E. & Browning, S. R. A groupwise association test for rare mutations  
1110 using a weighted sum statistic. *PLoS Genet* **5**, e1000384 (2009).

1111 37. Morris, A. P. & Zeggini, E. An evaluation of statistical approaches to rare variant  
1112 analysis in genetic association studies. *Genet Epidemiol* **34**, 188–193 (2010).

1113 38. Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the  
1114 sequence kernel association test. *The American Journal of Human Genetics* **89**,  
1115 82–93 (2011).

1116 39. Sun, J., Zheng, Y. & Hsu, L. A unified mixed-effects model for rare-variant  
1117 association in sequencing studies. *Genet Epidemiol* **37**, 334–344 (2013).

1118 40. Chen, H., Meigs, J. B. & Dupuis, J. Sequence kernel association test for  
1119 quantitative traits in family samples. *Genet Epidemiol* **37**, 196–204 (2013).

1120 41. Brody, J. A. *et al.* Analysis commons, a team approach to discovery in a big-data  
1121 environment for genetic epidemiology. *Nat Genet* **49**, 1560–1563 (2017).

1122 42. Li, Z. *et al.* A framework for detecting noncoding rare-variant associations of  
1123 large-scale whole-genome sequencing studies. *Nat Methods* 1–13 (2022).

1124 43. Morrison, A. C. *et al.* Practical approaches for whole-genome sequence analysis of  
1125 heart-and blood-related traits. *The American Journal of Human Genetics* **100**, 205–  
1126 215 (2017).

1127 44. Li, X. *et al.* Powerful, scalable and resource-efficient meta-analysis of rare variant  
1128 associations in large whole genome sequencing studies. *Nat Genet* 1–11 (2022).

1129 45. Li, G. *et al.* Extensive promoter-centered chromatin interactions provide a  
1130 topological basis for transcription regulation. *Cell* **148**, 84–98 (2012).

1131 46. Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction  
1132 landscape of gene promoters. *Nature* **489**, 109–113 (2012).

1133 47. Jin, F. *et al.* A high-resolution map of the three-dimensional chromatin interactome  
1134 in human cells. *Nature* **503**, 290–294 (2013).

1135 48. Heidari, N. *et al.* Genome-wide map of regulatory interactions in the human  
1136 genome. *Genome Res* **24**, 1905–1917 (2014).

1137 49. Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals  
1138 principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).

1139 50. Turner, A. W. *et al.* Single-nucleus chromatin accessibility profiling highlights  
1140 regulatory mechanisms of coronary artery disease risk. *Nat Genet* **54**, 804–816  
1141 (2022).

1142 51. Zhou, T. *et al.* Lupus enhancer risk variant causes dysregulation of IRF8 through  
1143 cooperative lncRNA and DNA methylation machinery. *Nat Commun* **13**, 1–16  
1144 (2022).

1145 52. Nasser, J. *et al.* Genome-wide enhancer maps link risk variants to disease genes.  
1146 *Nature* **593**, 238–243 (2021).

1147 53. McCullagh, P. & Nelder, J. Generalized Linear Models Second edition Chapman  
1148 & Hall. Preprint at (1989).

1149 54. Lin, X. & Zhang, D. Inference in generalized additive mixed models by using  
1150 smoothing splines. *J R Stat Soc Series B Stat Methodol* **61**, 381–400 (1999).

1151 55. Chen, H. *et al.* Control for population structure and relatedness for binary traits in  
1152 genetic association studies via logistic mixed models. *The American Journal of  
1153 Human Genetics* **98**, 653–666 (2016).

1154 56. Hoffmann, T. J. *et al.* A large electronic-health-record-based genome-wide study  
1155 of serum lipids. *Nat Genet* **50**, 401–413 (2018).

1156 57. Peloso, G. M. *et al.* Association of low-frequency and rare coding-sequence  
1157 variants with blood lipids and coronary heart disease in 56,000 whites and blacks.  
1158 *The American Journal of Human Genetics* **94**, 223–232 (2014).

1159 58. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for  
1160 blood lipids. *Nature* **466**, 707–713 (2010).

1161 59. Kurano, M. *et al.* Genome-wide association study of serum lipids confirms  
1162 previously reported associations as well as new associations of common SNPs  
1163 within PCSK7 gene with triglyceride. *J Hum Genet* **61**, 427–433 (2016).

1164 60. Jiang, L. *et al.* A resource-efficient tool for mixed model association analysis of  
1165 large-scale data. *Nat Genet* **51**, 1749–1755 (2019).

1166 61. Wang, X. *et al.* Efficient gene–environment interaction tests for large biobank-  
1167 scale sequencing studies. *Genet Epidemiol* **44**, 908–923 (2020).

1168 62. Christakoudi, S., Evangelou, E., Riboli, E. & Tsilidis, K. K. GWAS of allometric  
1169 body-shape indices in UK Biobank identifies loci suggesting associations with  
1170 morphogenesis, organogenesis, adrenal cell renewal and cancer. *Sci Rep* **11**, 1–18  
1171 (2021).

1172 63. Lotta, L. A. *et al.* Association of genetic variants related to gluteofemoral vs  
1173 abdominal fat distribution with type 2 diabetes, coronary disease, and  
1174 cardiovascular risk factors. *JAMA* **320**, 2553–2563 (2018).

1175 64. Pulit, S. L. *et al.* Meta-analysis of genome-wide association studies for body fat  
1176 distribution in 694 649 individuals of European ancestry. *Hum Mol Genet* **28**, 166–  
1177 174 (2019).

1178 65. Westerman, K. E. *et al.* GEM: scalable and flexible gene–environment interaction  
1179 analysis in millions of samples. *Bioinformatics* **37**, 3514–3520 (2021).

1180 66. Winkler, T. W. *et al.* The influence of age and sex on genetic associations with  
1181 adult body size and shape: a large-scale genome-wide interaction study. *PLoS  
1182 Genet* **11**, e1005378 (2015).

1183 67. Wood, A. C. *et al.* Identification of genetic loci simultaneously associated with  
1184 multiple cardiometabolic traits. *Nutrition, Metabolism and Cardiovascular  
1185 Diseases* **32**, 1027–1034 (2022).

1186 68. Zhu, Z. *et al.* Shared genetic and experimental links between obesity-related traits  
1187 and asthma subtypes in UK Biobank. *Journal of Allergy and Clinical Immunology*  
1188 **145**, 537–549 (2020).

1189 69. Consortium, 1000 Genomes Project. A global reference for human genetic  
1190 variation. *Nature* **526**, 68 (2015).

1191 70. Mumbach, M. R. *et al.* HiChIP: efficient and sensitive analysis of protein-directed  
1192 genome architecture. *Nat Methods* **13**, 919–922 (2016).

1193 71. Mumbach, M. R. *et al.* Enhancer connectome in primary human cells identifies  
1194 target genes of disease-associated DNA elements. *Nat Genet* **49**, 1602–1612  
1195 (2017).

1196 72. Salameh, T. J. *et al.* A supervised learning framework for chromatin loop detection  
1197 in genome-wide contact maps. *Nat Commun* **11**, 1–12 (2020).

1198 73. Noordam, R. *et al.* Multi-ancestry sleep-by-SNP interaction analysis in 126,926  
1199 individuals reveals lipid loci stratified by sleep duration. *Nat Commun* **10**, 1–13  
1200 (2019).

1201 74. Ripatti, P. *et al.* Polygenic hyperlipidemias and coronary artery disease risk. *Circ  
1202 Genom Precis Med* **13**, e002725 (2020).

1203 75. Tamai, K. *et al.* LDL-receptor-related proteins in Wnt signal transduction. *Nature*  
1204 **407**, 530–535 (2000).

1205 76. Wojcik, G. L. *et al.* Genetic analyses of diverse populations improves discovery  
1206 for complex traits. *Nature* **570**, 514–518 (2019).

1207 77. Galinsky, K. J. *et al.* Fast principal-component analysis reveals convergent  
1208 evolution of ADH1B in Europe and East Asia. *The American Journal of Human  
1209 Genetics* **98**, 456–472 (2016).

1210 78. Agrawal, A., Chiu, A. M., Le, M., Halperin, E. & Sankararaman, S. Scalable  
1211 probabilistic PCA for large-scale genetic variation data. *PLoS Genet* **16**, e1008773  
1212 (2020).

1213 79. Pazokitoroudi, A. *et al.* Efficient variance components analysis across millions of  
1214 genomes. *Nat Commun* **11**, 1–10 (2020).

1215 80. Wu, Y. *et al.* Fast estimation of genetic correlation for biobank-scale data. *The  
1216 American Journal of Human Genetics* **109**, 24–32 (2022).

1217 81. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample  
1218 relatedness in large-scale genetic association studies. *Nat Genet* **50**, 1335–1341  
1219 (2018).

1220 82. Lumley, T., Brody, J., Peloso, G., Morrison, A. & Rice, K. FastSKAT: Sequence  
1221 kernel association tests for very large sets of markers. *Genet Epidemiol* **42**, 516–  
1222 527 (2018).

1223 83. Lee, S., Wu, M. C. & Lin, X. Optimal tests for rare variant effects in sequencing  
1224 association studies. *Biostatistics* **13**, 762–775 (2012).

1225 84. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of  
1226 human genetic variants. *Nat Genet* **46**, 310–315 (2014).

1227 85. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD:  
1228 predicting the deleteriousness of variants throughout the human genome. *Nucleic  
1229 Acids Res* **47**, D886–D894 (2019).

1230 86. Rogers, M. F. *et al.* FATHMM-XF: accurate prediction of pathogenic point  
1231 mutations via extended features. *Bioinformatics* **34**, 511–513 (2018).

1232 87. Li, X. *et al.* Dynamic incorporation of multiple in silico functional annotations  
1233 empowers rare variant association analysis of large whole-genome sequencing  
1234 studies at scale. *Nat Genet* **52**, 969–983 (2020).

1235 88. Zheng, J. *et al.* LD Hub: a centralized database and web interface to perform LD  
1236 score regression that maximizes the potential of summary level GWAS data for  
1237 SNP heritability and genetic correlation analysis. *Bioinformatics* **33**, 272–279  
1238 (2017).

1239 89. Regier, A. A. *et al.* Functional equivalence of genome sequencing analysis  
1240 pipelines enables harmonized variant calling across human genetics projects. *Nat  
1241 Commun* **9**, 1–8 (2018).

1242 90. LaVange, L. M. *et al.* Sample design and cohort selection in the Hispanic  
1243 Community Health Study/Study of Latinos. *Ann Epidemiol* **20**, 642–649 (2010).

1244 91. Sorlie, P. D. *et al.* Design and implementation of the Hispanic community health  
1245 study/study of Latinos. *Ann Epidemiol* **20**, 629–641 (2010).

1246 92. Baigent, C. Cholesterol Treatment Trialists' (CTT) Collaborators: Efficacy and  
1247 safety of cholesterol-lowering treatment: prospective meta-analysis of data from  
1248 90,056 participants in 14 randomised trials of statins. *Lancet* **366**, 1267–1278  
1249 (2005).

1250 93. Klarin, D. *et al.* Genetics of blood lipids among~ 300,000 multi-ethnic participants  
1251 of the Million Veteran Program. *Nat Genet* **50**, 1514–1523 (2018).

1252 94. Wright, J. D., Folsom, A. R., Coresh, J., Sharrett, A. R., Couper, D., Wagenknecht,  
1253 L. E., & Heiss, G. The ARIC (Atherosclerosis Risk in Communities) study: JACC  
1254 focus seminar 3/8. *Journal of the American College of Cardiology* **77**, 2939–2959  
1255 (2021).

1256 95. 1000 Genomes Project Consortium. A map of human genome variation from  
1257 population scale sequencing. *Nature* **467**, 1061 (2010).

1258 96. Sofer, T. *et al.* A fully adjusted two-stage procedure for rank-normalization in  
1259 genetic association studies. *Genet Epidemiol* **43**, 263–275 (2019).

1260 97. Bakshi, A. *et al.* Fast set-based association analysis using summary data from  
1261 GWAS identifies novel gene loci for human complex traits. *Sci Rep* **6**, 1–9 (2016).

1262 98. Madsen, B. E. & Browning, S. R. A groupwise association test for rare mutations  
1263 using a weighted sum statistic. *PLoS Genet* **5**, e1000384 (2009).

1264 99. Speed, D. *et al.* Reevaluation of SNP heritability in complex human traits. *Nat*  
1265 *Genet* **49**, 986–992 (2017).

1266 100. De Vlaming, R., Johannesson, M., Magnusson, P. K. E., Ikram, M. A. & Visscher,  
1267 P. M. Equivalence of LD-score regression and individual-level-data methods.  
1268 *Biorxiv* 211821 (2017).

1269

1270 **Acknowledgements**

1271 Molecular data for the Trans-Omics in Precision Medicine (TOPMed) program was  
1272 supported by the National Heart, Lung and Blood Institute (NHLBI). Whole genome  
1273 sequencing for “NHLBI TOPMed - NHGRI CCDG: Hispanic Community Health  
1274 Study/Study of Latinos (HCHS/SOL) (phs001395.v1.p1)” was performed at Baylor  
1275 College of Medicine Human Genome Sequencing Center (HHSN268201600033I). Whole  
1276 genome sequencing for “NHLBI TOPMed - NHGRI CCDG: Atherosclerosis Risk in  
1277 Communities (ARIC) (phs001211.v1.p1)” was performed at Baylor College of Medicine  
1278 Human Genome Sequencing Center (3U54HG003273-12S2; HHSN268201500015C) and  
1279 the Broad Institute Genomics Platform (3R01HL092577-06S1). Core support including  
1280 centralized genomic read mapping and genotype calling, along with variant quality metrics  
1281 and filtering were provided by the TOPMed Informatics Research Center (3R01HL-  
1282 117626-02S1; contract HHSN268201800002I). Core support including phenotype  
1283 harmonization, data management, sample-identity QC, and general program coordination  
1284 were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-  
1285 120393; contract HHSN268201800001I). We gratefully acknowledge the studies and  
1286 participants who provided biological samples and data for TOPMed.

1287

1288 The Genome Sequencing Program (GSP) was funded by the National Human Genome  
1289 Research Institute (NHGRI), the National Heart, Lung, and Blood Institute (NHLBI), and  
1290 the National Eye Institute (NEI). The GSP Coordinating Center (U24 HG008956)  
1291 contributed to cross program scientific initiatives and provided logistical and general study  
1292 coordination. The Centers for Common Disease Genomics (CCDG) program was

1293 supported by NHGRI and NHLBI, and whole genome sequencing was performed at the  
1294 Baylor College of Medicine Human Genome Sequencing Center (UM1 HG008898).

1295

1296 The Hispanic Community Health Study/Study of Latinos is a collaborative study supported  
1297 by contracts from the National Heart, Lung, and Blood Institute (NHLBI) to the University  
1298 of North Carolina (HHSN268201300001I / N01-HC-65233), University of Miami  
1299 (HHSN268201300004I / N01-HC-65234), Albert Einstein College of Medicine  
1300 (HHSN268201300002I / N01-HC-65235), University of Illinois at Chicago –  
1301 HHSN268201300003I / N01-HC-65236 Northwestern Univ), and San Diego State  
1302 University (HHSN268201300005I / N01-HC-65237). The following  
1303 Institutes/Centers/Offices have contributed to the HCHS/SOL through a transfer of funds  
1304 to the NHLBI: National Institute on Minority Health and Health Disparities, National  
1305 Institute on Deafness and Other Communication Disorders, National Institute of Dental  
1306 and Craniofacial Research, National Institute of Diabetes and Digestive and Kidney  
1307 Diseases, National Institute of Neurological Disorders and Stroke, NIH Institution-Office  
1308 of Dietary Supplements.

1309

1310 The Atherosclerosis Risk in Communities study has been funded in whole or in part with  
1311 Federal funds from the National Heart, Lung, and Blood Institute, National Institutes of  
1312 Health, Department of Health and Human Services, under Contract nos.  
1313 (75N92022D00001, 75N92022D00002, 75N92022D00003, 75N92022D00004,  
1314 75N92022D00005). The authors thank the staff and participants of the ARIC study for their  
1315 important contributions.

1316

1317 This work was supported by NHLBI grant R01 HL145025.

## 1318 **Competing interests**

1319 The authors declare no competing interests.

1320 **Supplementary Note**

1321 **1. Approximating eigenvalues in variant set tests using singular values from StocSum.**

1322 For  $q$  variants ( $q < B$ ), the  $q \times q$  covariance matrix used in variant set tests is  
1323  $\tilde{\mathbf{V}} = \tilde{\mathbf{G}}^T \mathbf{P} \tilde{\mathbf{G}}$ . In the StocSum framework, we compute a  $q \times B$  matrix  $\tilde{\mathbf{U}} = \tilde{\mathbf{G}}^T \mathbf{R}$ , where  
1324 each column  $\mathbf{R}_b$  ( $1 \leq b \leq B$ ) of an  $N \times B$  random matrix  $\mathbf{R} = (\mathbf{R}_1 \mathbf{R}_2 \dots \mathbf{R}_B)$  is a length  
1325  $N$  random vector generated from a multivariate normal distribution with mean  $\mathbf{0}$  and  
1326 covariance matrix  $\mathbf{P}$ . Each column  $\tilde{\mathbf{U}}_b = \tilde{\mathbf{G}}^T \mathbf{R}_b$  of  $\tilde{\mathbf{U}}$  then follows a multivariate normal  
1327 distribution with mean  $\mathbf{0}$  and covariance matrix  $\tilde{\mathbf{V}}$ , and the  $B$  columns of  $\tilde{\mathbf{U}}$  are  
1328 independent and identically distributed. Therefore, when  $B$  is large,  $\frac{1}{B} \tilde{\mathbf{U}} \tilde{\mathbf{U}}^T$  converges to  
1329 the covariance matrix  $\tilde{\mathbf{V}}$ . For  $\mathbf{\Xi}_{SKAT} = \mathbf{W} \tilde{\mathbf{V}} \mathbf{W}$  in SKAT, we can use  $\frac{1}{B} \mathbf{W} \tilde{\mathbf{U}} \tilde{\mathbf{U}}^T \mathbf{W}$  to estimate  
1330  $\mathbf{\Xi}_{SKAT}$ .

1331

1332 We compute the singular value decomposition  $\frac{1}{\sqrt{B}} \mathbf{W} \tilde{\mathbf{U}} = \mathbf{Q}_L \mathbf{D} \mathbf{Q}_R^T$ , where  $r \leq \min(q, B)$   
1333 is the rank of  $\frac{1}{\sqrt{B}} \mathbf{W} \tilde{\mathbf{U}}$ ,  $\mathbf{Q}_L$  and  $\mathbf{Q}_R$  are  $q \times r$  and  $B \times r$  semi-unitary matrices, respectively  
1334 ( $\mathbf{Q}_L^T \mathbf{Q}_L = \mathbf{Q}_R^T \mathbf{Q}_R = \mathbf{I}_r$ ), and  $\mathbf{D}$  is an  $r \times r$  diagonal matrix with elements being the  
1335 singular values of  $\frac{1}{\sqrt{B}} \mathbf{W} \tilde{\mathbf{U}}$ . As we use  $\frac{1}{B} \mathbf{W} \tilde{\mathbf{U}} \tilde{\mathbf{U}}^T \mathbf{W}$  to estimate  $\mathbf{\Xi}_{SKAT}$ , where  $\frac{1}{B} \mathbf{W} \tilde{\mathbf{U}} \tilde{\mathbf{U}}^T \mathbf{W} =$   
1336  $\mathbf{Q}_L \mathbf{D} \mathbf{Q}_R^T \mathbf{Q}_R \mathbf{D} \mathbf{Q}_L^T = \mathbf{Q}_L \mathbf{D} \mathbf{D} \mathbf{Q}_L^T$ , elements in the  $r \times r$  diagonal matrix  $\mathbf{D} \mathbf{D}$  (the square of  
1337 the singular values of  $\frac{1}{\sqrt{B}} \mathbf{W} \tilde{\mathbf{U}}$ ) can be used to estimate the eigenvalues of  $\mathbf{\Xi}_{SKAT}$  when  $r =$   
1338  $q$ . If  $r < q$  (for example, when testing a large genomic region with  $q > B$ ), we could only  
1339 estimate the top  $r$  (which is usually equal to  $B$  when  $q > B$ ) eigenvalues of  $\mathbf{\Xi}_{SKAT}$  using  
1340 the singular values of  $\frac{1}{\sqrt{B}} \mathbf{W} \tilde{\mathbf{U}}$ .

1341

1342

1343 **2. Approximating eigenvalues in the efficient hybrid variant set test using singular**  
1344 **values from StocSum.**

1345

1346 In the efficient hybrid variant set test to combine the burden test and SKAT, the adjusted  
 1347 SKAT statistic asymptotically follows a weighted sum of independent chi-square  
 1348 distributions with 1 df, where the weights are the eigenvalues of

1349

$$1350 \quad \boldsymbol{\Xi}_{SKAT|Burden} = \boldsymbol{\Xi}_{SKAT} - \boldsymbol{\Xi}_{Burden} = \boldsymbol{\Xi}_{SKAT} - \boldsymbol{\Xi}_{SKAT} \mathbf{1}_q (\mathbf{1}_q^T \boldsymbol{\Xi}_{SKAT} \mathbf{1}_q)^{-1} \mathbf{1}_q^T \boldsymbol{\Xi}_{SKAT}.$$

1351 As we use  $\frac{1}{B} \mathbf{W} \tilde{\mathbf{U}} \tilde{\mathbf{U}}^T \mathbf{W}$  to estimate  $\boldsymbol{\Xi}_{SKAT}$  (**Supplementary Note 1**), let  $\tilde{\mathbf{u}} = \tilde{\mathbf{U}}^T \mathbf{W} \mathbf{1}_q$  be a  
 1352 length  $B$  vector denoting the column sum of  $\mathbf{W} \tilde{\mathbf{U}}$ , and define  $\tilde{\mathbf{U}}_{Burden} =$   
 1353  $\mathbf{W} \tilde{\mathbf{U}} \tilde{\mathbf{u}} (\tilde{\mathbf{u}}^T \tilde{\mathbf{u}})^{-1} \tilde{\mathbf{u}}^T$ ,  $\tilde{\mathbf{U}}_{SKAT|Burden} = \mathbf{W} \tilde{\mathbf{U}} - \tilde{\mathbf{U}}_{Burden} = \mathbf{W} \tilde{\mathbf{U}} - \mathbf{W} \tilde{\mathbf{U}} \tilde{\mathbf{u}} (\tilde{\mathbf{u}}^T \tilde{\mathbf{u}})^{-1} \tilde{\mathbf{u}}^T$  (see  
 1354 **Methods**), it follows that

$$1355 \quad \boldsymbol{\Xi}_{SKAT|Burden} \approx \frac{1}{B} \mathbf{W} \tilde{\mathbf{U}} \tilde{\mathbf{U}}^T \mathbf{W} \mathbf{1}_q (\mathbf{1}_q^T \mathbf{W} \tilde{\mathbf{U}} \tilde{\mathbf{U}}^T \mathbf{W} \mathbf{1}_q)^{-1} \mathbf{1}_q^T \mathbf{W} \tilde{\mathbf{U}} \tilde{\mathbf{U}}^T \mathbf{W}$$

1356

$$1357 \quad = \frac{1}{B} \mathbf{W} \tilde{\mathbf{U}} \tilde{\mathbf{U}}^T \mathbf{W} - \frac{1}{B} \mathbf{W} \tilde{\mathbf{U}} \tilde{\mathbf{u}} (\tilde{\mathbf{u}}^T \tilde{\mathbf{u}})^{-1} \tilde{\mathbf{u}}^T \tilde{\mathbf{U}}^T \mathbf{W}$$

$$1358 \quad = \frac{1}{B} (\mathbf{W} \tilde{\mathbf{U}} - \mathbf{W} \tilde{\mathbf{U}} \tilde{\mathbf{u}} (\tilde{\mathbf{u}}^T \tilde{\mathbf{u}})^{-1} \tilde{\mathbf{u}}^T) (\mathbf{W} \tilde{\mathbf{U}} - \mathbf{W} \tilde{\mathbf{U}} \tilde{\mathbf{u}} (\tilde{\mathbf{u}}^T \tilde{\mathbf{u}})^{-1} \tilde{\mathbf{u}}^T)^T$$

$$1359 \quad = \frac{1}{B} \tilde{\mathbf{U}}_{SKAT|Burden} \tilde{\mathbf{U}}_{SKAT|Burden}^T.$$

1360 Therefore, similar to **Supplementary Note 1**, the eigenvalues of the  $q \times q$  matrix  
 1361  $\boldsymbol{\Xi}_{SKAT|Burden}$  can be estimated using the square of the single values of the  $q \times B$  matrix

$$1362 \quad \frac{1}{\sqrt{B}} \tilde{\mathbf{U}}_{SKAT|Burden}.$$

1363

1364

### 1365 3. Derivation of the adjusted correlation coefficient in the StocSum framework

1366

1367 Let  $r_{jk}$  be the Pearson correlation coefficient between variants  $j$  and  $k$ , the sample  
 1368 correlation coefficient  $\hat{r}_{jk}$  can be estimated using individual-level centered and rescaled

1369 genotypes (with mean 0 and variance 1), namely,  $\hat{r}_{jk} = \frac{w_j \mathbf{G}_{.j}^T \mathbf{L} \mathbf{G}_{.k} w_k}{N-1}$ , where  $\mathbf{G}_{.j}$  and  $\mathbf{G}_{.k}$  are  
 1370 the  $j$ th and  $k$ th columns of the full genotype matrix  $\mathbf{G}$ , representing variants  $j$  and  $k$ ,  $w_j$   
 1371 and  $w_k$  are rescaling weights that standardize genotypes to a unit variance, and  $\mathbf{L} =$

1372  $(\mathbf{I}_N - \mathbf{1}_N(\mathbf{1}_N^T \mathbf{1}_N)^{-1} \mathbf{1}_N^T)$  is an  $N \times N$  idempotent projection matrix that centers the  
 1373 genotypes (see **Methods**). The asymptotic distribution of  $\hat{r}_{jk}$  is given by

1374 
$$\sqrt{N}(\hat{r}_{jk} - r_{jk}) \rightarrow N(0, (1 - r_{jk}^2)^2).$$

1375 Therefore,

1376 
$$E(\hat{r}_{jk}^2) = E(\hat{r}_{jk})^2 + Var(\hat{r}_{jk}) = r_{jk}^2 + \frac{(1 - r_{jk}^2)^2}{N} \approx r_{jk}^2 + \frac{1 - r_{jk}^2}{N}.$$

1377 In LD score regression, the higher order term is ignored and the adjusted squared  
 1378 correlation coefficient is computed as  $\hat{r}_{adj_{jk}}^2 = \hat{r}_{jk}^2 - \frac{1 - \hat{r}_{jk}^2}{N-2}$  to reduce the bias (Bulik-  
 1379 Sullivan et al., 2015).

1380

1381 In the StocSum framework, we compute the  $M \times B$  stochastic summary statistic matrix  
 1382  $\mathbf{U} = \mathbf{W}\mathbf{G}^T\mathbf{R}$ , where  $\mathbf{W} = diag\{w_j\}$  is an  $M \times M$  diagonal weight matrix, and  $\mathbf{G}$  is an  
 1383  $N \times M$  genotype matrix for all  $M$  genetic variants on the whole genome (or one  
 1384 chromosome). We use  $\mathbf{U}_j$  and  $\mathbf{U}_k$  to denote length  $B$  row vectors from  $\mathbf{U}$  for variants  $j$   
 1385 and  $k$ , respectively. Then we can use  $\frac{1}{B}\mathbf{U}_j\mathbf{U}_k^T$  to estimate  $w_j \mathbf{G}_{j\cdot}^T \mathbf{L} \mathbf{G}_{k\cdot} w_k$ , and therefore  
 1386  $\tilde{r}_{jk} = \frac{\tilde{\mathbf{U}}_j \cdot \tilde{\mathbf{U}}_k^T}{B(N-1)}$  converges to  $\hat{r}_{jk} = \frac{w_j \mathbf{G}_{j\cdot}^T \mathbf{L} \mathbf{G}_{k\cdot} w_k}{N-1}$  when  $B$  is large. Given  $\hat{r}_{jk}$ , the asymptotic  
 1387 distribution of  $\tilde{r}_{jk} | \hat{r}_{jk}$  follows

1388 
$$\sqrt{B}(\tilde{r}_{jk} - \hat{r}_{jk}) \rightarrow N(0, (1 - \hat{r}_{jk}^2)^2).$$

1389

1390 Therefore,

1391 
$$E(\tilde{r}_{jk}) = E\{E(\tilde{r}_{jk} | \hat{r}_{jk})\} = E(\hat{r}_{jk}) = r_{jk},$$

1392 and ignoring the higher order terms in the variance, we have

1393 
$$Var(\tilde{r}_{jk}) = E\{Var(\tilde{r}_{jk} | \hat{r}_{jk})\} + Var\{E(\tilde{r}_{jk} | \hat{r}_{jk})\} \approx E\left\{\frac{1 - \hat{r}_{jk}^2}{B}\right\} + Var\{\hat{r}_{jk}\}$$
  
 1394 
$$= \frac{1 - r_{jk}^2 - \frac{1 - r_{jk}^2}{N}}{B} + \frac{1 - \rho_{jk}^2}{N}.$$

1395 Hence,

1396 
$$E(\tilde{r}_{jk}^2) = E(\tilde{r}_{jk})^2 + Var(\tilde{r}_{jk}) = r_{jk}^2 + \frac{1-r_{jk}^2 - \frac{1-r_{jk}^2}{N}}{B} + \frac{1-r_{jk}^2}{N} \approx r_{jk}^2 + \frac{1-r_{jk}^2}{B} + \frac{1-r_{jk}^2}{N}.$$

1397 The term  $\frac{1-r_{jk}^2}{NB}$  is ignored as both  $N$  and  $B$  are large. Following the same adjustment in  
1398 LDSC (Bulik-Sullivan et al., 2015), we calculate adjusted correlation coefficient  $\tilde{r}_{adj_{jk}}$  for  
1399  $\tilde{r}_{jk}$  from StocSum using

1400 
$$\tilde{r}_{adj_{jk}}^2 = \tilde{r}_{jk}^2 - \frac{1 - \tilde{r}_{jk}^2}{B - 2} - \frac{1 - \tilde{r}_{jk}^2}{N - 2}.$$

1401 **Supplementary Tables**

1402

1403 **Table S1.** Significant association regions with LDL cholesterol levels from single-variant  
1404 tests in HCHS/SOL. Only variants with MAF > 0.5% were included. Genome coordinates  
1405 presented were based on GRCh38.

1406

1407 **Table S2.** Regions showing suggestive evidence of gene-sex interactions or genetic  
1408 associations accounting for gene-sex interactions on WHR in HCHS/SOL. Only variants  
1409 with  $P$  values  $< 5 \times 10^{-7}$  and MAF > 0.5% were included. Previously reported marginal  
1410 genetic effects, gene-sex interactions, or joint effects within 1Mb flanking regions were  
1411 shown. Genome coordinates presented were based on GRCh38.

1412

1413 **Table S3.** Significant association regions with LDL cholesterol levels from variant set tests  
1414 in a 20kb sliding window analysis in HCHS/SOL. Genome coordinates presented were  
1415 based on GRCh38.

1416

1417 **Table S4.** Significant association regions with LDL cholesterol levels from single-variant  
1418 meta-analysis combining stochastic summary statistics from HCHS/SOL, ARIC EA and  
1419 ARIC AA. Only variants with MAF > 0.5% were included. Genome coordinates presented  
1420 were based on GRCh38.

1421

1422 **Table S5.** Significant association regions with LDL cholesterol levels from variant set  
1423 meta-analysis in a 20kb sliding window analysis after combining stochastic summary  
1424 statistics from HCHS/SOL, ARIC EA and ARIC AA. Genome coordinates presented were  
1425 based on GRCh38.

1426

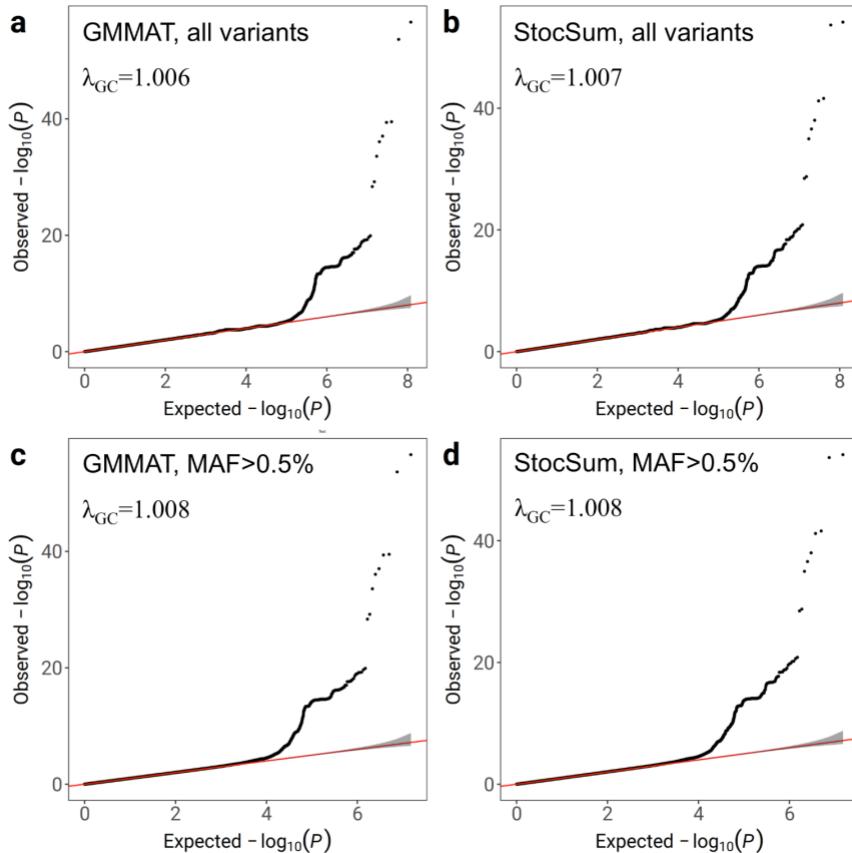
1427 **Table S6.** Regions excluded from LD score regression due to long-range LD on the human  
1428 genome. Genome coordinates presented were based on GRCh38.

1429

Chromosome	Start Position (Mb)	End Position (Mb)
1	45.5	52
1	72.5	73.5
1	174	175
1	24.61	24.63
2	85.9	100.1
2	133.5	137.5
2	182	189.5
3	47.4	51.3
3	89	98.5
3	162	163.6
4	33.5	34.5
4	97.5	98.2
4	119	120
4	143	144.2
5	44.4	51.2

5	98.5	101.5
5	129.6	133
5	136.2	139.2
6	25.3	33.5
6	57.7	64.3
6	139	142.5
7	54.9	66.9
7	119	120
8	8	12.5
8	42	49
8	110	114
10	36.5	43.2
11	46	58
11	88	91.2
12	33	41.3
12	109	111.6
14	66.1	67.5
17	45	47
19	23.5	28
20	33.9	41.3

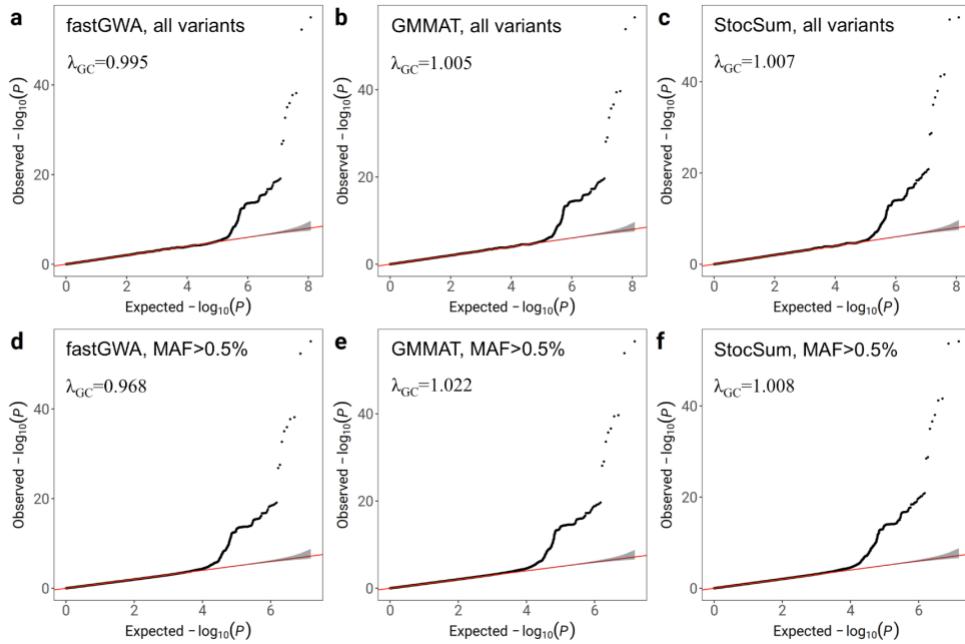
1431 **Supplementary Figures**



1432

1433 Figure S1 Quantile-quantile (Q-Q) plots of  $P$  values from single-variant tests on LDL  
1434 cholesterol levels using GMMAT and StocSum in HCHS/SOL. The number of random  
1435 vector replicates  $B$  in StocSum was set to 1,000. a, GMMAT  $P$  values from all variants. b,  
1436 StocSum  $P$  values from all variants. c, GMMAT  $P$  values from variants with MAF > 0.5%.  
1437 d, StocSum  $P$  values from variants with MAF > 0.5%. The gray shaded areas in the Q-Q  
1438 plots represent 95% confidence intervals under the null hypothesis of no genetic  
1439 associations.

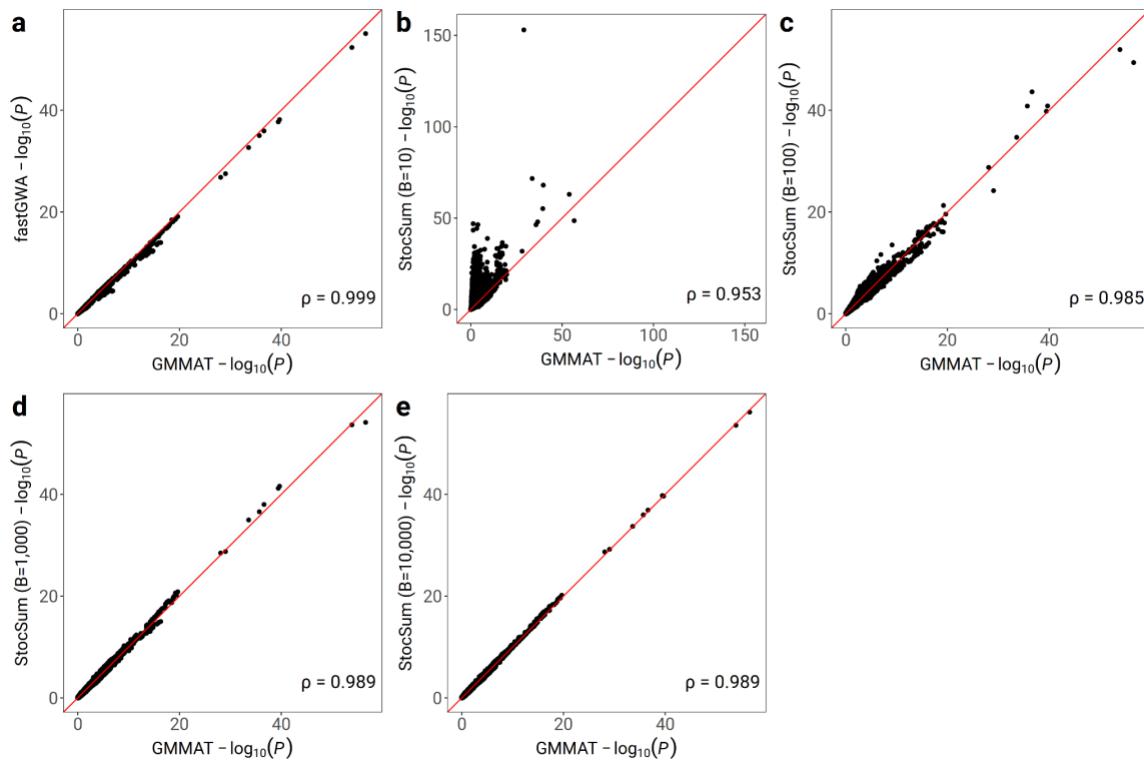
1440



1441

1442 Figure S2. Quantile-quantile (Q-Q) plots of  $P$  values from single-variant tests on LDL  
1443 cholesterol levels using fastGWA, GMMAT, and StocSum in HCHS/SOL. The number of  
1444 random vector replicates  $B$  in StocSum was set to 1,000. a, fastGWA  $P$  values from all  
1445 variants. b, GMMAT  $P$  values from all variants. c, StocSum  $P$  values from all variants. d,  
1446 fastGWA  $P$  values from variants with MAF > 0.5%. e, GMMAT  $P$  values from variants  
1447 with MAF > 0.5%. f, StocSum  $P$  values from variants with MAF > 0.5%. The gray shaded  
1448 areas in the Q-Q plots represent 95% confidence intervals under the null hypothesis of no  
1449 genetic associations.

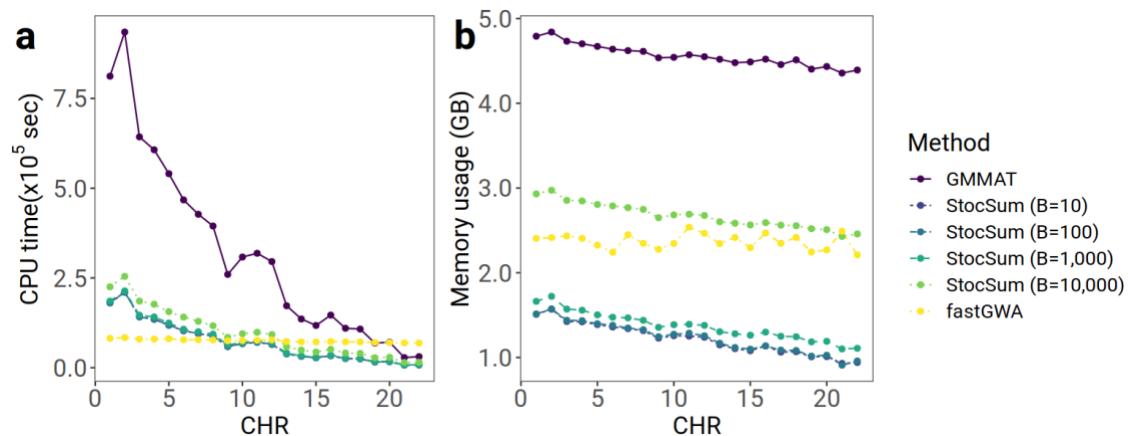
1450



1451

1452 Figure S3 Comparison of  $P$  values from single-variant tests on LDL cholesterol levels  
 1453 using fastGWA, GMMAT, and StocSum in HCHS/SOL. a, comparison of  $P$  values from  
 1454 GMMAT and fastGWA. b-e, comparisons of  $P$  values from GMMAT and StocSum with  
 1455 the number of random vector replicates  $B$  being equal to 10 (b), 100 (c), 1,000 (d), and  
 1456 10,000 (e). The red line denotes the reference line of equality. Spearman's rank correlation  
 1457 coefficients are shown at the bottom right. The data used in this test consisted of 120M  
 1458 variants from 7,297 individuals in HCHS/SOL.

1459



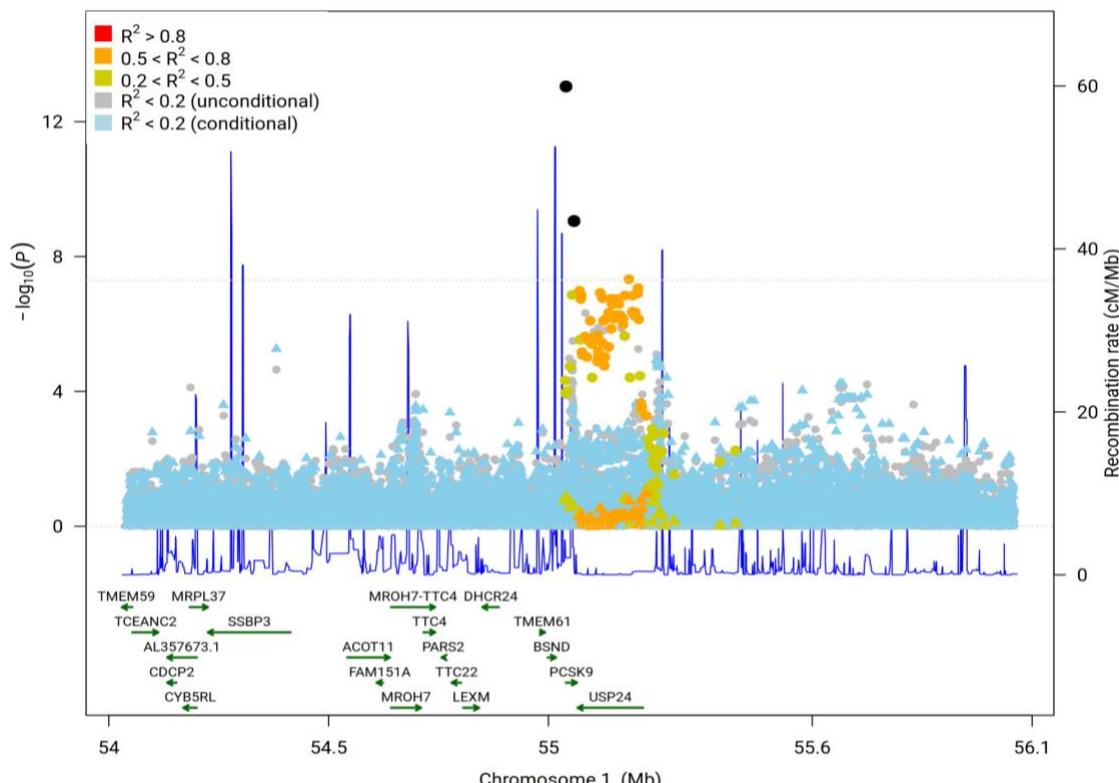
1460

1461 Figure S4 Comparison of CPU time and memory usage from fastGWA, GMMAT and  
1462 StocSum in single-variant tests. a, CPU time. The x axis represents the chromosome  
1463 numbers and the y axis represents the CPU time in  $10^5$  seconds. For GMMAT, the CPU  
1464 time consists of fitting the null model and conducting the association test. For StocSum,  
1465 the CPU time is the sum of four steps: fitting the null model, generating the random vectors,  
1466 computing the single-variant score statistics and the stochastic summary statistics, and  
1467 computing the  $P$  values. b, Memory usage. The x axis represents the chromosome numbers  
1468 and the y axis represents the memory footprint per thread in GB. The data used in this test  
1469 consisted of 120M variants from 7,297 individuals in HCHS/SOL. All tests were  
1470 performed on a high-performance computing server, with 64 threads running in parallel.

1471

1472

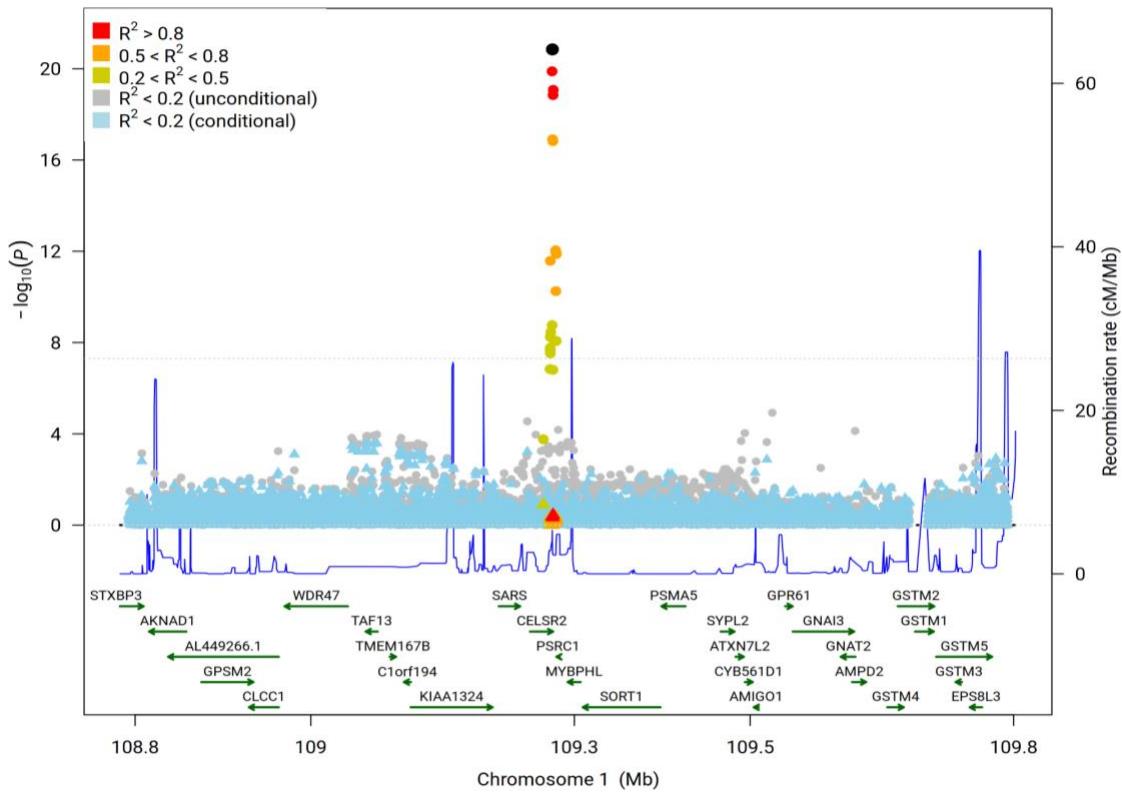
1473 a



1474

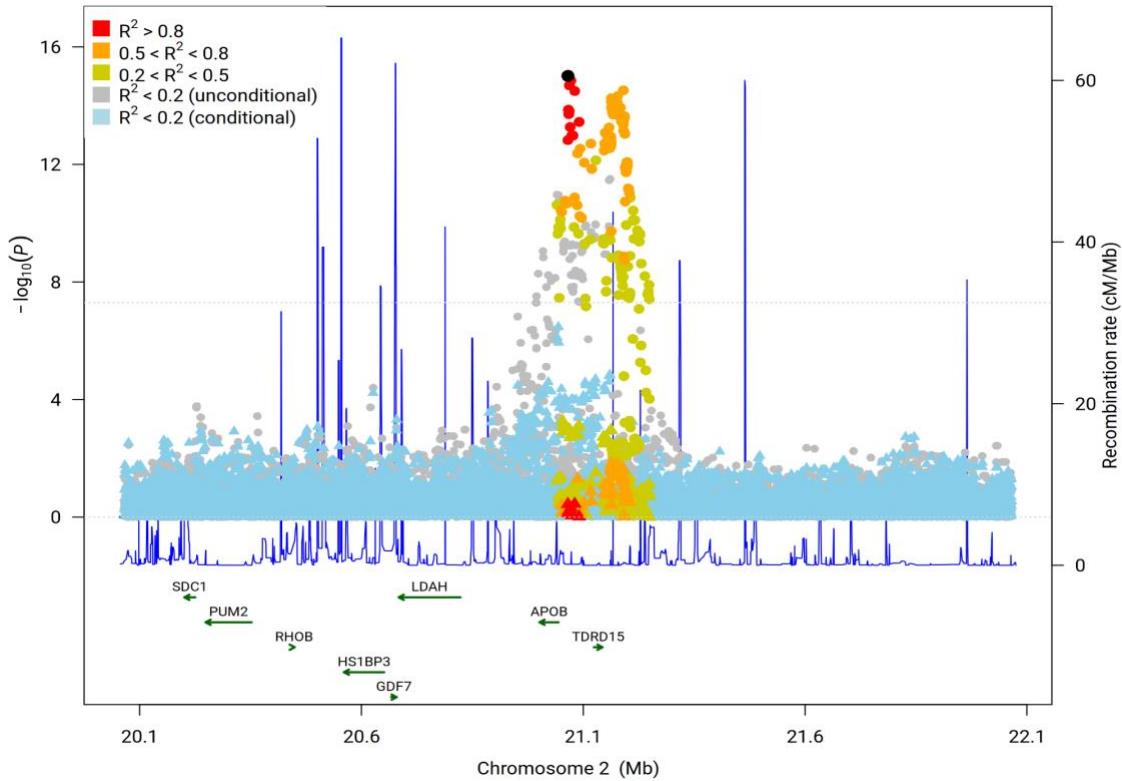
1475

1476 b



1477

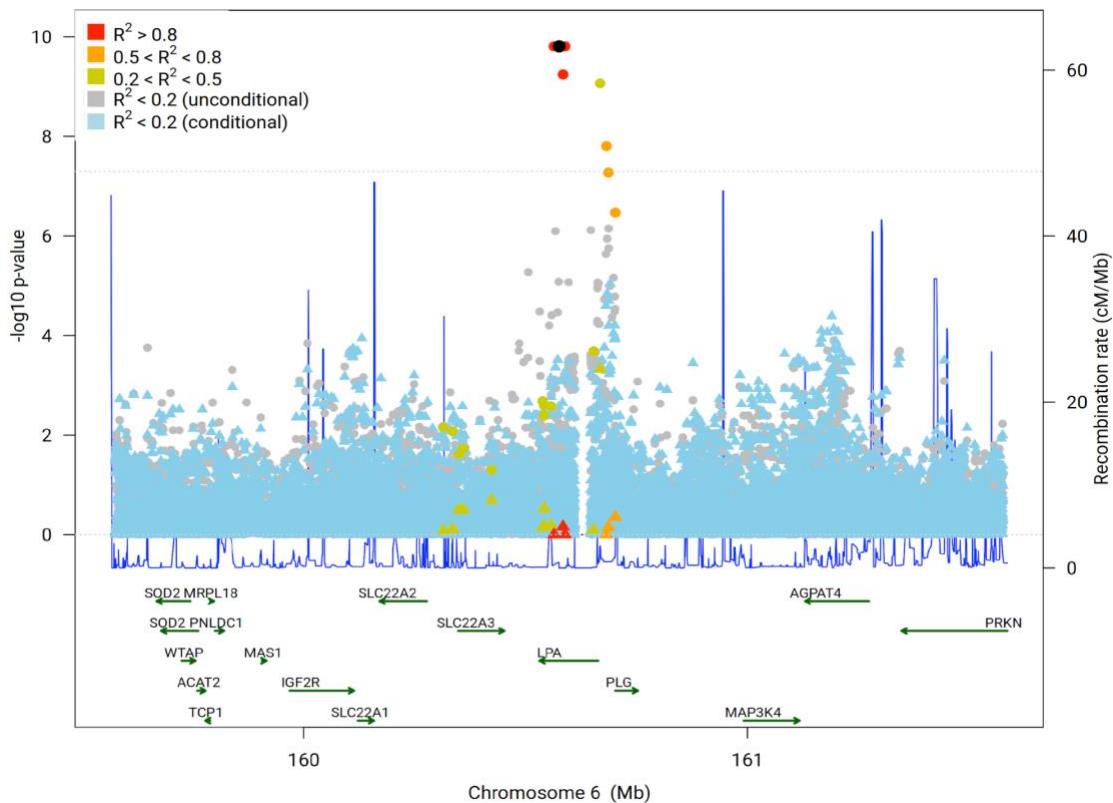
1478 c



1479

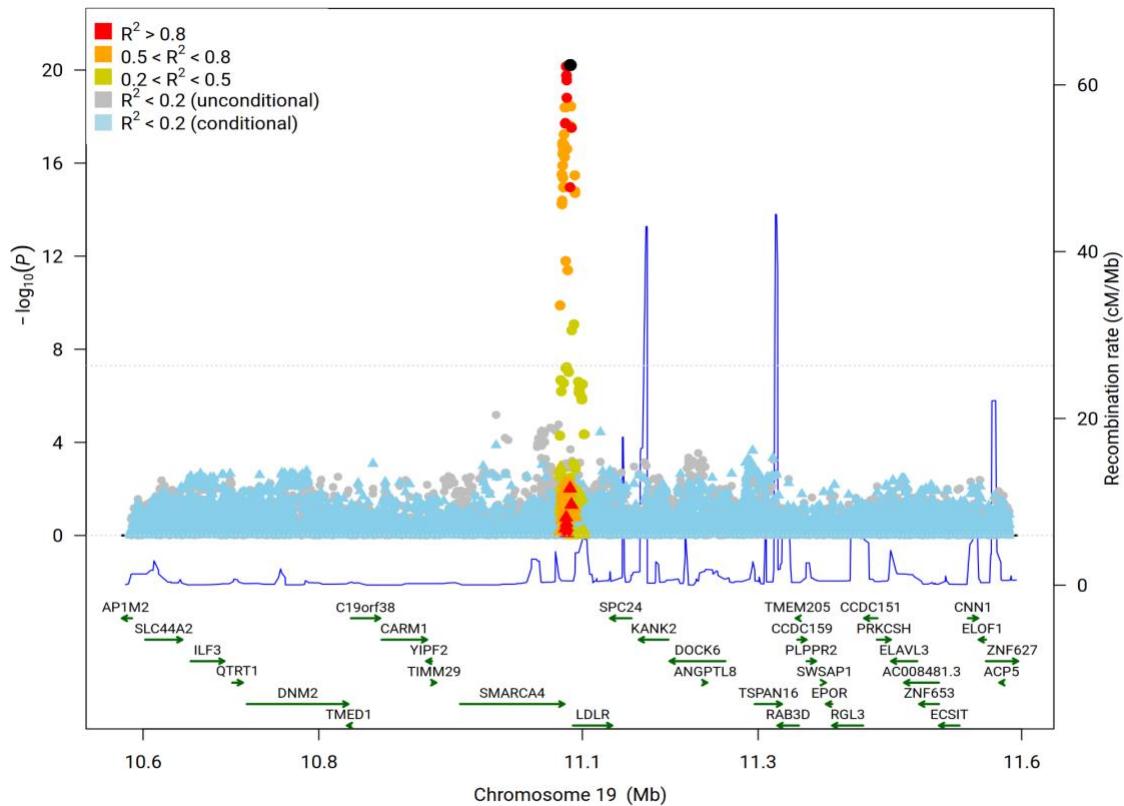
1480

1481 d



1482

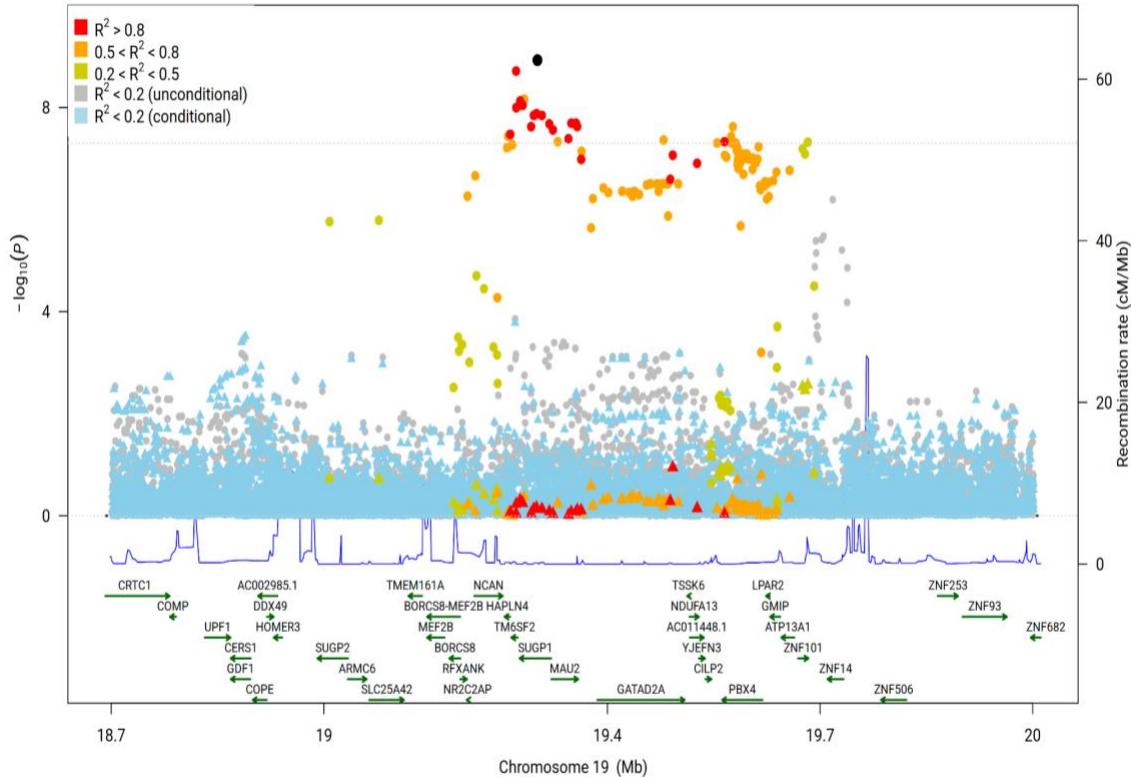
1483 e



1484

1485

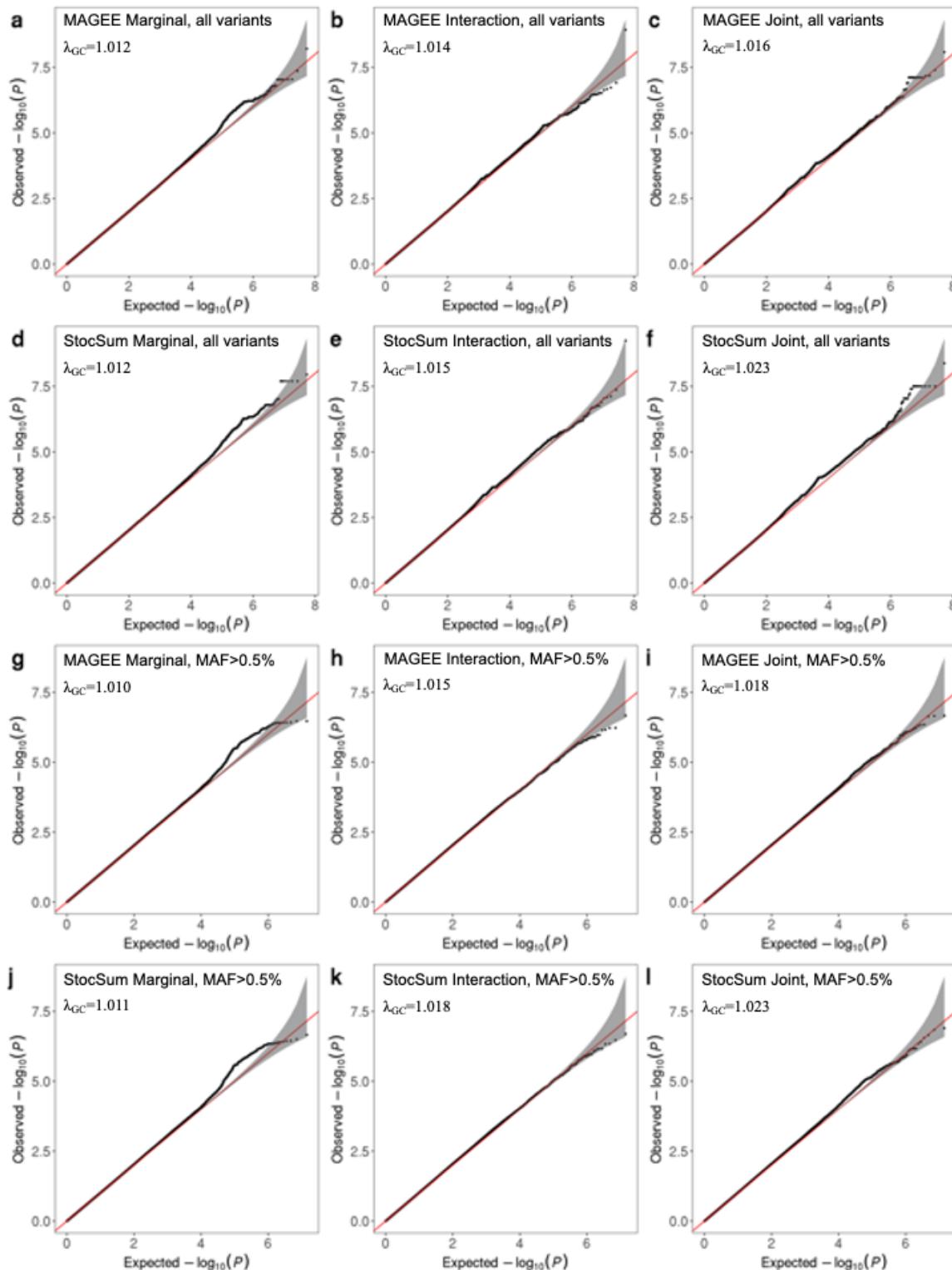
1486 f



1487

1488 Figure S5 Regional plots of StocSum conditional association test results. a, *PCSK9* gene  
1489 region with association variants chr1:55039974 (rs28362263) and chr1:55058182  
1490 (rs28362263). b, *CELSR2* gene region with the sentinel variant chr1:109274968  
1491 (rs562338). c, *APOB* gene region with the sentinel variant chr2:21065449 (rs562338). d,  
1492 *LPA* gene region with the sentinel variant chr6:160576086 (rs10455872). e, *LDLR* gene  
1493 region with the sentinel variant chr19:11086210 (rs8106503). f, *SUGP1* gene region with  
1494 the sentinel variant chr19:19301236 (rs57915152). Association variants are highlighted in  
1495 black dots. Original single-variant test *P* values are shown in dots and conditional *P* values  
1496 are shown in triangles. Variants in four LD categories are shown in different colors based  
1497 on the maximum squared correlation to the sentinel variant and the secondary association  
1498 variant calculated in HCHS/SOL if there are two association variants (a), or the squared  
1499 correlation to the sentinel variant in HCHS/SOL if there is only one sentinel association  
1500 variant (b-f). The horizontal line indicates the genome-wide significance level on the log  
1501 scale,  $-\log_{10}(5 \times 10^{-8})$ . The blue curve shows recombination rates from all populations in  
1502 the 1000 Genome Project.

1503

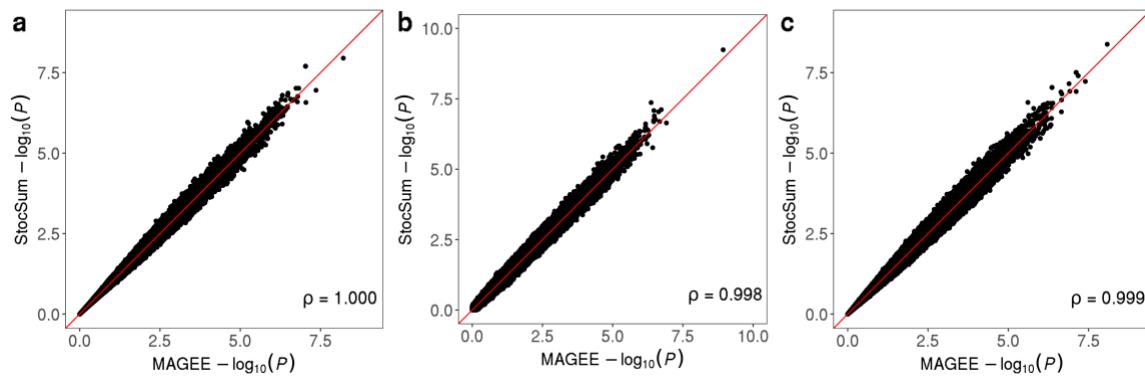


1504

1505 Figure S6 Quantile-quantile (Q-Q) plots of  $P$  values from gene-sex interaction tests on  
 1506 WHR using MAGEE and StocSum in HCHS/SOL. The number of random vector  
 1507 replicates  $B$  in StocSum was set to 1,000. a, Marginal  $P$  values for all variants from

1508 MAGEE. b, Interaction  $P$  values for all variants from MAGEE. c, Joint  $P$  values for all  
1509 variants from MAGEE. d, Marginal  $P$  values for all variants from StocSum. e, Interaction  
1510  $P$  values for all variants from StocSum. f, Joint  $P$  values for all variants from StocSum. g,  
1511 Marginal  $P$  values for variants with MAF > 0.5% from MAGEE. h, Interaction  $P$  values  
1512 for variants with MAF > 0.5% from MAGEE. i, Joint  $P$  values for variants with MAF >  
1513 0.5% from MAGEE. j, Marginal  $P$  values for variants with MAF > 0.5% from StocSum.  
1514 k, Interaction  $P$  values for variants with MAF > 0.5% from StocSum. l, Joint  $P$  values for  
1515 variants with MAF > 0.5% from StocSum. The gray shaded areas in the Q-Q plots represent  
1516 95% confidence intervals under the null hypothesis of no genetic associations and/or gene-  
1517 sex interactions.

1518

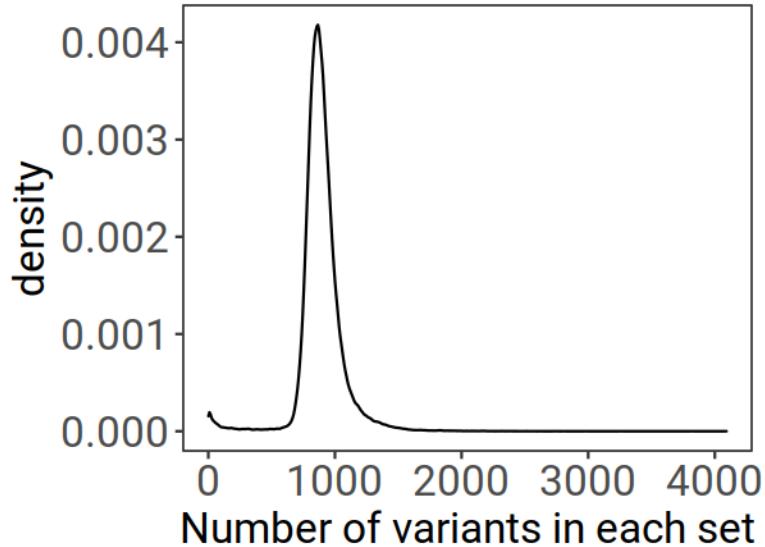


1519

1520 Figure S7 Comparison of  $P$  values from single-variant gene-sex interaction tests on WHR  
1521 using MAGEE and StocSum in HCHS/SOL. a, comparison of marginal genetic effect test  
1522  $P$  values. b, comparison of gene-sex interaction test  $P$  values. c, comparison of joint test  $P$   
1523 values. The x axis and the y axis represent  $-\log_{10}(P)$  using MAGEE and StocSum,  
1524 respectively. The red line denotes the reference line of equality. Spearman's rank  
1525 correlation coefficients are shown at the bottom right.

1526

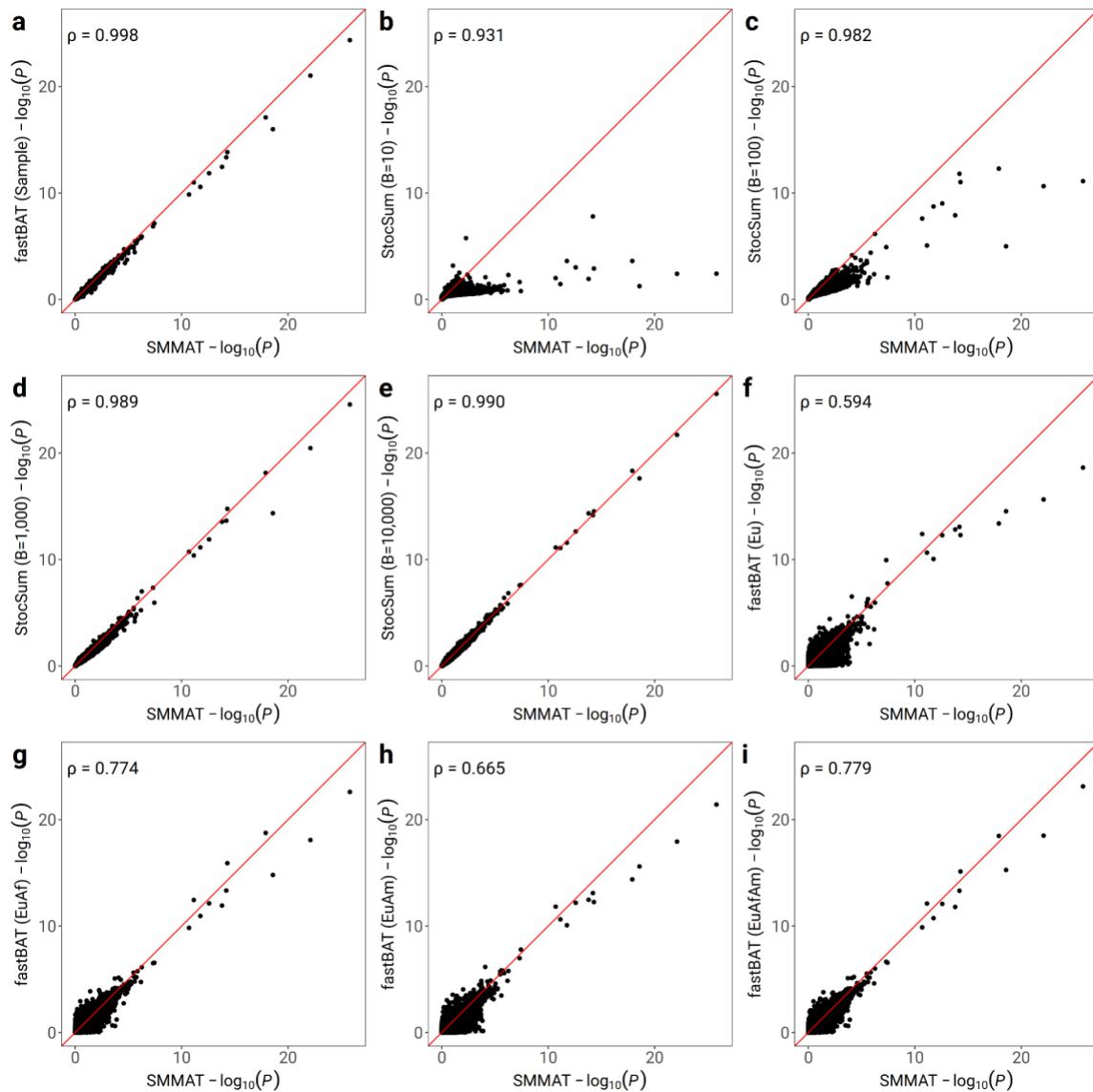
1527



1528

1529 Figure S8 A density plot showing the distribution of variant numbers in each set in a 20 kb  
1530 sliding window analysis on LDL cholesterol levels in HCHS/SOL.

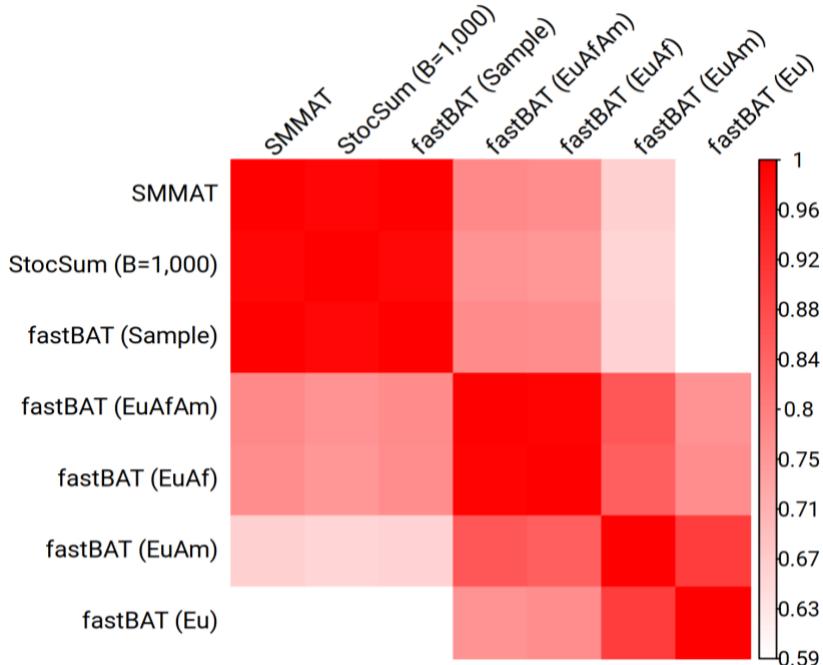
1531



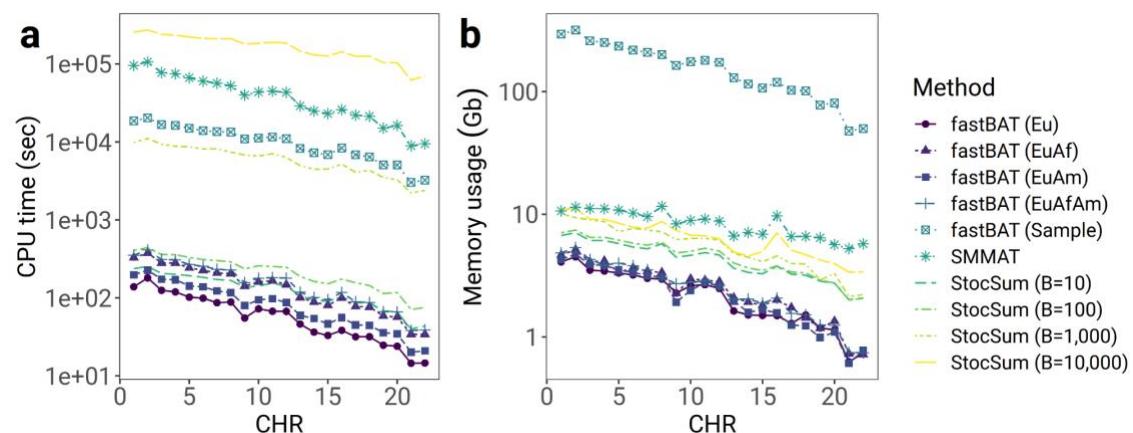
1532

1533 Figure S9 Comparison of  $P$  values from variant set tests in a 20 kb sliding window analysis  
 1534 on LDL cholesterol levels using fastBAT, SMMAT, and StocSum in HCHS/SOL. The x  
 1535 axis represents the  $-\log_{10}(P)$  from variant set tests using SMMAT on individual-level data  
 1536 and the y axis represents the  $-\log_{10}(P)$  from variant set tests using StocSum or fastBAT.  
 1537 a, fastBAT with an internal reference panel using the HCHS/SOL study samples (fastBAT  
 1538 (Sample)). b-e, StocSum with the number of random vector replicates  $B$  being equal to 10  
 1539 (b), 100 (c), 1,000 (d) and 10,000 (e). f-i, fastBAT with external reference panels from  
 1540 1000 Genomes using European (fastBAT (Eu)) (f), European and African (fastBAT  
 1541 (EuAf)) (g), European and American (fastBAT (EuAm)) (h), and European, African, and  
 1542 American (fastBAT (EuAfAm)) (i) populations. The red line denotes the reference line of

1543 equality. The data used in this test consisted of 120M variants from 7,297 individuals in  
1544 HCHS/SOL. Spearman's rank correlation coefficients are shown at the top left.  
1545

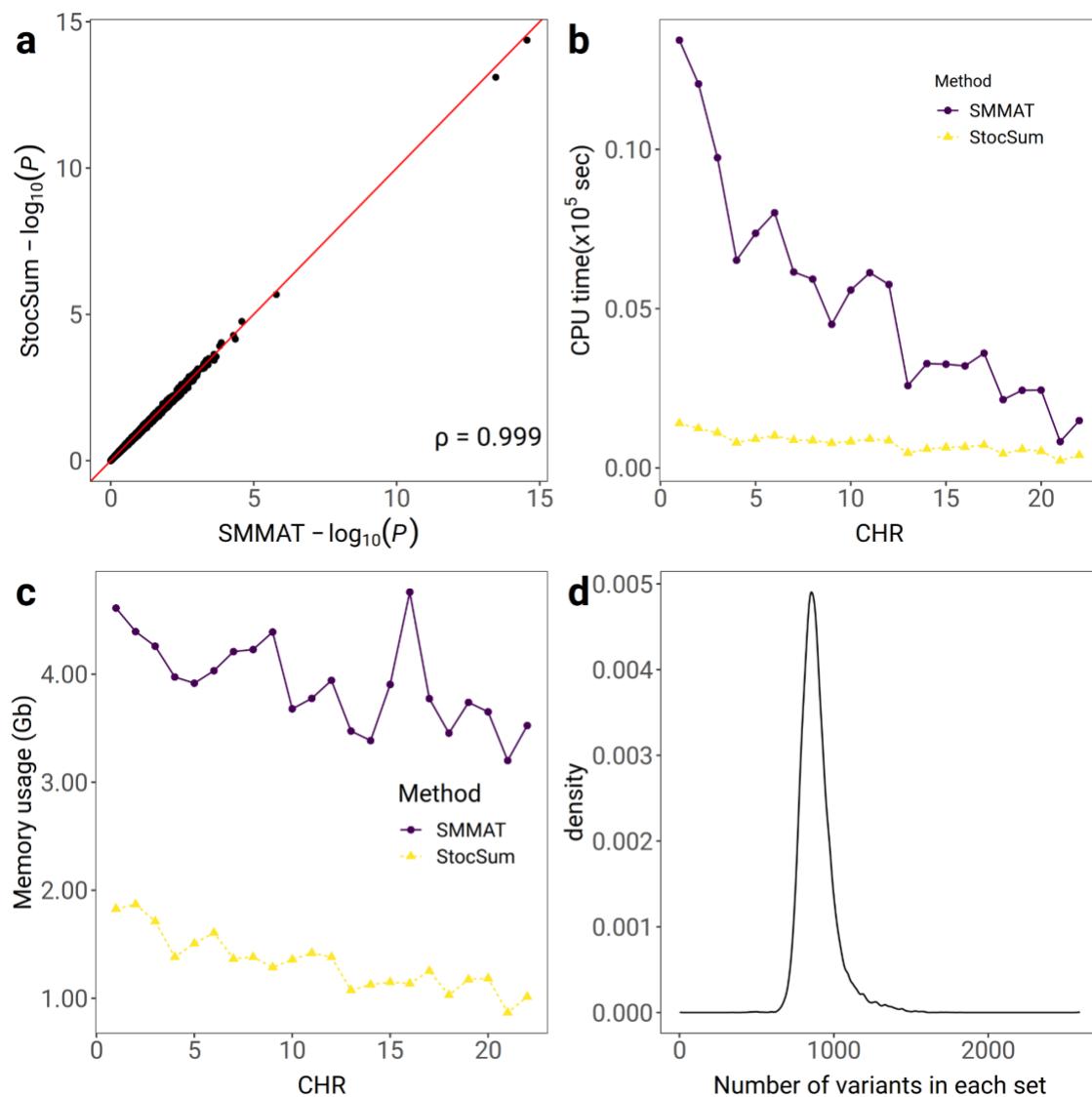


1546  
1547 Figure S10 Heatmap showing Spearman's rank correlation coefficients of  $P$  values from  
1548 variant set tests in a 20 kb sliding window analysis on LDL cholesterol levels using  
1549 fastBAT, SMMAT, and StocSum in HCHS/SOL. For fastBAT, we used an internal  
1550 reference panel using the HCHS/SOL study samples (fastBAT (Sample)), as well as four  
1551 external reference panels from 1000 Genomes (Eu, EuAf, EuAm, EuAfAm).  
1552



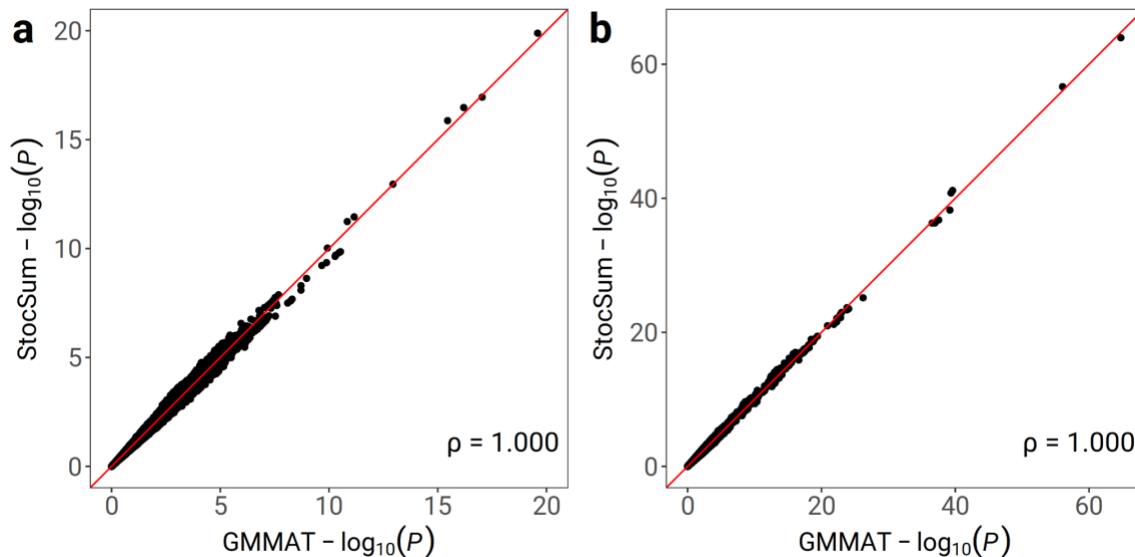
1553

1554 Figure S11 Comparison of CPU time and memory usage from fastBAT, SMMAT, and  
1555 StocSum in variant set tests in a 20 kb sliding window analysis on LDL cholesterol levels  
1556 in HCHS/SOL. a, CPU time. The x axis represents the chromosome numbers and the y axis  
1557 represents the CPU time on the logarithmic scale. The CPU time only includes the step of  
1558 computing the  $P$  values, assuming corresponding summary statistics have been computed  
1559 in single-variant tests. b, Memory usage. The x axis represents the chromosome numbers  
1560 and the y axis represents the memory footprint per thread in GB on the logarithmic scale.  
1561 The data used in this test consisted of 120M variants from 7,297 individuals in HCHS/SOL.  
1562 All tests were performed on a high-performance computing server, with a single thread for  
1563 each chromosome.



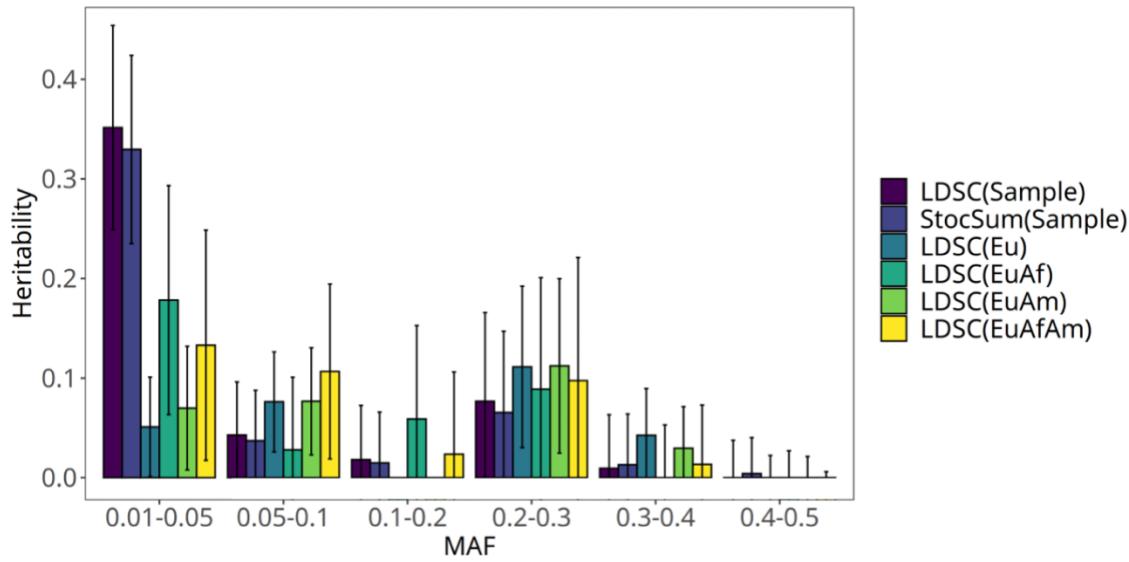
1564

1565 Figure S12 Comparison of SMMAT and StocSum variant set tests in a non-sliding-window  
1566 analysis on LDL cholesterol levels in HCHS/SOL. The variant sets were defined by  
1567 merging chromatin loops of H3K27ac HiChIP interaction in the GM12878 cell line. There  
1568 are a total of 17,224 paired regions, each as a variant set, including two 10kb windows  
1569 which may not be located in close proximity on the primary structure of DNA and not  
1570 typically covered using fixed-size sliding windows. a, comparison of  $P$  values from  
1571 SMMAT and StocSum with the number of random vector replicates  $B$  being equal to 1,000.  
1572 The x axis and the y axis represent the  $-\log_{10}(P)$  from variant set tests using SMMAT and  
1573 StocSum, respectively. The red line denotes the reference line of equality. b, comparison  
1574 of CPU time between SMMAT and StocSum. The x axis represents the chromosome  
1575 numbers and the y axis represents the CPU time in  $10^5$  seconds. For SMMAT and StocSum,  
1576 the CPU time only includes the step of computing the  $P$  values, assuming corresponding  
1577 summary statistics have been computed in single-variant tests. c, comparison of memory  
1578 usage between SMMAT and StocSum. The x axis represents the chromosome numbers and  
1579 the y axis the memory footprint per thread in GB. d, a density plot showing the distribution  
1580 of variant numbers in each set.



1581  
1582 Figure S13 Comparison of  $P$  values from single-variant tests on longitudinal LDL  
1583 cholesterol levels using GMMAT and StocSum in ARIC AA (a) and ARIC EA (b). The  
1584 ARIC AA data used in this test consisted of 70M variants and 7,514 observations from  
1585 2,045 individuals. The ARIC EA data used in this test consisted of 92M variants and 26,668

1586 observations from 6,327 individuals. The x axis and the y axis represent the  $-\log_{10}(P)$  from  
1587 single-variant tests using GMMAT and StocSum with the number of random vector  
1588 replicates  $B$  being equal to 1,000. The red line denotes the reference line of equality.  
1589 Spearman's rank correlation coefficients are shown at the bottom right.  
1590



1591  
1592 Figure S14 LDL heritability estimates by stratified LDSC and StocSum for different MAF  
1593 bins. The error bars show point estimates  $\pm$  standard errors. Negative heritability estimates  
1594 reported from stratified LDSC were truncated at 0. LD scores for different MAF bins were  
1595 estimated from LDSC (Sample) and StocSum (Sample) using HCHS/SOL study samples,  
1596 or LDSC on external reference panels using European, African and/or American  
1597 populations from the 1000 Genomes Project: LDSC (Eu), LDSC (EuAf), LDSC (EuAm),  
1598 and LDSC (EuAfAm).  
1599