StocSum: stochastic summary statistics for whole genome sequencing studies

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

[1]Human Genetics Center, Department of Epidemiology, Human Genetics and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, USA.

[2]Department of Epidemiology & Population Health, Albert Einstein College of Medicine, Bronx, NY, USA.

[3]The Saban Research Institute, Children's Hospital Los Angeles, Los Angeles, California.

[4]Department of Pediatrics, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA

[5]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA.

[6]Department of Epidemiology, School of Public Health, University of Washington, 3980 15th Ave NE, Seattle, WA, USA.

[7]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA.

[8]Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA.

Correspondence: Han Chen (Han.Chen.2@uth.tmc.edu)

Abstract

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

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

Main

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

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

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

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

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

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

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

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

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

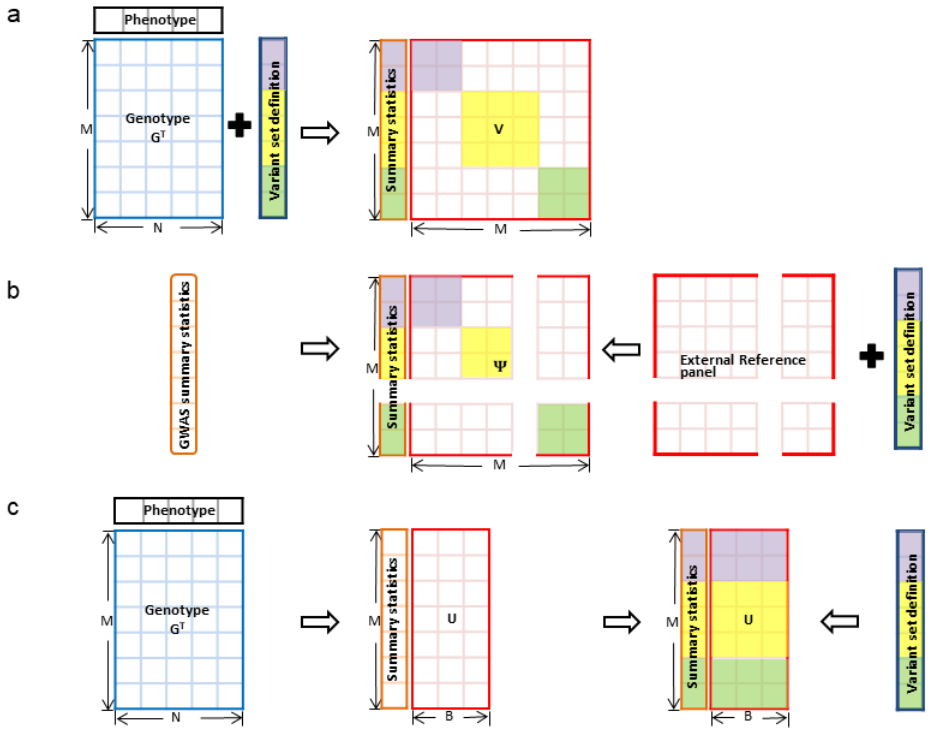


Figure 1: The StocSum framework. **a)** Traditional methods calculate the correlation or LD matrix $\boldsymbol{V}$ from individual-level genotype data. To reduce the computational burden, the full $M \times M$ matrix is usually not computed in practice, but rather replaced by a block-diagonal or banded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

Results

Overview of the method

We describe StocSum under the generalized linear mixed model (GLMM) framework. It can also be applied to simpler statistical models such as generalized linear models[53] and extended to more complex models such as generalized additive mixed models[54]. The GLMM can be written as:

$$\mathbb{g}(\mu_i) = \boldsymbol{X}_i\boldsymbol{\alpha} + \widetilde{\boldsymbol{G}}_i\boldsymbol{\beta} + b_i \quad (1)$$

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

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

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

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

Single-variant tests

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

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

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

Using StocSum, we ran a WGS study of LDL cholesterol in HCHS/SOL and identified seven genome-wide significant (*P* values $< 5\times10^{-8}$) regions mapped to genes *PCSK9* and *CELSR2* on chromosome 1, *APOB* on chromosome 2, *LPA* on chromosome 6, *LDLR*, *SUGP1*, and *APOE* on chromosome 19 (**Fig. 2g**, **Table S1**), all of which had been previously reported to be associated with LDL[4,56–59].

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

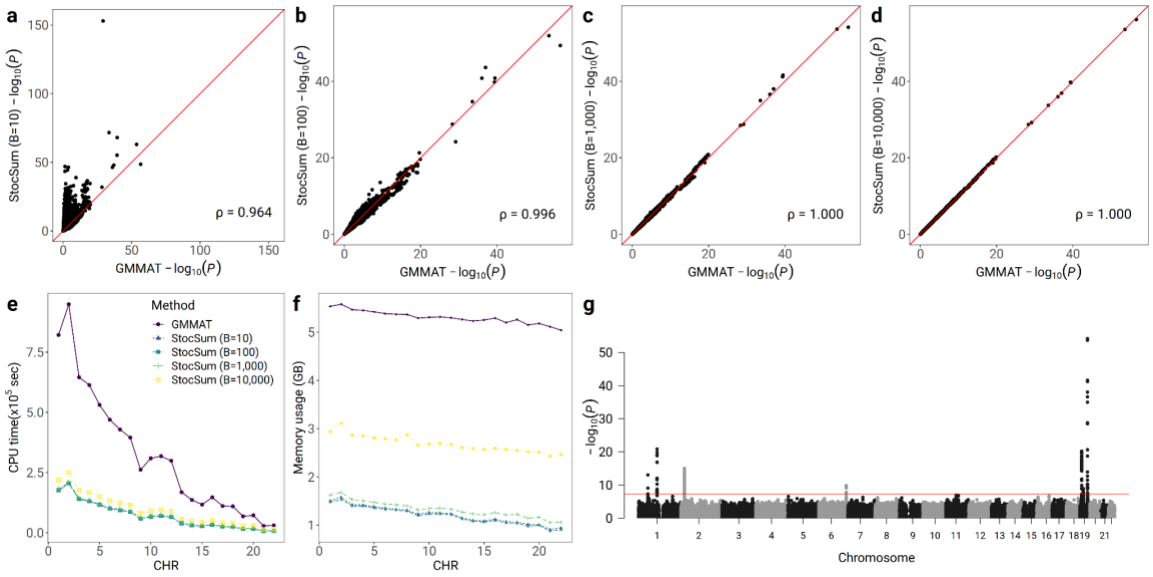


Figure 2: StocSum in single-variant tests. a-d) comparison of *P* values from GMMAT and StocSum with the number of random vector replicates *B* being equal to 10 **(a)**, 100 **(b)**, 1,000 **(c)** and 10,000 **(d)**. The x axis and the y axis represent $-\log_{10}(P)$ from single-variant tests using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

were included in the Manhattan plot. The red line indicates the genome-wide significance level on the log scale, $-\log_{10}(5\times10^{-8})$.

Conditional association tests

We implemented StocSum for conditional association tests and applied it to the seven genome-wide significant regions identified in **Fig. 2g**. The sentinel variant in the *APOE* gene region is chr19: 44908822 (rs7412) with $P = 7.1\times10^{-55}$. There are 26 common variants with MAF > 0.5% close to this sentinel variant in this region, with a *P* value less than 5×10^{-8} (**Fig. 3**). After conditioning on the sentinel variant, we identified a secondary association variant chr19: 44908684 (rs429358) with conditional $P = 8.2\times10^{-15}$. After conditioning on both rs7412 and rs429358, all other variants in the region had *P* values > 1.8×10^{-3}, indicating that no additional independent associations exist in this region. We also observed similar patterns in the other six regions (**Table S1**), after conditioning on either one or two top associated variants in each region (**Fig. S5 a-f**).

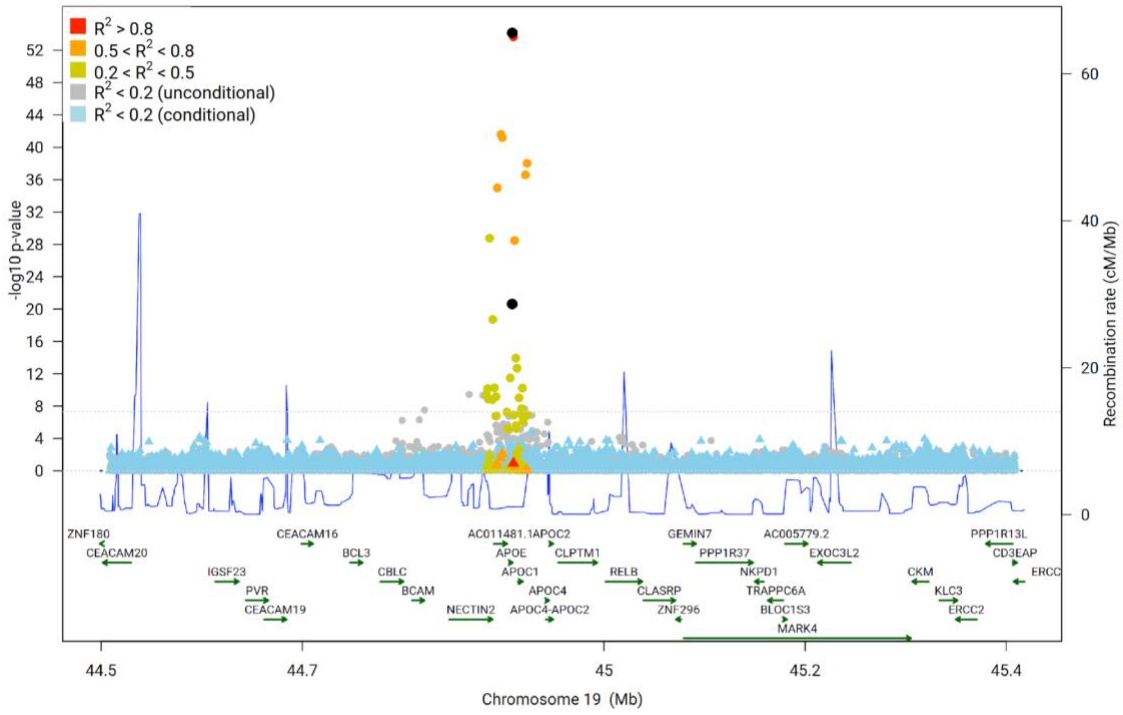


Figure 3: A regional plot of StocSum conditional association test results in the *APOE* region. Variants with MAF > 0.5% in a 1Mb window near association variants rs7412 and rs429358 (highlighted in black dots). Original single variant test *P* values are shown in dots

and conditional *P* values are shown in triangles. Variants in four LD categories are shown in different colors based on the maximum squared correlation to the sentinel variant rs7412 and the secondary association variant rs429358 calculated in HCHS/SOL. The horizontal 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 Project.

Gene-environment interaction tests

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

Variant set tests

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

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

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

StocSum variant set tests are computationally efficient (**Figs. 4e-f**). It only took StocSum (B=1,000) 2.7 CPU hours to finish variant set tests on chromosome 1 using 20 kb sliding windows, which was 9.7-fold faster than SMMAT using individual-level data. Across the autosomes, there were a total of 134,739 non-overlapping 20 kb windows containing at least one variant. On average, the StocSum (B=1,000) CPU time was about 14.3% of the SMMAT CPU time. Meanwhile, StocSum (B=1,000) only required about 68.1% of the memory compared to SMMAT. StocSum with B=10,000 utilized more CPU time than SMMAT since B was larger than the sample size (N=7,297), making the $M \times B$ stochastic summary statistics matrix $\boldsymbol{U}$ even larger in size compared to the $N \times M$ genotype matrix $\boldsymbol{G}$. In this 20 kb sliding window analysis using StocSum variant set tests, we identified four regions associated with LDL levels in HCHS/SOL[4,56–59], at the significance level of 0.05/134,739=3.7x10^{-7} (**Fig. 4g, Table S3**).

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

panel using individual-level genotypes from the original study sample (called fastBAT (Sample)), as well as four external reference panels from the 1000 Genomes Project[69]: European populations (fastBAT (Eu)), European and African populations (fastBAT (EuAf)), European and American populations (fastBAT (EuAm)), and European, African and American populations (fastBAT (EuAfAm)). Variant set test *P* values from SMMAT, StocSum (*B*=1,000), and fastBAT (Sample) were highly concordant (**Figs. S9-10**), with pairwise Spearman correlation coefficients being greater than 0.99. However, fastBAT with external reference panels, i.e., fastBAT (Eu), fastBAT (EuAf), fastBAT (EuAm), fastBAT (EuAfAm), gave inaccurate variant set test *P* values compared to SMMAT using individual-level genotypes. The correlation coefficients of $\log_{10}(P)$ between SMMAT and fastBAT with Eu, EuAf, EuAm, EuAfAm reference panels were 0.59, 0.77, 0.66, and 0.78, respectively (**Fig. S10**). Since Hispanic/Latino adults are three-way admixed populations with European, African and Amerindian ancestries, it is not surprising that an external reference panel from only European populations could not represent the LD structure in HCHS/SOL samples accurately. Interestingly, although including African and American populations in the external reference panel did improve the concordance of fastBAT *P* values compared to SMMAT, fastBAT using the internal reference panel clearly outperformed all external reference panels that we investigated. In addition, when an external reference panel was used, variants not included in the panel would have to be excluded, leading to loss of unique variants in the study samples. This highlights the importance of choosing an accurate reference panel for fastBAT, and the best reference panel for study samples from underrepresented, admixed or isolated populations are the study samples themselves. StocSum represents the LD structure in any variant sets through a stochastic summary statistic matrix $\boldsymbol{U}$ directly derived from study samples rather than external reference panels, thus providing accurate variant set test results. Meanwhile, StocSum with *B*=1,000 was slightly faster (1.7-fold) than fastBAT (Sample) on the whole genome (**Fig. S11a**), with a dramatically reduced memory footprint (3.6%) compared to fastBAT (Sample) (**Fig. S11b**).

To illustrate StocSum variant set tests beyond sliding windows, we compared StocSum (B=1,000) with SMMAT when the variant sets composed of different regions that were

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

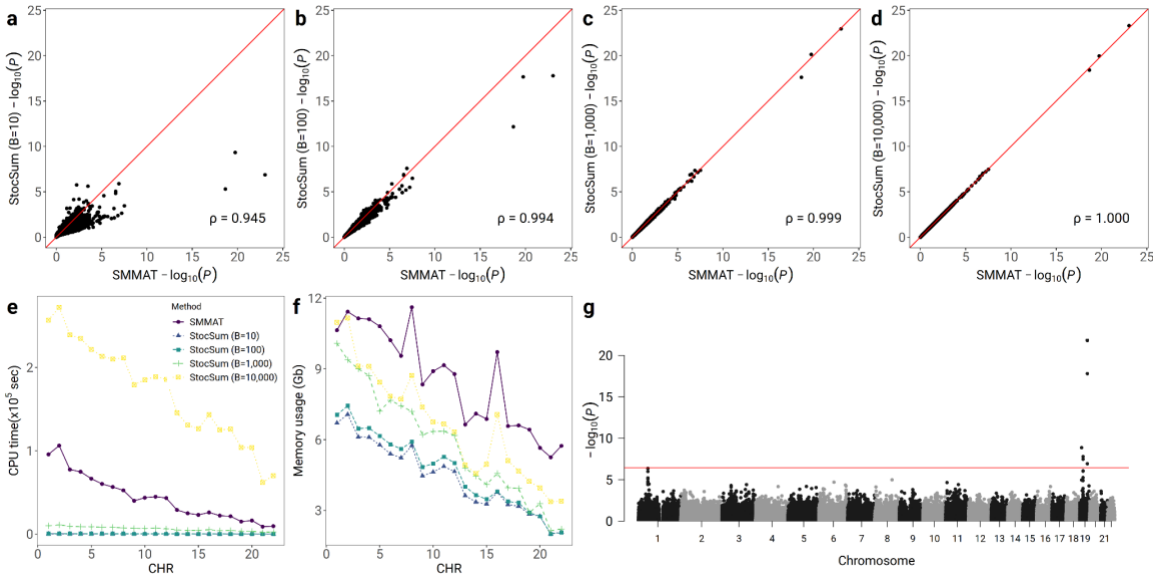


Figure 4: StocSum in variant set tests. Comparison of *P* values from SMMAT and StocSum with the number of random vector replicates *B* being equal to 10 (a), 100 (b), 1,000 (c) and 10,000 (d) in a 20 kb sliding window analysis on the whole genome. The x axis and the y axis represent $-\log_{10}(P)$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

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

Meta-analysis

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

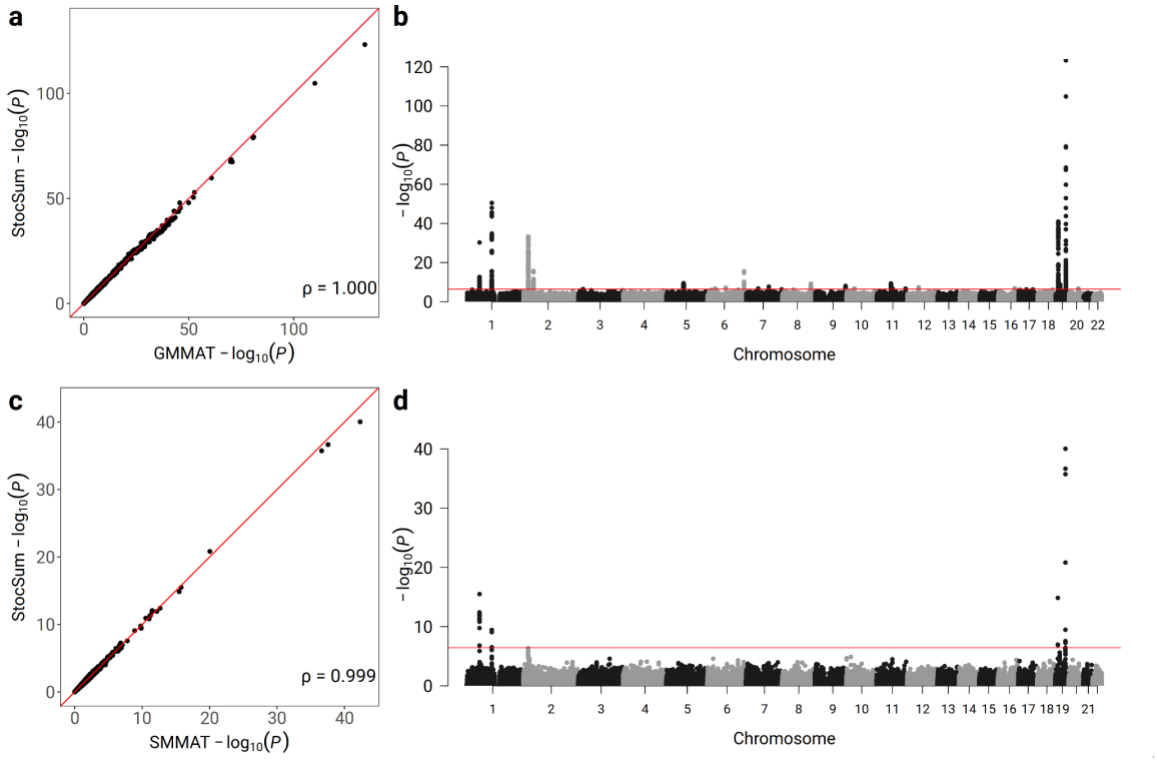


Figure 5: StocSum in meta-analysis. a, comparison of single-variant meta-analysis *P* values from GMMAT and StocSum with the number of random vector replicates *B* being equal to 1,000. The x axis and the y axis represent $-\log_{10}(P)$ from single-variant meta-analysis using GMMAT and StocSum, respectively. The red line denotes the reference line of equality. Spearman's rank correlation coefficients are shown at the bottom right. b, the Manhattan plot of single-variant tests on LDL in the meta-analysis of ARIC AA and EA, and HCHS/SOL studies using StocSum. The x-axis represents the physical chromosome and position of each variant and the y-axis represents $-\log_{10}(P)$ from the StocSum single-variant test. Only variants with MAF > 0.5% were included in the Manhattan plot. The red line indicates the genome-wide significance level on the log scale, $-\log_{10}(5\times10^{-8})$. c, comparison of variant set meta-analysis *P* values from SMMAT and StocSum with the number of random vector replicates *B* being equal to 1,000. The x axis and the y axis represent $-\log_{10}(P)$ from variant set meta-analysis using SMMAT and StocSum, respectively. The red line denotes the reference line of equality. Spearman's rank correlation coefficients are shown at the bottom right. d, the Manhattan plot of variant set tests on LDL in the meta-analysis of ARIC AA and EA, and HCHS/SOL studies using StocSum. The x-axis represents the start physical chromosome and position of each variant

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

LD score regression

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

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 seems to be an upward bias for many variants in LDSC (EuAf) and LDSC (EuAfAm) results, when African populations from the 1000 Genomes Project were included in the reference panel (**Fig. 6b,d**), highlighting the challenges in selecting appropriate external reference panels for LD score estimation in underrepresented, admixed or isolated populations. StocSum (Sample) required only 1.4% of CPU time used by LDSC (Sample) (**Fig. 6f**). It was also 6-fold to 42-fold faster than LDSC using different external reference panels.

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

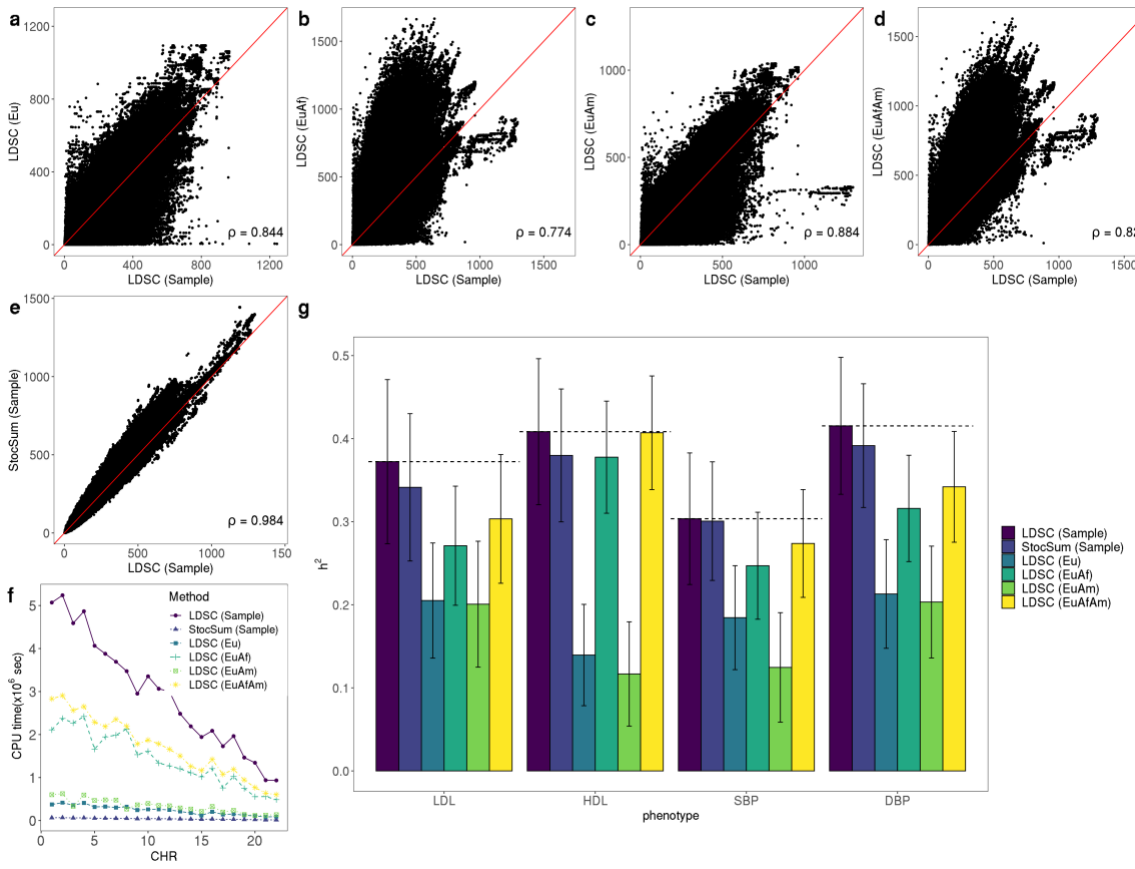


Figure 6: StocSum in LD score regression and heritability estimation. a-e, comparison of LD scores from LDSC (Sample) (x-axis) and different alternative methods (y-axis). a, LDSC (Eu). b, LDSC (EuAf). c, LDSC (EuAm). d, LDSC (EuAfAm). e, StocSum (Sample). Spearman's rank correlation coefficients are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

Discussion

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

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

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

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

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

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

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

Methods

Stochastic summary statistics

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

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

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

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

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

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

Single-variant tests

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

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

Conditional association tests

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

$$S_{\boldsymbol{g}|\dot{\boldsymbol{G}}} = S_{\boldsymbol{g}} - \boldsymbol{g}^T\boldsymbol{P}\dot{\boldsymbol{G}}\left(\dot{\boldsymbol{G}}^T\boldsymbol{P}\dot{\boldsymbol{G}}\right)^{-1}S_{\dot{\boldsymbol{G}}}.$$

The variance of the conditional score is $Var\left(S_{\boldsymbol{g}|\dot{\boldsymbol{G}}}\right) = \boldsymbol{g}^T\boldsymbol{P}\boldsymbol{g} - \boldsymbol{g}^T\boldsymbol{P}\dot{\boldsymbol{G}}\left(\dot{\boldsymbol{G}}^T\boldsymbol{P}\dot{\boldsymbol{G}}\right)^{-1}\dot{\boldsymbol{G}}^T\boldsymbol{P}\boldsymbol{g}$[17].

In the StocSum framework, $S_{\boldsymbol{g}}$ and $\boldsymbol{U_g}$ are the single-variant score and stochastic summary statistics corresponding to the variant of interest in the conditional association test and $\boldsymbol{S_{\dot{G}}}$ (a length c vector) and $\boldsymbol{U_{\dot{G}}}$ (a $c \times B$ matrix) are the single-variant score and stochastic summary statistics corresponding to the association variants to be conditioned on. The conditional score can be computed as

$$S_{\boldsymbol{g}|\dot{\boldsymbol{G}}} = S_{\boldsymbol{g}} - \boldsymbol{U_g U_{\dot{G}}^T}\left(\boldsymbol{U_{\dot{G}} U_{\dot{G}}^T}\right)^{-1}\boldsymbol{S_{\dot{G}}},$$

and the conditional stochastic summary statistics can be computed as

$$\boldsymbol{U_{g|\dot{G}}} = \boldsymbol{U_g} - \boldsymbol{U_g U_{\dot{G}}^T}\left(\boldsymbol{U_{\dot{G}} U_{\dot{G}}^T}\right)^{-1}\boldsymbol{U_{\dot{G}}}\,.$$

The variance $Var\left(S_{\boldsymbol{g}|\dot{\boldsymbol{G}}}\right)$ of the conditional score $S_{\boldsymbol{g}|\dot{\boldsymbol{G}}}$ can be estimated as $\frac{1}{B}\boldsymbol{U_{g|\dot{G}}}\boldsymbol{U_{g|\dot{G}}}^T$. The asymptotic P value is computed using the conditional score $S_{\boldsymbol{g}|\dot{\boldsymbol{G}}}$ and its variance estimated from the stochastic summary statistics $\boldsymbol{U_{g|\dot{G}}}$, for each variant of interest in the conditional association test.

Gene-environment interaction tests

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

$$\mathbb{g}(\mu_i) = \boldsymbol{X_i\alpha} + g_i\beta + \boldsymbol{H_i\gamma} + b_i, \tag{3}$$

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

The single-variant score for the marginal genetic effect is $S_{\boldsymbol{g}} = \frac{\boldsymbol{g}^T(\boldsymbol{y}-\hat{\boldsymbol{\mu}}_{\boldsymbol{0}})}{\hat{\phi}}$ and its variance is $Var\left(S_{\boldsymbol{g}}\right) = \boldsymbol{g}^T\boldsymbol{P}\boldsymbol{g}$. The single-variant score for the GEI effects is $S_{\boldsymbol{H}} = \frac{\boldsymbol{H}^T(\boldsymbol{y}-\hat{\boldsymbol{\mu}}_{\boldsymbol{0}})}{\hat{\phi}}$ and its $m \times m$ covariance matrix is $Var(S_{\boldsymbol{H}}) = \boldsymbol{H}^T\boldsymbol{P}\boldsymbol{H}$, where $\boldsymbol{H} = (\boldsymbol{H}_1^T\ \boldsymbol{H}_2^T\ \cdots\ \boldsymbol{H}_N^T)^T$ is a $N \times m$

matrix for the GEI terms. The score for GEI effects adjusting for the marginal genetic effect can be approximated by $S_{\boldsymbol{H}|\boldsymbol{g}} = S_{\boldsymbol{H}} - \boldsymbol{H}^T\boldsymbol{P}\boldsymbol{g}(\boldsymbol{g}^T\boldsymbol{P}\boldsymbol{g})^{-1}S_{\boldsymbol{g}}$ [61], with a covariance matrix $Var(S_{\boldsymbol{H}|\boldsymbol{g}}) = \boldsymbol{H}^T\boldsymbol{P}\boldsymbol{H} - \boldsymbol{H}^T\boldsymbol{P}\boldsymbol{g}(\boldsymbol{g}^T\boldsymbol{P}\boldsymbol{g})^{-1}\boldsymbol{g}^T\boldsymbol{P}\boldsymbol{H}$. The marginal genetic effect can be tested using the quadratic form ${S_{\boldsymbol{g}}}^T Var(S_{\boldsymbol{g}})^{-1} S_{\boldsymbol{g}}$, which follows a chi-square distribution with 1 degree of freedom under the null hypothesis of no marginal genetic effects. The GEI effects can be tested using ${S_{\boldsymbol{H}|\boldsymbol{g}}}^T Var(S_{\boldsymbol{H}|\boldsymbol{g}})^{-1} S_{\boldsymbol{H}|\boldsymbol{g}}$, which follows a chi-square distribution with m degrees of freedom under the null hypothesis of no gene-environment interactions. The joint test, which evaluates both marginal genetic effects and GEI effects, can be constructed by the sum of these two chi-square statistics, since $S_{\boldsymbol{H}}$ and $S_{\boldsymbol{H}|\boldsymbol{g}}$ are asymptotically independent. The joint test statistic follows a chi-square distribution with $1 + m$ degrees of freedom under the null hypothesis of no marginal genetic effects or gene-environment interactions.

In the StocSum framework, we first compute stochastic summary statistics for the marginal genetic effect $\boldsymbol{U}_{\boldsymbol{g}} = \boldsymbol{g}^T\boldsymbol{R}$ and GEI effects $\boldsymbol{U}_{\boldsymbol{H}} = \boldsymbol{H}^T\boldsymbol{R}$ using individual-level data. We can use $\frac{1}{B}\boldsymbol{U}_{\boldsymbol{g}}\boldsymbol{U}_{\boldsymbol{g}}^T$, $\frac{1}{B}\boldsymbol{U}_{\boldsymbol{H}}\boldsymbol{U}_{\boldsymbol{H}}^T$, and $\frac{1}{B}\boldsymbol{U}_{\boldsymbol{g}}\boldsymbol{U}_{\boldsymbol{H}}^T$ to estimate the variance of the marginal genetic effect score $Var(S_{\boldsymbol{g}})$, the covariance matrix of the GEI effect score $Var(S_{\boldsymbol{H}})$, and the covariance of $S_{\boldsymbol{g}}$ and $S_{\boldsymbol{H}}$, respectively. The adjusted scores can be constructed as $S_{\boldsymbol{H}|\boldsymbol{g}} = S_{\boldsymbol{H}} - \boldsymbol{U}_{\boldsymbol{H}}\boldsymbol{U}_{\boldsymbol{g}}^T\left(\boldsymbol{U}_{\boldsymbol{g}}\boldsymbol{U}_{\boldsymbol{g}}^T\right)^{-1}S_{\boldsymbol{g}}$, and its variance $Var(S_{\boldsymbol{H}|\boldsymbol{g}})$ can be approximated as $\frac{1}{B}\left\{\boldsymbol{U}_{\boldsymbol{H}}\boldsymbol{U}_{\boldsymbol{H}}^T - \boldsymbol{U}_{\boldsymbol{H}}\boldsymbol{U}_{\boldsymbol{g}}^T\left(\boldsymbol{U}_{\boldsymbol{g}}\boldsymbol{U}_{\boldsymbol{g}}^T\right)^{-1}\boldsymbol{U}_{\boldsymbol{g}}\boldsymbol{U}_{\boldsymbol{H}}^T\right\}$.

Variant set tests

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

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

The burden test statistic can be constructed as

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

where $\boldsymbol{W} = diag\{w_j\}$ is a pre-specified $q \times q$ diagonal weight matrix, and $\mathbf{1}_q$ is a length q vector of 1's. The weights can be a function of the MAF[36,38], or functional annotation scores such as CADD[84,85], FATHMM-XF[86], and annotation principal components from STAAR[87]. Under the null hypothesis, the statistic T_{Burden} asymptotically follows $\xi_{Burden}\chi_1^2$, where the scaling factor $\xi_{Burden} = \mathbf{1}_q^T \boldsymbol{W} \tilde{\boldsymbol{G}}^T \boldsymbol{P} \tilde{\boldsymbol{G}} \boldsymbol{W} \mathbf{1}_q = \mathbf{1}_q^T \boldsymbol{W} \tilde{\boldsymbol{V}} \boldsymbol{W} \mathbf{1}_q$ (where $\tilde{\boldsymbol{V}}$ is a $q \times q$ covariance matrix for the single-variant score vector $\tilde{\boldsymbol{S}}$), and χ_1^2 is a chi-square distribution with 1 df. In the StocSum framework, ξ_{Burden} can be estimated as $\frac{1}{B} \mathbf{1}_q^T \boldsymbol{W} \tilde{\boldsymbol{U}} \tilde{\boldsymbol{U}}^T \boldsymbol{W} \mathbf{1}_q = \frac{1}{B} \tilde{\boldsymbol{u}}^T \tilde{\boldsymbol{u}}$, where $\tilde{\boldsymbol{U}}$ is a $q \times B$ matrix (a subset of the $M \times B$ stochastic summary statistic matrix $\boldsymbol{U}$), and $\tilde{\boldsymbol{u}} = \tilde{\boldsymbol{U}}^T \boldsymbol{W} \mathbf{1}_q$ is a length B vector (i.e., column sum of $\boldsymbol{W} \tilde{\boldsymbol{U}}$).

The SKAT statistic can be constructed as

$$T_{SKAT} = \tilde{\boldsymbol{S}}^T \boldsymbol{W} \boldsymbol{W} \tilde{\boldsymbol{S}}.$$

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 independent chi-square distributions with 1 df, and ξ_{SKAT_j} are the eigenvalues of $\boldsymbol{\Xi}_{SKAT} = \boldsymbol{W} \tilde{\boldsymbol{G}}^T \boldsymbol{P} \tilde{\boldsymbol{G}} \boldsymbol{W} = \boldsymbol{W} \tilde{\boldsymbol{V}} \boldsymbol{W}$. In the StocSum framework, ξ_{SKAT_j} can be estimated as the square of the singular values of $\frac{1}{\sqrt{B}} \boldsymbol{W} \tilde{\boldsymbol{U}}$ (**Supplementary Note 1**).

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

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

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

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

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

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

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

Meta-analysis

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

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

LD score regression

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

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

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

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

Sample correlation coefficients $\hat{r}_{jk}$ can be estimated as $\frac{w_j \boldsymbol{G}_{.j}{}^T \boldsymbol{L} \boldsymbol{G}_{.k} w_k}{N-1}$, where $\boldsymbol{G}_{.j}$ is the jth column of the genotype matrix $\boldsymbol{G}$, representing variant j, $\boldsymbol{L} = \left(\boldsymbol{I}_N - \boldsymbol{1}_N\left({\boldsymbol{1}_N}^T \boldsymbol{1}_N\right)^{-1} {\boldsymbol{1}_N}^T\right)$ 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 j) is a weight that standardizes $\boldsymbol{G}_{.j}$ to a unit variance.

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

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

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

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

$$l_j = \sum_{k=1}^{M_j} \tilde{r}^2_{adj_{jk}} = \left\{ \sum_{k=1}^{M_j} \left(1 + \frac{1}{B-2} + \frac{1}{N-2}\right) {\tilde{r}_{jk}}^2 \right\} - \frac{M_j}{B-2} - \frac{M_j}{N-2}$$

$$= \left(1 + \frac{1}{B-2} + \frac{1}{N-2}\right) \left(\frac{\widetilde{\boldsymbol{U}}\,{\widetilde{\boldsymbol{U}}_{j.}}^T}{B(N-1)} \circ \frac{\widetilde{\boldsymbol{U}}\,{\widetilde{\boldsymbol{U}}_{j.}}^T}{B(N-1)} \right)^T 1_{M_j} - \frac{M_j}{B-2} - \frac{M_j}{N-2}.$$

in which $\circ$ denotes the Hadamard product, and $\widetilde{\boldsymbol{U}}_{j.}$ is the jth row of $\widetilde{\boldsymbol{U}}$.

Whole genome sequence and phenotype data

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

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

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

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

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

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

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

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

Statistical Analyses

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

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

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

Conditional association tests. We performed conditional analyses for the seven regions associated with LDL at the genome-wide significance level of 5×10^{-8} from the single-variant analysis in HCHS/SOL (**Table S1**). For each region, we started with a sentinel variant with the smallest *P* value, and computed conditional association test *P* values for all variants in the flanking region (1 Mb) after adjusting for the sentinel variant. If there were any variants in a region with a conditional $P < 5\times10^{-8}$, we then selected the variant

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

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

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

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

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

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

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

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

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

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

(EuAfAm), including 9,092,238, 14,296,986, 9,410,628, 13,819,023 common variants with MAF > 1% (1000 Genomes Project Consortium), from 503 Eu, 894 EuAf, 682 EuAm, and 1,073 EuAfAm samples, respectively. With the LD scores from these internal and external references, the LDSC program was used to estimate heritability. For both LDSC and StocSum, we used a 1 Mb window around each index variant to calculate its LD score.

To evaluate the performance of StocSum, we also compared heritability estimates from LDSC (Sample) and StocSum (Sample) partitioned by different MAF bins. Common variants from HCHS/SOL and external reference panels were divided into 6 MAF bins, i.e., $1\% < MAF \leq 5\%$, $5\% < MAF \leq 10\%$, $10\% < MAF \leq 20\%$, $20\% < MAF \leq 30\%$, $30\% < MAF \leq 40\%$, and $40\% < MAF \leq 50\%$. Partitioned LD scores for different MAF bins were calculated by LDSC and StocSum, i.e., LDSC (Sample), LDSC (Eu), LDSC (EuAf), LDSC(EuAm), LDSC (EuAfAm), and StocSum (Sample). Partitioned heritability was estimated by the LDSC program with summary statistics for the phenotype LDL and partitioned LD scores.

Reference

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

8. Graham, S. E. *et al.* The power of genetic diversity in genome-wide association studies of lipids. *Nature* **600**, 675–679 (2021).
9. Mikhaylova, A. V. *et al.* Whole-genome sequencing in diverse subjects identifies genetic correlates of leukocyte traits: The NHLBI TOPMed program. *The American Journal of Human Genetics* **108**, 1836–1851 (2021).
10. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
11. Halldorsson, B. V. *et al.* The sequences of 150,119 genomes in the UK Biobank. *Nature* **607**, 732–740 (2022).
12. Lin, D. Y. & Zeng, D. Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* **34**, 60–66 (2010).
13. Han, B. & Eskin, E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *The American Journal of Human Genetics* **88**, 586–598 (2011).
14. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291–295 (2015).
15. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**, 1228–1235 (2015).
16. Speed, D. & Balding, D. J. SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nat Genet* **51**, 277–284 (2019).
17. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* **44**, 369–375 (2012).
18. Lee, S., Teslovich, T. M., Boehnke, M. & Lin, X. General framework for meta-analysis of rare variants in sequencing association studies. *The American Journal of Human Genetics* **93**, 42–53 (2013).
19. Liu, D. J. *et al.* Meta-analysis of gene-level tests for rare variant association. *Nat Genet* **46**, 200–204 (2014).
20. Feng, S. *et al.* Methods for Association Analysis and Meta-Analysis of Rare Variants in Families. *Genet Epidemiol* **39**, 227–238 (2015).
21. Chen, H. *et al.* Efficient variant set mixed model association tests for continuous and binary traits in large-scale whole-genome sequencing studies. *The American Journal of Human Genetics* **104**, 260–274 (2019).
22. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet* **47**, 1091–1098 (2015).
23. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet* **48**, 245–252 (2016).
24. Zhu, X. *et al.* Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. *The American Journal of Human Genetics* **96**, 21–36 (2015).
25. Liu, Z. & Lin, X. Multiple phenotype association tests using summary statistics in genome-wide association studies. *Biometrics* **74**, 165–175 (2018).
26. Turley, P. *et al.* Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat Genet* **50**, 229–237 (2018).

27. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat Genet* **47**, 1236–1241 (2015).
28. Lu, Q. *et al.* A powerful approach to estimating annotation-stratified genetic covariance via GWAS summary statistics. *The American Journal of Human Genetics* **101**, 939–964 (2017).
29. Wang, Y. *et al.* The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. *Genome Biol* **19**, 1–12 (2018).
30. Zhang, Y., Qi, G., Park, J.-H. & Chatterjee, N. Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. *Nat Genet* **50**, 1318–1326 (2018).
31. Voorman, A., Brody, J., Chen, H., Lumley, T. & David, B. seqMeta: An R package for meta-analyzing region-based tests of rare DNA variants. Preprint at (2017).
32. Zhan, X., Hu, Y., Li, B., Abecasis, G. R. & Liu, D. J. RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. *Bioinformatics* **32**, 1423–1426 (2016).
33. Feng, S., Liu, D., Zhan, X., Wing, M. K. & Abecasis, G. R. RAREMETAL: fast and powerful meta-analysis for rare variants. *Bioinformatics* **30**, 2828–2829 (2014).
34. Morgenthaler, S. & Thilly, W. G. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **615**, 28–56 (2007).
35. Li, B. & Leal, S. M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics* **83**, 311–321 (2008).
36. Madsen, B. E. & Browning, S. R. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* **5**, e1000384 (2009).
37. Morris, A. P. & Zeggini, E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol* **34**, 188–193 (2010).
38. Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics* **89**, 82–93 (2011).
39. Sun, J., Zheng, Y. & Hsu, L. A unified mixed-effects model for rare-variant association in sequencing studies. *Genet Epidemiol* **37**, 334–344 (2013).
40. Chen, H., Meigs, J. B. & Dupuis, J. Sequence kernel association test for quantitative traits in family samples. *Genet Epidemiol* **37**, 196–204 (2013).
41. Brody, J. A. *et al.* Analysis commons, a team approach to discovery in a big-data environment for genetic epidemiology. *Nat Genet* **49**, 1560–1563 (2017).
42. Li, Z. *et al.* A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies. *Nat Methods* 1–13 (2022).
43. Morrison, A. C. *et al.* Practical approaches for whole-genome sequence analysis of heart-and blood-related traits. *The American Journal of Human Genetics* **100**, 205–215 (2017).

44. Li, X. *et al.* Powerful, scalable and resource-efficient meta-analysis of rare variant associations in large whole genome sequencing studies. *Nat Genet* 1–11 (2022).
45. Li, G. *et al.* Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**, 84–98 (2012).
46. Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109–113 (2012).
47. Jin, F. *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290–294 (2013).
48. Heidari, N. *et al.* Genome-wide map of regulatory interactions in the human genome. *Genome Res* **24**, 1905–1917 (2014).
49. Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
50. Turner, A. W. *et al.* Single-nucleus chromatin accessibility profiling highlights regulatory mechanisms of coronary artery disease risk. *Nat Genet* **54**, 804–816 (2022).
51. Zhou, T. *et al.* Lupus enhancer risk variant causes dysregulation of IRF8 through cooperative lncRNA and DNA methylation machinery. *Nat Commun* **13**, 1–16 (2022).
52. Nasser, J. *et al.* Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**, 238–243 (2021).
53. McCullagh, P. & Nelder, J. Generalized Linear Models Second edition Chapman & Hall. Preprint at (1989).
54. Lin, X. & Zhang, D. Inference in generalized additive mixed modelsby using smoothing splines. *J R Stat Soc Series B Stat Methodol* **61**, 381–400 (1999).
55. Chen, H. *et al.* Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *The American Journal of Human Genetics* **98**, 653–666 (2016).
56. Hoffmann, T. J. *et al.* A large electronic-health-record-based genome-wide study of serum lipids. *Nat Genet* **50**, 401–413 (2018).
57. Peloso, G. M. *et al.* Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. *The American Journal of Human Genetics* **94**, 223–232 (2014).
58. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
59. Kurano, M. *et al.* Genome-wide association study of serum lipids confirms previously reported associations as well as new associations of common SNPs within PCSK7 gene with triglyceride. *J Hum Genet* **61**, 427–433 (2016).
60. Jiang, L. *et al.* A resource-efficient tool for mixed model association analysis of large-scale data. *Nat Genet* **51**, 1749–1755 (2019).
61. Wang, X. *et al.* Efficient gene–environment interaction tests for large biobank-scale sequencing studies. *Genet Epidemiol* **44**, 908–923 (2020).
62. Christakoudi, S., Evangelou, E., Riboli, E. & Tsilidis, K. K. GWAS of allometric body-shape indices in UK Biobank identifies loci suggesting associations with morphogenesis, organogenesis, adrenal cell renewal and cancer. *Sci Rep* **11**, 1–18 (2021).

63. Lotta, L. A. *et al.* Association of genetic variants related to gluteofemoral vs abdominal fat distribution with type 2 diabetes, coronary disease, and cardiovascular risk factors. *JAMA* **320**, 2553–2563 (2018).
64. Pulit, S. L. *et al.* Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. *Hum Mol Genet* **28**, 166–174 (2019).
65. Westerman, K. E. *et al.* GEM: scalable and flexible gene–environment interaction analysis in millions of samples. *Bioinformatics* **37**, 3514–3520 (2021).
66. Winkler, T. W. *et al.* The influence of age and sex on genetic associations with adult body size and shape: a large-scale genome-wide interaction study. *PLoS Genet* **11**, e1005378 (2015).
67. Wood, A. C. *et al.* Identification of genetic loci simultaneously associated with multiple cardiometabolic traits. *Nutrition, Metabolism and Cardiovascular Diseases* **32**, 1027–1034 (2022).
68. Zhu, Z. *et al.* Shared genetic and experimental links between obesity-related traits and asthma subtypes in UK Biobank. *Journal of Allergy and Clinical Immunology* **145**, 537–549 (2020).
69. Consortium, 1000 Genomes Project. A global reference for human genetic variation. *Nature* **526**, 68 (2015).
70. Mumbach, M. R. *et al.* HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat Methods* **13**, 919–922 (2016).
71. Mumbach, M. R. *et al.* Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nat Genet* **49**, 1602–1612 (2017).
72. Salameh, T. J. *et al.* A supervised learning framework for chromatin loop detection in genome-wide contact maps. *Nat Commun* **11**, 1–12 (2020).
73. Noordam, R. *et al.* Multi-ancestry sleep-by-SNP interaction analysis in 126,926 individuals reveals lipid loci stratified by sleep duration. *Nat Commun* **10**, 1–13 (2019).
74. Ripatti, P. *et al.* Polygenic hyperlipidemias and coronary artery disease risk. *Circ Genom Precis Med* **13**, e002725 (2020).
75. Tamai, K. *et al.* LDL-receptor-related proteins in Wnt signal transduction. *Nature* **407**, 530–535 (2000).
76. Wojcik, G. L. *et al.* Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514–518 (2019).
77. Galinsky, K. J. *et al.* Fast principal-component analysis reveals convergent evolution of ADH1B in Europe and East Asia. *The American Journal of Human Genetics* **98**, 456–472 (2016).
78. Agrawal, A., Chiu, A. M., Le, M., Halperin, E. & Sankararaman, S. Scalable probabilistic PCA for large-scale genetic variation data. *PLoS Genet* **16**, e1008773 (2020).
79. Pazokitoroudi, A. *et al.* Efficient variance components analysis across millions of genomes. *Nat Commun* **11**, 1–10 (2020).
80. Wu, Y. *et al.* Fast estimation of genetic correlation for biobank-scale data. *The American Journal of Human Genetics* **109**, 24–32 (2022).

81. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet* **50**, 1335–1341 (2018).
82. Lumley, T., Brody, J., Peloso, G., Morrison, A. & Rice, K. FastSKAT: Sequence kernel association tests for very large sets of markers. *Genet Epidemiol* **42**, 516–527 (2018).
83. Lee, S., Wu, M. C. & Lin, X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13**, 762–775 (2012).
84. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310–315 (2014).
85. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* **47**, D886–D894 (2019).
86. Rogers, M. F. *et al.* FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. *Bioinformatics* **34**, 511–513 (2018).
87. Li, X. *et al.* Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nat Genet* **52**, 969–983 (2020).
88. Zheng, J. *et al.* LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* **33**, 272–279 (2017).
89. Regier, A. A. *et al.* Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. *Nat Commun* **9**, 1–8 (2018).
90. LaVange, L. M. *et al.* Sample design and cohort selection in the Hispanic Community Health Study/Study of Latinos. *Ann Epidemiol* **20**, 642–649 (2010).
91. Sorlie, P. D. *et al.* Design and implementation of the Hispanic community health study/study of Latinos. *Ann Epidemiol* **20**, 629–641 (2010).
92. Baigent, C. Cholesterol Treatment Trialists'(CTT) Collaborators: Efficacy and safety of cholesterol-lowering treatment: prospective meta-analysis of data from 90,056 participants in 14 randomised trials of statins. *Lancet* **366**, 1267–1278 (2005).
93. Klarin, D. *et al.* Genetics of blood lipids among~ 300,000 multi-ethnic participants of the Million Veteran Program. *Nat Genet* **50**, 1514–1523 (2018).
94. Wright, J. D., Folsom, A. R., Coresh, J., Sharrett, A. R., Couper, D., Wagenknecht, L. E., & Heiss, G. The ARIC (Atherosclerosis Risk in Communities) study: JACC focus seminar 3/8. *Journal of the American College of Cardiology* **77**, 2939-2959 (2021).
95. 1000 Genomes Project Consortium. A map of human genome variation from population scale sequencing. *Nature* **467**, 1061 (2010).
96. Sofer, T. *et al.* A fully adjusted two-stage procedure for rank-normalization in genetic association studies. *Genet Epidemiol* **43**, 263–275 (2019).
97. Bakshi, A. *et al.* Fast set-based association analysis using summary data from GWAS identifies novel gene loci for human complex traits. *Sci Rep* **6**, 1–9 (2016).

98. Madsen, B. E. & Browning, S. R. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* **5**, e1000384 (2009).
99. Speed, D. *et al.* Reevaluation of SNP heritability in complex human traits. *Nat Genet* **49**, 986–992 (2017).
100. De Vlaming, R., Johannesson, M., Magnusson, P. K. E., Ikram, M. A. & Visscher, P. M. Equivalence of LD-score regression and individual-level-data methods. *Biorxiv* 211821 (2017).

Acknowledgements

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

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

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

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

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

This work was supported by NHLBI grant R01 HL145025.

Competing interests

The authors declare no competing interests.

Supplementary Note

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

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

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

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

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

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

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

$$\boldsymbol{\Xi}_{SKAT|Burden} \approx \tfrac{1}{B}\tfrac{1}{B}\boldsymbol{W}\widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{U}}^T\boldsymbol{W}\mathbf{1}_q\left(\mathbf{1}_q^T\boldsymbol{W}\widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{U}}^T\boldsymbol{W}\mathbf{1}_q\right)^{-1}\mathbf{1}_q^T\boldsymbol{W}\widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{U}}^T\boldsymbol{W}$$

$$= \frac{1}{B}\boldsymbol{W}\widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{U}}^T\boldsymbol{W} - \frac{1}{B}\boldsymbol{W}\widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{u}}(\widetilde{\boldsymbol{u}}^T\widetilde{\boldsymbol{u}})^{-1}\widetilde{\boldsymbol{u}}^T\widetilde{\boldsymbol{U}}^T\boldsymbol{W}$$

$$= \frac{1}{B}(\boldsymbol{W}\widetilde{\boldsymbol{U}} - \boldsymbol{W}\widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{u}}(\widetilde{\boldsymbol{u}}^T\widetilde{\boldsymbol{u}})^{-1}\widetilde{\boldsymbol{u}}^T)(\boldsymbol{W}\widetilde{\boldsymbol{U}} - \boldsymbol{W}\widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{u}}(\widetilde{\boldsymbol{u}}^T\widetilde{\boldsymbol{u}})^{-1}\widetilde{\boldsymbol{u}}^T)^T$$

$$= \frac{1}{B}\widetilde{\boldsymbol{U}}_{SKAT|Burden}\widetilde{\boldsymbol{U}}_{SKAT|Burden}^{\;T}.$$

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

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

Let r_{jk} be the Pearson correlation coefficient between variants j and k, the sample correlation coefficient $\hat{r}_{jk}$ can be estimated using individual-level centered and rescaled genotypes (with mean 0 and variance 1), namely, $\hat{r}_{jk} = \frac{w_j\,\boldsymbol{G}_{\cdot j}^T\boldsymbol{L}\boldsymbol{G}_{\cdot k}w_k}{N-1}$, where $\boldsymbol{G}_{\cdot j}$ and $\boldsymbol{G}_{\cdot k}$ are the jth and kth columns of the full genotype matrix $\boldsymbol{G}$, representing variants j and k, w_j and w_k are rescaling weights that standardize genotypes to a unit variance, and $\boldsymbol{L} =$

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

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

Therefore,

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

In LD score regression, the higher order term is ignored and the adjusted squared 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-Sullivan et al., 2015).

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

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

Therefore,

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

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

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

$$= \frac{1 - r_{jk}^2 - \frac{1 - r_{jk}^2}{N}}{B} + \frac{1 - \rho_{jk}^2}{N}.$$

Hence,

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

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

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

Supplementary Tables

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

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

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

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

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

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

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

1430

Supplementary Figures

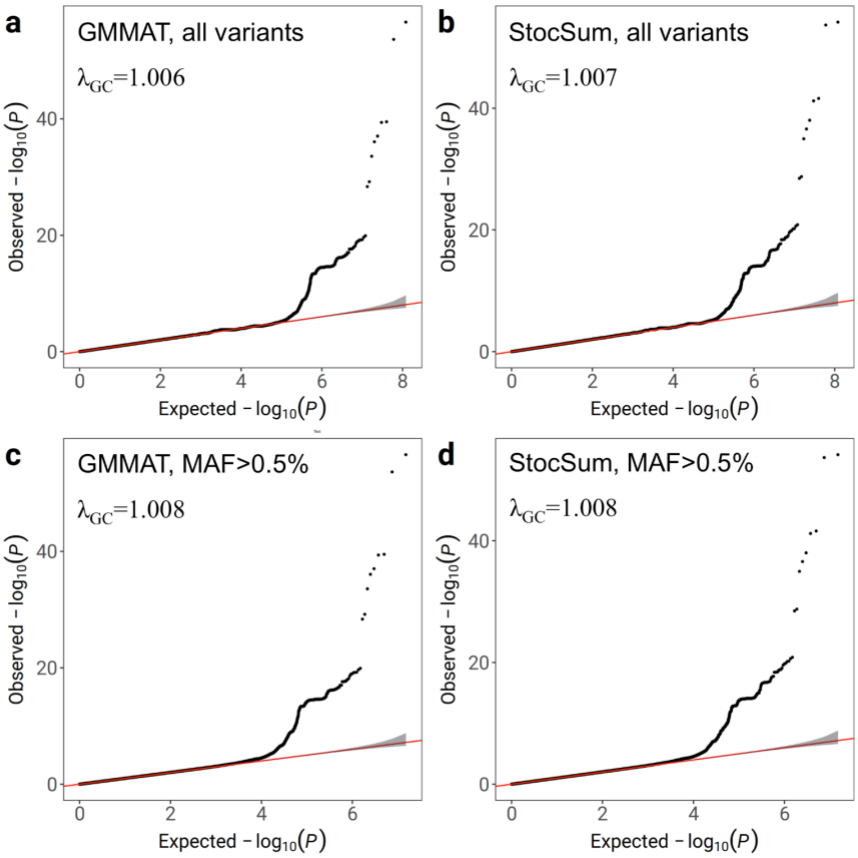


Figure S1 Quantile-quantile (Q-Q) plots of *P* values from single-variant tests on LDL cholesterol levels using GMMAT and StocSum in HCHS/SOL. The number of random vector replicates *B* in StocSum was set to 1,000. a, GMMAT *P* values from all variants. b, StocSum *P* values from all variants. c, GMMAT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

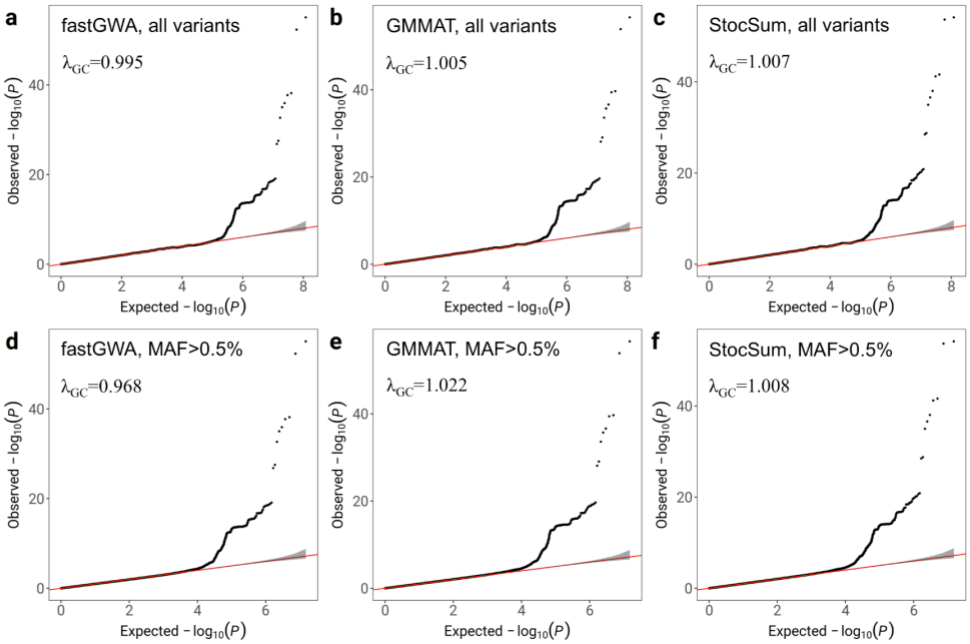


Figure S2. Quantile-quantile (Q-Q) plots of *P* values from single-variant tests on LDL cholesterol levels using fastGWA, GMMAT, and StocSum in HCHS/SOL. The number of random vector replicates *B* in StocSum was set to 1,000. a, fastGWA *P* values from all variants. b, GMMAT *P* values from all variants.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

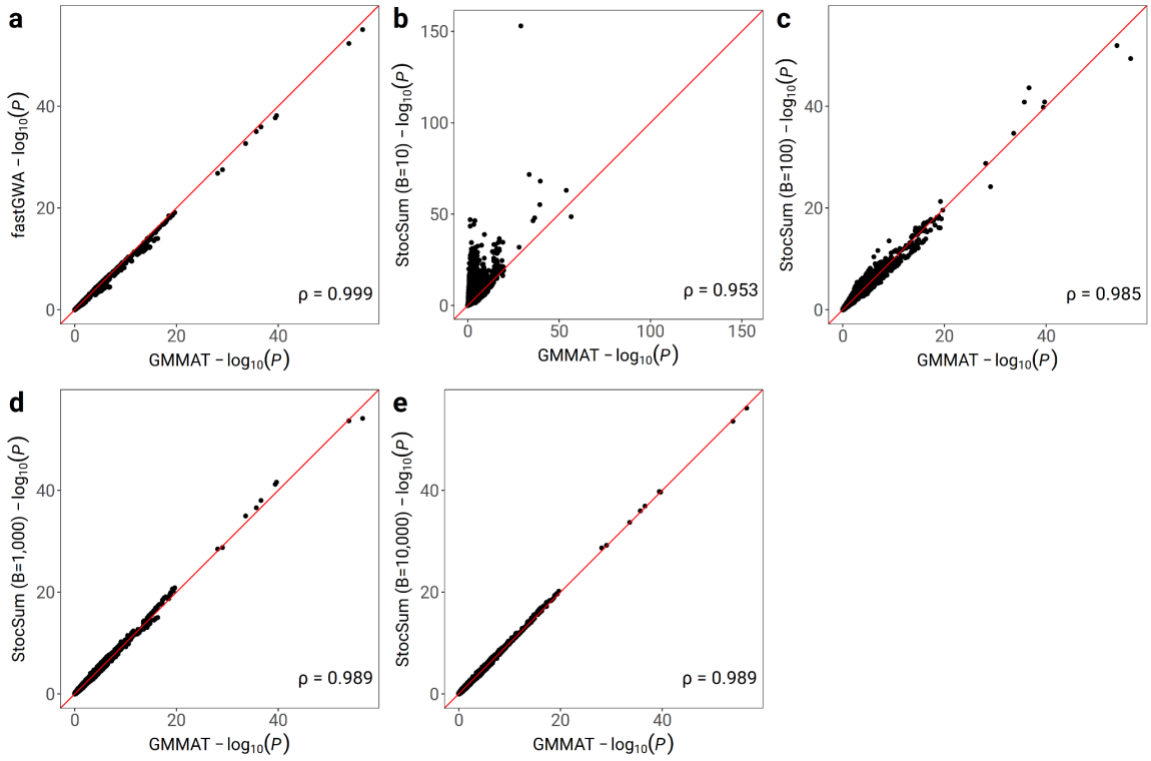


Figure S3 Comparison of *P* values from single-variant tests on LDL cholesterol levels using fastGWA, GMMAT, and StocSum in HCHS/SOL. a, comparison of *P* values from GMMAT and fastGWA. b-e, comparisons of *P* values from GMMAT and StocSum with the number of random vector replicates *B* being equal to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

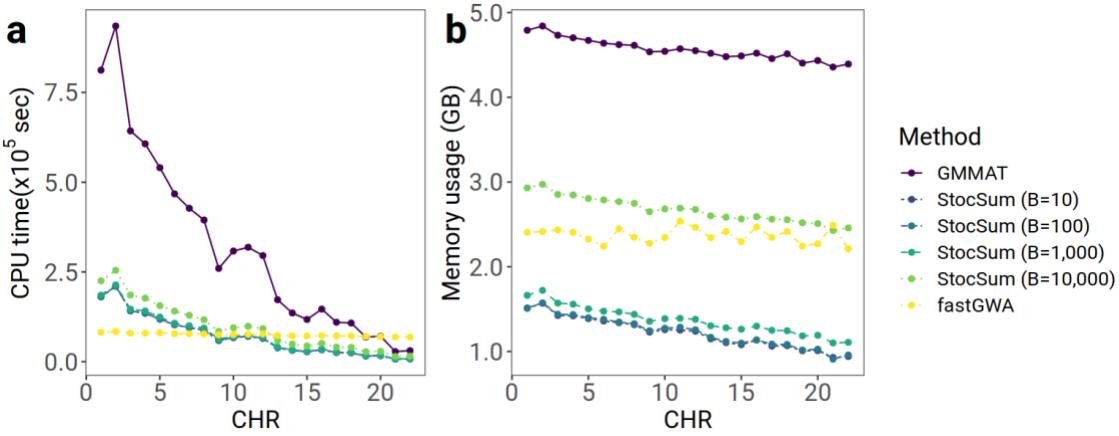
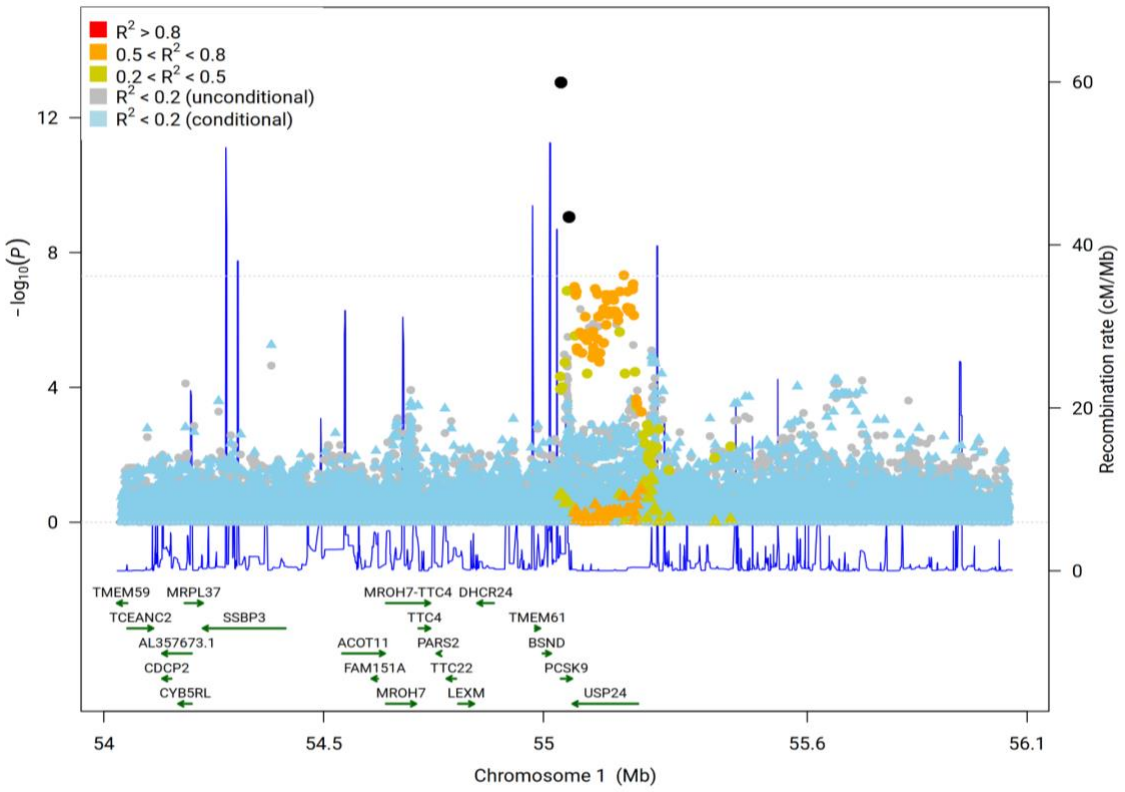
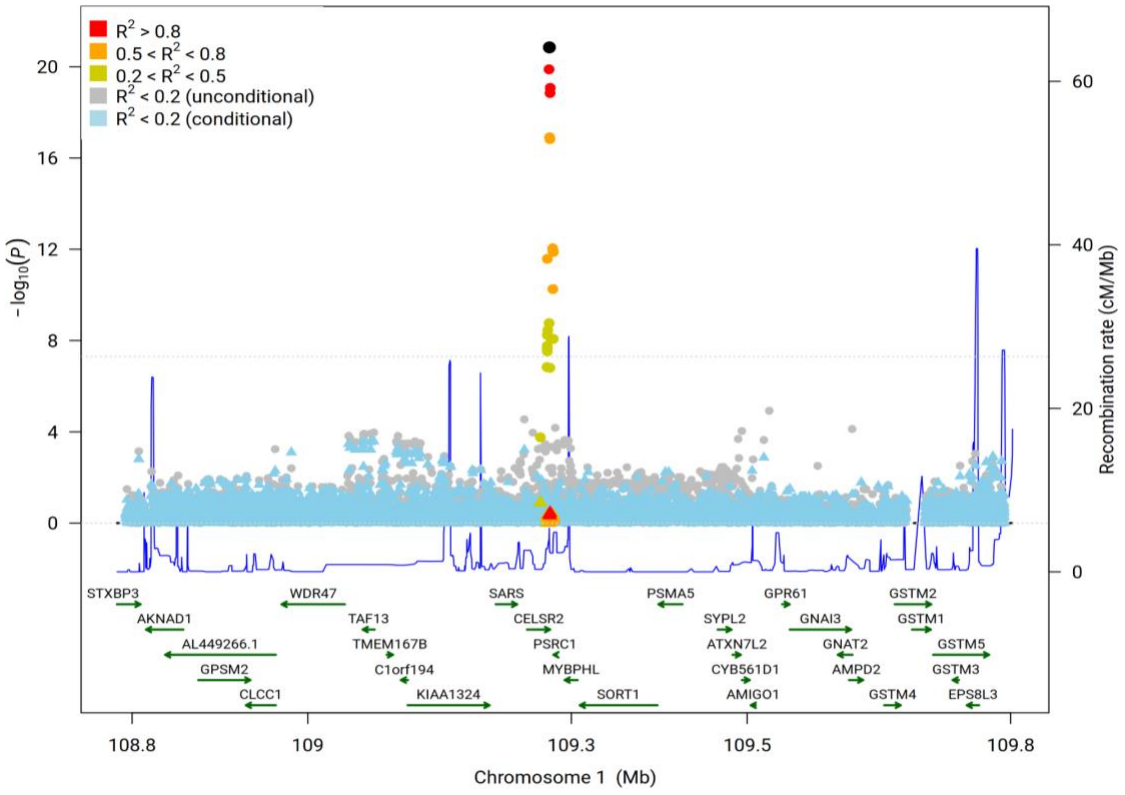


Figure S4 Comparison of CPU time and memory usage from fastGWA, GMMAT and StocSum in single-variant tests. a, CPU time. The x axis represents the chromosome numbers and the y axis represents the CPU time in 10^5 seconds. For GMMAT, the CPU time consists of fitting the null model and conducting the association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

a

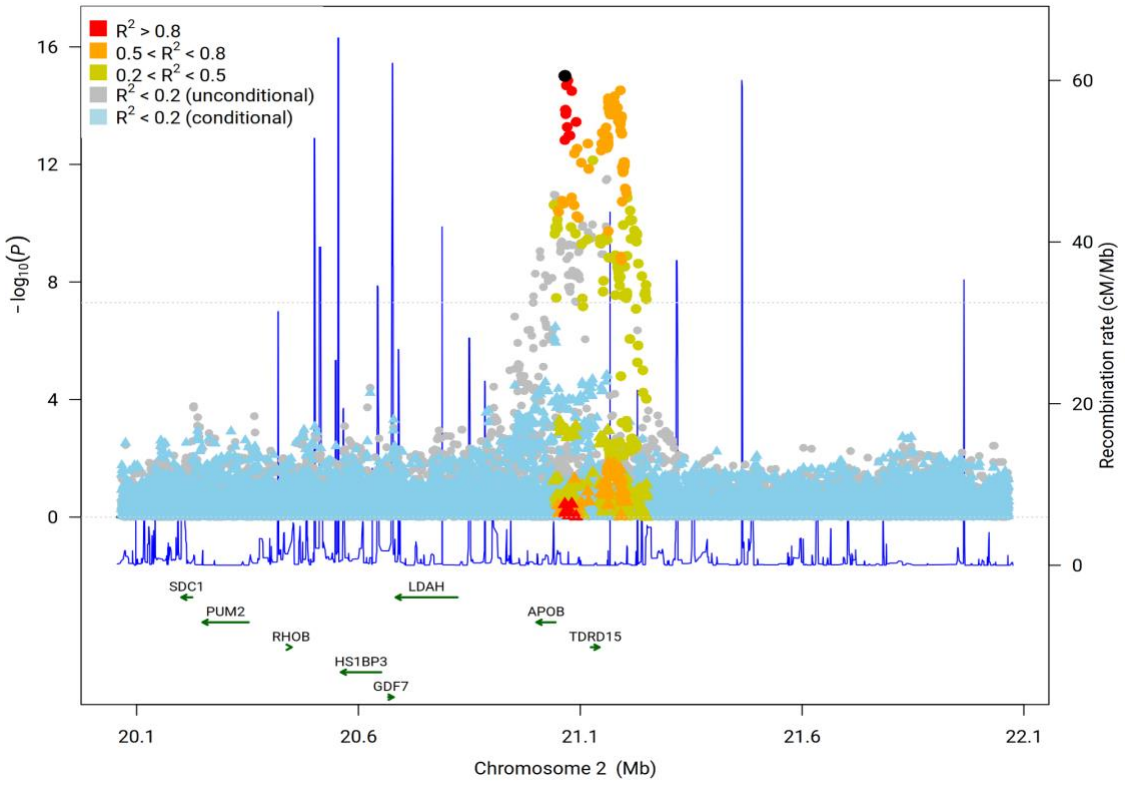


b



1477

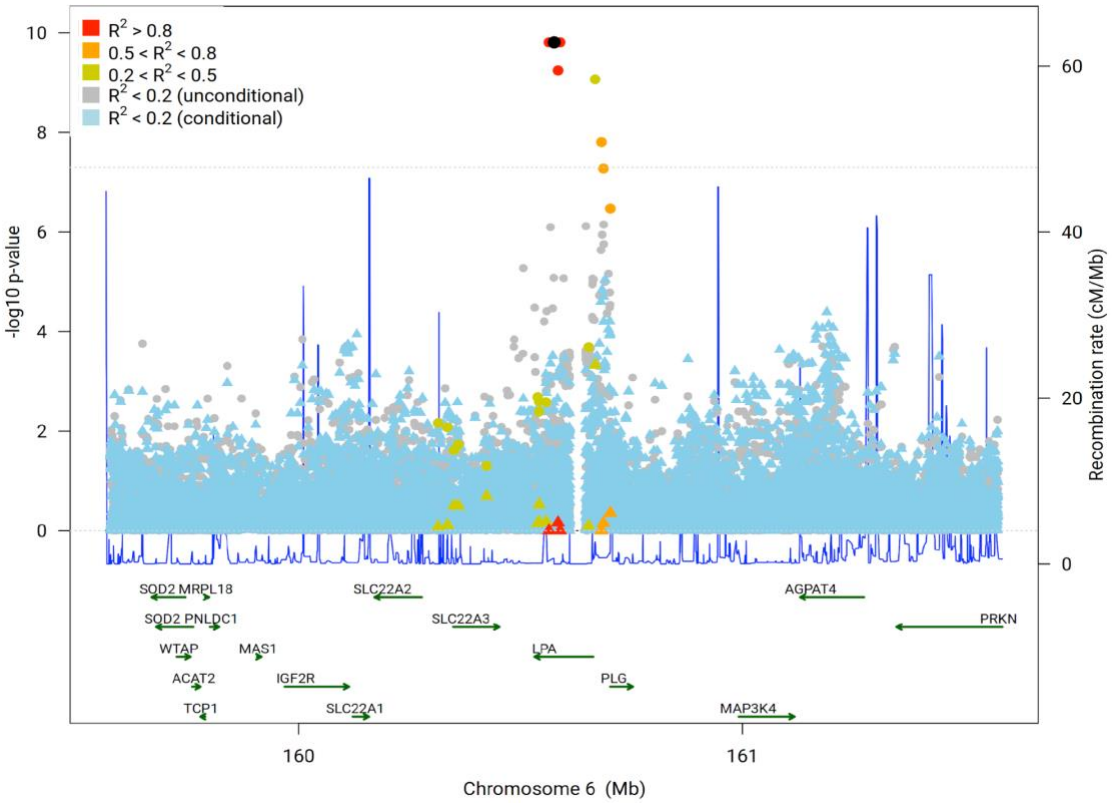
1478 c



1479

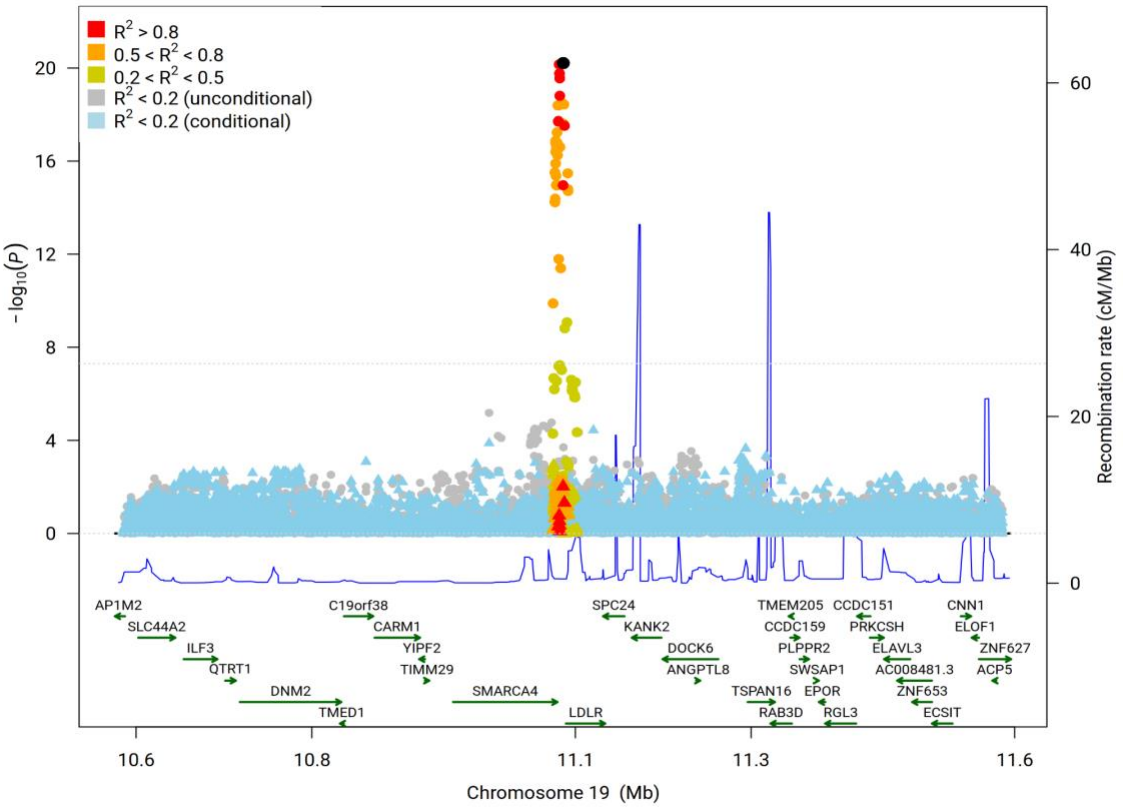
1480

1481 d



1482

1483 e



1484

1485

1486 f

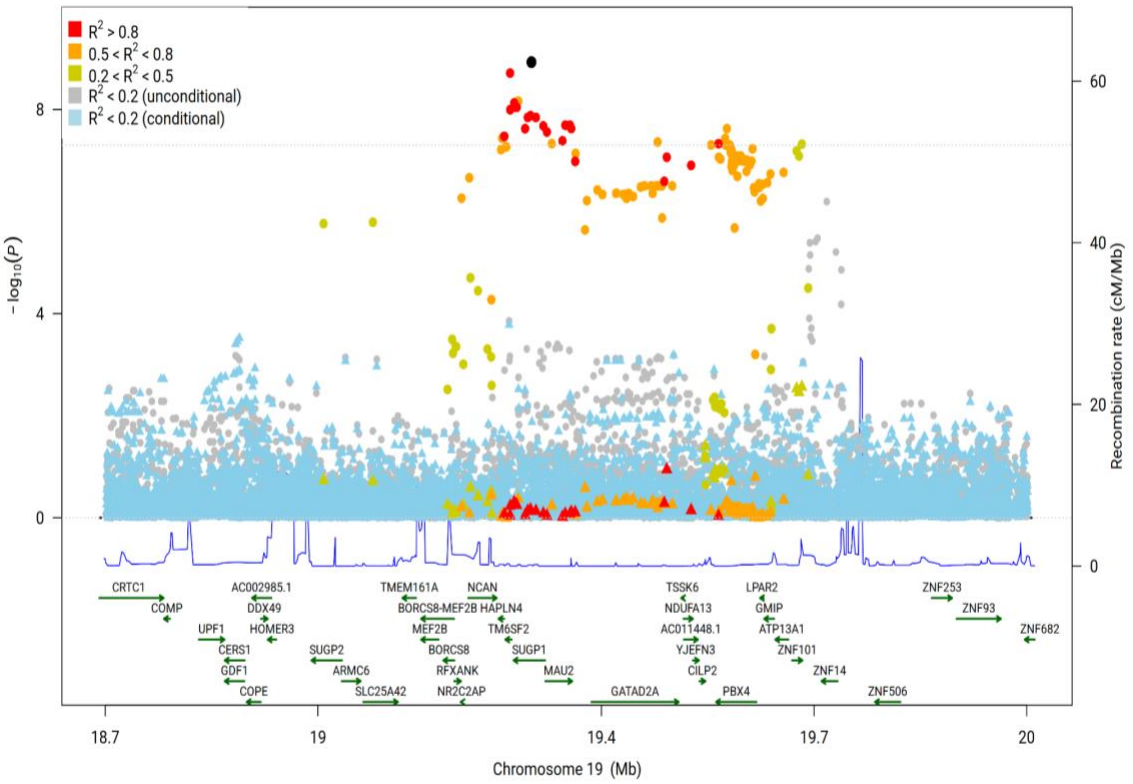


Figure S5 Regional plots of StocSum conditional association test results. a, *PCSK9* gene region with association variants chr1:55039974 (rs28362263) and chr1:55058182 (rs28362263). b, *CELSR2* gene region with the sentinel variant chr1:109274968 (rs562338). c, *APOB* gene region with the sentinel variant chr2:21065449 (rs562338). d, *LPA* gene region with the sentinel variant chr6:160576086 (rs10455872). e, *LDLR* gene region with the sentinel variant chr19:11086210 (rs8106503). f, *SUGP1* gene region with the sentinel variant chr19:19301236 (rs57915152). Association variants are highlighted in black dots. Original single-variant test *P* values are shown in dots and conditional *P* values are shown in triangles. Variants in four LD categories are shown in different colors based on the maximum squared correlation to the sentinel variant and the secondary association variant calculated in HCHS/SOL if there are two association variants (a), or the squared correlation to the sentinel variant in HCHS/SOL if there is only one sentinel association variant (b-f). The horizontal 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 Project.

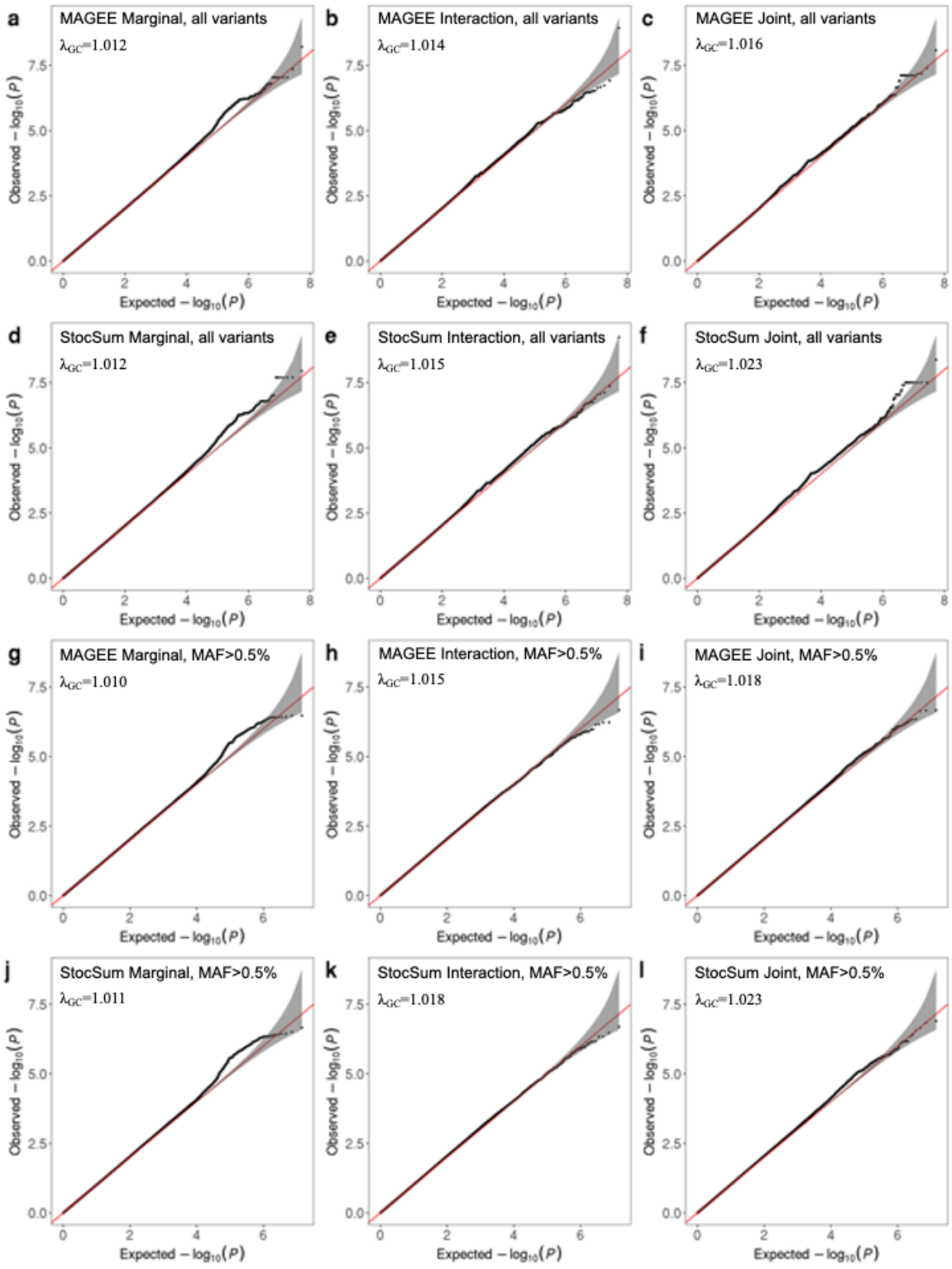


Figure S6 Quantile-quantile (Q-Q) plots of *P* values from gene-sex interaction tests on WHR using MAGEE and StocSum in HCHS/SOL. The number of random vector replicates *B* in StocSum was set to 1,000. a, Marginal *P* values for all variants from

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

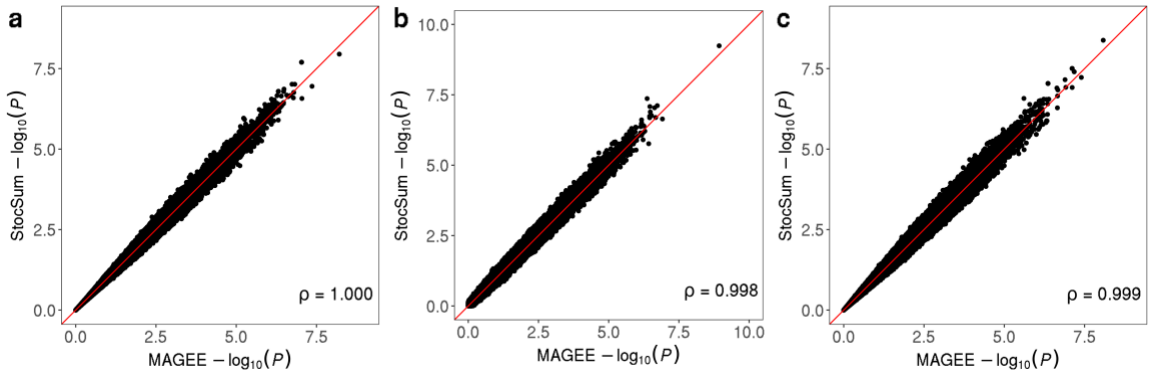
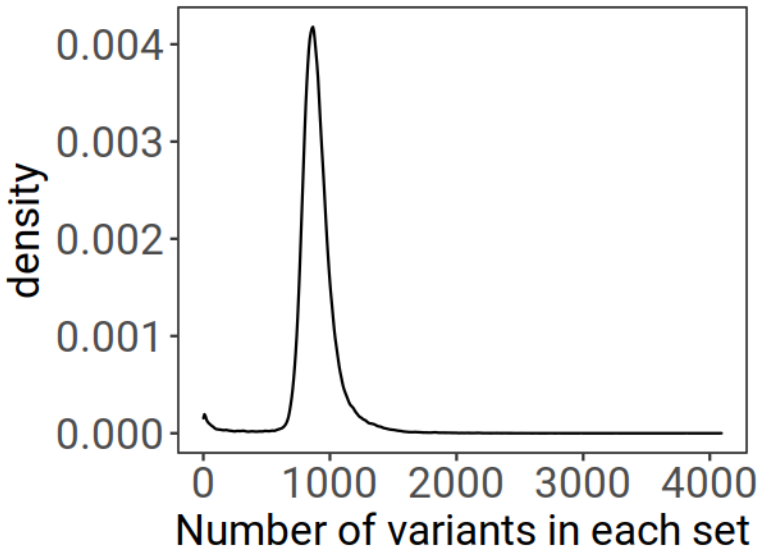


Figure S7 Comparison of *P* values from single-variant gene-sex interaction tests on WHR using MAGEE and StocSum in HCHS/SOL. a, comparison of marginal genetic effect test *P* values. b, comparison of gene-sex interaction test *P* values. c, comparison of joint test *P* values. The x axis and the y axis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



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

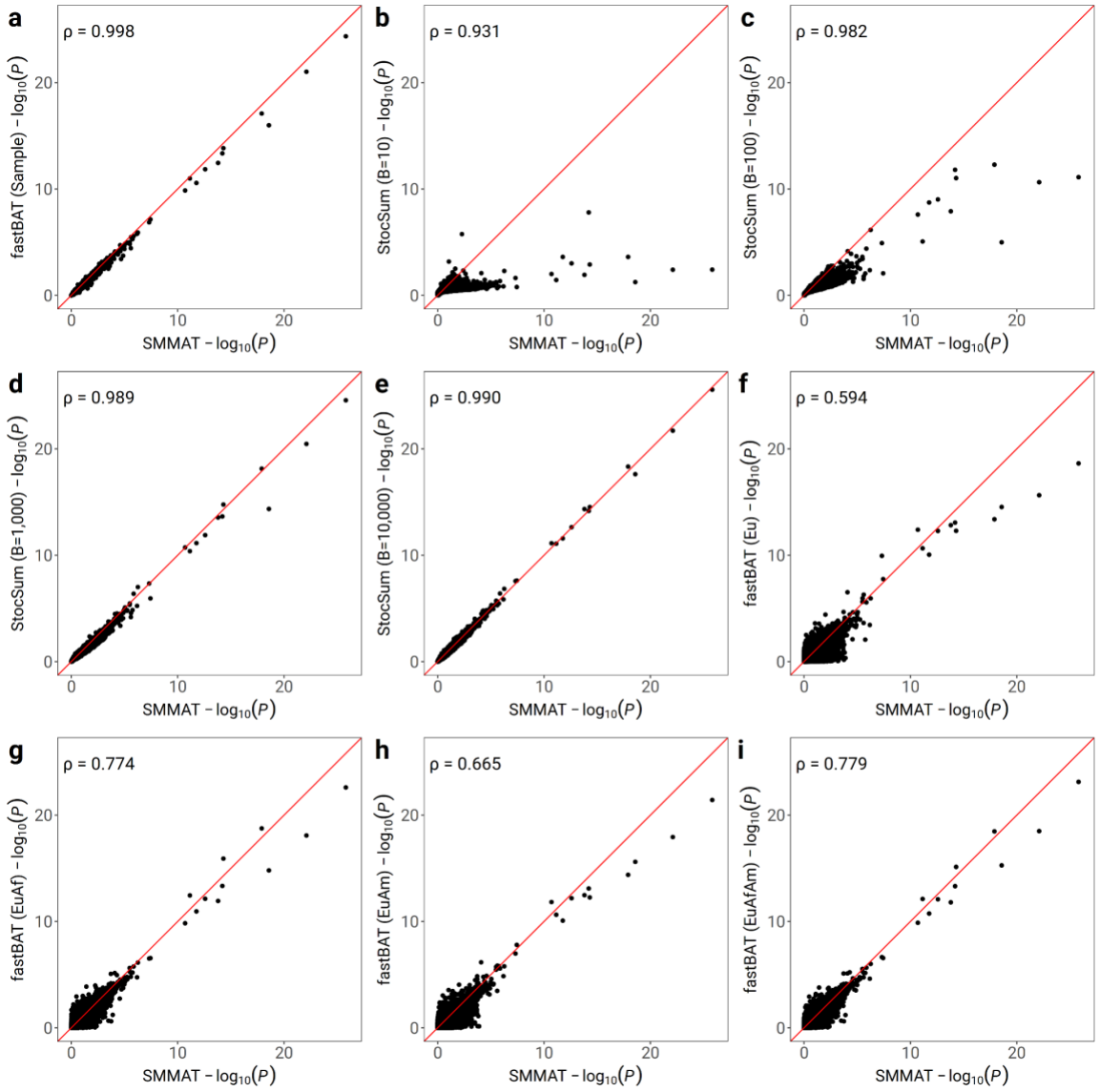


Figure S9 Comparison of *P* values from variant set tests in a 20 kb sliding window analysis on LDL cholesterol levels using fastBAT, SMMAT, and StocSum in HCHS/SOL. The x axis represents the –log10(*P*) from variant set tests using SMMAT on individual-level data and the y axis represents the –log10(*P*) from variant set tests using StocSum or fastBAT. a, fastBAT with an internal reference panel using the HCHS/SOL study samples (fastBAT (Sample)). b-e, StocSum with the number of random vector replicates *B* being equal to 10 (b), 100 (c), 1,000 (d) and 10,000 (e). f-i, fastBAT with external reference panels from 1000 Genomes using European (fastBAT (Eu)) (f), European and African (fastBAT (EuAf)) (g), European and American (fastBAT (EuAm)) (h), and European, African, and American (fastBAT (EuAfAm)) (i) populations. The red line denotes the reference line of

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

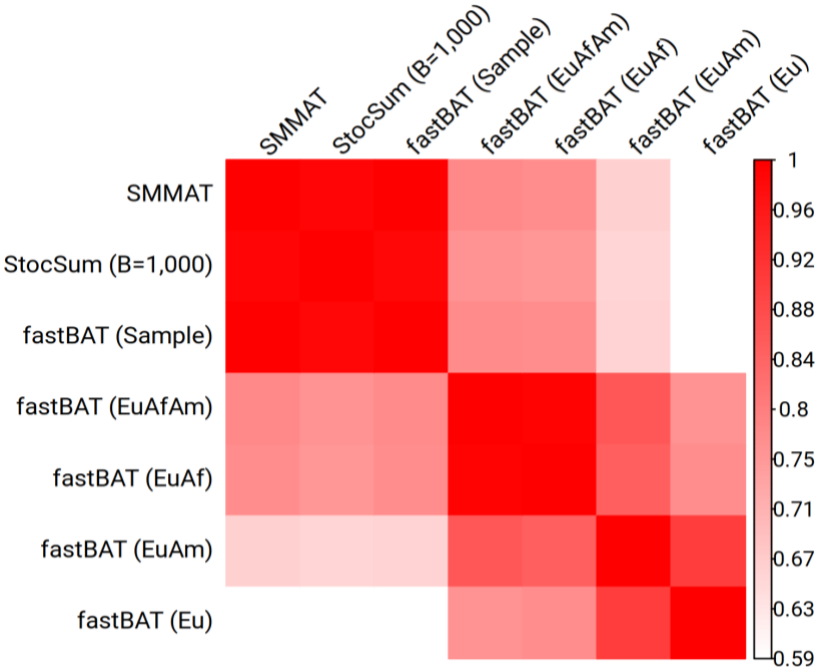


Figure S10 Heatmap showing Spearman's rank correlation coefficients of *P* values from variant set tests in a 20 kb sliding window analysis on LDL cholesterol levels using fastBAT, SMMAT, and StocSum in HCHS/SOL. For fastBAT, we used an internal reference panel using the HCHS/SOL study samples (fastBAT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

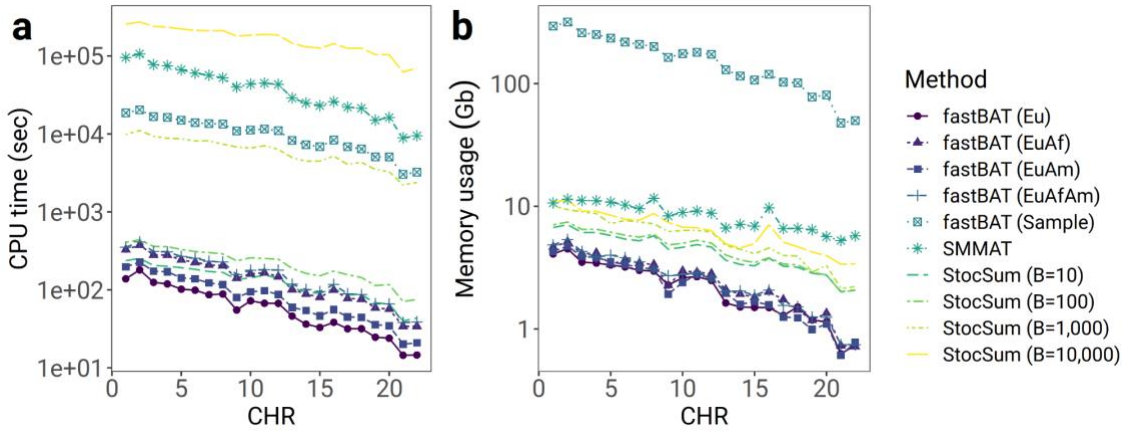


Figure S11 Comparison of CPU time and memory usage from fastBAT, SMMAT, and StocSum in variant set tests in a 20 kb sliding window analysis on LDL cholesterol levels in HCHS/SOL. a, CPU time. The x axis represents the chromosome numbers and the y axis represents the CPU time on the logarithmic scale.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

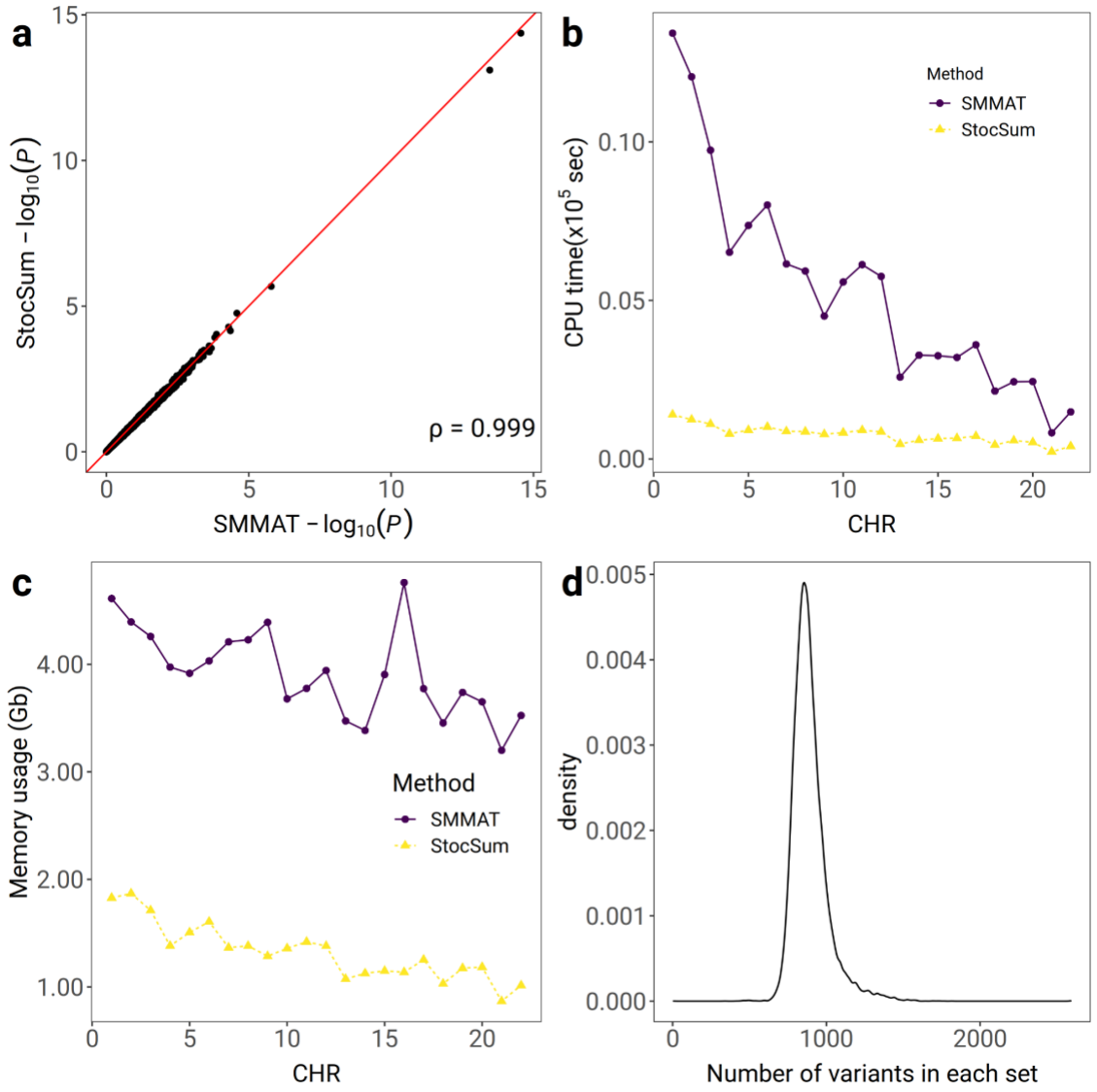


Figure S12 Comparison of SMMAT and StocSum variant set tests in a non-sliding-window analysis on LDL cholesterol levels in HCHS/SOL. The variant sets were defined by merging chromatin loops of H3K27ac HiChIP interaction in the GM12878 cell line. There are a total of 17,224 paired regions, each as a variant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

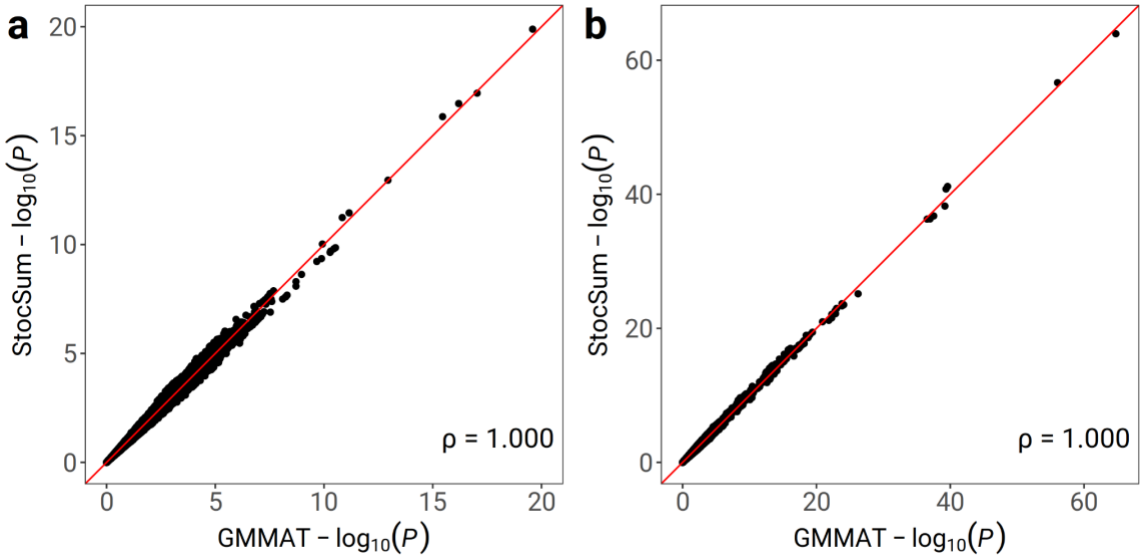


Figure S13 Comparison of *P* values from single-variant tests on longitudinal LDL cholesterol levels using GMMAT and StocSum in ARIC AA (a) and ARIC EA (b). The ARIC AA data used in this test consisted of 70M variants and 7,514 observations from 2,045 individuals. The ARIC EA data used in this test consisted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

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

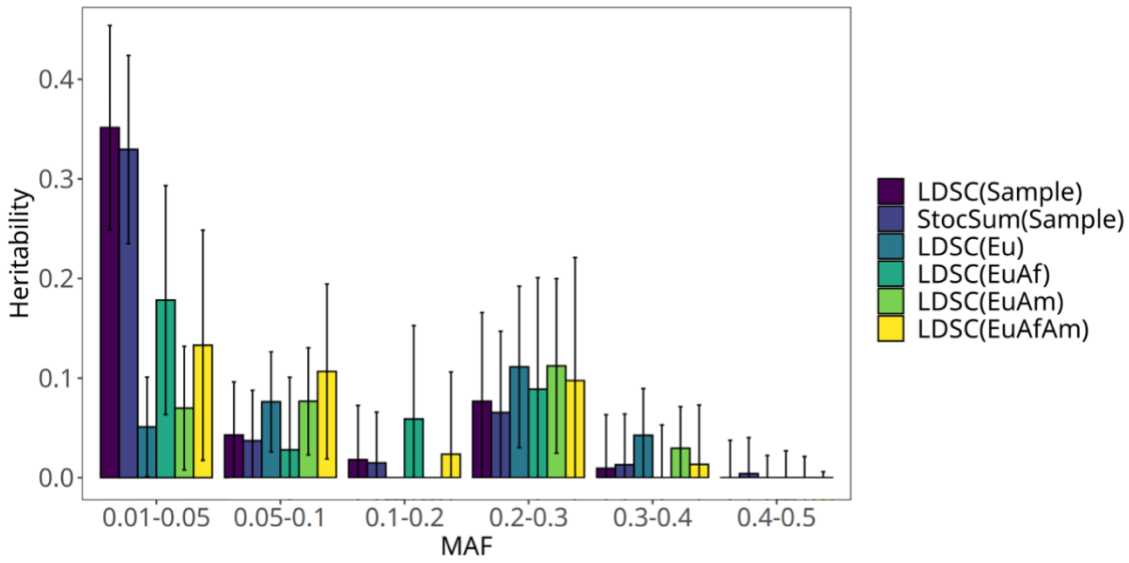


Figure S14 LDL heritability estimates by stratified LDSC and StocSum for different MAF bins. The error bars show point estimates ± standard errors. Negative heritability estimates reported from stratified LDSC were truncated at 0. LD scores for different MAF bins were estimated from LDSC (Sample) and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