

# Efficient variance components analysis across millions of genomes

Ali Pazokitoroudi<sup>1</sup>, Yue Wu<sup>1</sup>, Kathryn S. Burch<sup>2</sup>, Kangcheng Hou<sup>2</sup>, Aaron Zhou<sup>1</sup>, Bogdan Pasaniuc<sup>3,4,5</sup> & Sriram Sankararaman<sup>1,4,5</sup>  [✉](mailto:sriram@cs.ucla.edu)

While variance components analysis has emerged as a powerful tool in complex trait genetics, existing methods for fitting variance components do not scale well to large-scale datasets of genetic variation. Here, we present a method for variance components analysis that is accurate and efficient: capable of estimating one hundred variance components on a million individuals genotyped at a million SNPs in a few hours. We illustrate the utility of our method in estimating and partitioning variation in a trait explained by genotyped SNPs (SNP-heritability). Analyzing 22 traits with genotypes from 300,000 individuals across about 8 million common and low frequency SNPs, we observe that per-allele squared effect size increases with decreasing minor allele frequency (MAF) and linkage disequilibrium (LD) consistent with the action of negative selection. Partitioning heritability across 28 functional annotations, we observe enrichment of heritability in FANTOM5 enhancers in asthma, eczema, thyroid and autoimmune disorders.

<sup>1</sup>Department of Computer Science, UCLA, Los Angeles, CA 90095, USA. <sup>2</sup>Bioinformatics Interdepartmental Program, UCLA, Los Angeles, CA 90095, USA.

<sup>3</sup>Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, UCLA, Los Angeles, CA 90095, USA. <sup>4</sup>Department of Human Genetics, David Geffen School of Medicine, UCLA, Los Angeles, CA 90095, USA. <sup>5</sup>Department of Computational Medicine, David Geffen School of Medicine, UCLA, Los Angeles, CA 90095, USA. ✉email: [sriram@cs.ucla.edu](mailto:sriram@cs.ucla.edu)

Variance components analysis<sup>1</sup> has emerged as a versatile tool in human complex trait genetics, enabling studies of the genetic contribution to variation in a trait<sup>2</sup> as well as its distribution across genomic loci<sup>3,4</sup>, allele frequencies<sup>3</sup>, and functional annotations<sup>3,5,6</sup>. There is increasing interest in applying methods for variance components analysis to large-scale genetic datasets with the goal of uncovering novel insights into the genetic architecture of complex traits<sup>4,7</sup>. A prominent example of the utility of these methods is in the estimation of SNP heritability ( $h_{\text{SNP}}^2$ )<sup>2</sup>, the variance in a trait explained by a given set of genotyped SNPs. Variance components methods for estimating SNP heritability typically assume a genetic variance component that represents the fraction of phenotypic variation explained by the SNPs included in the study and a residual variance component. Recent studies have shown that these “single-component” methods yield biased estimates of SNP heritability due to the linkage disequilibrium (LD) and minor allele frequency (MAF)-dependent architecture of complex traits<sup>8,9</sup>. On the other hand, flexible models with multiple variance components<sup>3,4</sup> that allows for SNP effects to vary with MAF and LD, have been shown to yield more accurate SNP heritability estimates<sup>8,9</sup>. Recent work has shown that SNP heritability can be estimated with minimal assumptions about the genetic architecture<sup>10</sup>; however, this method cannot partition heritability across categories of SNPs of interest such as functional or population genomic annotations. Partitioning heritability requires fitting multiple variance components, thus creating the need for accurate and scalable methods that can fit tens or even hundreds of variance components to large-scale genomic data to obtain accurate and novel insights into genetic architecture.

While the ability to fit flexible variance component models to large-scale datasets is essential to obtain accurate and novel insights into genetic architecture, fitting such models requires scalable algorithms. Approaches for estimating variance components typically search for parameter values that maximize the likelihood or the restricted maximum likelihood (REML)<sup>11</sup>. Despite a number of algorithmic improvements<sup>2,4,12–16</sup>, computing REML estimates of the variance components on data sets such as the UK Biobank<sup>17</sup> (~500,000 individuals genotyped at nearly one million SNPs) remains challenging. The reason is that methods for computing these estimators typically perform repeated computations on the input genotypes.

We propose a method that can jointly estimate multiple variance components efficiently. Our proposed method, RHE-mc, is a randomized multi-component version of the classical Haseman–Elston regression for heritability estimation<sup>18,19</sup>. RHE-mc builds on our previously proposed method, RHE-reg<sup>20</sup>, which uses a randomized algorithm to estimate a single variance component. RHE-mc can simultaneously estimate multiple variance components, as well as variance components associated with continuous and overlapping annotations. Further, unlike REML estimation algorithms, RHE-mc requires only a single pass over the input genotypes that results in a highly memory efficient implementation. The resulting computational efficiency permits RHE-mc to jointly fit 300 variance components in less than an hour on a dataset of about 300,000 individuals and 500,000 SNPs, about two orders of magnitude faster than state-of-the-art methods. On a dataset of one million individuals and one million SNPs, RHE-mc can fit 100 variance components in about 12 h.

To demonstrate its utility, we first show that RHE-mc can accurately estimate genome-wide and partitioned SNP heritability under realistic genetic architectures (the functional dependence of SNP effect sizes on MAF and LD). We applied RHE-mc to 22 traits measured across 291,273 individuals genotyped at 459,792 common SNPs (MAF > 1%) in the UK Biobank to obtain estimates of genome-wide SNP heritability. We then used RHE-mc

to partition heritability for the 22 traits across seven million imputed SNPs (MAF > 0.1%) into 144 bins defined based on MAF and LD. We observe that the per-allele squared effect size tends to increase with lower MAF and LD across the traits considered. Finally, we partitioned heritability for SNPs with MAF > 0.1% across 28 functional annotations. We recover previously reported enrichment of heritability in annotations corresponding to conserved regions<sup>7</sup> and also document enrichment of heritability in FANTOM5 enhancers in eczema, asthma, autoimmune disorders, and thyroid disorders.

## Results

**Methods overview.** RHE-mc aims to fit a variance component model that relates phenotypes  $\mathbf{y}$  measured across  $N$  individuals to their genotypes over  $M$  SNPs  $\mathbf{X}$ :

$$\begin{aligned} \mathbf{y} | \boldsymbol{\epsilon}, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K &= \sum_{k=1}^K \mathbf{X}_k \boldsymbol{\beta}_k + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} &\sim \mathcal{D}(\mathbf{0}, \sigma_e^2 \mathbf{I}_N) \\ \boldsymbol{\beta}_k &\sim \mathcal{D}\left(\mathbf{0}, \frac{\sigma_k^2}{M_k} \mathbf{I}_{M_k}\right), k \in \{1, \dots, K\} \end{aligned}$$

where  $\mathcal{D}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  is an arbitrary distribution with mean  $\boldsymbol{\mu}$  and covariance  $\boldsymbol{\Sigma}$ . Each of the  $M$  SNPs is assigned to one of  $K$  non-overlapping categories so that  $\mathbf{X}_k$  is the  $N \times M_k$  matrix consisting of standardized genotypes of SNPs belonging to category  $k$  (note that the expected heritability is constant within categories when we use standardized genotypes).  $\boldsymbol{\beta}_k$  denotes the effect sizes of SNPs assigned to category  $k$  which are drawn from a zero-mean distribution with covariance parameter  $\frac{\sigma_k^2}{M_k} \mathbf{I}_{M_k}$  (the variance component of category  $k$ ) while  $\sigma_e^2$  is the residual variance.

In this model, the genome-wide SNP heritability is defined as:

$$h_{\text{SNP}}^2 = \frac{\sum_{k=1}^K \sigma_k^2}{\sum_{k=1}^K \sigma_k^2 + \sigma_e^2} \text{ while the SNP heritability of category } k \text{ is}$$

defined as:  $h_k^2 = \frac{\sigma_k^2}{\sum_{k=1}^K \sigma_k^2 + \sigma_e^2}$ . By choosing categories to represent genomic annotations of interest, e.g., chromosomes, allele frequencies, or functional annotations, these models can be used to estimate the phenotypic variation that can be attributed to the relevant annotation.

The key inference problem in this model is the estimation of the variance components:  $(\sigma_1^2, \dots, \sigma_K^2, \sigma_e^2)$ . These parameters are typically estimated by maximizing the likelihood or the restricted likelihood. Instead, RHE-mc uses a scalable method-of-moments estimator, i.e., finding values of the variance components such that the population moments match the sample moments<sup>18,19,21–23</sup>. RHE-mc uses a randomized algorithm that avoids explicitly computing  $N \times N$  genetic relatedness matrices that are required by method-of-moments estimators. Instead, it operates on a smaller matrix formed by multiplying the input genotype matrix with a small number of random vectors (see “Methods” section). The application of a randomized algorithm for SNP heritability estimation using a single variance component was proposed in our previous work, RHE-reg<sup>20</sup>. RHE-mc extends our previous work in several directions. RHE-mc can efficiently fit multiple variance components (both non-overlapping and overlapping) and can also handle continuous annotations. The resulting algorithm has scalable runtime as it only requires operating on the genotype matrix one time. Further, RHE-mc uses a streaming implementation that does not require all the genotypes to be stored in memory leading to scalable memory requirements (Supplementary Notes). Finally, RHE-mc uses an efficient implementation of a block Jackknife to estimate standard errors with little computational overhead (Supplementary Notes).

**Accuracy of genome-wide SNP heritability estimates in simulations.** We assessed the accuracy of RHE-mc in estimating genome-wide SNP heritability as previous attempts at estimating SNP heritability have been shown to be sensitive to assumptions about how SNP effect size varies with MAF and LD<sup>8</sup>. Starting with genotypes of  $M = 593,300$  array SNPs over  $N = 337,205$  unrelated white British individuals in the UK Biobank, we simulated phenotypes according to 64 MAF and LD-dependent architectures by varying the SNP heritability, the proportion of variants that have non-zero effects (causal variants or CVs), the distribution of CVs across minor allele frequencies (CVs distributed across all minor allele frequency bins or CVs restricted to either common or low-frequency bins), and the form of coupling between the SNP effect size and MAF as well as LD. For RHE-mc, we partitioned the SNPs into 24 variance components based on six MAF bins as well as four LD bins (see “Methods” section). The key parameter in applying RHE-mc is the number of random vectors  $B$  which we set to 10. RHE-mc estimates were relatively insensitive when we increased the number of random vectors  $B$  to 100 (Supplementary Figs. 1 and 2, Supplementary Table 1). Across these 64 architectures, RHE-mc is relatively unbiased (a two-sided  $t$ -test of the hypothesis of no bias is not rejected across any of the architectures at a  $p$ -value  $< 0.05$ ) with the largest relative bias observed to be 0.5% of the true SNP heritability (Supplementary Fig. 3). We used a block Jackknife (number of blocks = 100) to estimate the standard errors of RHE-mc and confirmed that the estimated standard errors are close to the true SE (Supplementary Table 2).

We compared the accuracy of RHE-mc to state-of-the-art methods for heritability estimation that can be applied to large datasets (across architectures where the true SNP heritability was fixed at 0.5). These methods, LDSC<sup>24</sup>, SumHer<sup>25</sup>, S-LDSC<sup>26</sup>, and GRE<sup>10</sup>, all leverage summary statistics while RHE-mc requires individual genotype data. We found that estimates from the summary-statistic methods tend to be sensitive to the underlying genetic architecture: across 16 architecture relative biases range from  $-31\%$  to  $27\%$  for LDSC,  $-27\%$  to  $5\%$  for S-LDSC, and  $-5\%$  to  $9\%$  for SumHer (Fig. 1). We also compared to a recently proposed method (GRE<sup>10</sup>) that only estimates genome-wide SNP heritability (without partitioning by MAF/LD) and observed that relative biases ranged from  $1\%$  to  $1.4\%$  for GRE and from  $-1.5\%$  to  $0.5\%$  for RHE-mc. We also considered architectures in which only rare variants are causal and found RHE-mc is accurate relative to other methods (Supplementary Fig. 4). These results further emphasize that RHE-mc can accurately estimate SNP-heritability through fitting multiple variance components.

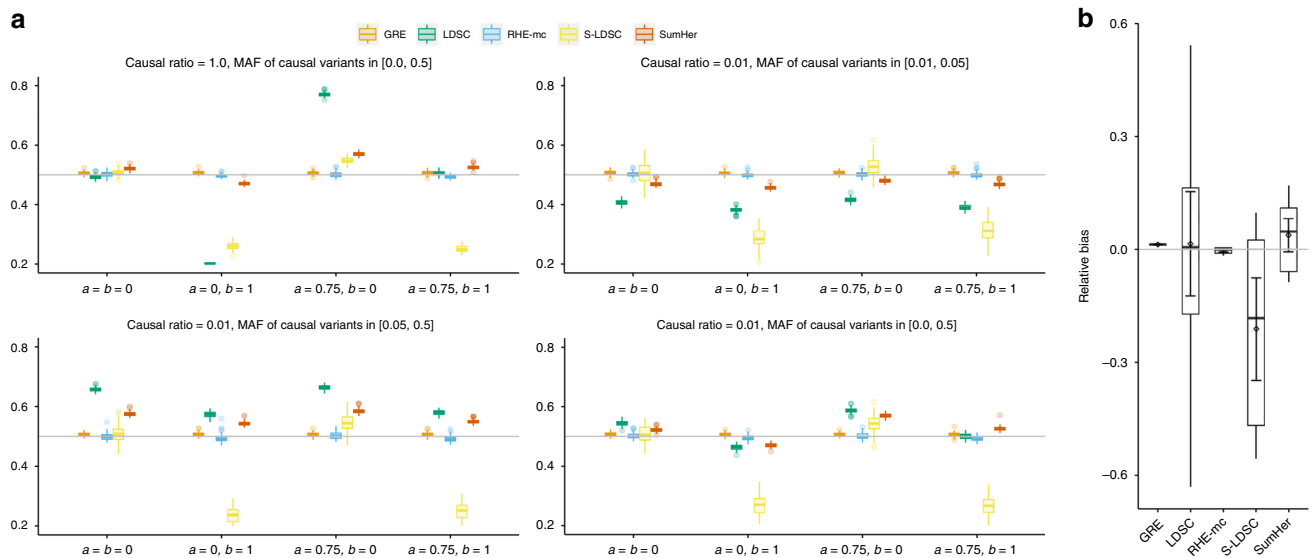
We compared RHE-mc to the state-of-the-art REML-based variance component estimation method, GCTA-mc (multi-component GREML<sup>8,27,28</sup>) and to exact multi-component Haseman-Elston Regression (HE-mc) as implemented in GCTA<sup>27</sup>. We ran each of these methods by partitioning SNPs into 24 variance components (6 MAF bins by 4 LD bins, see “Methods” section). To make these experiments computationally feasible, we simulated phenotypes starting from a smaller set of genotypes ( $M = 593,300$  array SNPs and  $N = 10,000$  white British individuals). Across 16 architectures where the true SNP heritability was fixed at 0.25, the relative biases for RHE-mc range from  $-3.2\%$  to  $3.6\%$ , and from  $-3.2\%$  to  $5\%$  for GCTA-mc (Fig. 2). On average, RHE-mc has standard errors that are 1.1 times larger than GCTA-mc (which range from 0.97 to 1.24) and 1.08 times larger than HE-mc (which range from 1.00 to 1.21).

**Accuracy of heritability partitioning in simulations.** We also evaluated the accuracy of RHE-mc in partitioning SNP heritability in both small-scale ( $M = 593,300$  SNPs,  $N = 10,000$

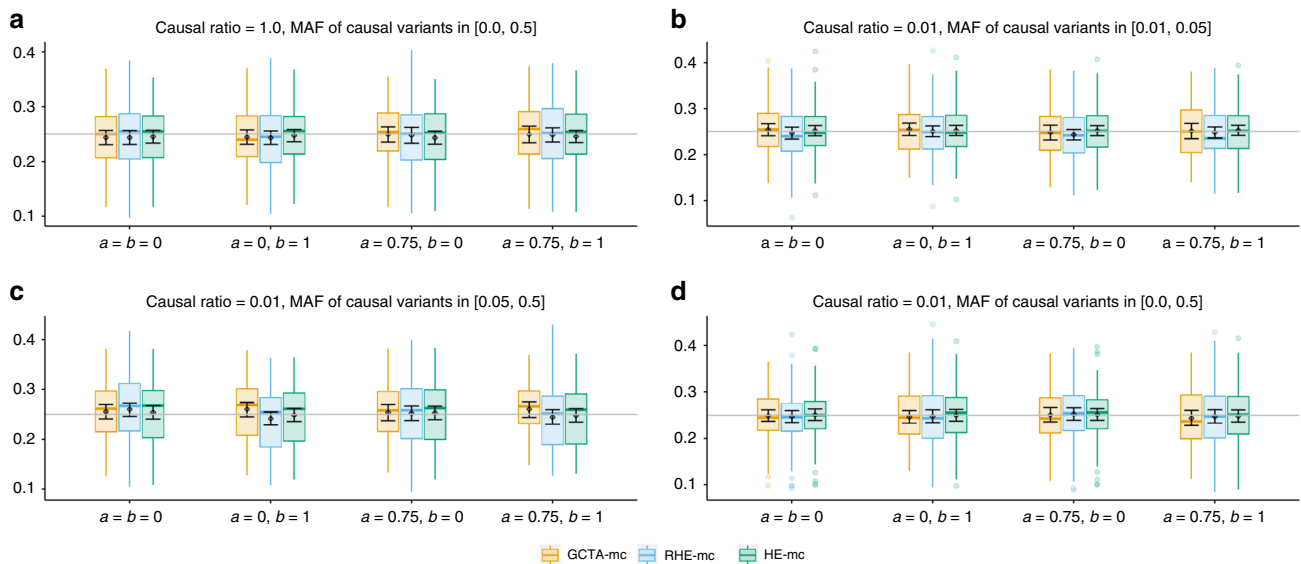
individuals) (Supplementary Fig. 5) and large-scale settings ( $M = 593,300$  SNPs,  $N = 337,205$  individuals) (see Supplementary Fig. 6). For these experiments, we restrict our attention to architectures for which the CVs are chosen to lie within a narrow range of MAF. Since the variance components correspond to bins of MAF and LD, a subset of the variance components would have no causal SNPs and hence have a heritability of zero. We assess the accuracy of estimates of heritability aggregated over these components (termed the non-causal bin) as well as the heritability aggregated over the remaining genetic components (termed the causal bin). For example, variance components that correspond to  $MAF \in [0.01, 0.05]$  would be included in the causal bin for an architecture that restricts the MAF of CVs to lie in the range  $[0.01, 0.05]$ . For the small-scale simulations, we compared RHE-mc to GCTA-mc. We ran both methods by partitioning the SNPs into 24 variance components based on six MAF bins as well as four LD bins defined by quartiles of the measure of LDK weight at a SNP (see “Methods” section). Across the genetic architectures tested, estimates of heritability within each of the causal and non-causal bins are highly concordant between RHE-mc and GCTA-mc (Supplementary Fig. 5, Supplementary Table 3): for the causal bin, the relative bias ranges from  $-4\%$  to  $0.4\%$  for RHE-mc and  $-3.6\%$  to  $2\%$  for GCTA-mc while, for the non-causal bin, the bias ranges from  $0$  to  $0.7\%$  for RHE-mc and  $0$  to  $1.4\%$  for GCTA-mc (Supplementary Table 3). For the large-scale settings, RHE-mc remains accurate: the relative bias ranges from  $-2.6\%$  to  $3.2\%$  (causal bin) and  $-0.5\%$  to  $0.2\%$  (non-causal bin) over the genetic architectures considered (Supplementary Fig. 6, Supplementary Table 4).

Heritability partitioning has been used to estimate heritability attributed to functional genomic annotations<sup>7</sup>. However, some of these annotations (such as FANTOM5 enhancers) are quite small covering  $< 1\%$  of the genome. We explored the ability of RHE-mc to accurately estimate heritability as a function of the size of the annotation. To this end, we performed simulations using  $N = 291,273$  unrelated white British individuals and  $M = 459,792$  common SNPs. We defined eight annotations (four MAF bins and two LD bins) in which we fixed the enrichment of a selected bin and varied the proportion of SNPs in the selected category. RHE-mc obtained accurate estimates of enrichment even when the selected bin only contained 0.4% of the genome-wide SNPs (comparable to the size of FANTOM5 enhancers). RHE-mc estimates are well-calibrated: when the bin has zero enrichment, RHE-mc rejected the null hypothesis of no enrichment in 5% of the simulations, while attaining high power to reject the null hypothesis even when the bin contained  $< 1\%$  of the SNPs (Supplementary Notes).

**Computational efficiency.** We benchmarked the runtime and memory usage of RHE-mc as a function of number of individuals, SNPs and variance components (Fig. 3, Table 1). We ran RHE-mc with  $B = 10$  random vectors and 22 variance components where each chromosome forms a distinct component. On a dataset of  $\approx 300,000$  individuals and  $\approx 500,000$  SNPs, RHE-mc can fit 22 variance components in less than an hour and  $\approx 300$  variance components (corresponding to bins of size 10 Mb) with little increase in its runtime. On a dataset of one million individuals and one million SNPs, RHE-mc can fit 100 variance components in a few hours. Further, due to its use of a streaming implementation that only requires the genotypes to be operated on once, the memory requirement of RHE-mc is modest: all experiments required  $< 60$  GB. We compared the run time and memory usage of RHE-mc with REML-based methods (GCTA<sup>27</sup> and BOLT-REML<sup>4</sup>) on the UK Biobank genotypes consisting of around 500,000 SNPs over varying sample sizes and observed that

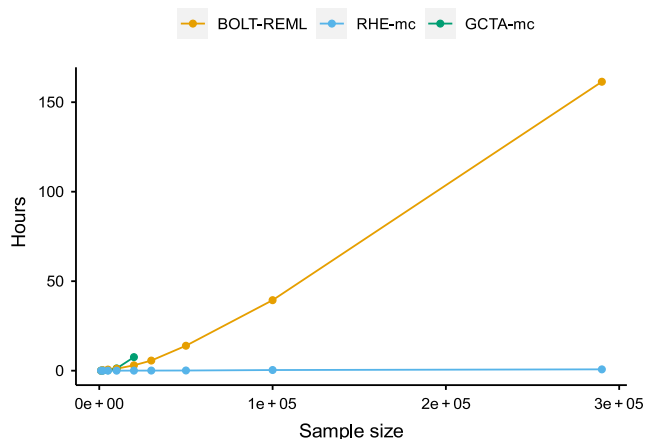


**Fig. 1 Comparison of estimates of genome-wide SNP heritability from RHE-mc with LDSC, GRE, S-LDSC, and SumHer in large-scale simulations ( $N = 337,205$  unrelated individuals,  $M = 593,300$  array SNPs).** **a** We compared methods for heritability estimation under 16 different genetic architectures. We set true heritability to 0.5 and varied the MAF range of causal variants (MAF of CV), the coupling of MAF with effect size ( $a = 0$  indicates no coupling with MAF and  $a = 0.75$  indicates coupling with MAF), and the effect of local LD on effect size ( $b = 0$  indicates no dependence on LDAK weights and  $b = 1$  indicates dependence on LDAK weights) (see “Methods” section). Each boxplot represents estimates from 100 simulations. **b** Relative bias of each method (as a percentage of the true  $h^2$ ) across 16 distinct MAF-dependent and LD-dependent architectures. Each boxplot contains 16 points; each point is the relative bias estimated from 100 simulations under a single genetic architecture. Points and error bars represent the mean and  $\pm 2$  SE. In **a** and **b**, boxplot whiskers extend to the minimum and maximum estimates located within  $1.5 \times$  interquartile range (IQR) from the first and third quartiles, respectively. Here, we run RHE-mc using 24 bins formed by the combination of six bins based on MAF as well as four bins based on quartiles of the LDAK score of a SNP (see “Methods” section). We run S-LDSC with only 10 MAF bins (see Supplementary Table 5). To do a fair comparison, for every method, we computed LD scores and LDAK weights by using in-sample LD, and in all simulations we aim to estimate the SNP-heritability explained by the same set of  $M$  SNPs.



**Fig. 2 Comparison of SNP heritability estimates from RHE-mc with GCTA-mc (GCTA with multiple variance components) and HE-mc (HE with multiple variance components) ( $N = 10,000$  unrelated individuals,  $M = 593,300$  array SNPs).** In **a–d** we compared heritability estimates from these methods under 16 different genetic architectures. We varied the MAF range of causal variants (MAF of CV), the coupling of MAF with effect size ( $a$ ), and the effect of local LD on effect size ( $b = 0$  indicates no dependence on LDAK weights and  $b = 1$  indicates dependence on LDAK weights) (see “Methods” section). We ran 100 replicates where the true heritability of the phenotype is 0.25. We run RHE-mc, HE-mc, and GCTA-mc using 24 bins formed by the combination of six bins based on MAF as well as four bins based on quartiles of the LDAK score of a SNP (see “Methods” section). Across all different genetic architectures, the relative biases range from  $-3.2\%$  to  $3.6\%$  for RHE-mc, and from  $-3.2\%$  to  $5\%$  for GCTA-mc, and from  $-2.6\%$  to  $1.45\%$  for HE-mc. On average, RHE-mc has SEs that are 1.1 and 1.08 times larger than GCTA-mc and HE-mc, respectively. Black points and error bars represent the mean and  $\pm 2$  SE. Each boxplot represents estimates from 100 simulations. Boxplot whiskers extend to the minimum and maximum estimates located within  $1.5 \times$  interquartile range (IQR) from the first and third quartiles, respectively. The SEs are computed from 100 simulations (note that GCTA-mc did not run successfully on all 100 simulations).





**Fig. 3 Comparison of running time of RHE-mc, GCTA-mc, and BOLT-REML.** We compared runtime of RHE-mc, GCTA-mc, and BOLT-REML with increasing sample size  $N$  (for a fixed number of SNPs  $M = 459,792$  and components  $K = 22$ ).

**Table 1 Comparison of running time of RHE-mc, GCTA-mc, and BOLT-REML.**

Parameters			Running time (h)		
$N$	$M$	$K$	RHE-mc	GCTA-mc	BOLT-REML
10,000	459,792	22	<1	1.3	1
100,000	459,792	22	<1	-	40
291,273	459,792	22	<1	-	162
291,273	459,792	300	<1	-	-
291,273	4,824,392	8	3.2	-	-
1,000,000	1,000,000	8	3	-	-
1,000,000	1,000,000	100	12.4	-	-

Here  $M$ ,  $N$ , and  $K$  are the number of SNPs, individuals, and variance components, respectively. RHE-mc can run efficiently even on datasets with one million individuals and SNPs as well as efficiently computing hundreds of variance components. All comparisons were performed on an Intel(R) Xeon(R) CPU 2.10 GHz server with 128 GB RAM.

RHE-mc achieves several orders-of-magnitude reduction in runtime. Summary-statistic methods such as S-LDSC requires pre-computed inputs which depend on the runtimes of other softwares making a direct comparison of speed difficult. Thus, we have restricted our comparison to individual-level methods where the benchmarking can be done in a comparable manner.

**Estimating total SNP heritability in the UK Biobank.** We applied RHE-mc to estimate genome-wide SNP heritability for 22 complex traits (6 quantitative and 16 binary traits) measured in the UK Biobank. We analyzed  $N = 291,273$  unrelated white British individuals and  $M = 459,792$  SNPs genotyped on the UK Biobank Axiom array (see “Methods” section). We ran RHE-mc with  $B = 10$  and with SNPs divided into eight bins based on two MAF bins ( $0.01 \leq \text{MAF} < 0.05$ ,  $\text{MAF} \geq 0.05$ ) and quartiles of the LD-scores. We compared the estimates from RHE-mc to those from LDSC, S-LDSC, SumHer, and GRE. Restricting our analysis to 18 traits for which the point estimate of genome-wide SNP heritability from RHE-mc is  $>0.05$ , the estimates from S-LDSC, GRE, SumHer, and LDSC were on average 2.5%, 10%, 25%, and 67% higher than RHE-mc (Fig. 4). Relative to the simulation results, the estimates from S-LDSC are generally consistent with those from RHE-mc. This is likely due to the fact that, in simulations, our application of S-LDSC used only MAF bins. On the

other hand, in real data, we used S-LDSC with the recommended baseline-LD annotations (including functional annotations).

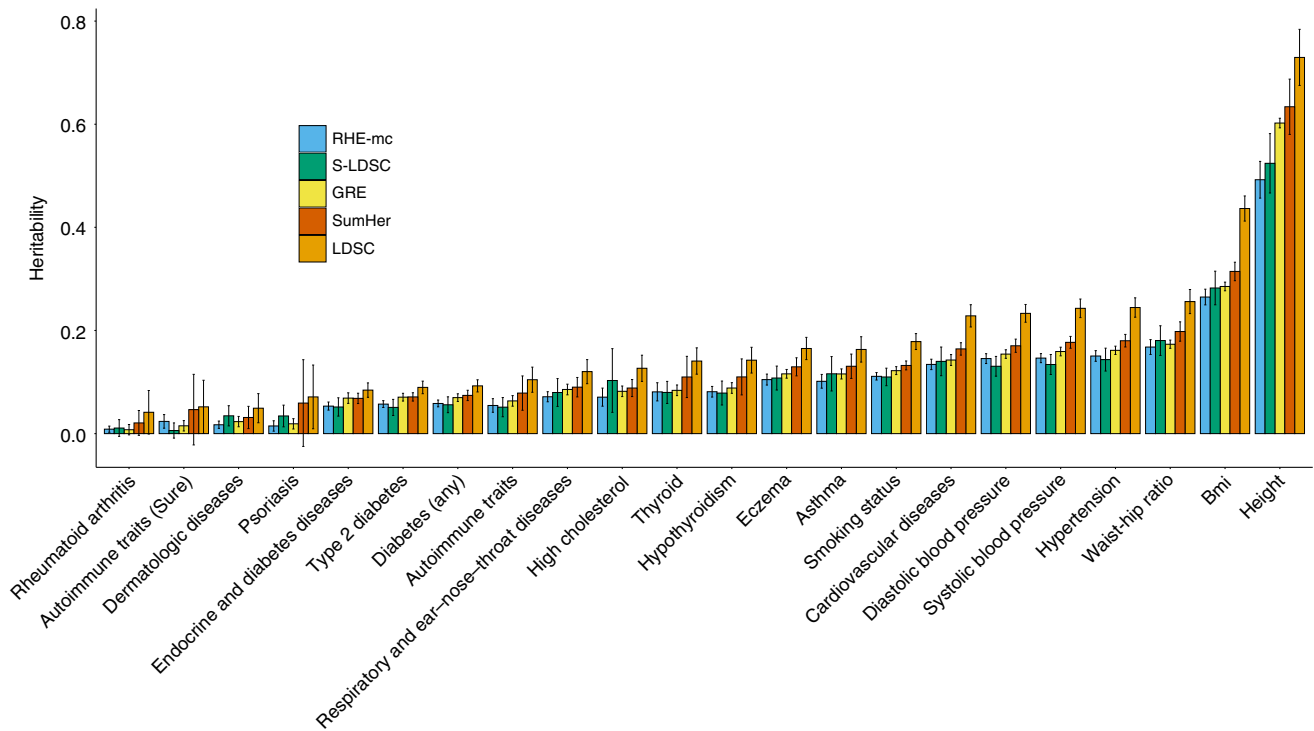
We then applied RHE-mc to estimate genome-wide heritability attributable to imputed variants. The genome-wide estimates of SNP heritability from RHE-mc on imputed SNPs ( $\text{MAF} > 1\%$ ) are concordant with the estimates from array SNPs (2.8% higher on average). We then analyzed  $M = 7,774,235$  imputed genotypes with  $\text{MAF} > 0.1\%$  using 144 bins formed by 4 LD bins and 36 MAF bins (see “Methods” section). Genome-wide SNP heritability estimates from RHE-mc on imputed SNPs ( $\text{MAF} > 0.1\%$ ) are 11.4% higher than RHE-mc on imputed SNPs ( $\text{MAF} > 1\%$ ) (Fig. 4, Supplementary Fig. 7). Following previous work<sup>10</sup>, we have removed the MHC region to enable a systematic comparison since the estimation of LD in the MHC region can be challenging; it would be of interest to compare methods when the MHC is included.

**Partitioning SNP heritability across allele frequency and LD bins.**

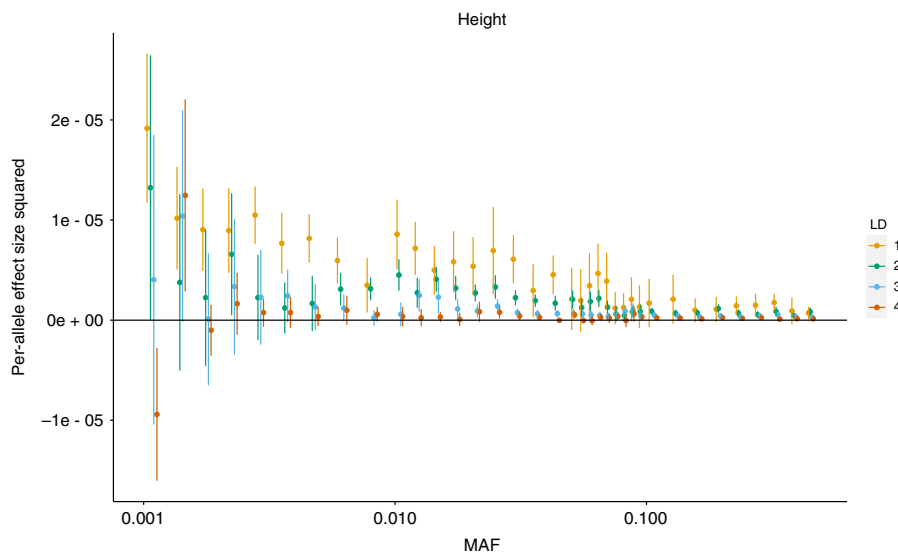
We used RHE-mc to partition SNP heritability of 22 complex traits across MAF and LD bins. We analyzed  $M = 7,774,235$  imputed SNPs with  $\text{MAF} > 0.1\%$ . We used 144 bins formed by 4 LD bins and 36 MAF bins (see “Methods” section). We compute the per-allele squared effect size of SNPs in bin  $k$  as  $\frac{h_k^2}{2f_k(1-f_k)M_k}$ , where  $h_k^2$  is the heritability estimated in bin  $k$ ,  $f_k$  is the mean MAF in bin  $k$ , and  $M_k$  is the number of SNPs in bin  $k$ . We observe that allelic effect size increases with lower MAF and LD. For height, in the lowest quartile of LD scores, SNPs with  $\text{MAF} \approx 0.1\%$  have allelic effect sizes  $\approx 27x \pm 8$  larger than SNPs with  $\text{MAF} \approx 50\%$ . Similarly, among SNPs with  $\text{MAF} \approx 50\%$ , SNPs in the lowest quartile of LD scores have allelic effect sizes  $\approx 5x \pm 1$  larger than SNPs in the highest quartile (Fig. 5 for height; other traits in Supplementary Fig. 9). While these trends have been observed in previous studies<sup>9,29,30</sup>, the ability of RHE-mc to jointly fit multiple variance components allows us to estimate effect sizes at SNPs with MAF as low as 0.1%. We caution that negative heritability estimates in bins of lowest MAF and high LD score could arise due to one or more of the following factors: low number of SNPs in this bin (we did not constrain our variance components estimates to be non-negative), the inadequacy of the assumed heritability model, and errors in the imputed genotypes used for the analysis.

**Partitioning heritability by functional annotations.**

The ability of RHE-mc to estimate variance components associated with a large number of overlapping annotations enables us to explore the contribution of a variety of functional genomic annotations to trait heritability using individual-level data in the UK Biobank. We applied RHE-mc to jointly partition heritability of 22 complex traits across 28 functional annotations as defined in ref. 7 restricting our analysis to  $N = 291,273$  unrelated white British individuals and  $M = 5,670,959$  imputed SNPs (we restrict to SNPs with  $\text{MAF} > 0.1\%$  which are also present in 1000 Genomes Project). We grouped the traits into five categories (auto-immune, diabetes, respiratory, anthropometric, cardiovascular); for a representative trait from each category, we report enrichment of each of the 28 functional annotations in Fig. 6 (see “Methods” section; for all traits see Supplementary Fig. 8). Our results are largely concordant with previous studies<sup>7,9</sup>: we observe enrichment of heritability across traits in conserved regions ( $Z$ -score  $> 3$  in 15 traits). We also observe enrichment of heritability at FANTOM5 enhancers (labeled Enhancer\_Andersson in Fig. 6) in asthma, eczema, autoimmune disorders (broad), hypothyroidism, and thyroid disorders ( $Z$ -score  $> 3$ ) even though these annotations cover only 0.4% of the analyzed SNPs.



**Fig. 4 Estimates of genome-wide SNP heritability from RHE-mc, LDSC, S-LDSC, GRE, and SumHer for 22 complex traits and diseases in the UK Biobank.** We restricted our analysis to  $N = 291,273$  unrelated white British individuals. We applied all methods to  $M = 459,792$  array SNPs ( $MAF > 1\%$ ). We ran S-LDSC with baseline-LD model. For every method, LD scores or LDK weights are computed using in-sample LD among the SNPs, and we aim to estimate the SNP-heritability explained by the same set of SNPs. RHE-mc was applied to array SNPs with eight MAF/LD bins. Black error bars mark  $\pm 2$  standard errors centered on the estimated heritability. We used a block Jackknife (number of blocks = 100) to estimate the standard errors. In Supplementary Fig. 7, we also report RHE-mc estimates of genome-wide SNP heritability on  $M = 4,824,392$  imputed SNPs ( $MAF > 1\%$ ) with 8 MAF/LD bins and  $M = 7,774,235$  imputed SNPs ( $MAF > 0.1\%$ ) with 144 MAF/LD bins (see “Methods” section).

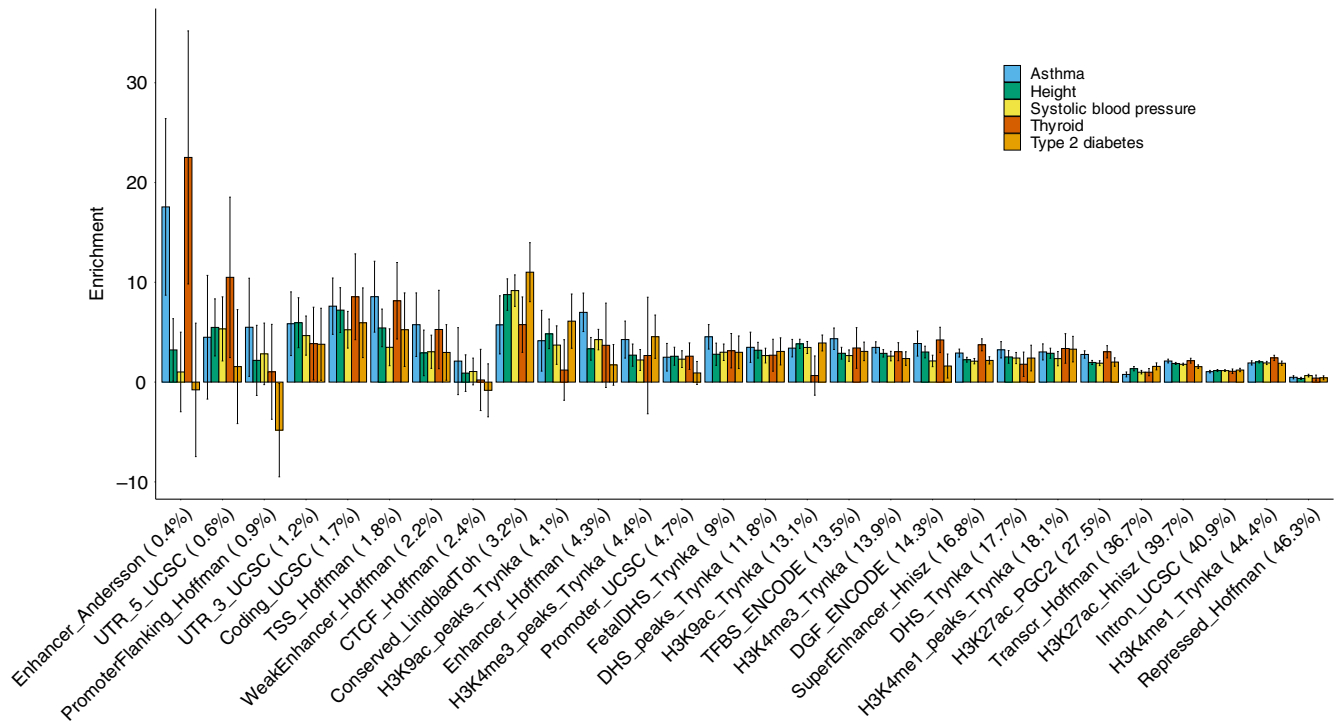


**Fig. 5 Per-allele squared effect size of height as a function of MAF.** We applied RHE-mc to  $N = 291,273$  unrelated white British individuals and  $M = 7,774,235$  imputed SNPs. SNPs were partitioned into 144 bins based on LD score (4 bins based on quartiles of the LD score with  $i$  denoting the  $i$ th quartile) and MAF (36 MAF bins) (see “Methods” section). Per-allele effect size squared for bin  $k$  is defined as  $\frac{h_k^2}{M_k * 2f_k * (1-f_k)}$ , where  $h_k^2$  is the heritability attributed to bin  $k$ ,  $M_k$  is the number of SNPs in bin  $k$ , and  $f_k$  is the average MAF in bin  $k$ . Each point represents the estimated allelic effect size. Bars mark  $\pm 2$  standard errors centered on the estimated allelic effect size. See Supplementary Fig. 9 for results on all 22 traits.

**Discussion**

We have presented RHE-mc, an algorithm that can efficiently estimate multiple variance components on large-scale genotype data. In light of increasing evidence for SNP effect sizes that vary

as a function of covariates, such as MAF and LD and the bias associated with methods that fit only a single variance component<sup>8</sup>, the ability to define flexible models endowed with multiple variance components is important to obtain unbiased estimates of



**Fig. 6 Enrichment of heritability across 28 functional annotations.** We applied RHE-mc to  $N = 291,273$  unrelated white British individuals and  $M = 5,670,959$  imputed SNPs ( $MAF > 0.1\%$  and present in 1000 Genomes Project). SNPs were partitioned based on 28 functional annotations that were defined in a previous study<sup>7</sup>. We grouped 22 traits in the UK Biobank into five categories (autoimmune, diabetes, respiratory, anthropometric, cardiovascular). Here we plot enrichment of five traits (one representative trait per category). Each bar represents the estimated enrichment. Black error bars mark  $\pm 2$  standard errors centered on the estimated enrichment. Annotations are ordered by the proportions of SNPs in that annotation (given in parentheses). See Supplementary Fig. 8 for results on all 22 traits.

fundamental quantities such as SNP heritability. We confirm that RHE-mc yields accurate genome-wide SNP heritability estimates under diverse genetic architectures. In applications to 22 complex traits in the UK Biobank, RHE-mc yields heritability estimates on array SNPs that are lower on average relative to S-LDSC and SumHer. We have explored the utility of RHE-mc in heritability partitioning analyses. These analyses show that per-allele squared effect sizes tend to increase with a decrease in MAF and LD consistent with previous studies<sup>9</sup>. We also partitioned heritability across functional annotations to reveal enrichment of heritability at FANTOM5 enhancers in specific traits such as asthma and eczema.

We discuss several limitations of RHE-mc as well as directions for future work. First, the method-of-moments estimator underlying RHE-mc tends to yield slightly larger standard errors, on average, relative to REML estimators. The relative performance of the two methods likely depends on a number of aspects of the study design such as sample size, number of SNPs, the LD structure, relatedness patterns, and the underlying genetic architecture. Nevertheless, our method is designed to be applicable to massive datasets for which the heritability estimates are relatively precise. Developing scalable variance components estimators that are as efficient as REML-based methods is an important direction for future work. Second, this work has primarily explored the partitioning of heritability across discrete annotations. While we have shown how the methodology can be extended to continuous-valued annotations (see “Methods” section and Supplementary Notes), it would be of interest to explore variation in trait heritability as a function of the value of an annotation. On the other hand, the ability of RHE-mc to fit many annotations allows the annotation to be divided into a sufficiently

large number of bins. Third, we have applied RHE-mc to binary traits available in the UK Biobank treating these traits as continuous. Methods that explicitly model binary traits as well as the underlying ascertainment involved in case-control studies are likely to lead to more accurate heritability estimates<sup>23,31</sup>. For example, the PCGC method<sup>23</sup> is an extension of HE regression and it would be of interest to develop a scalable randomized PCGC estimator. Fourth, RHE-mc requires access to individual-level genotype and phenotype data. Methods that only require summary statistic data (GRE<sup>10</sup>, LDSC<sup>24</sup>, and SumHer<sup>25</sup>) have the advantage of being applicable to datasets where acquiring access to individual-level data can be challenging<sup>10</sup>. Finally, our method could potentially lead to improvements in association testing, trait prediction, and understanding of polygenic selection.

**Methods**

**Multi-component linear mixed model.** RHE-mc attempts to fit the following variance component model:

$$\begin{aligned}
 y | \epsilon, \beta_1, \dots, \beta_K &= \sum_{k=1}^K X_k \beta_k + \epsilon \\
 \epsilon &\sim \mathcal{D}(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_N) \\
 \beta_k &\sim \mathcal{D}\left(\mathbf{0}, \frac{\sigma_k^2}{M_k} \mathbf{I}_{M_k}\right), k \in \{1, \dots, K\}
 \end{aligned}
 \tag{1}$$

Here  $y$  is a  $N$ -vector of centered phenotypes and each of the  $M$  SNPs is assigned to one of  $K$  non-overlapping categories. Each category  $k$  contains  $M_k$  SNPs,  $k \in \{1, \dots, K\}$ ,  $\sum_k M_k = M$ .  $X_k$  is a  $N \times M_k$  matrix, where  $x_{k,n,m}$  denotes the standardized genotype for individual  $n$  at SNP  $m$  in category  $k$ . We have  $\sum_n x_{k,n,m} = 0$  and  $\sum_n x_{k,n,m}^2 = N$  for  $m \in \{1, 2, \dots, M_k\}$ .  $\beta_k$  denote the  $M_k$ -vector of SNP effect sizes for the  $k$ th category where  $\mathcal{D}(\mu, \Sigma)$  is an arbitrary distribution with mean  $\mu$  and covariance  $\Sigma$ . In the above model,  $\sigma_\epsilon^2$  is the residual variance, and  $\sigma_k^2$  is the variance

component of the  $k$ th category. The total SNP heritability is defined as

$$h_{\text{SNP}}^2 = \frac{\sum_{k=1}^K \sigma_k^2}{\sum_{k=1}^K \sigma_k^2 + \sigma_e^2} \quad (2)$$

The SNP heritability of category  $k$  is defined as

$$h_k^2 = \frac{\sigma_k^2}{\sum_{k=1}^K \sigma_k^2 + \sigma_e^2}, k \in \{1, \dots, K\} \quad (3)$$

Enrichment in bin  $k$  is defined as

$$e_k = \frac{h_k^2/h_{\text{SNP}}^2}{M_k/M}, k \in \{1, \dots, K\} \quad (4)$$

**Method-of-moments for estimating multiple variance components.** To estimate the variance components, RHE-mc uses a Method-of-Moments (MoM) estimator that searches for parameter values so that the population moments are close to the sample moments<sup>32</sup>. Since  $\mathbb{E}[\mathbf{y}] = \mathbf{0}$ , we derived the MoM estimates by equating the population covariance to the empirical covariance. The population covariance is given by

$$\text{cov}(\mathbf{y}) = E[\mathbf{y}\mathbf{y}^T] - E[\mathbf{y}]E[\mathbf{y}^T] = \sum_k \sigma_k^2 \mathbf{K}_k + \sigma_e^2 \mathbf{I}_N \quad (5)$$

Here  $\mathbf{K}_k = \frac{\mathbf{X}_k \mathbf{X}_k^T}{M_k}$  is the genetic relatedness matrix (GRM) computed from all SNPs of  $k$ th category. Using  $\mathbf{y}\mathbf{y}^T$  as our estimate of the empirical covariance, we need to solve the following least-squares problem to find the variance components.

$$(\tilde{\sigma}_1^2, \dots, \tilde{\sigma}_K^2, \tilde{\sigma}_e^2) = \underset{(\sigma_1^2, \dots, \sigma_K^2, \sigma_e^2)}{\text{argmin}} \|\mathbf{y}\mathbf{y}^T - \left( \sum_k \sigma_k^2 \mathbf{K}_k + \sigma_e^2 \mathbf{I} \right)\|_F^2 \quad (6)$$

The MoM estimator satisfies the following normal equations:

$$\begin{bmatrix} \mathbf{T} & \mathbf{b} \\ \mathbf{b}^T & N \end{bmatrix} \begin{bmatrix} \tilde{\sigma}_g^2 \\ \tilde{\sigma}_e^2 \end{bmatrix} = \begin{bmatrix} \mathbf{c} \\ \mathbf{y}^T \mathbf{y} \end{bmatrix} \quad (7)$$

Here  $\tilde{\sigma}_g^2 = \begin{bmatrix} \tilde{\sigma}_1^2 \\ \vdots \\ \tilde{\sigma}_K^2 \end{bmatrix}$ ,  $\mathbf{T}$  is a  $K \times K$  matrix with entries  $T_{k,l} = \text{tr}(\mathbf{K}_k \mathbf{K}_l)$ ,  $k, l \in \{1, \dots, K\}$ ,

$\mathbf{b}$  is a  $K$ -vector with entries  $b_k = \text{tr}(\mathbf{K}_k) = N$  (because  $\mathbf{X}_k$  is standardized), and  $\mathbf{c}$  is a  $K$ -vector with entries  $c_k = \mathbf{y}^T \mathbf{K}_k \mathbf{y}$ . Each GRM  $\mathbf{K}_k$  can be computed in time  $\mathcal{O}(N^2 M_k)$  and  $\mathcal{O}(N^2)$  memory. Given  $K$  GRMs, the quantities  $T_{k,b}$ ,  $c_k$ ,  $k, l \in \{1, \dots, K\}$ , can be computed in  $\mathcal{O}(K^2 N^2)$ . Given the quantities  $T_{k,b}$ ,  $c_k$ , the normal Eq. (7) can be solved in  $\mathcal{O}(K^3)$ . Therefore, the total time complexity for estimating the variance components is  $\mathcal{O}(N^2 M + K^2 N^2 + K^3)$ .

**RHE-mc: Randomized estimator of multiple variance components.** The key bottleneck in solving the normal Eq. (7) is the computation of  $T_{k,b}$ ,  $k, l \in \{1, \dots, K\}$  which takes  $\mathcal{O}(N^2 M)$ . Instead of computing the exact value of  $T_{k,b}$  we use an unbiased estimator of the trace<sup>33</sup> based on the following identity: for a given  $N \times N$  matrix  $\mathbf{C}$ ,  $\mathbf{z}^T \mathbf{C} \mathbf{z}$  is an unbiased estimator of  $\text{tr}(\mathbf{C})$  ( $E[\mathbf{z}^T \mathbf{C} \mathbf{z}] = \text{tr}(\mathbf{C})$ ), where  $\mathbf{z}$  be a random vector with mean zero and covariance  $\mathbf{I}_N$ . Hence, we can estimate the values  $T_{k,b}$ ,  $k, l \in \{1, \dots, K\}$  as follows:

$$T_{k,l} = \text{tr}(\mathbf{K}_k \mathbf{K}_l) \approx \widehat{T}_{k,l} = \frac{1}{B} \frac{1}{M_k M_l} \sum_b \mathbf{z}_b^T \mathbf{X}_k \mathbf{X}_k^T \mathbf{X}_l \mathbf{X}_l^T \mathbf{z}_b \quad (8)$$

Here  $\mathbf{z}_1, \dots, \mathbf{z}_B$  are  $B$  independent random vectors with zero mean and covariance  $\mathbf{I}_N$ . We draw these random vectors independently from a standard normal distribution. Computing  $T_{k,l}$  using the unbiased estimator involves four multiplications of sub-matrices of the genotype matrix with a vector, repeated  $B$  times. Therefore, the total running time for estimating the matrix  $\mathbf{T}$  is  $\mathcal{O}(NMB + K^2 NB)$ .

Moreover, we can leverage the structure of the genotype matrix which only contains entries in  $\{0, 1, 2\}$ . For a fixed genotype matrix  $\mathbf{X}_k$ , we can improve the per iteration time complexity of matrix-vector multiplication from  $\mathcal{O}(NM)$  to  $\mathcal{O}(\frac{NM}{\max(\log_2 N, \log_3 M)})$  by using the Mailman algorithm<sup>34</sup>. Solving the normal equations takes  $\mathcal{O}(K^3)$  time so that the overall time complexity of our algorithm is  $\mathcal{O}(\frac{NMB}{\max(\log_2(N), \log_3(M))} + K^2(K + NB))$ .

RHE-mc uses a block Jackknife to estimate standard errors. In Supplementary Notes, we show how the block Jackknife estimates can be computed with little additional computational overhead. Further, we also show how covariates can be efficiently included in the model (Supplementary Notes).

**Multi-component LMM with overlapping annotations.** RHE-mc can also be applied in the setting where annotations overlap. Following ref.<sup>7</sup>, the heritability of

SNPs belong to annotation  $k$  is defined as

$$h_k^2 = \frac{\sum_{i \in S_k} \sum_{j \in S_k} \frac{\sigma_{ij}^2}{M_k}}{\sum_{k=1}^K \sigma_k^2 + \sigma_e^2}, k \in \{1, \dots, K\} \quad (9)$$

where  $S_k$  is the set of SNPs in  $k$ th annotation and  $M_k = |S_k|$ . Enrichment in bin  $k$  is defined as  $e_k = \frac{h_k^2/h_{\text{SNP}}^2}{M_k/M}$ .

**Multi-component LMM with continuous annotations.** We have described the derivation of RHE-mc using binary annotations. Following ref.<sup>29</sup>, we can extend RHE-mc to support continuous-value annotations as follows:

$$\begin{aligned} \mathbf{y} | \boldsymbol{\epsilon}, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K &= \sum_{k=1}^K \mathbf{X}_k \boldsymbol{\beta}_k + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} &\sim \mathcal{D}(\mathbf{0}, \sigma_e^2 \mathbf{I}_N) \\ \boldsymbol{\beta}_k &\sim \mathcal{D}\left(\mathbf{0}, \frac{\sigma_k^2}{M_k} \text{diag}(\mathbf{a}_k)\right), k \in \{1, \dots, K\} \end{aligned} \quad (10)$$

This model is similar to the model in Eq. (1) except that here we assume that the variance of effect sizes depend on continuous-valued annotation. Let  $\mathbf{a}_k$  be a  $M_k$ -vector where  $a_{k,m}$  is the value of  $k$ th annotation at SNP  $m$  (the elements of  $\mathbf{a}_k$  must be non-negative). Let  $S_k$  be the set of SNPs belong to annotation  $k$ . In this model, the SNP heritability of annotation  $k$  is defined as:

$$h_k^2 = \frac{\sum_{i \in S_k} \frac{\sigma_k^2}{M_k} a_{k,i}}{\sum_{k=1}^K \sum_{i \in S_k} \frac{\sigma_k^2}{M_k} a_{k,i} + \sigma_e^2}, k \in \{1, \dots, K\} \quad (11)$$

To estimate the variance components of this new model, we only need to replace  $\mathbf{X}_k$  with  $\mathbf{X}_k \text{diag}(\sqrt{\mathbf{a}_k})$  in Eq. (5) for every annotation  $k$ . We assessed the accuracy of RHE-mc in estimating variance components with continuous annotation in Supplementary Notes.

**Simulations.** We performed simulations to compare the performance of RHE-mc with several state-of-the-art methods for heritability estimation that cover the spectrum of methods that have been proposed.

We considered two simulation settings. In the large-scale simulation setting, we simulated phenotypes for the full set of UK Biobank genotypes consisting of  $M = 593,300$  array SNPs and  $N = 337,205$  individuals. We obtained the individuals by keeping unrelated white British individuals which are >3rd degree relatives (defined as pairs of individuals with kinship coefficient  $< 1/2^{(9/2)}$ )<sup>17</sup>, and removing individuals with putative sex chromosome aneuploidy. The small-scale setting was designed so that we could compare the accuracies of RHE-mc to REML methods. In this setting, we simulated phenotypes from a subsampled set of genotypes from the UK Biobank data genotypes used in large-scale simulation<sup>35</sup>. Specifically, we randomly chose a subset of  $N = 10,000$  individuals from the large-scale data so that we have  $M = 593,300$  array SNPs and  $N = 10,000$  individuals. We simulated phenotypes from genotypes using the following model which is used in refs.<sup>8,10</sup>.

$$\begin{aligned} \sigma_m^2 &= S c_m w_m^b [f_m (1 - f_m)]^a \\ (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \dots, \boldsymbol{\beta}_m)^T &\sim \mathcal{N}(\mathbf{0}, \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_m^2)) \\ \mathbf{y} | \boldsymbol{\beta} &\sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, (1 - h^2)\mathbf{I}_N) \end{aligned} \quad (12)$$

where  $S$  is a normalizing constant chosen so that  $\sum_{m=1}^M \sigma_m^2 = h^2$ . Here  $h^2 \in [0, 1]$ ,  $a \in \{0, 0.75\}$ ,  $b \in \{0, 1\}$ .  $\beta_m$ ,  $f_m$ , and  $w_m$  are the effect size, the minor allele frequency, and LDK score of  $m$ th SNP, respectively. Let  $c_m \in \{0, 1\}$  be an indicator variable for the causal status of SNP  $m$ . The LD score of a SNP is defined to be the sum of the squared correlation of the SNP with all other SNPs that lie within a specific distance, and the LDK score of a SNP is computed based on local levels of LD such that the LDK score tends to be higher for SNPs in regions of low LD<sup>36</sup>. The above models relating genotype to phenotype are commonly used in methods for estimating SNP heritability: the GCTA Model (when  $a = b = 0$  in Eq. (12)), which is used by the software GCTA<sup>27</sup> and LD Score regression (LDSC)<sup>24</sup>, and the LDK Model (where  $a = 0.75$ ,  $b = 1$  in Eq. (12)) used by software LDK<sup>36</sup>. Moreover, under each model, we varied the proportion and minor allele frequency (MAF) of CVs. Proportion of CVs were set to be either 100% or 1%, and MAF of CVs drawn uniformly from  $[0, 0.5]$  or  $[0.01, 0.05]$  or  $[0.05, 0.5]$  to consider genetic architectures that are either infinitesimal or sparse, as well genetic architectures that include a mixture of common and rare SNPs as well as ones that consist of only rare or common SNPs. The true heritability were chosen from  $\{0.1, 0.25, 0.5, 0.8\}$ .

We generated 100 sets of simulated phenotypes for each setting of parameters and report accuracies averaged over these 100 sets.

**Comparisons.** For the large-scale simulations, we compared RHE-mc to methods that rely on summary statistics for estimating heritability. Among the summary statistic methods, LD score regression (LDSC)<sup>24</sup> uses the slope from the GWAS  $\chi^2$  statistics regressed on the LD scores to estimate heritability. Stratified LD score regression (S-LDSC)<sup>7</sup> is an extension of LDSC for partitioning heritability from



summary statistics. SumHer is the summary statistic analog of LDAK<sup>25</sup>. We ran S-LDSC with 10 binary MAF bin annotations defined such that each bin contains exactly 10% of the typed SNPs; this is intended to mirror the 10 MAF bin annotations in the S-LDSC “baseline-LD model”<sup>29</sup> (see Supplementary Table 5). To run SumHer, we used the LDAK software to compute the default “LDAK weights” using in-sample LD<sup>25,36,37</sup>. We then computed “LD tagging” using 1-Mb windows centered on each SNP as recommended<sup>25</sup>. To do a fair comparison we computed LD scores for LDSC, S-LDSC, GRE, and SumHer by using in-sample LD among the M SNPs, and in all simulations we aim to estimate the SNP-heritability explained by the same set of M SNP. We described the parameter settings of summary statistic methods in Supplementary Notes.

For the small-scale simulations, we compared RHE-mc to GCTA-mc and HE-mc<sup>27</sup>. GCTA-mc and HE-mc are the extensions of GCTA and HE to a multi-component LMM, respectively, where the variance components are typically defined by binning SNPs according to their MAF as well as local LD<sup>8</sup>. We ran GCTA-mc, HE-mc and RHE-mc using 24 bins formed by the combination of six bins based on MAF ( $MAF \leq 0.01$ ,  $0.01 < MAF \leq 0.02$ ,  $0.02 < MAF \leq 0.03$ ,  $0.03 < MAF \leq 0.04$ ,  $0.04 < MAF \leq 0.05$ ,  $MAF > 0.05$ ) as well as four bins based on quartiles of the LDAK score of a SNP. We ran both GCTA-mc and RHE-mc allowing for estimates of a variance component to be negative.

For comparisons of runtime, we compared RHE-mc to GCTA<sup>27</sup> and BOLT-REML<sup>4</sup> which is a computationally efficient approximate method to compute the REML estimator. We ran all methods with 22 components (one for each chromosome). We also ran RHE-mc with  $\approx 300$  components (corresponding to 10 Mb bins) on the UK Biobank genotype (Supplementary Fig. 10). To create our largest dataset, we replicate individuals from the UK Biobank and a subset of the imputed SNPs to obtain a dataset with one million individuals and SNPs. We use the latest versions of BOLT-REML (Version 2.3.2) and GCTA (Version 1.92.1) in our comparison. All comparisons are performed on an Intel(R) Xeon(R) CPU 2.10 GHz server with 128 GB RAM.

**Heritability estimates in the UK Biobank.** We estimated SNP-heritability for 22 complex traits (6 quantitative, 16 binary) in the UK Biobank<sup>17</sup>. In this study, we restricted our analysis to SNPs that were present in the UK Biobank Axiom array used to genotype the UK Biobank. SNPs with >1% missingness and minor allele frequency <1% were removed. Moreover, SNPs that fail the Hardy–Weinberg test at significance threshold  $10^{-7}$  were removed. We restricted our study to self-reported British white ancestry individuals who are >3rd degree relatives defined as pairs of individuals with kinship coefficient  $< 1/2^{(9/2)}$ <sup>17</sup>. Furthermore, we removed individuals who are outliers for genotype heterozygosity and/or missingness. Finally, we obtained a set of  $N = 291,273$  individuals and  $M = 459,792$  SNPs to use in the real data analyses. We included age, sex, and the top 20 genetic principal components (PCs) as covariates in our analysis for all traits. We used PCs recomputed by the UK Biobank from a superset of 488,295 individuals. Additional covariates were used for waist-to-hip ratio (adjusted for BMI) and diastolic/systolic blood pressure (adjusted for cholesterol-lowering medication, blood pressure medication, insulin, hormone replacement therapy, and oral contraceptives).

**Heritability partitioning.** In our initial analysis, we removed SNPs with >1% missingness and minor allele frequency <1%. Moreover, we removed SNPs that fail the Hardy–Weinberg test at significance threshold  $10^{-7}$  as well as SNPs that lie within the MHC region (Chr6: 25–35 Mb) to obtain 4,824,392 SNPs. We restricted our study to self-reported British white ancestry individuals who are >3rd degree relatives defined as pairs of individuals with kinship coefficient  $< 1/2^{(9/2)}$ <sup>17</sup>. Furthermore, we removed individuals who are outliers for genotype heterozygosity and/or missingness. Finally, we obtained 291,273 individuals. We partitioned SNPs into eight bins based on two MAF bins ( $MAF \leq 0.05$ ,  $MAF > 0.05$ ) and quartiles of the LD-scores. For each bin  $k$ , we computed the heritability enrichment as the ratio of the percentage of heritability explained by SNPs in bin  $k$  to the percentage of SNPs in bin  $k$ .

We considered an additional analysis in which we included SNPs with  $MAF > 0.1\%$  resulting in  $N = 291,273$  unrelated white British individuals and  $M = 7,774,235$  imputed SNPs ( $MAF > 0.1\%$ ). We defined 144 bins based on 4 LD bins and 36 MAF bins. The 4 LD bins are defined based on quartile of LD-scores, and 36 MAF bins are defined based on 9-quantile of the following four intervals:  $0.001 \leq MAF \leq 0.01$ ,  $0.01 < MAF \leq 0.05$ ,  $0.05 \leq MAF \leq 0.10$ ,  $0.10 < MAF \leq 0.50$ .

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

Access to the UK Biobank resource is available via application at: <http://www.ukbiobank.ac.uk>.

## Code availability

RHE-mc software is open-source software freely available at: <https://github.com/sriramlab/RHE-mc>

Received: 4 April 2020; Accepted: 3 July 2020;

Published online: 11 August 2020

## References

- McCulloch, C. E. & Searle, S. R. *Generalized, Linear, and Mixed Models* (John Wiley & Sons, 2004).
- Yang, J. et al. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565 (2010).
- Yang, J. et al. Genome partitioning of genetic variation for complex traits using common snps. *Nat. Genet.* **43**, 519 (2011).
- Loh, P.-R. et al. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat. Genet.* **47**, 1385 (2015).
- Lee, S. H. et al. Estimating the proportion of variation in susceptibility to schizophrenia captured by common snps. *Nat. Genet.* **44**, 247 (2012).
- Gusev, A. et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* **95**, 535–552 (2014).
- Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228 (2015).
- Evans, L. M. et al. Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nat. Genet.* **50**, 737 (2018).
- Gazal, S. et al. Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations. *Nat. Genet.* **50**, 1600–1607 (2018).
- Hou, K. et al. Accurate estimation of snp-heritability from biobank-scale data irrespective of genetic architecture. *Nat. Genet.* <https://doi.org/10.1038/s41588-019-0465-0>. <https://www.biorxiv.org/content/early/2019/01/23/526855.full.pdf> (2019).
- Patterson, H. D. & Thompson, R. Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**, 545–554 (1971).
- Kuk, A. Y. & Cheng, Y. W. The Monte Carlo Newton–Raphson algorithm. *J. Stat. Comput. Simul.* **59**, 233–250 (1997).
- Liu, J. S. & Wu, Y. N. Parameter expansion for data augmentation. *J. Am. Stat. Assoc.* **94**, 1264–1274 (1999).
- Gilmour, A. R., Thompson, R. & Cullis, B. R. Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics* **51**, 1440–1450 (1995).
- Matilainen, K., Mäntyselä, E. A., Lidauer, M. H., Strandén, I. & Thompson, R. Employing a Monte Carlo algorithm in Newton-type methods for restricted maximum likelihood estimation of genetic parameters. *PLoS ONE* **8**, e80821 (2013).
- Runcie, D. E. & Crawford, L. Fast and exible linear mixed models for genome-wide genetics. *PLoS Genet.* **15**, e1007978 (2019).
- Bycroft, C. et al. The uk biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
- Haseman, J. & Elston, R. The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.* **2**, 3–19 (1972).
- Zhou, X. A unified framework for variance component estimation with summary statistics in genomewide association studies. *Ann. Appl. Stat.* **11**, 2027 (2017).
- Wu, Y. & Sankararaman, S. A scalable estimator of snp heritability for biobank-scale data. *Bioinformatics* **34**, i187–i194 (2018).
- Ge, T., Chen, C.-Y., Neale, B. M., Sabuncu, M. R. & Smoller, J. W. Phenome-wide heritability analysis of the uk biobank. *PLoS Genet.* **13**, e1006711 (2017).
- Visscher, P. M. et al. Statistical power to detect genetic (co) variance of complex traits using snp data in unrelated samples. *PLoS Genet.* **10**, e1004269 (2014).
- Golan, D., Lander, E. S. & Rosset, S. Measuring missing heritability: inferring the contribution of common variants. *Proc. Natl Acad. Sci. USA* **111**, E5272–E5281 (2014).
- Bulik-Sullivan, B. K. et al. Ld score regression distinguishes confounding from polygenicity in genomewide association studies. *Nat. Genet.* **47**, 291 (2015).
- Speed, D. & Balding, D. J. SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nat. Genet.* **51**, 277–284 (2019).
- Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228 (2015).
- Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. Gcta: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
- Yang, J. et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* **47**, 1114 (2015).
- Gazal, S. et al. Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421 (2017).

30. Wainschtein, P. et al. Recovery of trait heritability from whole genome sequence data. Preprint at 588020 (2019).
31. Weissbrod, O., Flint, J. & Rosset, S. Estimating snp-based heritability and genetic correlation in casecontrol studies directly and with summary statistics. *Am. J. Hum. Genet.* **103**, 89–99 (2018).
32. Henderson, C. R. Estimation of variance and covariance components. *Biometrics* **9**, 226–252 (1953).
33. Hutchinson, M. A stochastic estimator of the trace of the inuence matrix for Laplacian smoothing splines. *Commun. Stat.-Simul. Comput.* **18**, 1059–1076 (1989).
34. Liberty, E. & Zucker, S. W. The mailman algorithm: a note on matrix–vector multiplication. *Inf. Process. Lett.* **109**, 179–182 (2009).
35. Sudlow, C. et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
36. Speed, D., Hemani, G., Johnson, M. R. & Balding, D. J. Improved heritability estimation from genomewide SNPs. *Am. J. Hum. Genet.* **91**, 1011–1021 (2012).
37. Speed, D. et al. Reevaluation of SNP heritability in complex human traits. *Nat. Genet.* **49**, 986 (2017).

## Acknowledgements

This research was conducted using the UK Biobank Resource under applications 33127 and 33297. We thank the participants of UK Biobank for making this work possible. We thank Rob Brown, Steven Gazal, and members of the Sankararaman and Pasaniuc labs for feedback on this manuscript. This work was funded by NIH grants R01HG009120 (B. P. and K.S.B.), R35GM125055 (S.S.), an Alfred P. Sloan Research Fellowship (S.S.), and a NSF grant III-1705121 (A.P., Y.W., and S.S.).

## Author contributions

A.P. and S.S. conceived and designed the experiments. A.P. performed the experiment and statistical analyses. Y.W., K.S.B., and K.H. collected and managed the data. Y.W., K.S.B., K.H., and A.Z. assisted with the experiments. B.P. consulted on analysis and interpretation of the data. A.P., K.S.B., B.P., and S.S. wrote the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41467-020-17576-9>.

**Correspondence** and requests for materials should be addressed to S.S.

**Peer review information** *Nature Communications* thanks Doug Speed, Bjarni Vilhjalmsson, and the other, anonymous reviewer(s) for their contribution to the peer review of this work.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020