Method

Scalable summary-statistics-based heritability estimation method with individual genotype level accuracy

Moonseong Jeong,[1] Ali Pazokitoroudi,[1,2,3] Zhengtong Liu,[1] and Sriram Sankararaman[1,4,5]

[1]*Department of Computer Science, University of California, Los Angeles, Los Angeles, California 90095, USA;* [2]*Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts 02115, USA;* [3]*Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA;* [4]*Department of Human Genetics,* [5]*Department of Computational Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, California 90095, USA*

SNP heritability, the proportion of phenotypic variation explained by genotyped SNPs, is an important parameter in understanding the genetic architecture underlying various diseases and traits. Methods that aim to estimate SNP heritability from individual genotype and phenotype data are limited by their ability to scale to Biobank-scale data sets and by the restrictions in access to individual-level data. These limitations have motivated the development of methods that only require summary statistics. Although the availability of publicly accessible summary statistics makes them widely applicable, these methods lack the accuracy of methods that utilize individual genotypes. Here we present a SUMmary-statistics-based Randomized Haseman-Elston regression (SUM-RHE), a method that can estimate the SNP heritability of complex phenotypes with accuracies comparable to approaches that require individual genotypes, while exclusively relying on summary statistics. SUM-RHE employs Genome-Wide Association Study (GWAS) summary statistics and statistics obtained on a reference population, which can be efficiently estimated and readily shared for public use. Our results demonstrate that SUM-RHE obtains estimates of SNP heritability that are substantially more accurate compared with other summary statistic methods and on par with methods that rely on individual-level data.

[Supplemental material is available for this article.]

The exponentially decreasing cost of genotyping and sequencing technologies has led to an increase in the number and size of biobanks (Bycroft et al. 2018; Johnson et al. 2023; Kurki et al. 2023), covering a wide range of populations. With large samples of phenotype and genotype data now available in these biobanks, one of the major analyses often performed is estimating heritability, defined as the phenotypic variance explained by the variance in the genotype (Falconer and Mackay 1996). Heritability estimates in these large data sets have allowed researchers to better delineate the scope of the role genetics play in complex traits, ranging from schizophrenia (Niarchou et al. 2020) to height (Yang et al. 2015), and have assisted investigations into their genetic architectures (Lappalainen et al. 2024). Most heritability estimation methods fit linear mixed models (LMMs) (Yang et al. 2011; Loh et al. 2015a,b) to map the variation in genotypes measured at single-nucleotide polymorphisms (SNPs) to the variation in phenotypes and thereby estimate the SNP heritability, that is, the proportion of phenotypic variance explained by genotyped SNPs. Given the high dimensionality of the genotypes and the large sample sizes of biobanks, fitting or parameter estimation in LMMs is computationally prohibitive. Many methods have been proposed to reduce computational complexity while retaining statistical accuracy (Yang et al. 2011; Zhou and Stephens 2012; Loh et al. 2015a,b; Zhou 2017; Wu and Sankararaman 2018; Pazokitoroudi et al. 2020). These methods, although highly accurate, generally take hours or days to run and require access to individual genotypes and phenotypes.

The rise of large-scale biobanks has also brought increased attention to the issue of genomic privacy owing to a surge in security breaches (Frizzo-Barker et al. 2016; Savatt et al. 2019; Akyüz et al. 2021). Consequently, additional measures have been implemented to safeguard individual information throughout its processing, storage, and sharing (Wan et al. 2022). Gaining access to raw individual-level data is now more challenging, exemplified by the UK Biobank's decision to restrict access to its WGS data to its cloud server (Deflaux et al. 2023). Given these developments, there is a growing preference for summary-statistics-based methods due to their portability and speed, even though they may sacrifice some statistical power compared with methods that use individual-level data (Bulik-Sullivan et al. 2015; Shi et al. 2016; Zhou 2017; Hou et al. 2019; Speed and Balding 2019). Such a loss in statistical power is particularly pronounced in smaller sample sizes and may result in inflated estimates of heritability owing to underestimation of linkage disequilibrium (LD) (Zhou 2017), even if correct reference summary statistics were used.

To address these challenges of heritability estimation in large biobanks, we propose SUMmary-statistics Randomized Haseman–Elston regression (SUM-RHE), by extending our previous work, Randomized Haseman–Elston regression (RHE) (Wu and Sankararaman 2018; Pazokitoroudi et al. 2020), to work exclusively on summary statistics. This adaptation leverages the observation

Corresponding authors: bronsonj@cs.ucla.edu, sriram@cs.ucla.edu
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.279207.124. Freely available online through the *Genome Research* Open Access option.

34:1286–1293 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/24; www.genome.org

that the trace estimates of the squared genetic relatedness matrix (GRM), which are needed to compute the method-of-moments (MoM) estimator underlying RHE, can be related to population-level parameters. By combining these trace estimates from a reference sample with Genome-Wide Association Study (GWAS) summary statistics from a target sample (consisting of individuals sampled from the same population as the reference sample), we can reconstruct the MoM estimates for the target sample without access to the individual data. In comprehensive simulations across various genetic architectures and scenarios, we show that SUM-RHE estimates are on par with methods that rely on individual-level data, and are substantially more accurate than popular summary statistic-based methods, all while exclusively utilizing summary statistics.

Methods

Background on heritability estimation and LMMs

Early attempts to calculate SNP heritability of complex traits by aggregating SNPs identified as GWAS-significant have revealed the issue of missing heritability, as this estimate of heritability was significantly lower than the narrow-sense heritability estimated in other studies (e.g., twin studies) (Manolio et al. 2009). The seminal work by Yang et al. (2010) reduced this discrepancy by jointly modeling all the SNPs, such that their effect sizes come from a distribution of some fixed variance that quantifies the genetic variation. In this LMM framework, the standardized phenotype vector $\boldsymbol{y}$ is modeled as a linear combination of SNP effect sizes $\boldsymbol{\beta}$ multiplied by the standardized genotype matrix $\boldsymbol{X}$ of M SNPs and N individuals with uniform noise $\boldsymbol{\epsilon}$:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\beta} \sim \mathcal{D}\left(\boldsymbol{0}, \frac{\sigma_g^2}{M}\mathbf{I}_M\right), \quad \boldsymbol{\epsilon} \sim \mathcal{D}(\boldsymbol{0}, \sigma_e^2\mathbf{I}_N), \tag{1}$$

where the additive effect sizes $\boldsymbol{\beta}$ are drawn from an arbitrary distribution $\mathcal{D}$ with mean zero and variance of $\frac{\sigma_g^2}{M}\mathbf{I}_M$, and the environmental/noise effects $\boldsymbol{\epsilon}$ are drawn from a distribution with variance σ_e^2. In the original work by Yang et al. (2010) and GCTA (Yang et al. 2011), the distribution $\mathcal{D}$ was chosen as a normal distribution. The SNP heritability is then defined as the proportion of genetic variance over total phenotypic variance, $h^2_{SNP} = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$.

One approach to estimating the variance components σ_g^2, σ_e^2 is to find the maximum likelihood estimator (MLE) and its variants, such as restricted maximum likelihood (REML) estimators (Yang et al. 2011; Loh et al. 2015b). These methods often rely on iterative optimizations, which tend to be inefficient, and could lead to biased estimates owing to the normality assumption (Zhou 2017; Wu and Sankararaman 2018). On the other hand, MoM approaches such as the Haseman–Elston regression (HE) (Haseman and Elston 1972; Sham and Purcell 2001), RHE (Wu and Sankararaman 2018; Pazokitoroudi et al. 2020), MQS (Zhou 2017), or LDSC (Bulik-Sullivan et al. 2015), only require solving the normal equations and do not make any assumptions on the underlying distribution $\mathcal{D}$. Here, we briefly discuss the MoM estimator (HE), which sets the foundation for our work.

Heritability estimation from individual genotype data using MoM and randomized MoM

The HE MoM estimator of the parameters σ_g^2, σ_e^2 can be obtained by minimizing the discrepancy between the sample covariance $\boldsymbol{y}\boldsymbol{y}^\top$ and the population covariance matrices. The population covariance is given as

$$cov[y] = \mathbb{E}[\boldsymbol{y}\boldsymbol{y}^\top] - \mathbb{E}[\boldsymbol{y}]\,\mathbb{E}[\boldsymbol{y}^\top] = \sigma_g^2\boldsymbol{K} + \sigma_e^2\mathbf{I}_N, \tag{2}$$

where $\boldsymbol{K} = \frac{1}{M}\boldsymbol{X}\boldsymbol{X}^\top$ is defined as the GRM. We want to find the estimates of the parameters $\widehat{\sigma_g^2}$, $\widehat{\sigma_e^2}$ that minimize the Frobenius norm (the measure of discrepancy) between the two covariance matrices. This is equivalent to solving the following normal equations:

$$\begin{bmatrix} tr(\boldsymbol{K}^\top\boldsymbol{K}) & tr(\boldsymbol{K}) \\ tr(\boldsymbol{K}) & N \end{bmatrix}\begin{bmatrix} \widehat{\sigma_g^2} \\ \widehat{\sigma_e^2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}^\top\boldsymbol{K}\boldsymbol{y} \\ \boldsymbol{y}^\top\boldsymbol{y} \end{bmatrix}, \tag{3}$$

where $tr(\boldsymbol{K}) = N$ and $\boldsymbol{y}^\top\boldsymbol{y} = N$, given both $\boldsymbol{X}$ and $\boldsymbol{y}$ are standardized. Equation 3 has the analytical solution for the variance components $\widehat{\sigma_g^2}$, $\widehat{\sigma_e^2}$:

$$\widehat{\sigma_g^2} = \frac{\boldsymbol{y}^\top\boldsymbol{K}\boldsymbol{y} - \boldsymbol{y}^\top\boldsymbol{y}}{tr(\boldsymbol{K}^\top\boldsymbol{K}) - N}, \quad \widehat{\sigma_e^2} = \frac{\boldsymbol{y}^\top\boldsymbol{y}}{N} - \widehat{\sigma_g^2}, \tag{4}$$

giving the MoM estimate for heritability $\widehat{h^2}$:

$$\widehat{h^2} = \frac{\frac{\boldsymbol{y}^\top\boldsymbol{K}\boldsymbol{y}}{\boldsymbol{y}^\top\boldsymbol{y}} - 1}{\frac{tr(\boldsymbol{K}^2)}{N} - 1} = \frac{\frac{1}{M}\frac{\boldsymbol{y}^\top\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{y}}{\boldsymbol{y}^\top\boldsymbol{y}} - 1}{\frac{tr(\boldsymbol{K}^2)}{N} - 1} = \frac{\frac{1}{M}\left(\frac{\boldsymbol{X}^\top\boldsymbol{y}}{\sqrt{\boldsymbol{y}^\top\boldsymbol{y}}}\right)^\top\frac{\boldsymbol{X}^\top\boldsymbol{y}}{\sqrt{\boldsymbol{y}^\top\boldsymbol{y}}} - 1}{\frac{tr(\boldsymbol{K}^2)}{N} - 1}. \tag{5}$$

The biggest bottleneck in the equation is calculating $tr(\boldsymbol{K}^\top\boldsymbol{K})$. An exact calculation of the trace involves forming the matrix $\boldsymbol{K}^\top\boldsymbol{K}$, which has a computational complexity of $\mathcal{O}(MN^2)$. Given $M \approx$ 1,000,000 and $N \approx$ 1,000,000, this is not tractable in modern biobanks. One of the main contributions of RHE-reg (Wu and Sankararaman 2018) and RHE-mc (Pazokitoroudi et al. 2020) is the efficient estimation of $tr(\boldsymbol{K}^\top\boldsymbol{K})$ by leveraging the fact that the trace of $\boldsymbol{K}^\top\boldsymbol{K}$ can be approximated by a stochastic trace estimator (Girard 1989; Hutchinson 1989):

$$\widehat{tr(\boldsymbol{K}^\top\boldsymbol{K})} = \frac{1}{B}\sum_{b=1}^{B}\boldsymbol{z}_b^\top\boldsymbol{K}^\top\boldsymbol{K}\boldsymbol{z}_b, \tag{6}$$

where $\boldsymbol{z}_b$ are independent and identically distributed random vectors such that $\mathrm{E}[\boldsymbol{z}_b] = 0$ and $\mathrm{E}[\boldsymbol{z}_b\boldsymbol{z}_b^\top] = \mathbf{I}_N$. In both RHE-reg and RHE-mc, random vectors sampled from the standard normal distribution are used. Numerically, it was found that using $B \approx 100$ can estimate the trace of the squared GRM matrix $\boldsymbol{K}^\top\boldsymbol{K}$ with high accuracy for moderate sample sizes of $N \approx 5000$ and $B \approx 10$ for large samples sizes (Wu and Sankararaman 2018; Pazokitoroudi et al. 2020). This stochastic trace estimator reduces the computational complexity to $\mathcal{O}(MNB)$. Additional optimizations, such as the mailman algorithm (Liberty and Zucker 2009) and implementation of a streaming version of the algorithm, reduce the computational complexity to $\mathcal{O}\left(\frac{NMB}{\max(\log_3 N, \log_3 M)}\right)$ and the memory complexity to $\mathcal{O}(NB)$, thus allowing estimation of heritability across millions of SNPs and individuals.

Extensive benchmarking of the RHE methods has shown that their performance is on par with other methods that require individual-level data, such as GCTA or BOLT-REML (Wu and Sankararaman 2018; Pazokitoroudi et al. 2020). RHE offers a

distinct advantage over these likelihood-based methods, which tend to scale poorly, as well as other MoM approaches in terms of computational and memory efficiency (e.g., HE) or statistical efficiency (LDSC) (Pazokitoroudi et al. 2020). However, RHE is still limited in that it requires access to individual-level genotype and phenotype data, which restricts its applicability in cases in which such data are unavailable or only GWAS summary statistics are available.

Heritability estimation from summary statistics

In this work, we further extend RHE to work exclusively with GWAS summary statistics. Our key observation is the fact that the left-hand side (LHS) of the normal equations (Equation 3) is related to the LD in the population (Bulik-Sullivan 2015; Zhou 2017) and not the phenotype. Thus, if we can summarize the trace estimate for a reference sample drawn from a population, we can use these trace estimates to reconstruct the corresponding trace estimates for a target sample drawn from the same population. Indeed, we find that the expected value of the trace of $\boldsymbol{K}^\top\boldsymbol{K}$ can be related to the LD scores of the SNPs. Furthermore, the RHS of the normal equations can be computed from GWAS summary statistics obtained on the target population.

The LD score of a variant j is defined as the sum of squared correlation with all the variants,

$$r_j^2 = \sum_{k=1}^{M} r_{jk}^2, \tag{7}$$

then $tr(\boldsymbol{K}^\top\boldsymbol{K})$ is

$$\begin{aligned} tr(\boldsymbol{K}^\top\boldsymbol{K}) &= tr\left(\frac{\boldsymbol{X}\boldsymbol{X}^\top}{M}\frac{\boldsymbol{X}\boldsymbol{X}^\top}{M}\right) \\ &= \frac{1}{M^2}tr(\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{X}), \quad \text{Cyclic property of trace} \\ &= \frac{N^2}{M^2}tr\left(\frac{\boldsymbol{X}^\top\boldsymbol{X}}{N}\frac{\boldsymbol{X}^\top\boldsymbol{X}}{N}\right) \\ &= \frac{N^2}{M^2}tr(\boldsymbol{R}^2) \\ &= \frac{N^2}{M^2}\sum_{j=1}^{M}\sum_{k=1}^{M} r_{jk}^2 \end{aligned}, \tag{8}$$

where $\boldsymbol{R}$ is the $M\times M$ correlation matrix (or the LD matrix), and $r_{jk} = \frac{1}{N}\sum_{n=1}^{N} x_{nj}x_{nk}$ is the sample correlation between variants j, k. This gives

$$\begin{aligned} r_{jk}^2 &= \left(\frac{1}{N}\sum_{n=1}^{N} x_{nj}x_{nk}\right)\left(\frac{1}{N}\sum_{n=1}^{N} x_{nj}x_{nk}\right) \\ &= \frac{1}{N^2}\left(\sum_{n=1}^{N} x_{nj}^2x_{nk}^2 + \sum_{n=1}^{N}\sum_{i\neq n}^{N} x_{nj}x_{nk}x_{ij}x_{ik}\right) \end{aligned}.$$

For large sample sizes (N), we have

$$\begin{aligned} r_{jk}^2 &\approx \frac{1}{N^2}(N\mathbb{E}[X_j^2X_k^2] + (N^2-N)\mathbb{E}[X_jX_k]) \\ &= \frac{1}{N^2}(N\mathbb{E}[X_j^2X_k^2] + (N^2-N)\rho_{jk}^2) \end{aligned},$$

where ρ_{jk} is the expected correlation or population LD between SNPs j and k. Assuming (X_j, X_k) are normally distributed with mean zero and covariance $\begin{bmatrix} 1 & \rho_{jk} \\ \rho_{jk} & 1 \end{bmatrix}$, we can use Isserlis' theorem to compute $\mathbb{E}[X_j^2X_k^2] = \mathbb{E}[X_j]^2\mathbb{E}[X_k]^2 + 2\mathbb{E}[X_jX_k]^2 = 1 + 2\rho_{jk}^2$. We then have

$$\begin{aligned} r_{jk}^2 &\approx \frac{1}{N^2}(N(1+2\rho_{jk}^2) + (N^2-N)\rho_{jk}^2) \\ &= \frac{1}{N^2}(N + (N^2+N)\rho_{jk}^2) \\ &= \frac{1}{N} + \frac{N+1}{N}\rho_{jk}^2 \end{aligned}. \tag{9}$$

Substituting Equation 9 into Equation 8,

$$\begin{aligned} tr(\boldsymbol{K}^\top\boldsymbol{K}) &= N + \frac{N(N+1)}{M^2}\sum_{j=1}^{M}\sum_{k=1}^{M}\rho_{jk}^2 \\ &= N + \frac{N(N+1)}{M^2}\sum_{j=1}^{M} l_j \\ &\approx N + \frac{N^2}{M^2}\sum_{j=1}^{M} l_j \\ &= N + \frac{N^2}{M^2}S \\ &= N + N^2\rho \end{aligned}, \tag{10}$$

where $l_j = \sum_{k=1}^{M}\rho_{jk}^2$ is the expected or population-level LD score associated with SNP j, whereas ρ can be interpreted as the average (expected) LD across all SNPs in genotype $\boldsymbol{X}$.

Let $\widehat{\beta}_j$ denote the GWAS effect size estimates for SNP j obtained by linear regression. Given the observed count N_j of the SNP, we have $\widehat{\beta}_j = \frac{\boldsymbol{X}^\top\boldsymbol{y}}{N_j}$ because the genotypes are standardized. The standard error of the GWAS estimate is $s_j = \sqrt{\frac{MSE}{N_j}} \approx \sqrt{\frac{\boldsymbol{y}^\top\boldsymbol{y}}{NN_j}}$. Thus, if we define the vector of adjusted z-scores,

$$\boldsymbol{z} = \left\{\frac{\widehat{\beta}_j}{s_j}\sqrt{\frac{N_j}{N}}\right\},$$

we have that

$$\boldsymbol{z} = \frac{\boldsymbol{X}^\top\boldsymbol{y}}{\sqrt{\boldsymbol{y}^\top\boldsymbol{y}}}. \tag{11}$$

Substituting Equations 10 and 11 into Equation 5 gives us

$$\widehat{h^2} \approx \frac{\frac{\boldsymbol{z}^\top\boldsymbol{z}}{M} - 1}{N\rho}, \tag{12}$$

where $\rho = \frac{1}{N}\left(\frac{tr(K^2)}{N} - 1\right)$. Given ρ, $\widehat{h^2}$ can be computed using summary statistics using Equation 12. However, computing ρ requires the exact $\kappa = tr(K^2)$, which is computationally intractable. Instead, we approximate ρ by using the stochastic trace estimates $\hat{\kappa} = \widehat{tr(K^2)}$ described in Equation 6, such that $\hat{\rho} = \frac{1}{N}\left(\frac{\hat{\kappa}}{N} - 1\right)$. This gives us the randomized MoM estimator of h^2 that can be calculated from summary statistics:

$$\widehat{h^2}_{MOM} = \frac{\frac{\boldsymbol{z}^\top\boldsymbol{z}}{M} - 1}{N\hat{\rho}}. \tag{13}$$

We propose releasing $\hat{\rho}$ as "trace summaries," which can then be combined with phenotype-specific GWAS summary statistics $\boldsymbol{z}$ to estimate heritability. The estimator in Equation 14 assumes that the $\hat{\rho}$ was computed on the same genotypes used to generate the GWAS summary statistics. In settings in which $\hat{\rho}$ cannot be

Table 1. Inputs for the different methods evaluated

Method	BOLT/GCTA-GREML, RHE	SUM-RHE	LDSC	SumHer
Inputs	Individual genotype	Ref. trace	Ref. LD scores	Ref. SNP tag
	Individual phenotype	GWAS summary	GWAS summary	GWAS summary

LDSC and SUM-RHE rely only on the summary statistics, whereas GCTA-GREML, BOLT-REML, and RHE require individual data for target genotypes and phenotypes.

computed on the same genotypes, we can use $\hat{\rho}$ computed on a reference genotype data set drawn from a population that is similar to the population that was used to generate GWAS summary statistics (such as The 1000 Genomes Project). This is under the assumption that the LD structures of similar populations will also be related.

Estimating standard errors

To calculate the standard error of our estimator, we perform SNP-level block jackknife resampling, as done in RHE-mc (Pazokitoroudi et al. 2020). When generating the trace summaries with individual genotypes, we report the $\hat{\rho}_{jack}$ values estimated from jackknife subsamples. Excluding the same SNPs from the PLINK GWAS summary statistics, we can compute the denominator in Equation 14, $\boldsymbol{z}_{jack}^{\top}\boldsymbol{z}_{jack}/M_{jack}$, to get the jackknife replicate of $\widehat{h^2}$:

$$\widehat{h^2}_{jack} = \frac{1}{N\hat{\rho}_{jack}}\left(\frac{\boldsymbol{z}_{jack}^{\top}\boldsymbol{z}_{jack}}{M_{jack}} - 1\right).$$

Following the execution on all SNP blocks, we employ jackknife resampling to obtain SE estimates:

$$\widehat{\mathrm{Var}}(\widehat{h^2}) = \frac{n-1}{n}\sum_{i=1}^{n}[\widehat{h_i^2} - \mu(\widehat{h^2})], \quad \widehat{\mathrm{SE}}(\widehat{h^2}) := \sqrt{\widehat{\mathrm{Var}}(\hat{\sigma}_g^2)}. \qquad (14)$$

Results

Simulations under varied genetic architectures

We assessed the performance of SUM-RHE against methods that require individual genotypes—RHE (RHE-mc run with a single component), GCTA-GREML, and BOLT-REML—and methods that can work with summary statistics—LDSC and SumHer (Speed and Balding 2019)—on the task of estimating genome-wide heritability. We applied all methods to unrelated White British individuals genotyped on $M = 454{,}207$ common SNPs (MAF > 0.01 excluding SNPs in the MHC region) typed on the UK Biobank Axiom array. Because of the computational scalability of GCTA-GREML and BOLT-REML, we tested these and other methods in a small-scale setting in which the number of individuals in the target data set was set to $N_{target} = 10{,}060$ (we term this the 10k sample). In addition, we compared all the remaining methods in a large-scale setting in which the number of individuals in the target data set was set to $N_{target} = 50{,}112$ (termed the 50k sample). For the summary statistic methods, the GWAS summary statistics were computed on the target data sets. These methods also require population statistics, calculated in the remainder of the $N = 291{,}273$ unrelated White British individuals as the reference data set. For the small-scale simulation, the reference set has $N_{ref} = 281{,}213$, and for the large-scale simulation, we set $N_{ref} = 241{,}161$. On each of the reference sets, we generated SUM-RHE trace summary statistics, LDSC reference LD scores, and LDAK SNP taggings.

We then simulated phenotypes corresponding to nine different genetic architectures: $h^2_{SNP} = 0.1, 0.25, 0.4$ and causal ratio $P = 1.0, 0.1, 0.01$ (where the causal ratio represents the proportion of variants with non-zero effects), each with 100 replicates. Table 1 summarizes the inputs for each method: For the calculation of LDSC LD scores, we used the entire N_{ref} reference sample with a window size of 2 Mb. SUM-RHE trace summaries were calculated by aggregating the trace estimates of 25 runs on the reference set with $B = 100$ (equivalent to stochastic trace estimation with $B' = 2500$ random vectors) (for more information, see Supplemental Fig. S2) and 1000 jackknife blocks, yielding a single trace summary statistic with 1000 jackknife estimates of $\hat{\rho}$. SumHer was run assuming the GCTA model to calculate the SNP taggings (consistent with the genetic architecture assumed in our simulations). RHE was run with $B = 100$ and 1000 jackknife blocks as well. BOLT-REML/GCTA-GREML was run with default parameter settings. GWAS summary statistics for each simulated phenotype were generated using PLINK 2.0. Figure 1 summarizes the heritability estimates on the target data. Across the 18 different settings (genetic architectures and sample sizes) we tested, we found that the

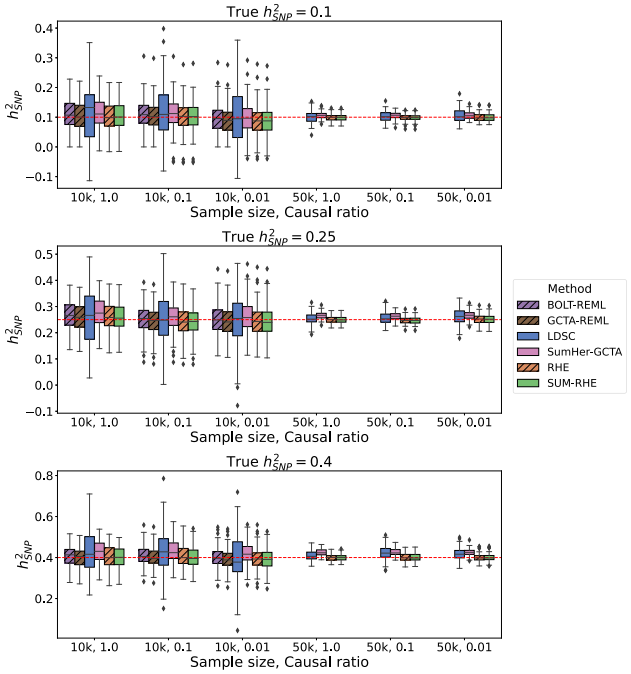

Figure 1. Comparison of SNP heritability estimates across methods. SUM-RHE heritability estimates are comparable to those from RHE or BOLT-REML/GCTA-GREML and are significantly more accurate than those of LDSC and SumHer. Because of computational limitations, BOLT-REML/GCTA-GREML was not run on the $N = 50$k simulations. Dashed hatches denote individual-level data methods.

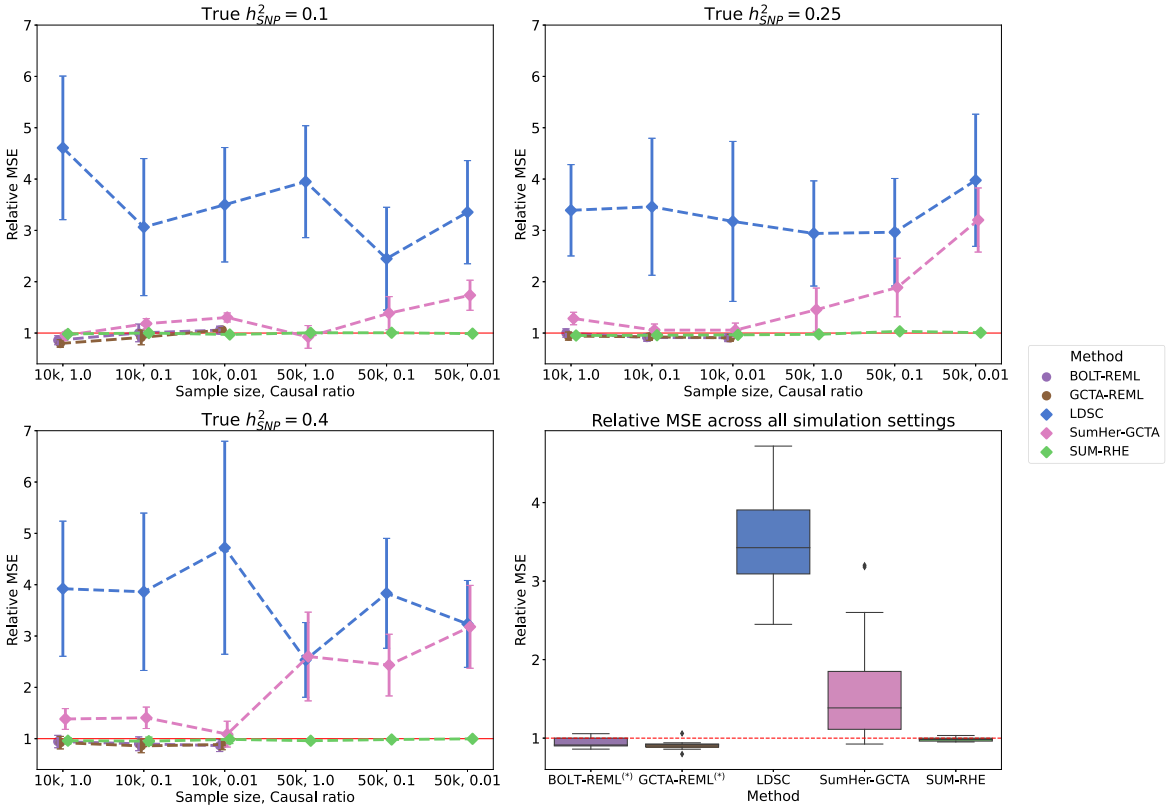

Figure 2. Mean squared error (MSE) of heritability estimates of each method relative to RHE. The dot and error bar denote the relative MSE and the 95% CI calculated based on bootstrap resampling (using 10,000 bootstrap samples), respectively. Although the MSE of SUM-RHE is within ±5% of the MSE of RHE, the MSE of LDSC ranges from 245% to 472%, whereas the MSE of SumHer-GCTA ranges from 92% to 320%. (Diamond) BOLT-REML and GCTA-GREML have relative MSE in the range of 80% and 106% for the 10k samples and were not benchmarked on the 50k samples.

accuracy of SUM-RHE was comparable to RHE (Fig. 1) with the mean-squared error of SUM-RHE close to one relative to RHE, despite relying only on the summary statistics (Fig. 2; for the MSE of each method, see Supplemental Fig. S1).

SUM-RHE has substantially improved accuracy over other summary-statistic-based methods: LDSC exhibits MSE ranging from 244% to 478% relative to that of SUM-RHE (mean: 356%), whereas SumHer-GCTA has MSE ranging from 94% to 331% relative to that of SUM-RHE (mean: 167%). SumHer-GCTA has lower MSE than SUM-RHE for low heritability ($h^2 = 0.1$) and high polygenicity (causal ratio, $P = 1.0$). The improvement in MSE is particularly pronounced with smaller sample sizes ($N = 10{,}060$).

We also tested the calibration of SUM-RHE by simulating phenotypes with $h^2_{SNP} = 0$ and testing the hypothesis of $h^2_{SNP} = 0$ with a rejection threshold of $\alpha = 0.05$ for both $N = 10{,}060$ and $N = 50{,}112$ to find that SUM-RHE is well calibrated (Table 2).

Simulations with a mixture model

We further test the robustness of our method by simulating phenotypes with a combination of large- and small-effect SNPs. We selected the first $\pi = 0.05$ of the SNPs to account for $\gamma = 0.25$ of the total SNP heritability, whereas the rest of the SNPs accounted for the remainder. The causal SNPs were then selected at random

Table 2. Calibration of the methods

Method	Sample	BOLT-REML	GCTA-GREML	LDSC	SumHer-GCTA	SUM-RHE	RHE
Bias	10k	0.0247	0.0219	0.0084	0.0095	0.0058	0.0047
	50k	—	—	0.0019	0.0005	−0.0002	−0.0002
SE	10k	0.0020	0.0020	0.0101	0.0031	0.0045	0.0045
	50k	—	—	0.0019	0.0006	0.0010	0.0010
FPR	10k	0.15	0.14	0.04	0.06	0.06	0.05
	50k	—	—	0.03	0.05	0.06	0.06

We report the bias, SE, and the false-positive rate (FPR) of each method in the setting in which $h^2_{SNP} = 0$. Because of computational limitations, GCTA-GREML and BOLT-REML were only run on 10k samples.

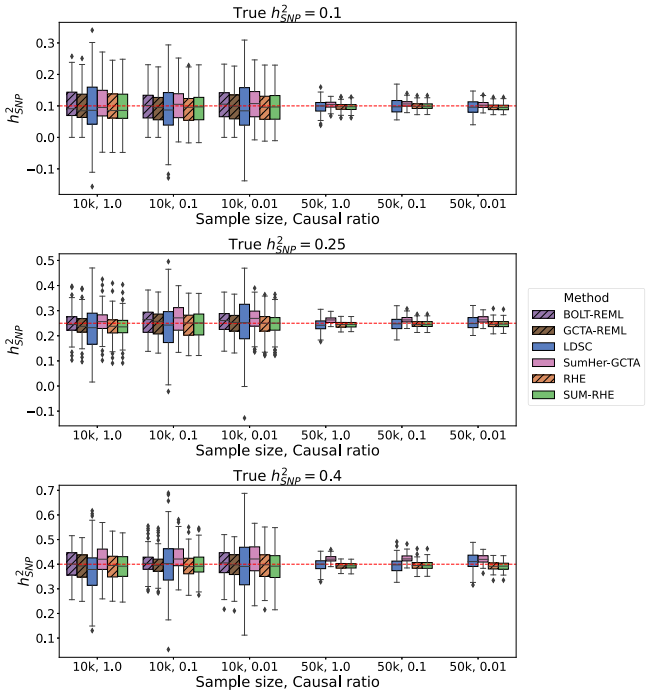

Figure 3. Comparison of SNP heritability estimates across methods on simulations with mixtures of large and small genetic effects.

with fixed probability α, such that the effect sizes for the large-effect SNPs were sampled from a different distribution than for the small-effect SNPs. Specifically, the effect sizes were sampled from two distributions:

$$\beta_j \sim \begin{cases} \mathcal{N}\left(0, \frac{\gamma h^2}{\pi M \alpha}\right), & \text{for} \quad j \in \text{Large effect} \\ \mathcal{N}\left(0, \frac{(1-\gamma)h^2}{(1-\pi)M\alpha}\right), & \text{for} \quad j \in \text{Small effect} \end{cases}.$$

Figure 3 shows the boxplots of estimates from LDSC, SumHer-GCTA, SUM-RHE, and RHE, whereas Figure 4 plots the SE and MSE of the three summary-statistics methods (LDSC, SumHer-GCTA, SUM-RHE) relative to that of RHE. We observe that both LDSC and SumHer-GCTA show larger MSEs than SUM-RHE, similar to our previous simulations. LDSC has a MSE in the range of 266% to 444% relative to that of SUM-RHE (mean: 348%), and SumHer has a MSE in 102% to 304% (mean: 151%). These results indicate that SUM-RHE is robust under the mixture model simulations and attains accuracy comparable to individual-level methods across the scenarios we tested.

Runtime measurements

We compared the runtime of SUM-RHE to other methods (Table 3). The heritability estimation step for all the summary statistic methods is computationally efficient irrespective of sample size. The generation of the reference statistics will depend on the size

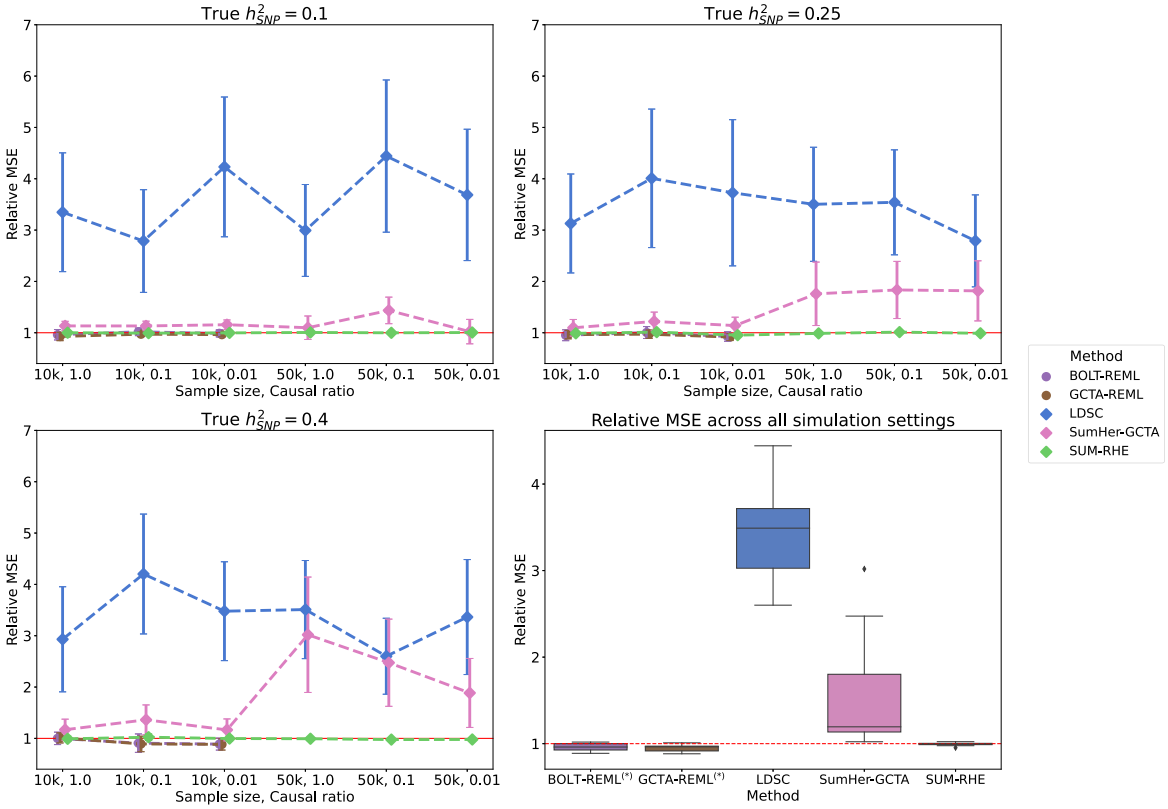

Figure 4. Comparison of MSE and SE of the summary-statistics-based methods against RHE on simulations with mixtures of large and small genetic effects. Here we report the relative MSE of the five methods on the mixture-of-effect simulations. Their performances are similar to those in the previous simulations (Fig. 2). SUM-RHE has an MSE within ±5% relative to RHE.

Table 3. Runtime estimates of the six methods

Step	Sample	BOLT-REML	GCTA-GREML	RHE	SUM-RHE	LDSC	SumHer
Reference statistic estimation	281k	—	—	—	14,286.4	3120.8	697.3
GWAS summary	10k	—	—	—	7.0	7.0	7.0
Heritability estimation	10k	589.6	847.3	590.6	1.6	4.0	2.9
Reference statistic estimation	241k	—	—	—	14,686.0	2699.9	1180.7
GWAS summary	50k	—	—	—	17.9	17.9	17.9
Heritability estimation	50k	—	—	3032.6	1.3	4.0	1.6

We ran each method on 10 replicates to measure wall clock time. For methods or tools that allow multithreading (BOLT/GCTA-GREML, SumHer, PLINK 2.0), we used six threads, run on the UCLA Hoffman2 computing nodes. SUM-RHE trace summaries were estimated by running the original RHE-mc codes (which does not support multithreading). PLINK 2.0 was used for calculating the GWAS summary statistics. All measurements are in seconds.

of the reference data set but is typically a one-time computation that is relatively efficient even for data sets with hundreds of thousands of individuals with access to a compute cluster.

Application to traits in the UK Biobank

Finally, we applied three of the summary statistic methods as well as one of the scalable individual genotype-based method (RHE) to real UK Biobank phenotypes measured on $N = 291{,}273$ unrelated White British individuals paired with genotypes assayed on 454,207 common array SNPs (we ran SumHer with both GCTA and LDAK SNP taggings for real phenotypes). Here we plot the 15 quantitative traits with the highest z-scores of SUM-RHE heritability estimates, from overall health ($z = 34.3$) to albumin ($z = 17.9$), ordered by the heritability estimates. For the summary-statistics-based methods (LDSC, SUM-RHE, SumHer-LDAK/GCTA), we use in-sample statistics.

As expected, SUM-RHE has estimates that agree well with RHE estimates (Fig. 5). SUM-RHE estimates tend to lie in between those from LDSC and SumHer (with both GCTA and LDAK SNP taggings), consistent with our previous work (Pazokitoroudi et al. 2020).

Discussion

Here we propose a summary-statistics-based heritability estimation method, SUM-RHE, that has performance comparable to that of individual genotype-based methods. SUM-RHE is accurate, fast, and highly portable. It uses a trace summary statistic calculated by aggregating stochastic trace estimates and PLINK GWAS statistics. In the era of large biobanks, SUM-RHE will be a useful tool in estimating heritability while maintaining the privacy of the patients.

We conclude with a discussion of limitations and directions for future work. First, heritability estimates from SUM-RHE are accurate under the assumption that the summary statistics are free of confounding owing to population stratification and cryptic relatedness. In a setting in which the summary statistics are affected by confounders, LDSC could potentially be more robust (as confounding would affect the intercept of LDSC whereas the slope would provide a robust estimator of heritability). Second, our preliminary experiments suggest that SUM-RHE retains its accuracy even when reference trace estimates are computed using a smaller number of random vectors on a smaller number of individuals (as

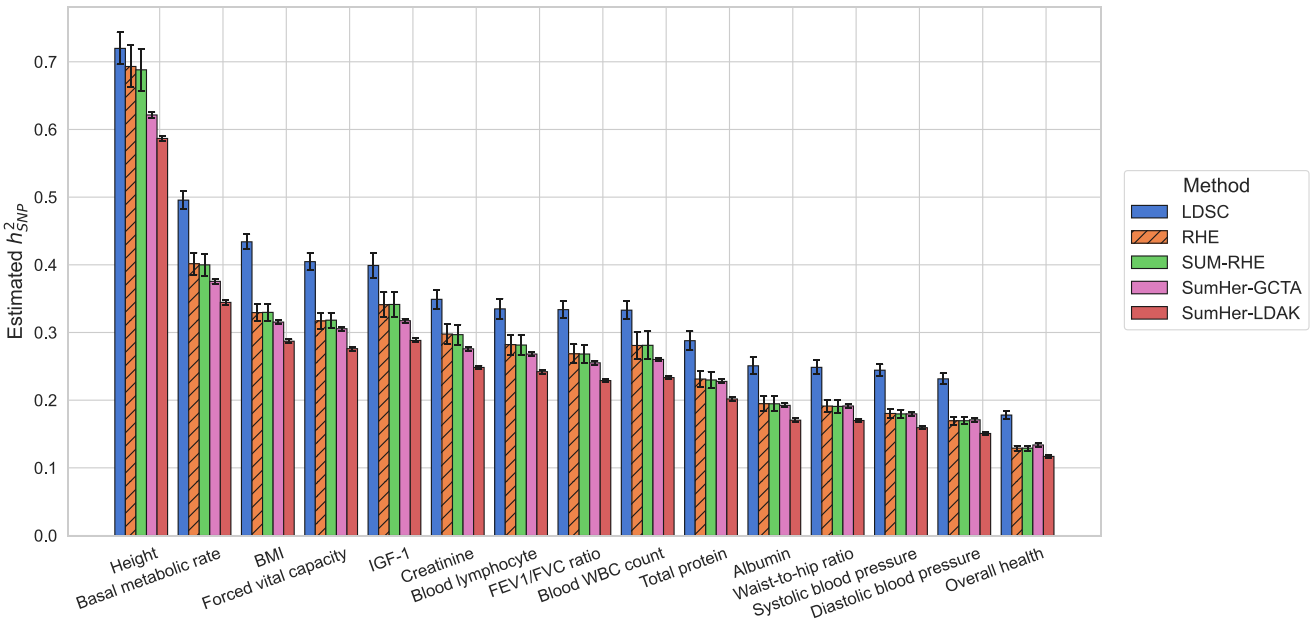

Figure 5. Application to UK Biobank phenotypes.

low as N_{ref} = 30k and B = 1000) (see Supplemental Fig. S2). These results suggest that the computation of trace summaries can be even more efficient. Third, SUM-RHE is not applicable to the setting of MAF and LD-dependent architectures, nor does it estimate partitioned or local heritability. These applications will require computation and release of partitioned trace summaries. We view this as a promising direction for future work.

Software availability

SUM-RHE source code is available at GitHub (https://github.com/sriramlab/SUMRHE) and as Supplemental Code. Access to the UK Biobank data (genotype and phenotypes measured at baseline) requires an approved application. Details on the application and approval process can be found at https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

This research was conducted using the UK Biobank Resource under application 33127. We thank the participants of UK Biobank for making this work possible. This work was supported by the National Institutes of Health (GM125055, HG006399) and the National Science Foundation (CAREER-1943497).

Author contributions: S.S. conceived and supervised the project. M.J., A.P., and S.S. developed the methods. M.J. and S.S. wrote the manuscript. M.J. and Z.L. wrote the software code and performed the analyses. All authors read, reviewed, and approved the final manuscript.

References

Akyüz K, Chassang G, Goisauf M, Kozera L, Mezinska S, Tzortzatou O, Mayrhofer MT. 2021. Biobanking and risk assessment: a comprehensive typology of risks for an adaptive risk governance. *Life Sci Soc Policy* **17:** 10. doi:10.1186/s40504-021-00117-7

Bulik-Sullivan B. 2015. Relationship between LD score and Haseman-Elston regression. bioRxiv doi:10.1101/018283

Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, of the Psychiatric Genomics Consortium SWG, Patterson N, Daly MJ, Price AL, Neale BM. 2015. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47:** 291–295. doi:10.1038/ng.3211

Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, et al. 2018. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562:** 203–209. doi:10.1038/s41586-018-0579-z

Deflaux N, Selvaraj MS, Condon HR, Mayo K, Haidermota S, Basford MA, Lunt C, Philippakis AA, Roden DM, Denny JC, et al. 2023. Demonstrating paths for unlocking the value of cloud genomics through cross cohort analysis. *Nat Commun* **14:** 5419. doi:10.1038/s41467-023-41185-x

Falconer DS, Mackay TFC. 1996. *Introduction to quantitative genetics*, 4th ed. Longmans Green, Harlow, Essex, UK.

Frizzo-Barker J, Chow-White PA, Charters A, Ha D. 2016. Genomic big data and privacy: challenges and opportunities for precision medicine. *Comput Support Coop Work* **25:** 115–136. doi:10.1007/s10606-016-9248-7

Girard A. 1989. A fast 'Monte-Carlo cross-validation' procedure for large least squares problems with noisy data. *Numer Math* **56:** 1–23. doi:10.1007/BF01395775

Haseman J, Elston R. 1972. The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet* **2:** 3–19. doi:10.1007/BF01066731

Hou K, Burch KS, Majumdar A, Shi H, Mancuso N, Wu Y, Sankararaman S, Pasaniuc B. 2019. Accurate estimation of SNP-heritability from biobank-scale data irrespective of genetic architecture. *Nat Genet* **51:** 1244–1251. doi:10.1038/s41588-019-0465-0

Hutchinson MF. 1989. A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Commun Stat -Simul Comput* **18:** 1059–1076. doi:10.1080/03610918908812806

Johnson R, Ding Y, Bhattacharya A, Knyazev S, Chiu A, Lajonchere C, Geschwind DH, Pasaniuc B. 2023. The UCLA ATLAS community health initiative: promoting precision health research in a diverse biobank. *Cell Genom* **3:** 100243. doi:10.1016/j.xgen.2022.100243

Kurki MI, Karjalainen J, Palta P, Sipilä TP, Kristiansson K, Donner KM, Reeve MP, Laivuori H, Aavikko M, Kaunisto MA, et al. 2023. FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613:** 508–518. doi:10.1038/s41586-022-05473-8

Lappalainen T, Li YI, Ramachandran S, Gusev A. 2024. Genetic and molecular architecture of complex traits. *Cell* **187:** 1059–1075. doi:10.1016/j.cell.2024.01.023

Liberty E, Zucker SW. 2009. The Mailman algorithm: a note on matrix–vector multiplication. *Inf Process Lett* **109:** 179–182. doi:10.1016/j.ipl.2008.09.028

Loh PR, Bhatia G, Gusev A, Finucane HK, Bulik-Sullivan BK, Pollack SJ, Schizophrenia Working Group of the Psychiatric Genomics Consortium, de Candia TR, Lee SH, Wray NR, et al. 2015a. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat Genet* **47:** 1385–1392. doi:10.1038/ng.3431

Loh PR, Tucker G, Bulik-Sullivan BK, Vilhjálmsson BJ, Finucane HK, Salem RM, Chasman DI, Ridker PM, Neale BM, Berger B, et al. 2015b. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet* **47:** 284–290. doi:10.1038/ng.3190

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. 2009. Finding the missing heritability of complex diseases. *Nature* **461:** 747–753. doi:10.1038/nature08494

Niarchou M, Byrne EM, Trzaskowski M, Sidorenko J, Kemper KE, McGrath JJ, O'Donovan MC, Owen MJ, Wray NR. 2020. Genome-wide association study of dietary intake in the UK Biobank study and its associations with schizophrenia and other traits. *Transl Psychiatry* **10:** 51. doi:10.1038/s41398-020-0688-y

Pazokitoroudi A, Wu Y, Burch KS, Hou K, Zhou A, Pasaniuc B, Sankararaman S. 2020. Efficient variance components analysis across millions of genomes. *Nat Commun* **11:** 4020. doi:10.1038/s41467-020-17576-9

Savatt J, Pisieczko CJ, Zhang Y, Lee MT, Faucett WA, Williams JL. 2019. Biobanks in the era of genomic data. *Curr Genet Med Rep* **7:** 153–161. doi:10.1007/s40142-019-00171-w

Sham PC, Purcell S. 2001. Equivalence between Haseman-Elston and variance-components linkage analyses for sib pairs. *Am J Hum Genet* **68:** 1527–1532. doi:10.1086/320593

Shi H, Kichaev G, Pasaniuc B. 2016. Contrasting the genetic architecture of 30 complex traits from summary association data. *Am J Hum Genet* **99:** 139–153. doi:10.1016/j.ajhg.2016.05.013

Speed D, Balding DJ. 2019. SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nat Genet* **51:** 277–284. doi:10.1038/s41588-018-0279-5

Wan Z, Hazel JW, Clayton EW, Vorobeychik Y, Kantarcioglu M, Malin BA. 2022. Sociotechnical safeguards for genomic data privacy. *Nat Rev Genet* **23:** 429–445. doi:10.1038/s41576-022-00455-y

Wu Y, Sankararaman S. 2018. A scalable estimator of SNP heritability for biobank-scale data. *Bioinformatics* **34:** i187–i194. doi:10.1093/bioinformatics/bty253

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, et al. 2010. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42:** 565–569. doi:10.1038/ng.608

Yang J, Lee SH, Goddard ME, Visscher PM. 2011. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88:** 76–82. doi:10.1016/j.ajhg.2010.11.011

Yang J, Bakshi A, Zhu Z, Hemani G, Vinkhuyzen AA, Lee SH, Robinson MR, Perry JR, Nolte IM, van Vliet-Ostaptchouk JV, et al. 2015. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat Genet* **47:** 1114–1120. doi:10.1038/ng.3390

Zhou X. 2017. A unified framework for variance component estimation with summary statistics in genome-wide association studies. *Ann Appl Stat* **11:** 2027. doi:10.1214/17-AOAS1052

Zhou X, Stephens M. 2012. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* **44:** 821–824. doi:10.1038/ng.2310

Received February 26, 2024; accepted in revised form July 12, 2024.

Scalable summary-statistics-based heritability estimation method with individual genotype level accuracy

Moonseong Jeong, Ali Pazokitoroudi, Zhengtong Liu, et al.

Genome Res. 2024 34: 1286-1293 originally published online July 22, 2024
Access the most recent version at doi:10.1101/gr.279207.124