Estimating SNP heritability in presence of population substructure in biobank-scale datasets

Zhaotong Lin[1], Souvik Seal[1], and Saonli Basu[1]

[1]Division of Biostatistics, University of Minnesota, Minneapolis, Minnesota

Abstract

SNP heritability of a trait is measured by the proportion of total variance explained by the additive effects of genome-wide single nucleotide polymorphisms (SNPs). Linear mixed models are routinely used to estimate SNP heritability for many complex traits. The basic concept behind this approach is to model genetic contribution as a random effect, where the variance of this genetic contribution attributes to the heritability of the trait. This linear mixed model approach requires estimation of 'relatedness' among individuals in the sample, which is usually captured by estimating a genetic relationship matrix (GRM). Heritability is estimated by the restricted maximum likelihood (REML) or method of moments (MOM) approaches, and this estimation relies heavily on the GRM computed from the genetic data on individuals. Presence of population substructure in the data could significantly impact the GRM estimation and may introduce bias in heritability estimation. The common practice of accounting for such population substructure is to adjust for the top few principal components of the GRM as covariates in the linear mixed model. Here we propose an alternative way of estimating heritability in multi-ethnic studies. Our proposed approach is a MOM estimator derived from the Haseman-Elston regression and gives an asymptotically unbiased estimate of heritability in presence of population stratification. It introduces adjustments for the population stratification in a second-order estimating equation and allows for the total phenotypic variance vary by ethnicity. We study the performance of different MOM and REML approaches in presence of population stratification through extensive simulation studies. We estimate the heritability of height, weight and other anthropometric traits in the UK Biobank cohort to investigate the impact of subtle population substructure on SNP heritability estimation.

Keywords: Haseman-Elston Regression; population substructure; SNP-heritability estimation; method of moments; UK Biobank

1 Introduction

Fundamental to the study of the inheritance is the partitioning of the total phenotypic variation into genetic and environmental components (Visscher et al., 2008). Using twin studies, the phenotypic variance can be partitioned to include the variance of an additive genetic effect, shared and non-shared environmental effects. The ratio of the genetic variance component to the total phenotypic variance is the proportion of genetically controlled variation and is termed as the 'narrow-sense heritability'. As shown in the recent review of more than 17,000 twin studies (Polderman et al., 2015), heritability provides useful information to estimate familial recurrence risk of disease, to inform about the genetic architecture of the trait, and to generate an upper bound for disease risk prediction.

In recent years, the genome-wide association studies (GWAS) are gaining momentum with the availability of whole genome sequencing data. Heritability is routinely being estimated from the genome-wide data on variants (single nucleotide polymorphisms or SNPs), which is often termed as 'SNP heritability'. Traditionally, SNP heritability is estimated by fitting variance components models with restricted maximum likelihood (REML) approach. These approaches partition the phenotypic covariance matrix of all individuals into a genetic similarity matrix and a random variation matrix (Yang et al., 2010; Lee et al., 2011, 2012; Ripke et al., 2013; AR et al., 2014; Locke et al., 2015).

However, with the large sample size, for example, biobanks that assay hundreds of thousands of individuals (UK Biobank (Biobank, 2014), Precision Medicine cohort (Ashley, 2015), Millions Veterans Program (Gaziano et al., 2016)), existing heritability estimation methods such as REML-based methods become computationally expensive and memory intensive, and thus can be difficult to apply.

Alternatively, there are method-of-moments (MOM) estimators for heritability. LD-score regression approach (Bulik-Sullivan et al., 2015) estimates heritability by regressing the summary statistics from single variant association analysis in a GWAS on linkage disequilibrium (LD) scores. A version of Haseman-Elston approach (Haseman, 1972) for heritability estimation provides a method of moments estimator for the heritability parameter by associating phenotypic covariance values with genetic covariance estimates. There are several recent work on extending these Method of Moments estimators (Ge et al., 2015; Schwartzman et al., 2019; Ma and Dicker, 2019; Hou et al., 2019) to make it more computationally feasible for large sample sizes and more robust to linkage disequilibrium.

Presence of population substructure can significantly bias the heritability estimation (Browning and Browning, 2011). Confounding can occur because of not accounting for the phenotypic differences among different sub-populations due to differences in environmental influences. Moreover, population substructure introduces differences in allele frequencies across sub-populations. Current heritability estimation methods primarily work well in samples from a homogeneous population. However, diverse populations are increasingly being used to conduct GWAS to improve fine-mapping of relevant variants. Recently, Conomos et al. (2016) performed an association study in the admixed Hispanic Community Health Study/Study of Latinos (HCHS/SOL) samples, where many biomedical traits in HCHS/SOL displayed heterogeneous variances across ethnic groups. Modeling this heteroscedasticity reduced genomic inflation. Conomos et al. (2016) estimated the underlying ethnic groups through multi-dimensional scaling and estimated distinct ethnic clusters to implement such correction. It is often desirable to implement such corrections on a continuous scale, for example, modeling heterogeneity in variances along the axes of genetic variation.

In this paper, we propose a strategy to correct for the impact of population stratification on heritability estimation with Haseman-Elston regression. Our approach does not require classifying individuals into discrete sub-populations, rather the corrections are implemented as a function of axes of genetic variation. Another huge advantage of our proposed approach is that it is a method of moments estimator and can provide computationally efficient estimates of heritability even for large biobank-scale datasets.

The rest of the paper is arranged as follows. We describe few existing approaches to estimate heritability. We propose our modified Haseman-Elston estimator and show the equivalency with the heritability estimator proposed by Ge et al. (2015). We further demonstrate that this estimator gives an unbiased estimate of heritability in presence of 2 discrete sub-populations. We explore the performance of the estimator under various alternative models and compare the performance with existing approaches. Finally we estimate heritability for a number of anthropometric traits on UK Biobank dataset.

2 Methods

Linear mixed models are emerging as the method of choice for association testing in genome-wide association studies (GWAS) because they account for both population stratification and cryptic relatedness and achieve increased statistical power by jointly modeling all genotyped markers.

2.1 Existing Approaches

Here we first introduce a general mixed-effect model to quantify how genes influence phenotypes. Suppose the data consists of P SNPs on N subjects. For a subject i ($i = 1, \ldots, N$), $\mathbf{y}_i$ is a normally distributed continuous outcome, $\mathbf{C}_i$ is the vector of covariates, $\boldsymbol{\beta}$ is a vector of fixed effects, $\mathbf{Z}_i$ is a $P \times 1$ vector of genetic variants from a GWAS. The outcome $\mathbf{y}_i$ depends on $\mathbf{Z}_i$ through the following mixed effect model, $\mathbf{y} = \mathbf{C}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}$, with $\text{var}(\mathbf{y}) = \boldsymbol{\Sigma} = \mathbf{Z}\mathbf{Z}'\sigma_g^2 + \mathbf{I}\sigma_e^2$, where $\mathbf{u}$ is a vector of SNP effects with $\mathbf{u} \sim N(0, \mathbf{I}\sigma_u^2)$, $\mathbf{I}$ is an $N \times N$ identity matrix, and $\boldsymbol{\epsilon}$ is a vector of residual effects with $\boldsymbol{\epsilon} \sim N(0, \mathbf{I}\sigma_e^2)$. $\mathbf{Z}$ is a standardized $N \times P$ genotype matrix with the is-th element $\mathbf{z}_{is} = (x_{is} - 2p_s)/\sqrt{2p_s(1-p_s)}$, where x_{is} is the number of copies of the reference allele for the s-th SNP of the i-th individual and p_s is the frequency of the reference allele. If we define $\mathbf{A} = \mathbf{Z}\mathbf{Z}'/P$ and define σ_g^2 as the variance explained by all the SNPs, i.e, $\sigma_g^2 = P\sigma_u^2$, with P being the number of SNPs, then the above linear model reduces to

$$\mathbf{y} = \mathbf{C}\boldsymbol{\beta} + \mathbf{g} + \boldsymbol{\epsilon}, \quad \text{with } \boldsymbol{\Sigma} = \mathbf{A}\sigma_g^2 + \mathbf{I}\sigma_e^2, \tag{1}$$

where $\mathbf{g}$ is an $N \times 1$ vector of the total genetic effects of the individuals with $\mathbf{g} \sim N(0, \mathbf{A}\sigma_g^2)$, and $\mathbf{A}$ is interpreted as the genetic relationship matrix (GRM) between individuals. Note that the genetic relatedness $\mathbf{A}_{ij}$ between the i-th individual and the j-th individual is measured by the dot product of their standardized genotypes and then divided by the number of markers, $\mathbf{A}_{ij} = \frac{z_i z_j}{P} = \frac{1}{P}\sum_{s=1}^{P}\Big(\big((x_{is} - 2p_s)/\sqrt{2p_s(1-p_s)}\big)\big(x_{js} - 2p_s)/\sqrt{2p_s(1-p_s)}\big)\Big)$. The heritability h^2 of the trait $\mathbf{y}$ is defined as $h^2 = \sigma_g^2/(\sigma_g^2 + \sigma_e^2)$. We are interested in estimating the parameters σ_g^2 and σ_e^2.

Maximum Likelihood Estimation

We assume $\mathbf{y} \sim N(\mathbf{C}\boldsymbol{\beta}, \boldsymbol{\Sigma} = \mathbf{A}\sigma_g^2 + \mathbf{I}\sigma_e^2)$. The likelihood of the data is given by:

$$l(\mu, \beta, \sigma_g^2, \sigma_e^2) \quad = \quad \frac{N}{2}\log(2\pi) + \frac{1}{2}\log\det\boldsymbol{\Sigma} - \frac{1}{2}(\mathbf{y} - \mathbf{C}\boldsymbol{\beta})\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{C}\boldsymbol{\beta}),$$

The estimation of the mixed effect model mentioned above is performed through maximum likelihood estimation. The software GCTA (Yang et al., 2011) uses the iterative restricted maximum likelihood (REML) algorithm to estimate the variance components σ_g^2 and σ_e^2 in the model 2 and gives an estimate of heritability by $\hat{h}^2 = \hat{\sigma}_g^2/\hat{\sigma}_p^2$, where $\hat{\sigma}_p^2$ is the estimated phenotypic total variance ($\sigma_p^2 = \sigma_g^2 + \sigma_e^2$).

However, this mixed-model methods can easily become computationally intractable as the sample size increases. Recently, there have been attempts to generate computationally scalable algorithms to implement this mixed models on biobank scale data (Loh et al., 2015). However, these approaches still encounter computational challenges on large biobanks. Moreover, even subtle population substructure could significantly impact the heritability estimation (Conomos et al., 2016).

Method of Moments Approach

The method of moments (Haseman-Elston regression (Haseman, 1972), LDscore regression (Bulik-Sullivan et al., 2015), MMHE (Ge et al., 2017)) approaches are another set of widely used methods for estimating heritability h^2 under Equation 1. We will next provide short overview of these different approaches.

Haseman-Elston (HE) Regression: Generally we assume that the GRM is normalized with its diagonal entries all equal 1 and $\mathbf{y}$ is centered and that Equation 1 holds. One of the classical moment estimators for h^2 comes from the least squares regression coefficient for regressing $\mathbf{y}_i\mathbf{y}_j$ on $\mathbf{A}_{ij}$ for all $i < j$. This is because Equation 1 implies that $E(\mathbf{y}_i\mathbf{y}_j|\mathbf{A}) = h^2\mathbf{A}_{ij}$ for $i \neq j$. The heritability can be estimated from the following equation:

$$\mathbf{y}_i\mathbf{y}_j = \beta^*\mathbf{A}_{ij} + \epsilon^*_{ij} \tag{2}$$

Note $\beta^* = h^2$ is the heritability parameter. The corresponding estimator for h^2 is $\hat{h}^2_{HE} = \left(\hat{Var}(\mathbf{A}_{ij})\right)^{-1}\hat{Cov}(y_iy_j, \mathbf{A}_{ij})$. Note that

$$\begin{aligned} \hat{Var}(\mathbf{A}_{ij}) &= \frac{2}{N(N-1)}\sum_{i<j}\mathbf{A}^2_{ij}, \\ \hat{Cov}(y_iy_j, \mathbf{A}_{ij}) &= \frac{2}{N(N-1)}\sum_{i<j}y_iy_j\mathbf{A}_{ij}. \end{aligned}$$

Henderson (1984) used least squares in this way to estimate $\hat{h}^2_{HE}$. This approach is also referred to as Haseman-Elston (HE) regression (Haseman, 1972).

A modification to the above approach is to consider two estimating equations for both parameters σ^2_g and σ^2_e:

$$E(\mathbf{y}_i\mathbf{y}_j) = \sigma^2_g\mathbf{A}_{ij},\ E(\mathbf{y}^2_i) = \sigma^2_g\mathbf{A}_{ii} + \sigma^2_e \tag{3}$$

Again, least square estimation can be used to produce unbiased estimates of σ_g^2 and σ_e^2.

Linkage Disequilibrium Score Regression: As an alternative method, linkage disequilibrium score (LDSC) regression (Bulik-Sullivan et al., 2015) has become a popular approach for estimating SNP heritability from summary statistics. LDSC estimates SNP heritability by regressing squared per-SNP univariate regression test statistics on corresponding "LD Scores", defined as estimates of the sum of squared correlations for a given SNP with all other SNPs within a region. The main advantage of this approach is that it can utilize the summary statistics generated from a GWAS, which are publicly available. The asymptotic equivalence between LDSC approach and the HE regression has been derived under certain assumptions (Chen, 2014; Bulik-Sullivan, 2015). However, while an effective and computationally efficient approach, LDSC relies on a number of assumptions, including independence of individuals to compute the summary statistics, binning of LD scores. This introduces some arbitrariness to consider the approach analytically and limit the assessment of its theoretical properties.

MMHE: Ge et al. (2016, 2017) proposed this MOM estimator, which is closely related to Haseman-Elston regression estimator (Haseman, 1972) and is equivalent to LD score regression estimator under certain situations (Bulik-Sullivan et al., 2015; Bulik-Sullivan, 2015). Specifically, in the presence of covariates, i.e., $\mathbf{y} = \mathbf{C}\boldsymbol{\beta} + \mathbf{g} + \boldsymbol{\epsilon}$, an $N \times (N-k)$ matrix $\mathbf{U}$ always exists, such that $\mathbf{U}^T\mathbf{U} = \mathbf{I}, \mathbf{U}\mathbf{U}^T = \mathbf{H}, \mathbf{U}^T\mathbf{C} = \mathbf{0}$ and $\mathbf{H} = \mathbf{I} - \mathbf{C}(\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T$.Applying $\mathbf{U}^T$ to both sides of the model gives $\mathbf{U}^T\mathbf{y} = \mathbf{U}^T\mathbf{g} + \mathbf{U}^T\boldsymbol{\epsilon}$. The covariance structure of the transformed trait is $cov[\mathbf{U}^T\mathbf{y}] = \sigma_g^2\mathbf{U}^T\mathbf{A}\mathbf{U} + \sigma_e^2\mathbf{I}_{(N-k)}$. Then by converting a matrix into a vector by stacking its columns, an ordinary least squares (OLS) estimator of $\hat{\sigma^*}_g^2, \hat{\sigma^*}_e^2$ can be obtained by solving the linear system:

$$\begin{bmatrix} \operatorname{Tr}\mathbf{HAHA} & \operatorname{Tr}\mathbf{HA} \\ \operatorname{Tr}\mathbf{HA} & N-k \end{bmatrix} \begin{bmatrix} \sigma_g^{*2} \\ \sigma_e^{*2} \end{bmatrix} = \begin{bmatrix} \mathbf{y^T HAHy} \\ \mathbf{y^T Hy} \end{bmatrix} \tag{4}$$

and obtain

$$c_1\hat{\sigma^*}_g^2 = (N-k)\mathbf{y^T HAHy} - (\operatorname{Tr}\mathbf{HA})\mathbf{y^T Hy} \tag{5}$$

$$c_1\hat{\sigma^*}_e^2 = -(\mathrm{Tr}\,\mathbf{HA})\mathbf{y^T HAHy} + (\mathrm{Tr}\,\mathbf{HAHA})\mathbf{y^T Hy} \tag{6}$$

$$\hat{h}^2_{\mathrm{MMHE}} = \frac{c_1\hat{\sigma^*}_g^2}{c_1(\hat{\sigma^*}_g^2 + \hat{\sigma^*}_e^2)} \tag{7}$$

where $\mathbf{H} = \mathbf{I} - \mathbf{P}$ and $\mathbf{P} = \mathbf{C}(\mathbf{C^T C})^{-1}\mathbf{C^T}$

2.2 Proposed Adjusted HE method

The MOM or the likelihood-based approaches generally assumes a homogeneous population. In a sample of diverse ancestry, these existing methods could produced very biased estimates of heritability. The proposed concept is motivated by the idea is that population substructure causes differences in allele frequencies as well as differences in trait distributions among the sub-populations. Not accounting for such differences could significantly introduce bias in the estimation of heritability. The standard approach is to adjust for principal components (PCs) estimated from the GRM $\mathbf{A}$ as covariates in Equation 1. The MOM-based approaches provide an useful alternative for large samples, but it is always not clear how such adjustments for substructure could be implemented in the approach. In this paper, we propose a two-step strategy to adjust for population substructure in estimating heritability using Haseman-Elston regression (Haseman, 1972). We first perform a regression on the mean level of the trait:

$$\mathbf{y}_i = \mathbf{C}_i\boldsymbol{\gamma} + \boldsymbol{\epsilon}_i, \tag{8}$$

by regressing out covariates $\mathbf{C}$, which might consist of k PCs and other covariates such as sex and age. We assume that the residuals $\mathbf{y}'$ still preserve the same information and structure of heritability, i.e., $\mathrm{var}(\mathbf{y}') = \mathbf{A}\sigma_g^2 + \mathbf{I}\sigma_e^2$. For the second step, we consider two different approaches to account for population stratification. One approach is to adjust the allele frequencies with PCs and recompute GRM $\mathbf{A}$ as implemented in PC-Relate (Conomos et al., 2016), and then apply the standard HE regression to obtain heritability estimate (referred as 'PC-Relate-HE'). Another novel way is that we introduce PC-based corrections in Equation 3. We will refer to the method as Adjusted-HE approach. The potential difference between PC-Relate-HE and Adjusted-HE is that, the former only adjusts the GRM entries for population substructure, whereas Adjusted-HE introduces correction to both GRM entries and to the total variance of outcome.

HE regression with PC-Relate adjusted GRM (PC-Relate-HE): PC-Relate (Conomos et al., 2016) is a PCA-based method for robust estimation of IBD-sharing probabilities and kinship coefficients that is applicable to general samples with population structure. Consider the linear regression $\mathbb{E}[\mathbf{x}_s|\mathbf{V}] = 1\alpha_0 + \mathbf{V}\boldsymbol{\alpha}_s$ where $\mathbf{x}_s$ is the vector of genotype values for all samples at SNP s and $\mathbf{V}$ is a matrix whose columns correspond to the top k PCs from PC-Air (Conomos et al., 2015). The fitted values from this regression can be used as prediction of individual-specific allele frequencies from the PCs: $\hat{\mu}_{is} = \frac{1}{2}\hat{\mathbb{E}}[x_{is}|V_i^1, \ldots, V_i^k]$. Then the PC-Relate estimator of the genetic relationship coefficient $\mathbf{A}_{ij}$ for individual i and j is

$$\hat{\mathbf{A}}_{ij} = \frac{\sum_{s=1}^{P}(x_{is} - 2\hat{\mu}_{is})(x_{js} - 2\hat{\mu}_{js})}{2\sum_{s=1}^{P}[\hat{\mu}_{is}(1-\hat{\mu}_{is})\hat{\mu}_{js}(1-\hat{\mu}_{js})]^{1/2}}$$

, where $\hat{\mu}_{is}$ and $\hat{\mu}_{js}$ are the estimated individual-specific genotype mean for individual i and j, respectively, at SNP s. Then we estimate the heritability with this PC-adjusted GRM with the standard HE regression.

Unstandardized-Adjusted-HE (UAdj-HE): In this method, we consider the following estimating equation:

$$\mathbb{E}(\mathbf{y}'\mathbf{y}'^{\mathbf{T}}) = \sigma_g^2\mathbf{A} + \sigma_e^2\mathbf{I} + \sum_{j=1}^{k} a_j\mathbf{PC_j}\mathbf{PC_j^T} \tag{9}$$

where $\mathbf{y}' = (\mathbf{I} - \mathbf{C}(\mathbf{C^TC})^{-1}\mathbf{C^T})\mathbf{y}$ is the residual of the regression in Equation 8 . One could use ordinary least square (OLS) approach to estimate σ_g^2 and σ_e^2 here. We have also derived a closed form estimator of SNP heritability using the following equations (see Appendix A),

$$\begin{bmatrix} \mathrm{Tr}\,\mathbf{A^2} - \sum_j s_j^2 & \mathrm{Tr}\,\mathbf{A} - \sum_j s_j \\ \mathrm{Tr}\,\mathbf{A} - \sum_j s_j & N-k \end{bmatrix} \begin{bmatrix} \sigma_g^2 \\ \sigma_e^2 \end{bmatrix} = \begin{bmatrix} \mathbf{y'^TAy'} - \sum_j t_j s_j \\ \mathbf{y'^Ty'} - \sum_j t_j \end{bmatrix} \tag{10}$$

and obtain

$$\begin{aligned}
c &= (N-k)(\operatorname{Tr}\mathbf{A^2} - \sum_{j=1}^{k} s_j^2) - (\operatorname{Tr}\mathbf{A} - \sum_{j=1}^{k} s_j)^2 \\
c\hat{\sigma}_g^2 &= (N-k)(\mathbf{y^T A y} - \sum_{j=1}^{k} s_j t_j) - (\operatorname{Tr}\mathbf{A} - \sum_{j=1}^{k} s_j)(\mathbf{y'^T y'} - \sum_{j=1}^{k} t_j) \\
c\hat{\sigma}_e^2 &= -(\operatorname{Tr}\mathbf{A} - \sum_{j=1}^{k} s_j)(\mathbf{y'^T A y'} - \sum_{j=1}^{k} s_j t_j) + (\operatorname{Tr}\mathbf{A^2} - \sum_{j=1}^{k} s_j^2)(\mathbf{y'^T y'} - \sum_{j=1}^{k} t_j) \\
\hat{h}^2_{\text{UAdj-HE}} &= \frac{\hat{\sigma}_g^2}{\hat{\sigma}_g^2 + \hat{\sigma}_e^2} = \frac{c\hat{\sigma}_g^2}{c(\hat{\sigma}_g^2 + \hat{\sigma}_e^2)} \qquad (11)
\end{aligned}$$

where $t_j = \mathbf{y'^T PC_j PC_j^T y'}$, $s_j = \mathbf{PC_j^T A PC_j}$.

Standardized-Adjusted-HE (SAdj-HE): If we have the residuals $\mathbf{y'}$ standardized by the sample mean and variance of Equation 8, then based on Equation 2 we have the following estimating equation:

$$\mathbb{E}(\mathbf{y'y'^T}) = h^2\mathbf{A} + (1-h^2)\mathbf{I} + \sum_{j=1}^{k} a_j \mathbf{PC_j PC_j^T} \qquad (12)$$

where $\mathbf{y'}$ is the standardized residual of the Equation 8.

Then we can obtain the estimate (derivation is shown in Appendix A)

$$\hat{h}^2_{\text{SAdj-HE}} = \frac{N - \operatorname{Tr}\mathbf{A} + \mathbf{y'^T A y'} - \mathbf{y'^T y'} - \sum_{j=1}^{k}(\mathbf{PC_j^T A PC_j} - 1)(\mathbf{y'^T PC_j PC_j^T y'} - 1)}{\operatorname{Tr}\mathbf{A}^2 - 2\operatorname{Tr}\mathbf{A} + N - \sum_{j=1}^{k}(\mathbf{PC_j^T A PC_j} - 1)^2} \qquad (13)$$

We have shown that, in the presence of two distinct sub-populations, this estimator with the first PC product adjustment can give us unbiased estimate (See Appendix B). To calculate the variance of Adj-HE estimators, we can make two similar assumptions as Ge et al. (2017) :(1) the off-diagonal elements in the empirical GRM matrix $\mathbf{A}$ are small and the diagonal elements are close to 1, such that $\mathbf{A} \approx \mathbf{I}$ and (2) the phenotypic variance can be estimated precisely. Therefore, we have $\operatorname{var}(\hat{h}^2_{\text{Adj-HE}}) \approx 2/(\operatorname{Tr}\mathbf{A}^2 - 2\operatorname{Tr}\mathbf{A} + N - \sum_{j=1}^{k}(\mathbf{PC_j^T A PC_j} - 1)^2)$. And with the assumption of independence among samples, we use the standard error of the OLS estimator derived from Equation 12 and Equation 9.

2.2.1 Relationship between MMHE and Adjusted-HE

Our proposed UAdj-HE is equivalent to the MMHE approach (Ge et al., 2017), when we adjust for the PCs in Equation 9 and Equation 4 are the same PCs computed from the entire GRM $\mathbf{A}$. Assume the set of covariates, $\mathbf{C}$ only consists of the PCs, i.e., $\mathbf{P} = \sum_j \mathbf{PC_j PC_j^T}$,

$$\begin{aligned} t_j &= \mathbf{y'^T PC_j PC_j^T y'} = \mathbf{y^T(I-P)PC_j PC_j^T (I-P)y} \\ &= \mathbf{y^T}(\mathbf{PC_j PC_j^T} - \sum_i \mathbf{PC_i PC_i^T PC_j PC_j^T})(\mathbf{I-P})\mathbf{y} = 0 \end{aligned}$$

Then Equation 10 reduces to

$$\begin{bmatrix} \operatorname{Tr}\mathbf{A^2} - \sum_j s_j^2 & \operatorname{Tr}\mathbf{A} - \sum_j s_j \\ \operatorname{Tr}\mathbf{A} - \sum_j s_j & N-k \end{bmatrix} \begin{bmatrix} \sigma_g^2 \\ \sigma_e^2 \end{bmatrix} = \begin{bmatrix} \mathbf{y^T HAHy} \\ \mathbf{y^T Hy} \end{bmatrix} \tag{14}$$

where $s_j = \mathbf{PC_j^T APC_j}$. Moreover, in Equation 4,

$$\begin{aligned} \operatorname{Tr}\mathbf{HA} &= \operatorname{Tr}(\mathbf{I-P})\mathbf{A} = \operatorname{Tr}\mathbf{A} - \sum_j \operatorname{Tr}\mathbf{PC_j PC_j^T A} = \operatorname{Tr}\mathbf{A} - \sum_j s_j \\ \operatorname{Tr}\mathbf{HAHA} &= \operatorname{Tr}(\mathbf{A-PA})(\mathbf{A-PA}) = \operatorname{Tr}\mathbf{A^2} - \operatorname{Tr}\mathbf{APA} - \operatorname{Tr}\mathbf{PA^2} + \operatorname{Tr}\mathbf{PAPA} \\ &= \operatorname{Tr}\mathbf{A^2} - 2\sum_j \mathbf{PC_j^T A^2 PC_j} + \sum_j s_j^2 + 2\sum_{i<j} v_{ij}^2, \quad \text{where } v_{ij} = \mathbf{PC_i^T APC_j} \end{aligned}$$

The only difference between Equation 4 and Equation 14 is $\operatorname{Tr}\mathbf{HAHA}$ and $\operatorname{Tr}\mathbf{A^2} - \sum_j s_j^2$. In the case of using the set of $\mathbf{PC}$s calculated from the same GRM matrix $\mathbf{A}$, we can use the fact that $\mathbf{APC_j} = \lambda_j \mathbf{PC_j}$, then $v_{ij} = 0$ and $\mathbf{PC_j^T A^2 PC_j} = s_j^2$. As a result, $\operatorname{Tr}\mathbf{HAHA} = \operatorname{Tr}\mathbf{A^2} - \sum_j s_j^2$ and two methods are equivalent.

However if the set of covariates contain other covariates such as age, sex or if the PCs are estimated by sampling an independent subset of markers from the given set of markers, the two methods are not equivalent and may produce different estimates for heritability. However, unless there is significant impact of other covariates on the variance and covariance of the trait, we do not expect the estimates to differ significantly.

3 Results

We conducted extensive simulation studies and real data analysis to evaluate the performance of Adjusted-HE, MMHE, PC-Relate-HE and GCTA-REML methods with and without principal components adjustment.

3.1 Simulation Studies

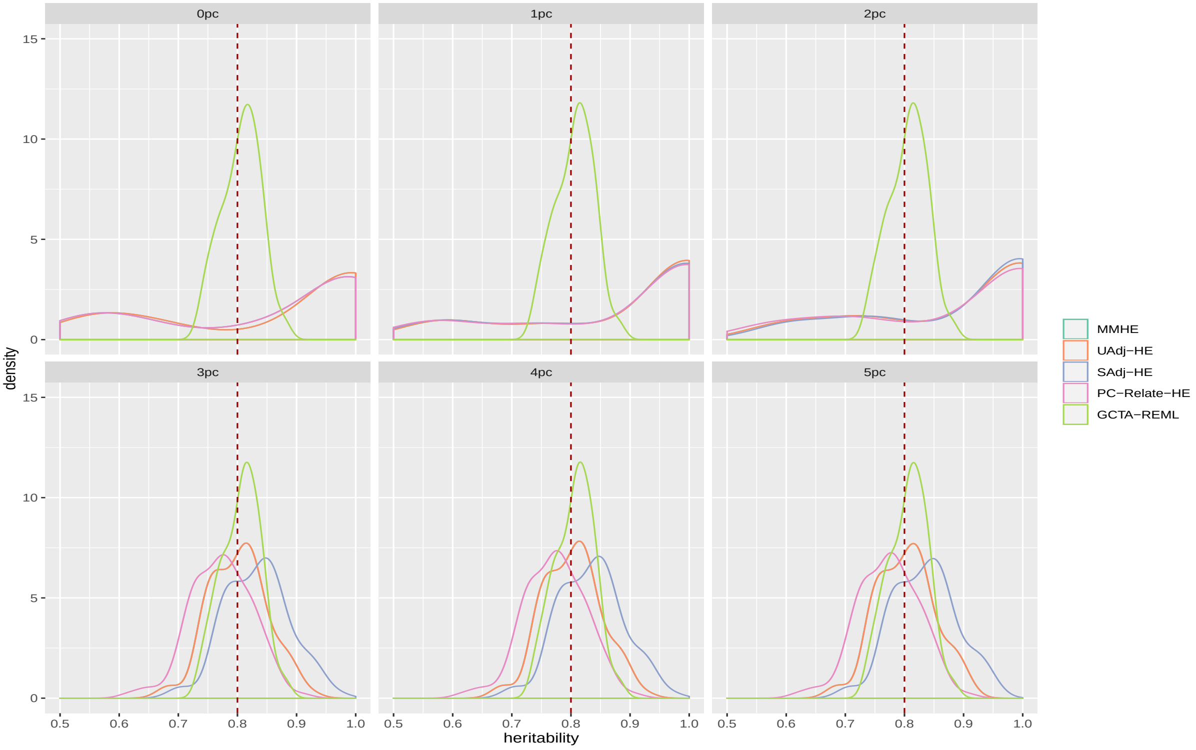
We considered 3 different simulation setup to assess the performance of these methods in presence of allele frequency differences and differences in trait distributions among the sub-populations. We simulated allele frequencies of 15,000 SNPs for each distinct population using the Balding-Nichols model (Balding and Nichols, 1995). The SNPs were assumed to be uncorrelated. For each SNP s, the allele frequency p_{0s} in the ancestral population was drawn from a uniform distribution on $[0.1, 0.9]$. In simulation 1 and 2, for each sub-population k, the allele frequency p_{ks} was generated from a beta distribution with parameters $p_{0s}(1-\theta_k)/\theta_k$ and $(1-p_{0s})(1-\theta_k)/\theta_k$. The parameter θ_k was set to a common value in simulation 1 and we varied θ_k across sub-populations in simulation 2. In simulation 3, the allele frequency of k-th population p_{ks} at SNP s was generated from a beta distribution with parameters $p_{(k-1)s}(1-\theta_k)/\theta_k$ and $(1-p_{(k-1)s})(1-\theta_k)/\theta_k$, where θ_k was set to a common value ($k = 1, 2, 3, 4$ and $s = 1, 2, \ldots, 15000$). Next, we simulated the genotypes $\mathbf{x}_{ks}$ of individuals in sub-population k from a binomial distribution $Bin(2, p_{ks})$ assuming Hardy-Weinberg equilibrium. We only considered SNPs with MAF > 0.05 and selected m (15,000 × p_{causal}) SNPs as causal variants with an effect size $\mathbf{u} \sim N(0, \frac{h^2}{m})$, where p_{causal} was the proportion of causal variants. Then the residual effects e_k were generated from a normal distribution with mean of 0 and variance of $\hat{\sigma}^2_{g_k}(1/h^2 - 1)$, where $\hat{\sigma}^2_{g_k}$ is the empirical variance of $\mathbf{X}_k\mathbf{u}$, $\mathbf{X}_k$ is a $n_k \times m$ unstandardized causal genotype matrix, $\mathbf{u}$ is a $m \times 1$ vector of causal effects and h^2 is the given heritability. Finally, we simulated phenotype y_{ki} of individual i in population k as $y_{ki} = \mathbf{X}_{ki}\mathbf{u} + e_{ki} + a_k$, where $\mathbf{X}_{ki}$ is a $1 \times m$ vector of causal SNPs of individual i in sub-population k and a_k is a population-intercept to make the means of sub-populations more different.

We considered 4 discrete populations, each with 1,000 samples, and $p_{causal} = 0.02$ in all simulations. The results were based on 100 replications for each setup. We used SAdj-HE, UAdj-HE, MMHE, PC-Relate-HE and GCTA-REML with no PC adjustment to 5 PCs adjustments for estimating heritability.

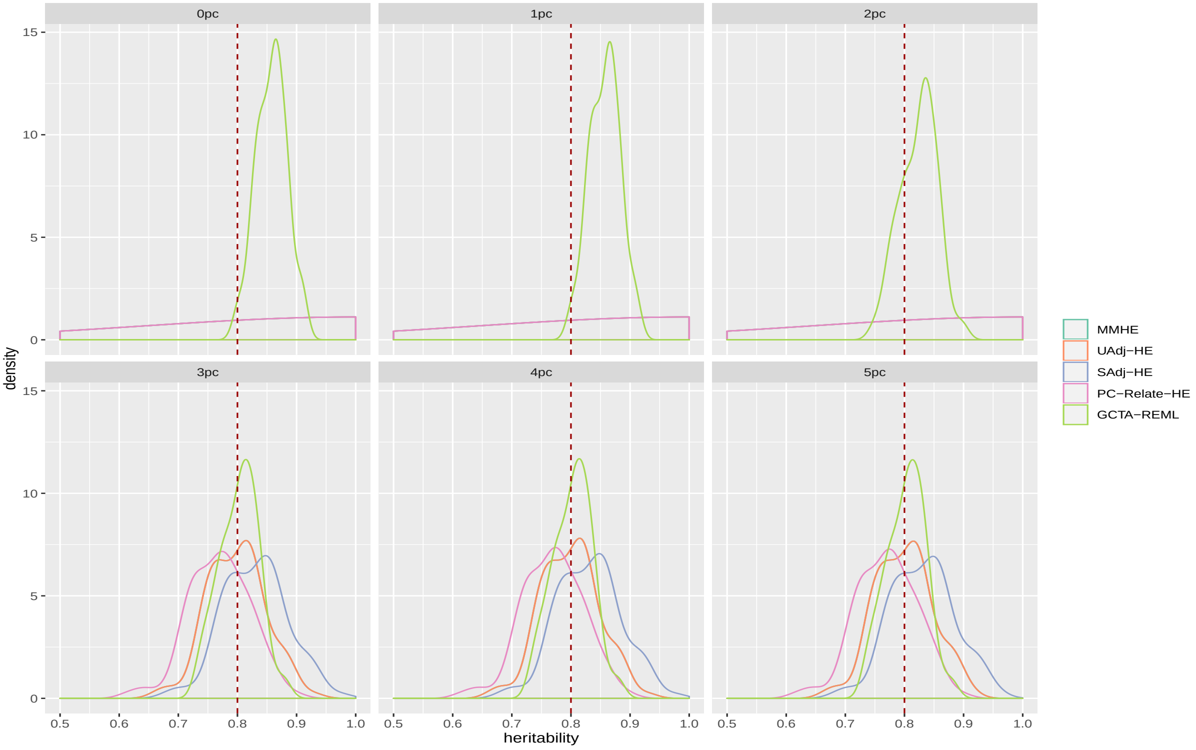
Simulation 1: We considered θ_k of 0.01 (closely related sub-populations) and 0.1 (more divergent populations) with h^2 set to 0.8. We also considered scenarios without or with population-intercept (i.e., $(a_1, a_2, a_3, a_4) = (0, 0, 0, 0)$ or $(a_1, a_2, a_3, a_4) = (0, 1, 2, 3)$). Figure 1 shows the results of scenario when $\theta_k = 0.1$ and $h^2 = 0.8$. The density curves of MMHE and UAdj-HE were identical, since they are equivalent methods as demonstrated in Section 2.2.1. As expected, the heritability estimation for MOM approaches stabilized after adjusting for 3 PCs as the sample had 4 different sub-populations. When the mean differences in $\mathbf{y}$ across sub-populations were small ($a_k = 0$), GCTA-REML handled the impact of population substructure well, even when there were no PC adjustments (Figure 1 top panel). But when population means were different, i.e., $(a_1, a_2, a_3, a_4) = (0, 1, 2, 3)$, GCTA-REML showed bias in heritability estimation when less than 3 PCs were used as covariates (Figure 1 bottom panel).

Figure 2 shows boxplots of heritability estimates over 100 replicates for different methods. We show results for all the methods adjusted for 3 PCs. When 4 populations were similar ($\theta_k = 0.01$ and $a_k = 0$), all methods performed well while GCTA-REML had the smallest variance. But when the populations were genetically similar ($\theta_k = 0.01$) but with different population intercept ($a_k \neq 0$), heritability was underestimated by all methods. This is possibly due to the fact that the PCs were not informative to distinguish between the sub-populations and hence PC adjustments could not account for the differences among the subpopulations. When the populations were more diverse ($\theta_k = 0.1$), PC-Relate-HE showed downward bias, whereas GCTA-REML, MMHE and Adj-HE estimates were biased upward.

Simulation 2: In this simulation setup, we varied θ_k to consider different pairwise similarities between each of the sub-populations with the ancestral population. The parameters $\theta_1, \theta_2, \theta_3, \theta_4$ were set to 0.05, 0.1, 0.15 and 0.2 respectively. As a result, the



(a) $(a1, a2, a3, a4) = (0, 0, 0, 0)$



(b) $(a1, a2, a3, a4) = (0, 1, 2, 3)$

Figure 1. **Simulation 1:** Heritability estimation of different methods with 0 to 5 PCs adjustment with $h^2 = 0.8$ and $\theta_k = 0.1$

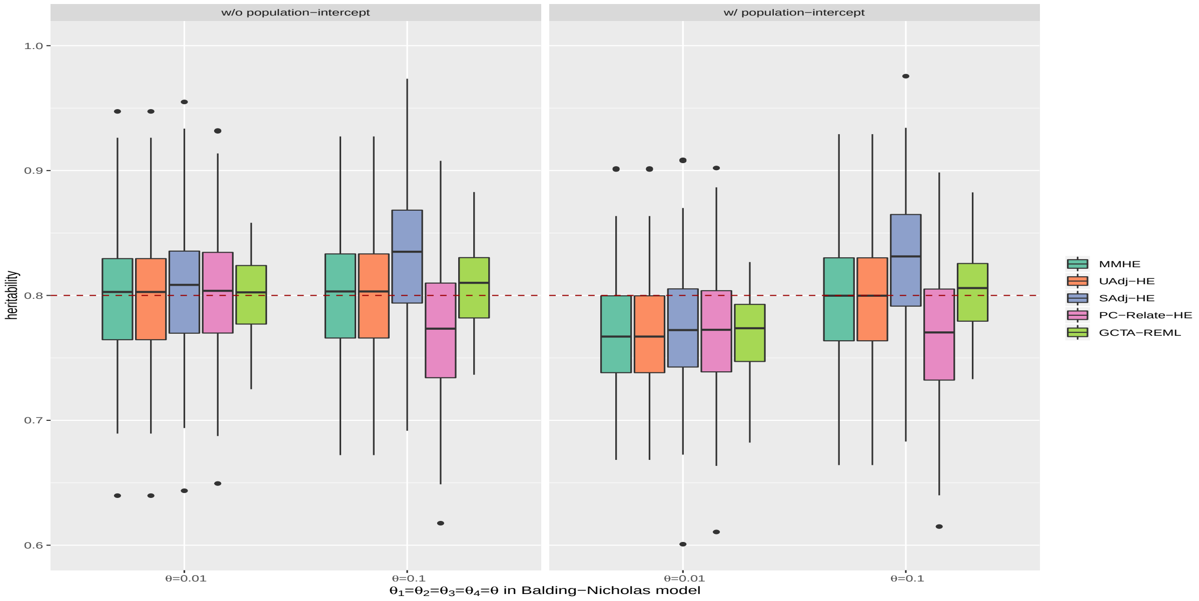


Figure 2. **Simulation 1:** Heritability estimation of different methods with 3 PCs adjustment across 100 replicates. Dashed line is the true $h^2 = 0.8$. Left: $a_k = 0$. Right: $(a_1, a_2, a_3, a_4) = (0, 1, 2, 3)$

variances of $\mathbf{y}$ in different sub-populations were different compared to simulation 1 with a common θ_k (Supplementary Figure S1). Figure 3 shows the result of scenarios without population-intercept a_k (left panel) and with population-intercept a_k (right panel). The methods MMHE, UAdj-HE and GCTA-REML showed marginal overestimation, but PC-Relate-HE significantly underestimated heritability.

Simulation 3: In this simulation, we used θ_ks to represent genetic similarity between population pair $(k-1, k), k = 2, 3, 4$. This simulation generated more diverse sub-populations as compared to Simulation 1 and Simulation 2. Similar to simulation 2, the variances of $\mathbf{y}$ in each sub-population were more different as θ_k increasing (Supplementary Figure S2). We estimated the pairwise Fst value between sub-populations using the empirical Bayes estimator in FinePop package (Kitada et al., 2007) (Supplementary Table S1). As we increased θ_k, PC-Relate-HE showed increasing downward bias and GCTA-REML and SAdj-HE biased upward more. The approaches while UAdj-HE and MMHE performed the best with smallest bias among the methods (Figure 4).

In general, in the presence of population substructure, all the MOM estimators performed better than the original HE regression without any PC adjustment, when adjusted for sufficient PCs. PC-Relate-HE showed underestimation, especially when sub-

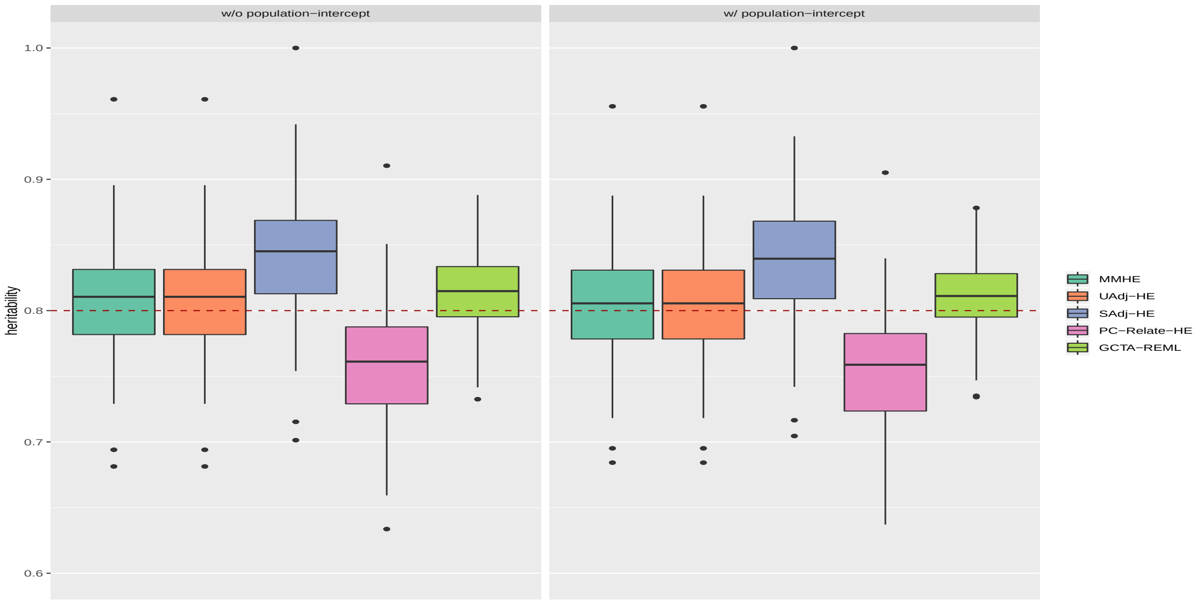


Figure 3. **Simulation 2:** Heritability estimation of different methods with 3 PCs adjustment with $h^2 = 0.8$ and $\theta_1, \theta_2, \theta_3, \theta_4 = (0.05, 0.1, 0.15, 0.2)$. Left: $a_k = 0$; Right: $(a_1, a_2, a_3, a_4) = (0, 1, 2, 3)$

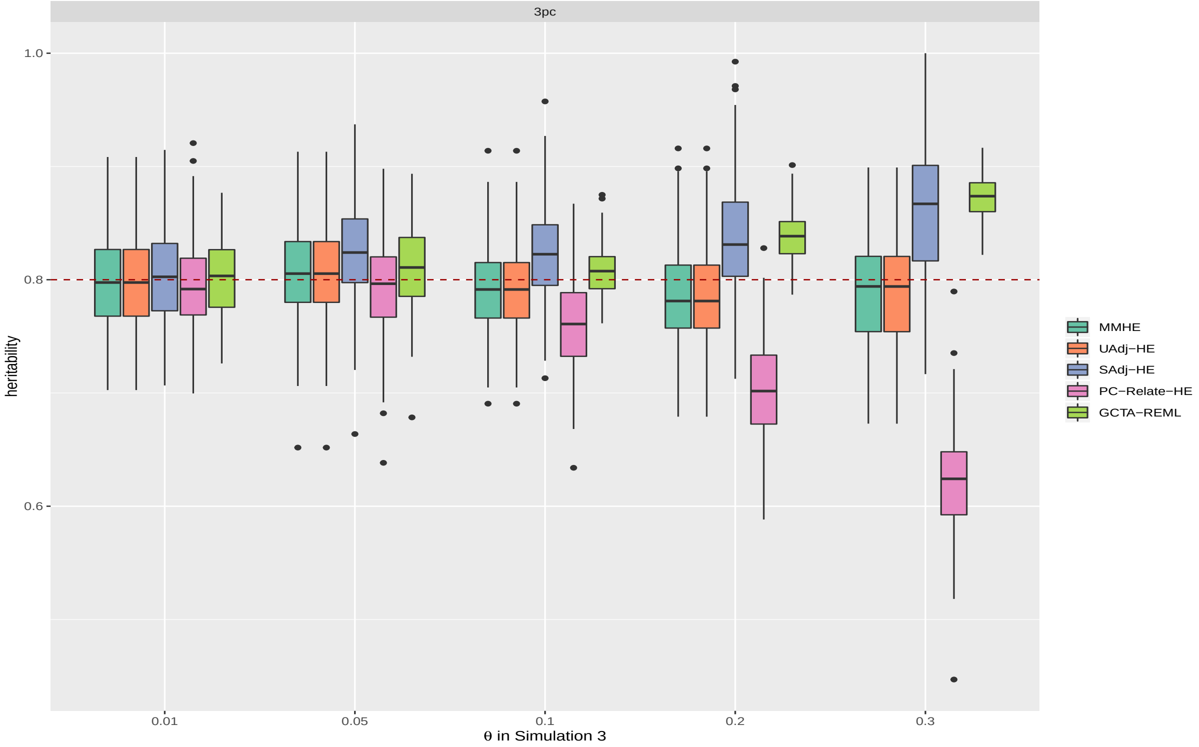


Figure 4. **Simulation 3:** Heritability estimation of different methods with 3 PCs adjustment with $h^2 = 0.8$ and $\theta_1 = \theta_2 = \theta_3 = \theta_4 = \theta$, $a_k = 0$.

populations were more diverse. Our proposed approach and REML showed slight overestimation in most scenarios. When the variances of phenotype differed across sub-populations, PC-Relate-HE underestimated severely. This is probably due to fact that PC-Relate-HE only adjusted the GRM for population substructure, whereas Adj-HE approaches corrected both the GRM and the total variance of $\mathbf{y}$ for population substructure and showed little bias in the estimation of heritability. As expected, GCTA-REML estimates had the smallest variance compared to the other MOM estimators, since the data was simulated from a normal distribution, however, it suffered from overestimation when there was heterogeneity of variances across sub-populations. SAdj-HE showed more bias in estimation over UAdj-HE, probably because of the inaccuracy in standardization of $\mathbf{y}$, especially when there was heterogeneity of variances across sub-populations. In terms of averaged computational time for each replicate of simulation, PC-Relate took about 1700 seconds to run PC-Air and create the PCs-adjusted GRM for 4,000 samples, whereas it only took less than 30 seconds totally to calculate the PCs and construct the standard GRM in PLINK and GCTA; and MMHE and GCTA-REML took about 3 seconds and 20 seconds in heritability estimation respectively whereas Adjusted-HE only took less than a second running on Haswell E5-2680v3 processors.

3.2 Real Data Analysis: UK Biobank

The UK Biobank data has approximately 800,000 markers and comprises 488,377 samples. We leveraged the QC information released by UK Biobank (Bycroft et al., 2018) and used SNPs that passed all QC tests in 106 batches. We removed the samples that had mismatch between inferred sex and self-reported sex, samples that were identified as outliers in heterozygosity and missing rates, and samples that were in the kinship table (Biobank, 2015). We further excluded SNPs that had high missing rate (>1.5%), low minor allele frequency (<1%) and subjects that had high missing genotype rate (>1%). 305,639 samples and 566,647 markers remained for the following analysis after QC. It is also worth mentioning that we did not restrict our analysis to subjects that were self-reported white British. The majority of the samples was British, but there were people with other ethnicities (Supplementary Table S3, S4). PCs were computed after LD pruning and removing long-range LD-regions (Abdellaoui et al., 2013). We pruned the SNPs after removing long-range LD-regions so that pairwise $r^2 < 0.2$ among the remaining markers for windows of 1000 markers and a step-size of 80 markers. We computed PCs using the pruned data with 247,135 markers.

We studied the performance of different approaches on a subset of the 305,639 samples, since GCTA-REML cannot be handle this large sample size. We sampled 45,510 subjects from 305,639 subjects who are self-reported White, Asian or Black, and applied GCTA-REML, LD score regression, MMHE, SAdj-HE and UAdj-HE on this sub-sample. We analyzed 7 quantitative phenotypes including height, weight, BMI, waist, levels of the diisobutyl phthalate (DiBP), systolic blood pressure (SysBP), hip and waist circumference. We adjusted for the top few PCs and other covariates such as age and height as recommended in Ge et al. (2017), except for UK Biobank assessment center (Table 1). For the Adjusted-HE methods, we first regressed out PCs and other covariates from each phenotype, then applied the closed-form formula or performed least-square estimation using Equation 12 to estimate heritability. For the MMHE method, we considered PCs and other covariates in the matrix **P**. We adjusted for covariates and PCs when conducting GWAS for LDscore regression, and for GCTA-REML, we included PCs and other covariates in Equation 1 while performing the REML estimation.

3.2.1 45k sample

45,510 subjects were sampled from 305,639 subjects who are self-reported White, Asian or Black (Supplementary Table S2). Similar to our simulation above, when all SNPs were used to compute PCs, UAdj-HE gave the same result as MMHE even when we adjusted for age and sex (results not shown here), which might indicate that other covariates such as sex and age did not affect the variance and covariance of the traits significantly. Table 1 shows the heritability estimation of different methods when 10 PCs (computed based on pruned SNPs) and appropriate covariates were adjusted. We can see that, Adjusted-HE estimates were slightly lower than MMHE, if PCs were not computed based on all SNPs that were used to generate the empirical GRM. However, Adj-HE methods were computationally more efficient (Table 3). LDSC regression method produced substantially smaller estimates compared to other approaches, which indicates severe underestimation by LDSC approach in presence of population stratification. GCTA-REML gave the highest estimates of heritability in most of the cases (except for height), which might also be overestimation due to heterogeneity of variances as shown in our simulation study. Also, as expected, in a large sample size, UAdj-HE and SAdj-HE gave similar results.

Table 1. A comparison of different methods of estimating heritability for 45k samples with 10 PCs to correct population substructure

	Height[1]	Weight[2]	BMI[1]	DiBP[1]	sysBP[1]	Hip[3]	Waist[3]
LDSC	0.523	0.196	0.197	0.119	0.136	0.099	0.047
GCTA-REML	0.648	0.346	0.344	0.180	0.196	0.142	0.165
MMHE	0.685	0.294	0.289	0.150	0.158	0.121	0.150
UAdj-HE	0.638	0.271	0.269	0.140	0.147	0.112	0.138
SAdj-HE	0.646	0.273	0.271	0.140	0.148	0.112	0.139

[1] Sex, age are included in other covariates.
[2] Sex, age and height are included in other covariates.
[3] Sex, age, height and weight are included in other covariates.

3.2.2 305k sample

As we mentioned before, this 305k UK Biobank cohort is a collection of samples from different ethnic backgrounds including White (British, Irish), Mixed (White and Black Caribbean, White and Black African, White and Asian), Asian (Indian, Pakistani, Bangladeshi, Chinese, Asian British) and Black (Caribbean, African, Black British) (Supplementary Figure S3). The estimated variability for different traits across ethnicity is shown in Supplementary Table S4, and it shows different outcome variances across sub-populations. We applied Adjusted-HE corrected for 10 PCs and other covariates on this cohort (Table 2). Compared to the Adjusted-HE results in the 45k sample, most of the estimations increased except for weight and BMI. The results also demonstrated a good consistency with previous results (Hou et al., 2019; Ge et al., 2017) and our approach was computationally very efficient. The MMHE approach needs to take block-columns GRM as input, which is not the standard GRM format provided by GCTA. The GRM file generated by

GCTA only stores the lower triangular and diagonal entries of the GRM. In contrast, our proposed method can take the standard GCTA file with whole GRM as input in a more efficient way and use formula (13) if the machine has sufficient memory, otherwise, it can also read part GRM files generated by GCTA and conducts the adjusted regression in parallel. The analysis for 10 PCs adjustment only took several minutes when jobs were paralleled on Haswell E5-2680v3 processors (Table 3).

Table 2. Heritability estimation of 305k samples with Adj-HE corrected for PCs and other covariates

Method	Height	Weight	BMI	DiBP	sysBP	Hip	Waist
SAdj-HE + 10 PCs	0.689	0.271	0.273	0.169	0.156	0.119	0.160
UAdj-HE + 10 PCs	0.684	0.271	0.272	0.169	0.155	0.119	0.159
Ge et al. (2017)[1]	0.685	0.277	0.274	0.184	0.156	0.106	0.155

[1] 108,158 self-reported white-British and 486,175 SNPs were used.

Table 3. Computational performance of MMHE, GCTA-REML and Adjusted-HE

Method	Computational time	Peak memory
MMHE, 45k[1]	522s	110GB
GCTA, 45k[1]	5160s	87GB
SAdj-HE, 45k[1]	89s	39GB
SAdj-HE, 305k[2]	479s	10GB

[1] Reported run times and memory are the average of seven runs for seven traits adjusted for 10 PCs and corresponding covariates using Haswell E5-2680v3 processors, and do not include estimating GRM and calculating PCs.

[2] Reported run times and memory are the maximum of 200 paralleled jobs for one trait adjusted for 10 PCs and corresponding covariates using Haswell E5-2680v3 processors, and do not include estimating GRM and calculating PCs.

4 Discussion

SNP-heritability, the proportion of variation in the phenotype attributable to the additive effects of a given set of SNPs, is a fundamental quantity in genetics and provides an upper bound to the risk explained by genetic prediction models. Traditionally, SNP-heritability is estimated by fitting variance components models with restricted maximum likelihood (REML). But these REML-based methods are not scalable to large biobank data. An alternative method is Haseman-Elston regression which is a moment-based method and is computationally much more efficient for large-scale datasets. In recent years, more GWAS are conducted on diverse population and the presence of population substructure can bias SNP-heritability estimation severely. For example, the difference in the outcome variance by ethnicity will impact the estimation. Another major impact of ethnicity is that the genetic relationship matrix $\mathbf{A}$ will be wrongly computed due to the difference in allele frequencies by ethnicity. Principal components estimated from GRM are usually used to account for population structure. A classical way to adjust for PCs in REML-based methods is to include them as fixed effects in the mixed linear model; and PCs can also be adjusted in the estimation of GRM using methods such as PC-Relate (Conomos et al., 2015). However, it is still unclear how to incorporate such corrections in different existing moment based approaches for estimating heritability.

In this paper, we proposed a computationally efficient MOM estimator of SNP-heritability in presence of population substructure, which can be easily applied on large scale biobank data. We have derived the estimator from the classical Haseman-Elston regression by adding product terms of PCs and have shown the equivalence with MMHE under specific conditions. We have also demonstrated the unbiasedness of our proposed estimator in presence of two discrete sub-populations. Another flexibility of our proposed approach is that it would be relatively easy to allow the heritability differ by ethnicity. One could incorporate multiple interaction terms in the Adj-HE approach (Equation 12 and Equation 9) to allow for such ethnic differences.

We conducted a number of simulations to study the performance of Adjusted-HE and other methods including GCTA-REML, MMHE and PC-Relate-HE for heritability estimation. In simulation studies under a variety of population substructure configurations, we showed that if not adjusting for PCs, MOM estimators are biased severely in the presence of population substructure; and the estimates stabilized after adjusting for PCs. When sub-populations had similar outcome mean and variances, GCTA-REML estimates were stable even without any PC adjustment, but it also needed PC adjustments to stabilize if outcome means or variances were different. We also noticed increasing downward bias for PC-Relate-HE and increasing upward bias for GCTA-REML, when the sub-populations were increasingly different in outcome variances. The UAdj-HE approach always maintained the smallest bias in these scenarios.

In the real data application on UK Biobank, we analyzed 7 quantitative traits including height, weight, BMI, systolic blood pressure, diastolic blood pressure, waist circumference and hip circumference. We compared Adjusted-HE to other widely used methods including GCTA-REML and LDSC regression on a small subset of 45k individuals, and also applied Adjusted-HE on a full sample of 305k individuals in a computationally efficient way. The results showed that LDSC regression tended to give underestimation. GCTA-REML gave higher heritability estimation than our methods for all traits, likely due to difference in trait variances by ethnicity. Our Adj-HE estimates were generally close to the estimates reported for anthropometric traits from other studies (Yang et al., 2015).

Despite the computational efficiency, our proposed method has a few limitations. Our two-step Adj-HE approach assumes that there is no impact of other covariates such as age and sex on the variance of the outcome. It also slightly underestimated heritability when the PCs were computed using a subset of markers in the GRM. Moreover, it is a bit unclear in terms of how many PC adjustments would be necessary to capture the impact of population stratification.

References

A. Abdellaoui, J.-J. Hottenga, P. De Knijff, M. G. Nivard, X. Xiao, P. Scheet, A. Brooks, E. A. Ehli, Y. Hu, G. E. Davies, et al. Population structure, migration, and diversifying selection in the netherlands. *European journal of human genetics*, 21(11):1277–1285, 2013.

W. AR et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genet.*, 46:1173–1186, 2014.

E. A. Ashley. The precision medicine initiative: a new national effort. *Jama*, 313(21): 2119–2120, 2015.

D. J. Balding and R. A. Nichols. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, 96(1-2):3–12, 1995.

U. Biobank. About uk biobank. *Available at h ttps://www. ukbiobank. ac. uk/a bout-biobank-uk*, 2014.

U. Biobank. Genotyping and quality control of uk biobank, a large-scale, extensively phenotyped prospective resource. *Available at biobank. ctsu. ox. ac. uk/crystal/docs/genotyping_ qc. pdf.*, 1:2016, 2015.

S. R. Browning and B. L. Browning. Population structure can inflate snp-based heritability estimates. *The American Journal of Human Genetics*, 89(1):191–193, 2011.

B. Bulik-Sullivan. Relationship between ld score and haseman-elston regression. *BioRxiv*, page 018283, 2015.

B. K. Bulik-Sullivan, P.-R. Loh, H. K. Finucane, S. Ripke, J. Yang, N. Patterson, M. J. Daly, A. L. Price, and B. M. Neale. Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics*, 47(3):291–295, 2015.

C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

G.-B. Chen. Estimating heritability of complex traits from genome-wide association studies using ibs-based haseman–elston regression. *Frontiers in genetics*, 5:107, 2014.

M. P. Conomos, M. B. Miller, and T. A. Thornton. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genetic epidemiology*, 39(4):276–293, 2015.

M. P. Conomos, C. A. Laurie, A. M. Stilp, S. M. Gogarten, C. P. McHugh, S. C. Nelson, T. Sofer, L. Fernández-Rhodes, A. E. Justice, M. Graff, et al. Genetic diversity and association studies in us hispanic/latino populations: applications in the hispanic community health study/study of latinos. *The American Journal of Human Genetics*, 98(1):165–184, 2016.

K. J. Galinsky, G. Bhatia, P.-R. Loh, S. Georgiev, S. Mukherjee, N. J. Patterson, and A. L. Price. Fast principal-component analysis reveals convergent evolution of adh1b in europe and east asia. *The American Journal of Human Genetics*, 98(3):456–472, 2016.

J. M. Gaziano, J. Concato, M. Brophy, L. Fiore, S. Pyarajan, J. Breeling, S. Whitbourne, J. Deen, C. Shannon, D. Humphries, et al. Million veteran program: A mega-biobank to study genetic influences on health and disease. *Journal of clinical epidemiology*, 70: 214–223, 2016.

T. Ge, T. E. Nichols, P. H. Lee, A. J. Holmes, J. L. Roffman, R. L. Buckner, M. R. Sabuncu, and J. W. Smoller. Massively expedited genome-wide heritability analysis (megha). *Proceedings of the National Academy of Sciences*, 112(8):2479–2484, 2015.

T. Ge, M. Reuter, A. M. Winkler, A. J. Holmes, P. H. Lee, L. S. Tirrell, J. L. Roffman, R. L. Buckner, J. W. Smoller, and M. R. Sabuncu. Multidimensional heritability analysis of neuroanatomical shape. *Nature communications*, 7(1):1–10, 2016.

T. Ge, C.-Y. Chen, B. M. Neale, M. R. Sabuncu, and J. W. Smoller. Phenome-wide heritability analysis of the uk biobank. *PLoS genetics*, 13(4):e1006711, 2017.

R. Haseman, JK amd Elston. The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet*, 2(1):3–19, 1972.

C. Henderson. Applications of linear models in animal breeding. 423 p. *University of Guelph, Guelph, Ontario*, 1984.

K. Hou, K. S. Burch, A. Majumdar, H. Shi, N. Mancuso, Y. Wu, S. Sankararaman, and B. Pasaniuc. Accurate estimation of snp-heritability from biobank-scale data irrespective of genetic architecture. *Nature genetics*, page 1, 2019.

S. Kitada, T. Kitakado, and H. Kishino. Empirical bayes inference of pairwise fst and its distribution in the genome. *Genetics*, 177(2):861–873, 2007.

S. Lee, N. Wray, M. Goddard, and P. Visscher. Estimating missing heritability for disease from genome-wide association studies. *Am J Hum Genet.*, 88(3):294–305, 2011.

S. Lee, T. DeCandia, S. Ripke, and J. Yang. Estimating the proportion of variation in susceptibility to schizophrenia captured by common snps. *Nat Genet*, 44(3):247–250, 2012.

A. Locke et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518:197–206, 2015.

P.-R. Loh, G. Tucker, B. K. Bulik-Sullivan, B. J. Vilhjálmsson, H. K. Finucane, R. M. Salem, D. I. Chasman, P. M. Ridker, B. M. Neale, B. Berger, et al. Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nature genetics*, 47 (3):284, 2015.

R. Ma and L. H. Dicker. The mahalanobis kernel for heritability estimation in genome-wide association studies: fixed-effects and random-effects methods. *arXiv preprint arXiv:1901.02936*, 2019.

N. Patterson, A. L. Price, and D. Reich. Population structure and eigenanalysis. *PLoS genetics*, 2(12), 2006.

T. J. Polderman, B. Benyamin, C. A. De Leeuw, P. F. Sullivan, A. Van Bochoven, P. M. Visscher, and D. Posthuma. Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature genetics*, 47(7):702–709, 2015.

S. Ripke et al. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat. Genet*, 45:1150–1159, 2013.

A. Schwartzman, A. J. Schork, R. Zablocki, W. K. Thompson, et al. A simple, consistent estimator of snp heritability from genome-wide association studies. *The Annals of Applied Statistics*, 13(4):2509–2538, 2019.

P. M. Visscher, W. G. Hill, and N. R. Wray. Heritability in the genomics era?concepts and misconceptions. *Nature reviews genetics*, 9(4):255, 2008.

J. Yang, B. Benyamin, B. McEvoy, S. Gordon, et al. Common snps explain a large proportion of the heritability for human height. *Nat. Genet.*, 42:565–569, 2010.

J. Yang, T. Manolio, L. Pasquale, E. Boerwinkle, et al. Genome partitioning of genetic variation for complex traits using common snps. *Nat. Genet.*, 43:519–525, 2011.

J. Yang, A. Bakshi, Z. Zhu, G. Hemani, A. A. Vinkhuyzen, S. H. Lee, M. R. Robinson, J. R. Perry, I. M. Nolte, J. V. van Vliet-Ostaptchouk, et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature genetics*, 47(10):1114, 2015.

Appendix A

Derivation of Adjusted-HE formula

UAdj-HE: For unstandardized $\mathbf{y}'$ (only mean-centered), we have

$$\mathbb{E}(\mathbf{y}'\mathbf{y}'^{\mathbf{T}}) = \sigma_g^2\mathbf{A} + \sigma_e^2\mathbf{I} + \sum_{j=1}^{k} a_j \mathbf{PC_j PC_j^T} \tag{9}$$

where $\mathbf{y}'$ is the residual after the regression step.

Using the method of moment and the Frobenius matrix norm to solve it,

$$\begin{aligned}(\hat{a_1}, \ldots, \hat{a_k}, \hat{\sigma_g^2}, \hat{\sigma_e^2}) &= \underset{a_j,\sigma_g^2,\sigma_e^2}{\arg\max}\, ||\mathbf{y}'\mathbf{y}'^{\mathbf{T}} - (\sigma_g^2\mathbf{A} + \sigma_e^2\mathbf{I} + \sum_{j=1}^{k} a_j \mathbf{PC_j PC_j^T})||_F \\ &= \underset{a_j,\sigma_g^2,\sigma_e^2}{\arg\max}\{(\sigma_g^2)^2 \operatorname{Tr}\mathbf{A^2} + (\sigma_e^2)^2 \operatorname{Tr}\mathbf{I} + \sum_{j=1}^{k} a_j^2 + 2\sigma_g^2\sigma_e^2 \operatorname{Tr}\mathbf{A} \\ &\quad + 2\sum_{j=1}^{k} a_j \mathbf{PC_j^T A PC_j}\sigma_g^2 + 2\sum_{j=1}^{k} a_j\sigma_e^2 \\ &\quad - 2\sigma_g^2\mathbf{y}'^{\mathbf{T}}\mathbf{A}\mathbf{y}' - 2\sigma_e^2\mathbf{y}'^{\mathbf{T}}\mathbf{y}' - 2\sum_{j=1}^{k} a_j \mathbf{y}'^{\mathbf{T}}\mathbf{PC_j PC_j^T}\mathbf{y}'\} \\ &= \underset{a_j,\sigma_g^2,\sigma_e^2}{\arg\max}\, Q(a_1, \ldots, a_k, \sigma_g^2, \sigma_e^2)\end{aligned}$$

Let $\frac{\partial Q}{\partial a_j} = 0$, $t_j = \mathbf{y}'^{\mathbf{T}}\mathbf{PC_j PC_j^T}\mathbf{y}'$, $s_j = \mathbf{PC_j^T A PC_j}$, we have $a_j = t_j - \sigma_g^2 s_j - \sigma_e^2$.

Let $\begin{cases} \frac{\partial Q}{\partial \sigma_g^2} &= 0 \\ \frac{\partial Q}{\partial \sigma_e^2} &= 0 \end{cases}$, we have

$$\begin{bmatrix} \operatorname{Tr}\mathbf{A^2} & \operatorname{Tr}\mathbf{A} \\ \operatorname{Tr}\mathbf{A} & N \end{bmatrix} \begin{bmatrix} \sigma_g^2 \\ \sigma_e^2 \end{bmatrix} = \begin{bmatrix} \mathbf{y}'^{\mathbf{T}}\mathbf{A}\mathbf{y}' - \sum_{j=1}^{k} a_j \mathbf{PC_j^T A PC_j} \\ \mathbf{y}'^{\mathbf{T}}\mathbf{y}' - \sum_{j=1}^{k} a_j \end{bmatrix}$$

Plug in a_j, it becomes

$$\begin{bmatrix} \operatorname{Tr}\mathbf{A^2} - \sum_j s_j^2 & \operatorname{Tr}\mathbf{A} - \sum_j s_j \\ \operatorname{Tr}\mathbf{A} - \sum_j s_j & N-k \end{bmatrix} \begin{bmatrix} \sigma_g^2 \\ \sigma_e^2 \end{bmatrix} = \begin{bmatrix} \mathbf{y'^T A y'} - \sum_j t_j s_j \\ \mathbf{y'^T y'} - \sum_j t_j \end{bmatrix} \tag{10}$$

Then

$$\hat{h}^2_{\text{UAdj-HE}} = \frac{\hat{\sigma}_g^2}{\hat{\sigma}_g^2 + \hat{\sigma}_e^2} = \frac{c\hat{\sigma}_g^2}{c\hat{\sigma}_g^2 + c\hat{\sigma}_e^2}$$

where

$$\begin{aligned} c &= (N-k)(\operatorname{Tr}\mathbf{A^2} - \sum_{j=1}^{k} s_j^2) - (\operatorname{Tr}\mathbf{A} - \sum_{j=1}^{k} s_j)^2 \\ c\hat{\sigma}_g^2 &= (N-k)(\mathbf{y^T A y} - \sum_{j=1}^{k} s_j t_j) - (\operatorname{Tr}\mathbf{A} - \sum_{j=1}^{k} s_j)(\mathbf{y'^T y'} - \sum_{j=1}^{k} t_j) \\ c\hat{\sigma}_e^2 &= -(\operatorname{Tr}\mathbf{A} - \sum_{i=1}^{k} s_j)(\mathbf{y'^T A y'} - \sum_{j=1}^{k} s_j t_j) + (\operatorname{Tr}\mathbf{A^2} - \sum_{j=1}^{k} s_j^2)(\mathbf{y'^T y'} - \sum_{j=1}^{k} t_j) \end{aligned}$$

SAdj-HE: For standardized $\mathbf{y'}$ (in both mean and variance), we have

$$\mathbb{E}(\mathbf{y' y'^T}) = h^2\mathbf{A} + (1-h^2)\mathbf{I} + \sum_{j=1}^{k} a_j \mathbf{PC_j PC_j^T} \tag{12}$$

where $\mathbf{y'}$ is the standardized residual after first step.

Using the same idea to solve Equation 12, we have

$$\begin{aligned} (\hat{a_1}, \ldots, \hat{a_k}, \hat{h^2}) &= \arg\max_{a_j, h^2} ||\mathbf{y' y'^T} - (h^2(\mathbf{A} - \mathbf{I}) + \mathbf{I} + \sum_j a_j \mathbf{PC_j PC_j^T})||_F \\ &= \arg\max_{a_j, h^2} \{(h^2)^2(\operatorname{Tr}\mathbf{A^2} - 2\operatorname{Tr}\mathbf{A} + N) + \sum_j a_j^2 + 2h^2(\operatorname{Tr}\mathbf{A} - N) \\ &\quad + 2\sum_j a_j + 2h^2 \sum_j a_j \mathbf{PC_j^T A PC_j} - 2h^2 \sum_j a_j \\ &\quad - 2h^2 \mathbf{y'^T A y'} + 2h^2 \mathbf{y'^T y'} - 2\sum_j a_j \mathbf{y'^T PC_j PC_j^T y'}\} \\ &= \arg\max_{a_j, h^2} Q(h^2, a_1, \ldots, a_k) \end{aligned}$$

Let $\begin{cases} \frac{\partial Q}{\partial a_j} & = 0 \\ \frac{\partial Q}{\partial h^2} & = 0 \end{cases}$, we have

$$\begin{bmatrix} \operatorname{Tr}\mathbf{A}^2-2\operatorname{Tr}\mathbf{A}+N & \mathbf{PC_1^TAPC_1}-1 & \mathbf{PC_2^TAPC_2}-1 & \dots & \mathbf{PC_k^TAPC_k}-1 \\ \mathbf{PC_1^TAPC_1}-1 & 1 & 0 & \dots & 0 \\ \mathbf{PC_2^TAPC_2}-1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{PC_k^TAPC_k}-1 & 0 & 0 & \dots & 1 \end{bmatrix} \begin{bmatrix} h^2 \\ a_1 \\ a_2 \\ \vdots \\ a_k \end{bmatrix} = \begin{bmatrix} N-\operatorname{Tr}\mathbf{A}+\mathbf{y'^TAy'}-\mathbf{y'^Ty'} \\ \mathbf{y'^TPC_1PC_1^Ty'}-1 \\ \mathbf{y'^TPC_2PC_2^Ty'}-1 \\ \vdots \\ \mathbf{y'^TPC_kPC_k^Ty'}-1 \end{bmatrix}$$

Then we can obtain

$$(\operatorname{Tr}\mathbf{A}^2 - 2\operatorname{Tr}\mathbf{A} + N)h^2 + \sum_{j=1}^{k}(\mathbf{PC_j^TAPC_j} - 1)a_j = N - \operatorname{Tr}\mathbf{A} + \mathbf{y'^TAy'} - \mathbf{y'^Ty'} \tag{15}$$

$$a_j = (\mathbf{y'^TPC_jPC_j^Ty'} - 1) - (\mathbf{PC_j^TAPC_j} - 1)h^2 \tag{16}$$

Substitute (16) into (15) we have

$$\hat{h}^2_{\text{SAdj-HE}} = \frac{N - \operatorname{Tr}\mathbf{A} + \mathbf{y'^TAy'} - \mathbf{y'^Ty'} - \sum_{j=1}^{k}(\mathbf{PC_j^TAPC_j} - 1)(\mathbf{y'^TPC_jPC_j^Ty'} - 1)}{\operatorname{Tr}\mathbf{A}^2 - 2\operatorname{Tr}\mathbf{A} + N - \sum_{j=1}^{k}(\mathbf{PC_j^TAPC_j} - 1)^2}$$

Appendix B

Expectation of the heritability estimates

Assuming that different clusters affect allele frequencies (GRM) only and not the mean directly, and the variance of $\mathbf{y}$ is estimated precisely by the sample variance. When we only consider PCs from the full GRM in the adjustment, formula 13 can be written as

$$\begin{aligned} \hat{h}^2_{\text{SAdj-HE}} &= \frac{N - \operatorname{Tr}\mathbf{A} + \mathbf{y^TAy} - \mathbf{y^Ty} - \sum_{j=1}^{k}(\mathbf{PC_j^TAPC_j} - 1)(\mathbf{y^TPC_jPC_j^Ty} - 1)}{\operatorname{Tr}\mathbf{A}^2 - 2\operatorname{Tr}\mathbf{A} + N - \sum_{j=1}^{k}(\mathbf{PC_j^TAPC_j} - 1)^2} \\ &= \frac{\operatorname{Tr}(\mathbf{A}-\mathbf{I})(\mathbf{I}-\mathbf{P})(\mathbf{yy^T}-\mathbf{I})(\mathbf{I}-\mathbf{P})}{\operatorname{Tr}(\mathbf{A}-\mathbf{I})(\mathbf{I}-\mathbf{P})(\mathbf{A}-\mathbf{I})(\mathbf{I}-\mathbf{P})} \end{aligned} \tag{17}$$

We have, $E(\mathbf{yy}^T) = h^2\mathbf{A}_{True} + (1-h^2)\mathbf{I}$ where $\mathbf{A}_{True}$ is the true GRM (where the standardisation of the genotype matrix has been done based on subclusters). Since, we do not know the subclusters, we do not know $\mathbf{A}_{True}$ either. Denote the GRM $\mathbf{A}$ which we generally work with as $\mathbf{A}_{usual}$ (where the genotype matrix is standardised overall). also denote $\mathbf{H} = \mathbf{I} - \mathbf{P}$. Under these notations the expression of heritability in Equation 17 above can be written as,

$$\hat{h}^2_{\text{SAdj-HE}} = \frac{\text{Tr}\,(\mathbf{A}_{usual} - \mathbf{I})\mathbf{H}(\mathbf{yy^T} - \mathbf{I})\mathbf{H}}{\text{Tr}\,(\mathbf{A}_{usual} - \mathbf{I})\mathbf{H}(\mathbf{A}_{usual} - \mathbf{I})\mathbf{H}}$$

and its expectation can be obtained as,

$$\begin{aligned} E(\hat{h}^2_{\text{SAdj-HE}}) &= \frac{\text{Tr}\,(\mathbf{A}_{usual} - \mathbf{I})\mathbf{H}(\mathbf{E}(\mathbf{yy^T}) - \mathbf{I})\mathbf{H}}{\text{Tr}\,(\mathbf{A}_{usual} - \mathbf{I})\mathbf{H}(\mathbf{A}_{usual} - \mathbf{I})\mathbf{H}} \\ &= \frac{\text{Tr}\,(\mathbf{A}_{usual} - \mathbf{I})\mathbf{H}(h^2\mathbf{A}_{True} + (1-h^2)\mathbf{I} - \mathbf{I})\mathbf{H}}{\text{Tr}\,(\mathbf{A}_{usual} - \mathbf{I})\mathbf{H}(\mathbf{A}_{usual} - \mathbf{I})\mathbf{H}} \\ &= \frac{\text{Tr}\,(\mathbf{A}_{usual} - \mathbf{I})(h^2\mathbf{H}\mathbf{A}_{True}\mathbf{H} - h^2\mathbf{H})}{\text{Tr}\,(\mathbf{A}_{usual} - \mathbf{I})\mathbf{H}(\mathbf{A}_{usual} - \mathbf{I})\mathbf{H}} \quad (\text{since}, \quad \mathbf{H}^2 = \mathbf{H}) \\ &= h^2 \frac{\text{Tr}\,(\mathbf{A}_{usual} - \mathbf{I})(\mathbf{H}\mathbf{A}_{True}\mathbf{H} - \mathbf{H})}{\text{Tr}\,(\mathbf{A}_{usual} - \mathbf{I})\mathbf{H}(\mathbf{A}_{usual} - \mathbf{I})\mathbf{H}} \end{aligned} \tag{18}$$

B.1 Theory with clusters

Suppose there are 2 clusters of individuals, i.e., $k = 1, 2$. Let the allele frequencies of s-th SNP for i-th individual from cluster k be p_{ks}. The total number of individuals is $N = n_1 + n_2$. Let $p_s = r_1 p_{1s} + r_2 p_{2s}$ with $r_k = \frac{n_k}{N}$. We make an assumption that $\sqrt{2p_{1s}(1 - p_{1s})} \approx \sqrt{2p_{2s}(1 - p_{2s})} \approx \sqrt{2p_s(1 - p_s)} = m_s$ (this is a reasonable assumption since even if p_{1s}, p_{2s} are much different $\sqrt{2p_{ks}(1 - p_{ks})}$'s are not, see Figure 5).

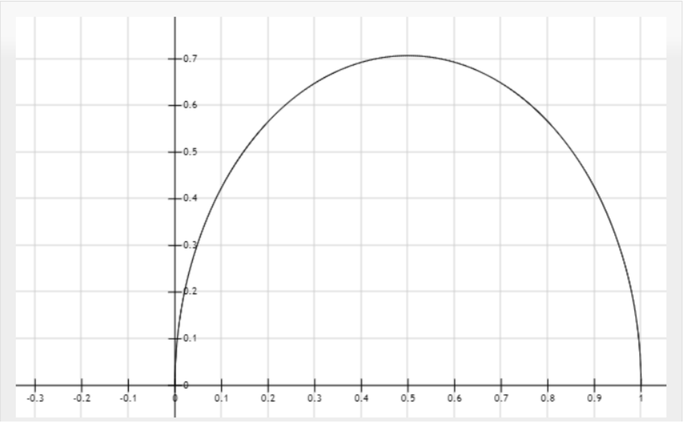


Figure 5. See how the variance function behaves for varying p, the function $\sqrt{2p(1 - p)}$ takes pretty close values even when p takes highly different values

The corresponding raw(unscaled) genotype matrix be $\mathbf{X} = \begin{bmatrix}\mathbf{X}_1^T & \mathbf{X}_2^T\end{bmatrix}^T$ where $\mathbf{X}_k$ is of dimension $n_k \times P$. Two following ways of standardizing elements of $\mathbf{X}$: for an individual i in cluster k,

$$\begin{aligned}
z^*_{kis} &= \frac{x_{is} - 2p_{ks}}{\sqrt{2p_{ks}(1-p_{ks})}} = \frac{x_{is} - 2p_{ks}}{m_s}, \quad \text{(True way)} \\
z_{is} &= \frac{x_{is} - 2p_s}{\sqrt{2p_s(1-p_s)}} = \frac{x_{ij} - 2p_s}{m_s} = z^*_{kis} + a_{ks}, \quad \text{(Usual way)} \\
a_{ks} &= \frac{2(p_{ks} - p_s)}{m_s}, \quad p_s = \frac{1}{N}(n_1 p_{1s} + n_2 p_{2s})
\end{aligned}$$

Thus, $\mathbf{Z}_k^* = \mathbf{Z}_k - \mathbf{a}_k$ with $\mathbf{a}_k = \begin{bmatrix} a_{k1}\mathbf{e}_{n_k} \ldots a_{kP}\mathbf{e}_{n_k}\end{bmatrix}$. Or, $\mathbf{Z}^* = \mathbf{Z} - \mathbf{a}$ with $\mathbf{a} = \begin{bmatrix}\mathbf{a}_1^T & \mathbf{a}_2^T\end{bmatrix}^T$. True GRM and usual GRM can be written as,

$$\mathbf{A}_{usual} = \frac{1}{P}\mathbf{Z}\mathbf{Z}^T; \mathbf{A}_{True} = \frac{1}{P}\mathbf{Z}^*\mathbf{Z}^{*T} = \mathbf{A}_{usual} - \frac{1}{P}(\mathbf{a}\mathbf{Z}^T + \mathbf{Z}\mathbf{a}^T) + \frac{1}{P}\mathbf{a}\mathbf{a}^T \tag{19}$$

With 2 distinct population subclusters, the first PC of $\mathbf{A}_{usual}$ would be a vector with v_1 for n_1 individuals from cluster 1 and $-v_2$ for n_2 individuals from cluster 2 with $v_k = \frac{1}{n_k}\sqrt{\frac{n_1 n_2}{N}}$ (Patterson et al., 2006; Galinsky et al., 2016). More formally, $\mathbf{PC}_1 = \begin{bmatrix} v_1\mathbf{e}_{n_1}^T & -v_2\mathbf{e}_{n_2}^T\end{bmatrix}^T$. We can write,

$$\mathbf{H}_1 = \mathbf{I}_N - \mathbf{PC}_1\mathbf{PC}_1^T = \mathbf{I}_N - \begin{bmatrix} v_1^2\mathbf{J}_{n_1,n_1} & -v_1v_2\mathbf{J}_{n_1,n_2} \\ -v_1v_2\mathbf{J}_{n_2,n_1} & v_2^2\mathbf{J}_{n_2,n_2}\end{bmatrix}$$

For the s-th column of $\mathbf{a}$ matrix,

$$\mathbf{H}_1 a_{.,s} = a_{.,s} - \begin{bmatrix} v_1^2\mathbf{J}_{n_1,n_1} & -v_1v_2\mathbf{J}_{n_1,n_2} \\ -v_1v_2\mathbf{J}_{n_2,n_1} & v_2^2\mathbf{J}_{n_2,n_2}\end{bmatrix}\begin{bmatrix} a_{1s}\mathbf{e}_{n_1} \\ a_{2s}\mathbf{e}_{n_2}\end{bmatrix}$$

$$
\begin{aligned}
&= a_{.,s} - \begin{bmatrix} (a_{1s}v_1^2 n_1 - a_{2s}v_1v_2n_2)\mathbf{e}_{n_1} \\ (-a_{1s}v_1v_2n_1 + a_{2s}v_2^2n_2)\mathbf{e}_{n_2} \end{bmatrix} \\
&= \begin{bmatrix} \left(a_{1s} - a_{1s}v_1^2 n_1 + a_{2s}v_1v_2n_2\right)\mathbf{e}_{n_1} \\ (a_{2s} + a_{1s}v_1v_2n_1 - a_{2s}v_2^2n_2)\mathbf{e}_{n_2} \end{bmatrix} \\
&= \begin{bmatrix} \left(a_{1s} - a_{1s}\frac{n_2}{N} + a_{2s}\frac{n_2}{N}\right)\mathbf{e}_{n_1} \\ \left(a_{2s} + a_{1s}\frac{n_1}{N} - a_{2s}\frac{n_1}{N}\right)\mathbf{e}_{n_2} \end{bmatrix} \\
&= \begin{bmatrix} \left(a_{1s}\frac{n_1}{N} + a_{2s}\frac{n_2}{N}\right)\mathbf{e}_{n_1} \\ \left(a_{2s}\frac{n_2}{N} + a_{1s}\frac{n_1}{N}\right)\mathbf{e}_{n_2} \end{bmatrix} \\
&= \begin{bmatrix} \left(\frac{2(p_{1s}-p_s)}{m_s}\frac{n_1}{N} + \frac{2(p_{2s}-p_s)}{m_s}\frac{n_2}{N}\right)\mathbf{e}_{n_1} \\ \left(\frac{2(p_{2s}-p_s)}{m_s}\frac{n_2}{N} + \frac{2(p_{1s}-p_s)}{m_s}\frac{n_1}{N}\right)\mathbf{e}_{n_2} \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix} \qquad \left(\text{plugging in } p_s = \frac{1}{N}(n_1p_{1s} + n_2p_{2s})\right)
\end{aligned}
$$

Thus, we find that premultiplying $\mathbf{a}$ by $\mathbf{H}_1$ gives $\mathbf{H}_1\mathbf{a} = \mathbf{0}$. Using this fact in the expressions of equation (2) we get,

$$
\mathbf{H}_1\mathbf{A}_{True}\mathbf{H}_1 = \mathbf{H}_1\left(\mathbf{A}_{usual} - \frac{1}{P}(\mathbf{a}\mathbf{Z}^T + \mathbf{Z}\mathbf{a}^T) + \frac{1}{P}\mathbf{a}\mathbf{a}^T\right)\mathbf{H}_1 = \mathbf{H}_1\mathbf{A}_{usual}\mathbf{H}_1. \tag{20}
$$

Plugging $\mathbf{H} = \mathbf{H}_1$ in equation (1) we get,

$$
\begin{aligned}
E(\hat{h}^2_{\text{SAdj-HE}}) &= h^2\frac{\operatorname{Tr}(\mathbf{A}_{usual} - \mathbf{I})(\mathbf{H}_1\mathbf{A}_{True}\mathbf{H}_1 - \mathbf{H}_1)}{\operatorname{Tr}(\mathbf{A}_{usual} - \mathbf{I})\mathbf{H}_1(\mathbf{A}_{usual} - \mathbf{I})\mathbf{H}_1} \\
&= h^2\frac{\operatorname{Tr}(\mathbf{A}_{usual} - \mathbf{I})(\mathbf{H}_1\mathbf{A}_{usual}\mathbf{H}_1 - \mathbf{H}_1)}{\operatorname{Tr}(\mathbf{A}_{usual} - \mathbf{I})\mathbf{H}_1(\mathbf{A}_{usual} - \mathbf{I})\mathbf{H}_1} \\
&= h^2\frac{\operatorname{Tr}(\mathbf{A}_{usual} - \mathbf{I})\mathbf{H}_1(\mathbf{A}_{usual} - \mathbf{I})\mathbf{H}_1}{\operatorname{Tr}(\mathbf{A}_{usual} - \mathbf{I})\mathbf{H}_1(\mathbf{A}_{usual} - \mathbf{I})\mathbf{H}_1} \\
&= h^2
\end{aligned}
$$

Thus, it shows why Haseman Elston regression with $\mathbf{H}_1\mathbf{A}_{usual}\mathbf{H}_1^T$ (MMHE) or our proposed Haseman Elston regression with the first PC product adjustment would give us asymptotically unbiased estimate of heritability. When there are more clusters, more PCs would be needed to be considered.