Kinship estimation bias carries over to heritability estimation bias using variance components

Zhuoran Hou[1], Alejandro Ochoa[1,2,*]

[1] Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27705, USA
[2] Duke Center for Statistical Genetics and Genomics, Duke University, Durham, NC 27705, USA
[*] Corresponding author: `alejandro.ochoa@duke.edu`

Abstract

Heritability is a fundamental parameter of diseases and other traits, quantifying the contribution of genetics to that trait as opposed to non-genetic environmental factors. Heritability is reflected in the extent to which relatives have similar phenotypes. The estimation of kinship matrices, also known as Genetic Relatedness Matrices or "GRMs", is required for heritability estimation with many common approaches based on variant components. However, the most common "standard" kinship estimator employed by these approaches, including the popular GCTA package, can be severely biased in structured populations. In this study, we first characterize the theoretically predicted heritability estimation bias in the GCTA model due to kinship bias. Heritability estimation in this model requires unbiased estimates of the random effect coefficient, which we find is biased when the standard kinship estimator is used, and this bias depends only on the mean kinship value and the true heritability. We conduct simulation studies to evaluate heritability estimation with various kinship matrices for scenarios such as admixture structure only and admixture plus family structure. The simulation results validate our theoretical prediction of the bias of the standard GCTA approach and show that our unbiased kinship matrices result in unbiased heritability estimates. A further complication is that upweighting rare variants in these kinship estimates, which is common practice, introduces an additional bias that has not been characterized in closed form. Then we apply various kinship matrices to several real datasets, such as the San Antonio Family Study and a Nephrotic Syndrome multiethnic cohort, to further characterize the source of biases and illustrate their extent in practice. We find that the heritability estimations using our unbiased kinship matrices lead to higher values, which are more consistent with previously published estimates in unstructured populations. Overall, we find that the most commonly used kinship estimator downwardly bias heritability estimation when there is population structure, and using an unbiased kinship estimator addresses this source of bias.

1 Introduction

Heritability is an important parameter of diseases and other traits, quantifying the contribution of genetics to that trait as opposed to non-genetic environmental factors (Lush et al., 1949). Heritability is reflected in the extent to which relatives have similar phenotypes (Visscher et al., 2008). In addition, heritability is closely related to the polygenic risk score (PRS) method, an estimate of genetic liability to a trait for individuals, as it defines the upper bound of the performance of PRS (Choi et al., 2020).

Heritability has long been estimated from close relatives, such as twins or siblings (Falconer, 1996), or more complex pedigrees (Almasy and Blangero, 1998). The variance component model of the SOLAR approach, which is based on estimating kinship from pedigrees, enables the use of more distant relatives (Almasy and Blangero, 1998). GCTA extended the last approach to population data, which was used to demonstrate that SNPs likely explain the majority of missing heritability for height (Yang et al., 2010; Yang et al., 2011). Alternative approaches for estimating heritability, such as LD Score Regression (Bulik-Sullivan et al., 2015; Luo et al., 2021) and SumHer (Speed and Balding, 2019), work with GWAS summary data and enable partitioning heritability within gene sets. We will consider LDSC in Aim 3, but not here since it is not based on kinship matrices and LMMs.

Accurate kinship estimation is crucial for heritability estimation based on LMMs, such as GCTA. However, the most common kinship estimator employed by these approaches can be severely biased in structured populations (Ochoa and Storey, 2021). In Aim 1 we showed, empirically and theoretically, that association tests are invariant to the use of common biased kinship estimators compared with an unbiased estimator (Hou and Ochoa, 2023). However, heritability estimation requires unbiased estimates of the random effect coefficient, which is biased when the standard kinship estimator is used.

In this study, we first characterize the theoretically predicted heritability estimation bias due to kinship bias, following the derivation in Aim 1. We conduct simulation studies to evaluate heritability estimation with various kinship matrices for scenarios such as admixture structure only and admixture plus family structure. Then we apply various kinship matrices to several real datasets to further characterize the source of biases and their extent.

2 Methods

2.1 Genetic model

Suppose that there are m biallelic loci and n diploid individuals. The genotype $x_{ij} \in \{0, 1, 2\}$ at a locus i of the individual j is encoded as the number of reference alleles, for a pre-selected but otherwise arbitrary reference allele per locus. φ_{jk} is the kinship coefficient of two individuals j and k, and p_i is the ancestral allele frequency at locus i, then under the kinship model (Malécot, 1948;

Wright, 1949; Jacquard, 1970; Astle and Balding, 2009; Ochoa and Storey, 2021) the expectation and covariance are given by

$$\mathrm{E}\left[\mathbf{x}_i\right] = 2p_i\mathbf{1}, \qquad \mathrm{Cov}\left(\mathbf{x}_i\right) = 4p_i\left(1-p_i\right)\mathbf{\Phi}, \tag{1}$$

where $\mathbf{x}_i = (x_{ij})$ is the vector of genotypes at locus i, $\mathbf{\Phi} = (\varphi_{jk})$ is the $n \times n$ kinship matrix, and $\mathbf{1}$ is a vector of ones.

The quantitative trait vector $\mathbf{y}$ for all individuals is assumed to follow a linear polygenic model,

$$\mathbf{y} = \mathbf{1}_n\alpha + \mathbf{X}'\beta + \epsilon, \tag{2}$$

where α is the intercept, $\beta = (\beta_i)$ is a vector of genetic effect coefficients for each locus i, and ϵ is a vector of non-genetic effects.

Let us shift the mean of genotypes to the intercept and denote $\mathbf{s} = \mathbf{X}'\beta$, then:

$$\begin{gathered}\mathbf{y} = \mathbf{1}_n\alpha + \mathbf{s} + \epsilon, \\ \mathbf{s} \sim \mathrm{Normal}\left(\mathbf{0}, 2\sigma_g^2\mathbf{\Phi}\right), \quad \epsilon \sim \mathrm{Normal}\left(\mathbf{0}, \sigma_e^2\mathbf{I}_n\right), \mathbf{s} + \epsilon \sim \mathrm{Normal}\left(\mathbf{0}, 2\sigma_g^2\mathbf{\Phi} + \sigma_e^2\mathbf{I}_n\right).\end{gathered}$$

To analyze the covariance structure of the trait, we treat α and β as fixed parameters, while $\mathbf{X}$ and ϵ are random. The non-genetic effects are assumed to be independent with variance $(1-h^2)\sigma^2$ given by the total trait variance scale σ^2 and the narrow-sense heritability h^2:

$$\begin{gathered}\mathrm{E}[\epsilon] = \mathbf{0}_n, \qquad \mathrm{Cov}(\epsilon) = \sigma_e^2\mathbf{I}_n, \\ \mathbf{X}'\beta + \epsilon \sim \mathrm{Normal}\left(\mathbf{0}, 2\sigma_g^2\mathbf{\Phi} + \sigma_e^2\mathbf{I}_n\right).\end{gathered}$$

Then the narrow-sense heritability h^2 is defined as:

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}.$$

2.2 Kinship estimation

Each estimator bias type has two locus weight types called *ratio-of-means* (ROM) and *mean-of-ratios* (MOR) (Bhatia et al., 2013; Ochoa and Storey, 2021). Only ROM estimators have closed-form limits. Below $\hat{p}_i = \frac{1}{2n}\mathbf{x}_i'\mathbf{1}$ is the standard ancestral allele frequency estimator, where the prime ($'$) denotes matrix transposition, and $\hat{\mathbf{\Phi}}^{\mathrm{name}} = (\hat{\varphi}_{jk}^{\mathrm{name}})$ relates the scalar and matrix formulas of each named kinship estimator.

2.2.1 Standard kinship estimator

The “standard” kinship estimator is the most widely used estimator in various applications of population structure (Astle and Balding, 2009; Speed and Balding, 2015; Wang et al., 2017), including heritability estimation (Speed et al., 2012; Speed and Balding, 2015; Speed et al., 2017) and genetic association tests based on PCA (Price et al., 2006), LMM (Astle and Balding, 2009; Zhou and Stephens, 2012; Loh et al., 2015; Sul et al., 2018), and other models (Rakovski and Stram, 2009; Thornton and McPeek, 2010). The popular heritability estimation approach GCTA (Yang et al., 2010; Yang et al., 2011) employs a variant of this estimator that is detailed in the next paragraph.

The ROM and MOR versions of the standard kinship estimator are, respectively,

$$\hat{\varphi}_{jk}^{\text{std-ROM}} = \frac{\sum_{i=1}^{m} \left(x_{ij} - 2\hat{p}_i\right)\left(x_{ik} - 2\hat{p}_i\right)}{\sum_{i=1}^{m} 4\hat{p}_i\left(1 - \hat{p}_i\right)}, \tag{3}$$

$$\hat{\varphi}_{jk}^{\text{std-MOR}} = \frac{1}{m}\sum_{i=1}^{m} \frac{\left(x_{ij} - 2\hat{p}_i\right)\left(x_{ik} - 2\hat{p}_i\right)}{4\hat{p}_i\left(1 - \hat{p}_i\right)}. \tag{4}$$

The ROM estimator has a biased limit, which is a function of the true kinship matrix (Ochoa and Storey, 2021):

$$\hat{\mathbf{\Phi}}^{\text{std-ROM}} \xrightarrow[m \to \infty]{\text{a.s.}} \frac{1}{1 - \bar{\varphi}}\left(\mathbf{\Phi} + \bar{\varphi}\mathbf{J} - \boldsymbol{\varphi}\mathbf{1}' - \mathbf{1}\boldsymbol{\varphi}'\right), \tag{5}$$

where $\mathbf{J} = \mathbf{1}\mathbf{1}'$ is the $n \times n$ matrix of ones, $\boldsymbol{\varphi} = \frac{1}{n}\mathbf{\Phi}\mathbf{1}$ is a vector of per-row mean kinship values, and $\bar{\varphi} = \frac{1}{n^2}\mathbf{1}'\mathbf{\Phi}\mathbf{1}$ is the scalar overall mean kinship. The MOR estimator does not have a closed-form limit, but it is well approximated by Eq. (5) in practice, especially when loci with small minor allele frequencies are excluded prior to calculating this estimate.

2.2.2 Popkin kinship estimator

The popkin (population kinship) estimator (Ochoa and Storey, 2021), generalized here to include locus weights w_i, is given by

$$\hat{\varphi}_{jk}^{\text{popkin}} = 1 - \frac{A_{jk}}{\hat{A}_{\text{min}}}, \qquad A_{jk} = \frac{1}{m}\sum_{i=1}^{m} w_i((x_{ij} - 1)(x_{ik} - 1) - 1), \tag{6}$$

where in this work $\hat{A}_{\text{min}} = \min_{j \neq k} A_{jk}$, and w_i must be positive but need not add to 1. We consider two broad forms for this estimator. The original ROM estimator has $w_i = 1$ and has an unbiased almost sure limit as the number of loci m goes to infinity,

$$\hat{\mathbf{\Phi}}^{\text{popkin-ROM}} \xrightarrow[m \to \infty]{\text{a.s.}} \mathbf{\Phi},$$

under the assumption that the true minimum kinship is zero. The MOR version, introduced here, outweighs rare variants using $w_i = \left(\hat{p}_i\left(1 - \hat{p}_i\right)\right)^{-1}$; although it does not have a closed-form limit, it is approximately unbiased as well and is connected to the most common estimator, the standard MOR (Hou and Ochoa, 2023).

2.3 Heritability estimation bias due to kinship bias

In our recent work, we characterized the theoretical relationship between variance component estimates of a biased kinship matrix and its unbiased counterpart (Hou and Ochoa, 2023). For standard ROM estimator,

$$\hat{\mathbf{\Phi}}^{\text{std-ROM}} = \frac{1}{1-\bar{\varphi}} \mathbf{C}\mathbf{\Phi}\mathbf{C},$$

where $\mathbf{C} = \mathbf{I} - \frac{1}{n}\mathbf{J}$ is the centering matrix. Let the estimates derived from the unbiased kinship matrix be $\hat{\sigma}_g^2, \hat{\sigma}_e^2$, so that the heritability estimate is

$$\hat{h}^2 = \frac{\hat{\sigma}_g^2}{\hat{\sigma}_g^2 + \hat{\sigma}_e^2}.$$

Now, denote the estimates derived from the biased kinship matrix as $\hat{\sigma}_g^{2,\text{biased}}, \hat{\sigma}_e^{2,\text{biased}}$. Then, the model based on the biased kinship matrix is:

$$\begin{aligned}
\mathbf{y} &= \mathbf{1}_n \alpha^{\text{biased}} + \mathbf{s}^{\text{biased}} + \epsilon, \\
\mathbf{s}^{\text{biased}} &= \mathbf{C}\mathbf{s} \sim \text{Normal}\left(\mathbf{0}, 2\hat{\sigma}_g^{2,\text{biased}} \hat{\mathbf{\Phi}}^{\text{std-ROM}}\right), \\
\mathbf{s}^{\text{biased}} &= \mathbf{s} - \mathbf{1}_n \bar{\mathbf{s}}, \quad \alpha^{\text{biased}} = \alpha + \bar{\mathbf{s}}, \quad \bar{\mathbf{s}} \sim \text{Normal}\left(0, \hat{\sigma}_g^2 \bar{\varphi}\right),
\end{aligned}$$

and we found that the relationship between these estimates is algebraically given by

$$\begin{aligned}
\hat{\sigma}_g^{2,\text{biased}} &= (1-\bar{\varphi})\hat{\sigma}_g^2, \\
\hat{\sigma}_e^{2,\text{biased}} &= \hat{\sigma}_e^2,
\end{aligned}$$

where $\bar{\varphi}$ is the mean value of the unbiased kinship matrix (Hou and Ochoa, 2023). Therefore, after noting that

$$\hat{\sigma}_e^2 = \left(\frac{1}{\hat{h}^2} - 1\right)\hat{\sigma}_g^2,$$

and substituting this and the values of the biased parameters, we find a form for the biased heritability estimate in terms of the unbiased estimate and the mean kinship:

$$\begin{aligned}
\hat{h}^{2,\text{biased}} &= \frac{\hat{\sigma}_g^{2,\text{biased}}}{\hat{\sigma}_g^{2,\text{biased}} + \hat{\sigma}_e^{2,\text{biased}}} \\
&= \frac{(1-\bar{\varphi})\hat{\sigma}_g^2}{(1-\bar{\varphi})\hat{\sigma}_g^2 + \left(\frac{1}{\hat{h}^2} - 1\right)\hat{\sigma}_g^2} \\
&= \hat{h}^2 \frac{1-\bar{\varphi}}{1-\bar{\varphi}\hat{h}^2}.
\end{aligned}$$

We further plot this relationship between the true heritability and biased estimates in Fig. 1. From this, we found that the bias becomes larger for the heritability ranging from 0.5 to 0.8 approximately while decreasing to 0 when the heritability is close to 0 or 1. In addition, the bias increases when $\bar{\varphi}$ increases.

2.4 Software

Kinship estimates based on the Popkin method are computed using the popkin R package, while standard MOR kinship estimates are obtained with GCTA (version 1.93.2beta). All other kinship estimators are calculated using the popkinsuppl R package. Plink (version 1.90 and 2.00a3LM) is used to process genotype data.

2.5 Simulations

To characterize the effect of kinship estimator bias in heritability estimation, we use simulated genotypes and traits and estimate kinship matrices from these genotypes. In particular, we consider the true kinship matrix of the simulation, the unbiased popkin estimator in Eq. (6), the standard kinship estimator in Eq. (3) and Eq. (4). Each scenario was replicated 50 times, in each case producing a new genotype matrix.

To further explore how rare variants affect kinship estimation and trait simulations, we use genotypes from the 1000 Genome project and simulate traits using different filters ($maf = \{0, 0.01, 0.05\}$), and estimate the kinship matrix using genotypes based on different filters ($maf = \{0, 0.01, 0.05\}$) as well.

2.5.1 Trait simulation algorithm

Suppose the genotype matrix $\mathbf{X}$ is available, and we have fixed values for the number of causal loci m_1, the trait mean, variance scale, and heritability (μ, σ^2, h^2). The goal is to choose the intercept α and draw random effect sizes β that result in the desired trait parameters. First we randomly select m_1 loci to be causal, and subset the genotype matrix $\mathbf{X}$ and ancestral allele frequency vector

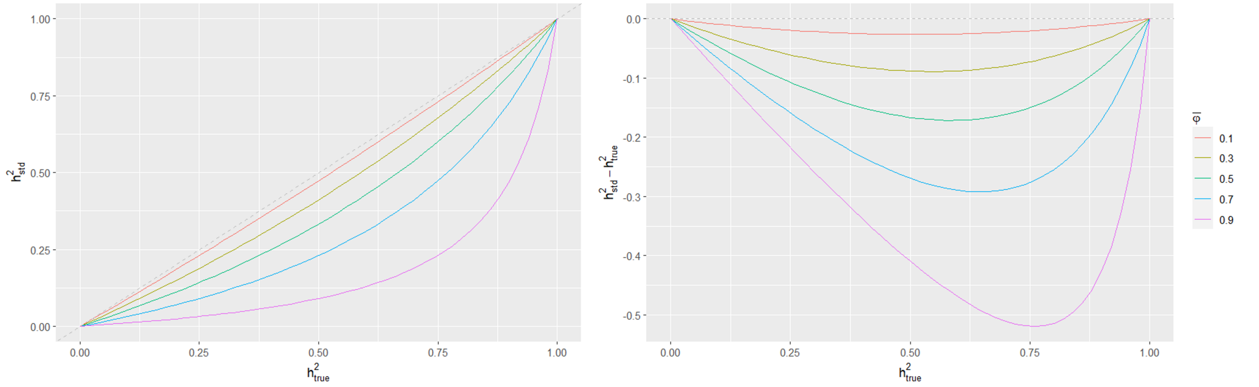

Figure 1: **Relationship between the true heritability and biased estimates.** The left one plots the relationship between biased estimates and the true heritability, while the right one plots the relationship between the bias and true heritability.

$\mathbf{p}$ so that from this point on they contain only those causal loci (they now have dimensions $m_1 \times n$ and length m_1, respectively).

Below we divide the algorithm into two steps: (1) scaling the effect sizes, and (2) centering the trait. Each step forks into two cases: whether the true ancestral allele frequencies $\mathbf{p}$ are known or not (the latter requires a known kinship matrix $\mathbf{\Phi}$).

Scaling effect sizes. The initial effect sizes β_i are drawn independently from a standard normal distribution:

$$\beta_i \sim \mathrm{N}(0, 1).$$

First we consider the simpler case of known ancestral allele frequencies $\mathbf{p} = (p_i)$. The initial genetic variance scale is

$$\sigma_0^2 = \sum_{i=1}^{m_1} 2p_i(1-p_i)\beta_i^2.$$

We obtain the desired variance by dividing each β_i by σ_0 (which results in a variance of 1) and then multiply by $h\sigma$ (which results in the desired variance of $h^2\sigma^2$). Combining both steps, the update is

$$\beta \leftarrow \beta \frac{h\sigma}{\sigma_0}.$$

Now we consider the case of unknown ancestral allele frequencies but known kinship matrix. First, sample estimates $\hat{\mathbf{p}} = (\hat{p}_i)$ of the ancestral allele frequencies are constructed from the genotype data as

$$\hat{p}_i = \frac{1}{2n} \mathbf{1}_n^\intercal \mathbf{x}_i.$$

Although this estimator is unbiased ($\mathrm{E}[\hat{\mathbf{p}}] = \mathbf{p}$), the resulting variance estimates of interest $\hat{p}_i \left(1 - \hat{p}_i\right)$ are downwardly biased (Ochoa and Storey, 2021):

$$\mathrm{E}\left[\hat{p}_i \left(1 - \hat{p}_i\right)\right] = p_i(1-p_i)(1-\bar{\varphi}),$$

where $\bar{\varphi} = \frac{1}{n^2} \mathbf{1}_n^\intercal \mathbf{\Phi} \mathbf{1}_n$ is the mean kinship coefficient in the data. Therefore the initial genetic variance scale, estimated as

$$\hat{\sigma}_0^2 = \sum_{i=1}^{m_1} 2\hat{p}_i(1-\hat{p}_i)\beta_i^2,$$

has an expectation of

$$\mathrm{E}\left[\hat{\sigma}_0^2\right] = \sigma_0^2(1-\bar{\varphi}).$$

Therefore, assuming that this additional factor $(1-\bar{\varphi})$ is known, the update

$$\beta \leftarrow \beta \frac{h\sigma\sqrt{1-\bar{\varphi}}}{\hat{\sigma}_0}$$

results in the desired variance.

Centering the trait. Here we consider the problem of selecting the intercept coefficient α that, together with the previous effect size coefficient vector β, result in the desired trait mean μ.

When ancestral allele frequencies are known, the trait can be centered precisely. Given our model, we obtain the desired overall trait mean μ by choosing the intercept coefficient to be

$$\alpha = \mu - 2\mathbf{p}^\intercal \beta.$$

When ancestral allele frequencies are unknown, the solution is to choose the intercept coefficient

$$\alpha = \mu - 2\hat{\bar{p}} \mathbf{1}_{m_1}^\intercal \beta, \qquad \hat{\bar{p}} = \frac{1}{m_1} \mathbf{1}_{m_1}^\intercal \hat{\mathbf{p}} = \frac{1}{2m_1 n} \mathbf{1}_{m_1}^\intercal \mathbf{X}^\intercal \mathbf{1}_n = \frac{1}{2}\bar{X},$$

where $\mathbf{1}_{m_1}$ is a length-m_1 column vector of ones. Note that this overal mean allele frequency $\hat{\bar{p}}$ is computed among causal loci only. This works very well in practice since β is drawn randomly, so it is uncorrelated to $\mathbf{p}$ and therefore

$$\frac{1}{m_1}\mathbf{p}^\intercal \beta = \frac{1}{m_1} \sum_{i=1}^{m_1} p_i \beta_i \approx \left(\frac{1}{m_1} \sum_{i=1}^{m_1} p_i \right) \left(\frac{1}{m_1} \sum_{i=1}^{m_1} \beta_i \right) = \frac{1}{m_1} \bar{p} \mathbf{1}_{m_1}^\intercal \beta$$

is a good approximation.

Now we discuss why the more obvious naive approach, which would be to center the trait using estimated ancestral allele frequencies as $\alpha = \mu - 2\hat{\mathbf{p}}^\intercal \beta$, does not work. This approach is equivalent to centering genotypes at each locus as

$$\mathbf{y} = \alpha \mathbf{1}_n + \sum_{i=1}^{m_1} (\mathbf{x}_i - 2\hat{p}_i \mathbf{1}_n)\beta_i + \epsilon.$$

However, this operation introduces a distortion in the covariance of the genotypes (Ochoa and Storey, 2021):

$$\mathrm{Cov}\left(\mathbf{x}_i - 2\hat{p}_i \mathbf{1}_n\right) = p_i(1-p_i)\left(\mathbf{\Phi} + \bar{\varphi}\mathbf{1}_n\mathbf{1}_n^\intercal - \varphi \mathbf{1}_n^\intercal - \mathbf{1}_n \varphi^\intercal\right),$$

where $\bar{\varphi}$ is the overall mean kinship, as before, and $\varphi = \frac{1}{n}\mathbf{\Phi}\mathbf{1}_n$ is a length-n column vector of per-row mean kinship values. These undesireable distortions propagate to the trait, which we confirmed in simulations (not shown). Note that the intercept version we chose instead does not induce this genotype centering, which prevents the undesireable distortions in the trait covariance.

2.5.2 Admixture simulation for genotype matrices

An admixed family is simulated following previous work (Yao and Ochoa, 2022), except here only $K = 3$ ancestries are simulated and $F_{ST} = 0.3$ for the admixed individuals, which more closely resembles Hispanics and African Americans. Briefly, our admixture model first simulates $n = 1000$ founder individuals with $m = 100,000$ loci. Random ancestral allele frequencies p_i, subpopulation allele frequencies $p_i^{S_u}$, individual-specific allele frequencies π_{ij}, and genotypes x_{ij} are drawn from this hierarchical model:

$$
\begin{aligned}
p_i &\sim \text{Uniform}(0.01, 0.5),\\
p_i^{S_u}|p_i &\sim \text{Beta}\left(p_i\left(\frac{1}{f_{S_u}}-1\right),(1-p_i)\left(\frac{1}{f_{S_u}}-1\right)\right),\\
\pi_{ij} &= \sum_{u=1}^{K} q_{ju}p_i^{S_u},\\
x_{ij}|\pi_{ij} &\sim \text{Binomial}(2,\pi_{ij}),
\end{aligned}
$$

where this Beta is the Balding-Nichols distribution (Balding and Nichols, 1995) with mean p_i and variance $p_i\,(1-p_i)\,f_{S_u}$. We also include family structure in the simulation. 20 generations are generated iteratively.

2.6 Real data analysis

We utilize the high-coverage NYGC version of the 1000 Genomes Project (Fairley et al., 2020) filter using `plink2` (Chang et al., 2015). We retain only autosomal biallelic SNP loci marked with the filter "PASS". The final dataset consists of $m = 91,784,660$ loci and $n = 2,504$ individuals.

San Antonio Family Study (SAMAFS) is a complex pedigree-based study designed to identify low frequency or rare variants influencing susceptibility to T2D, conducted in 20 Mexican American T2D-enriched pedigrees from San Antonio, Texas (Mitchell et al., 1996). We only use exome chip data for the SAMAFS. We use PLINK (version 1.90b4.9) `-autosome` to exclude all unplaced and non-autosomal variants, and further filter using `-maf 0.01 -hwe 1e-10` (version 2.00a3LM). The final dataset consists of $m = 36,293$ loci and $n = 914$ individuals. We also include pedigree information for the estimation of kinship matrix for this data set. We adjust for age and sex in the heritability estimate.

Hispanic Community Health Study / Study of Latinos (HCHS/SOL) is a multi-center study in Hispanic/Latino populations recruited through four centers in Miami, San Diego, Chicago, and the Bronx New York (Sorlie et al., 2010). Similarly, we filter genotypes using PLINK (version 2.00a3LM) `-maf 0.01 -hwe 1e-10`. The final dataset consists of $m = 1,656,020$ loci and $n = 11,721$ individuals. We adjust for age and sex in the heritability estimate.

Nephrotic Syndrome (NS) multiethnic cohort is a multi-ancestry study exploring the etiology of nephrotic syndrome (Cason et al., 2023). We imputed the genotype and filter genotypes using PLINK (version 2.00a3LM) `-mac 20`. The final dataset consists of $m = 16,605,628$ loci and $n = 1,981$ individuals. We adjust for sex in the heritability estimate.

3 Results

3.1 Simulation results

First, we considered a setting where there is only population structure (absence of family structure). We found that only popkin ROM results in unbiased heritability estimation while all estimates based on other kinship estimators are downwardly biased Fig. 2. To further illustrate the trend of the biases, we plot the biases for different kinship estimators. The trend is similar to the theoretical results in Fig. 1. Then we consider a simulation including both population and family structure, the trend is similar to the first scenario while the variability of heritability estimation is much smaller.

For the simulation based on the 1000 Genomes Project, we set the expected heritability to 0.8. When the traits are simulated based on all loci including rare ones, heritability estimations are all underestimated using different kinship estimators Fig. 3. When the traits are simulated based on loci after filtering out variants with allele frequency lower than 0.01 or 0.05, only heritability estimations using Popkin ROM generate unbiased results, while estimates based on Standard MOR underestimate the heritability. When kinship matrices are estimated filtering out more rare variants (maf=0.05), the heritability estimations differences is quite small using different kinship estimators.

3.2 Real data applications

To further validate our results on real datasets, we estimate heritability using the popkin ROM estimator (the unbiased kinship estimator) and standard kinship MOR estimator (the commonly used biased kinship estimator) on three real datasets: San Antonio Family Study (SAMAFS), Hispanic Community Health Study / Study of Latinos (HCHS/SOL) and Nephrotic Syndrome (NS) multiethnic cohort.

For both SAMAFS and HCHS/SOL, heritabilities estimated from popkin ROM estimator are higher for most traits while there are some exceptions (Figs. 4 and 5). This is because our theoretical justification in 2.3 holds only for estimators that weigh loci the same way (both ROM or both MOR), while there is some additional bias that has not yet been characterized when we compare results from popkin ROM to standard MOR (Fig. 7).

For NS, we find that all heritabilities estimated from popkin ROM estimator are larger than the standard kinship MOR estimator Fig. 6. This is explained by the fact that this dataset has higher mean kinship values, which leads to larger biases as shown in Fig. 1.

4 Discussion

Previous research found that commonly used kinship estimators are biased, and these kinship biases will not affect association test statistics. In this project, we explored how kinship estimation bias affects heritability, and found the bias carried away to heritability estimation.

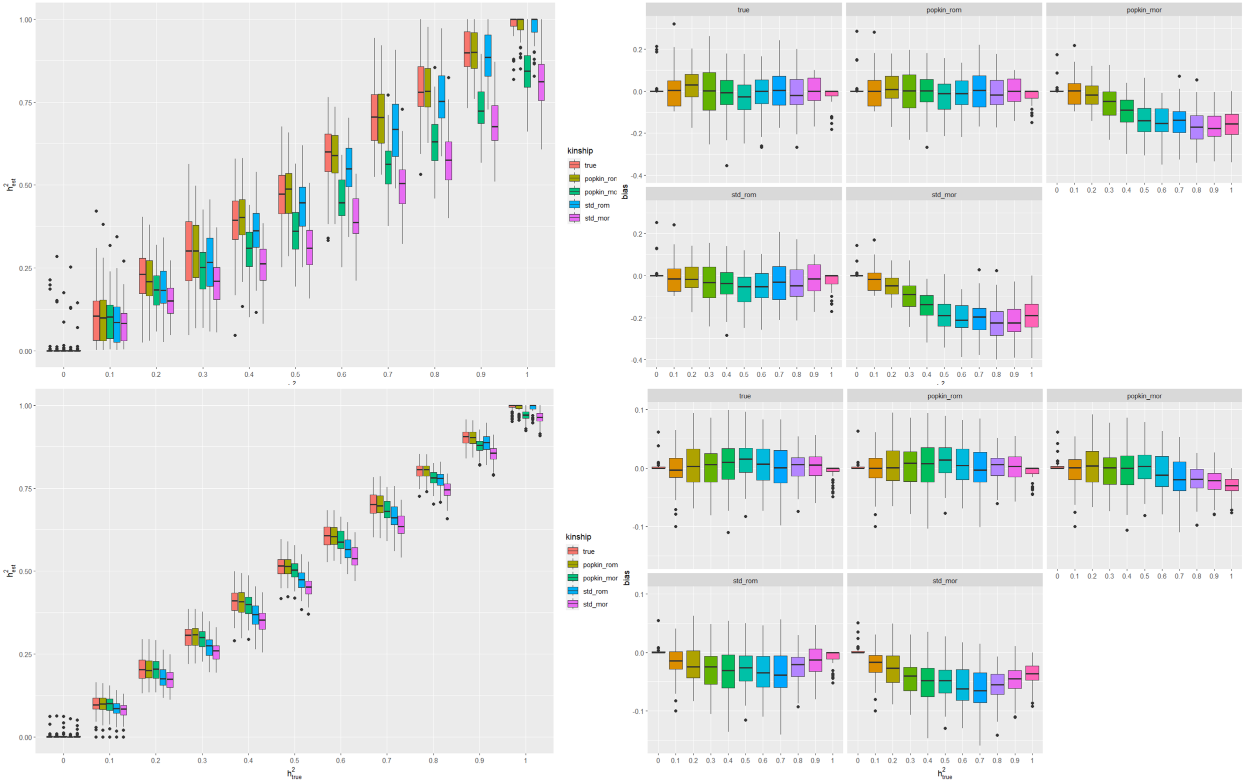

Figure 2: **Heritability estimation simulation by GCTA with various kinship matrices.** The upper and lower rows show simulation results for admixture structure only and admixture plus family structure, respectively.

...

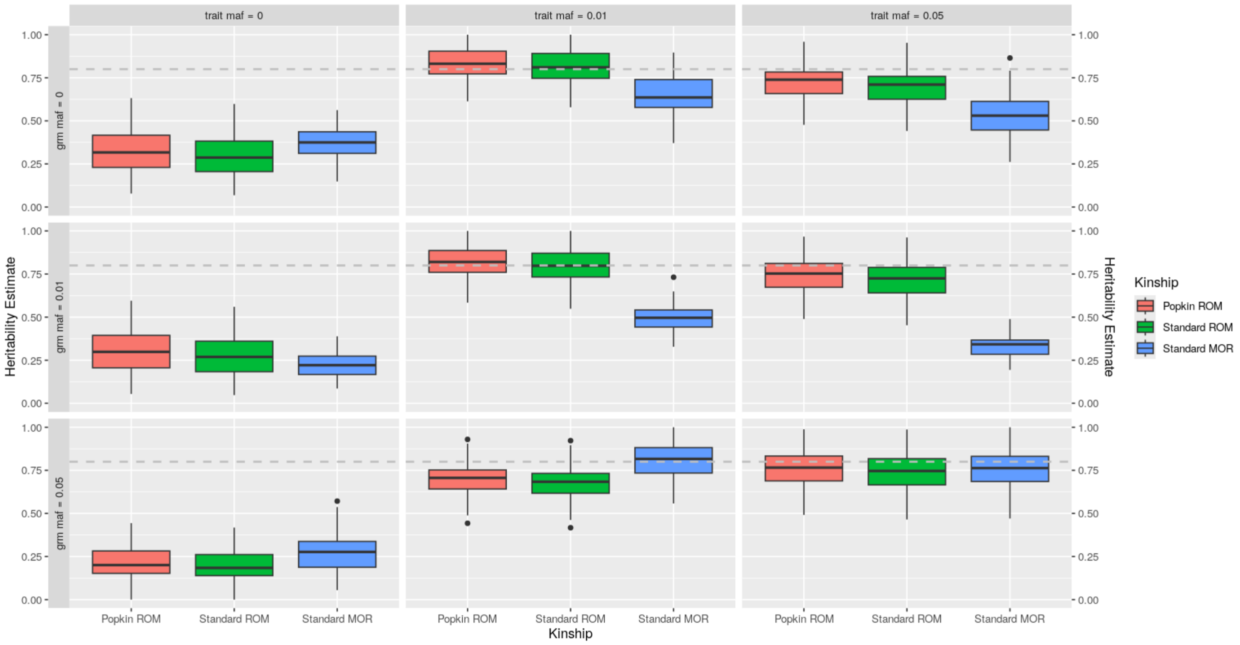

Figure 3: **Heritability estimation simulation based on 1000 Genomes Project with various kinship matrices using different rare variants filters.**

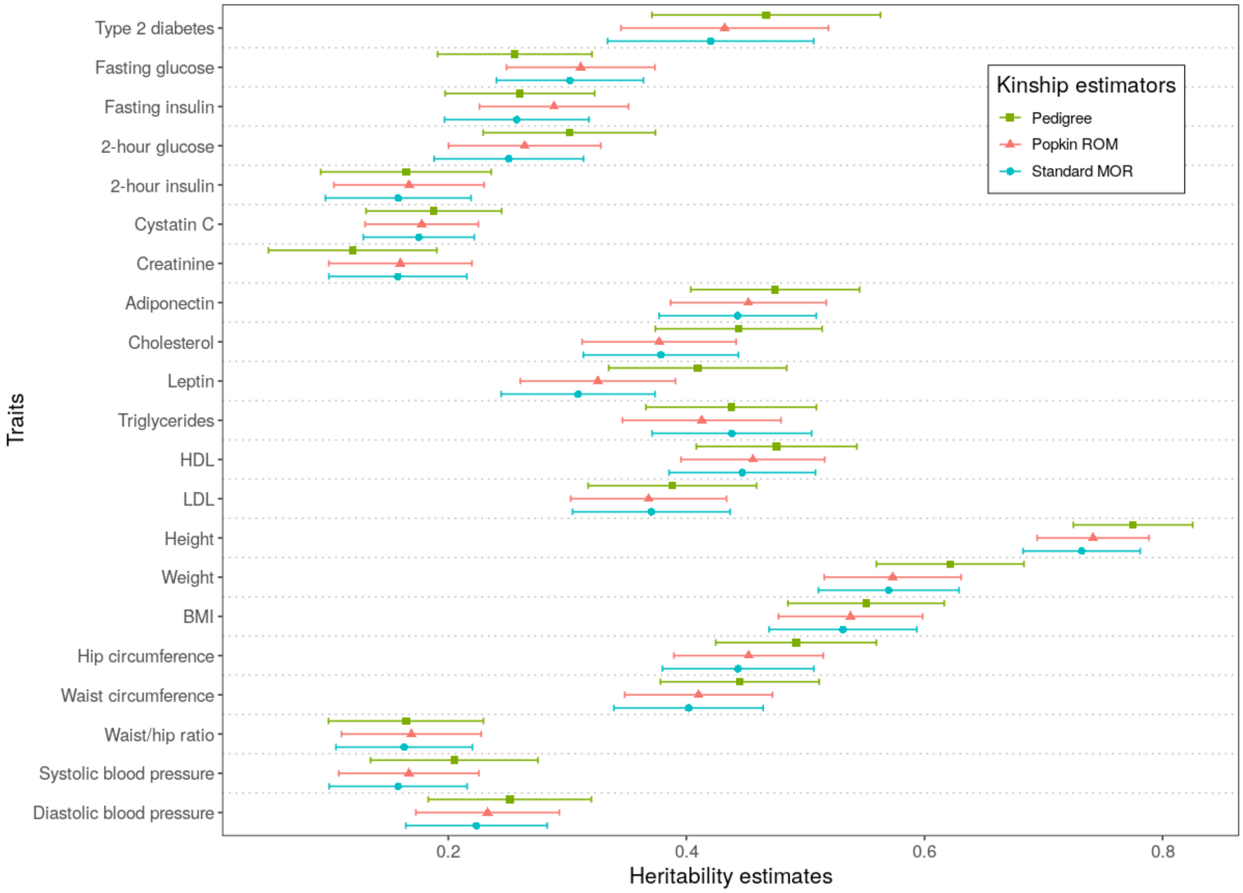

Figure 4: **Heritability estimation by standard kinship MOR estimator and popkin ROM estimator on real datasets-1.** The figure shows results from the San Antonio Family Study: Type 2 Diabetes (low mean kinship).

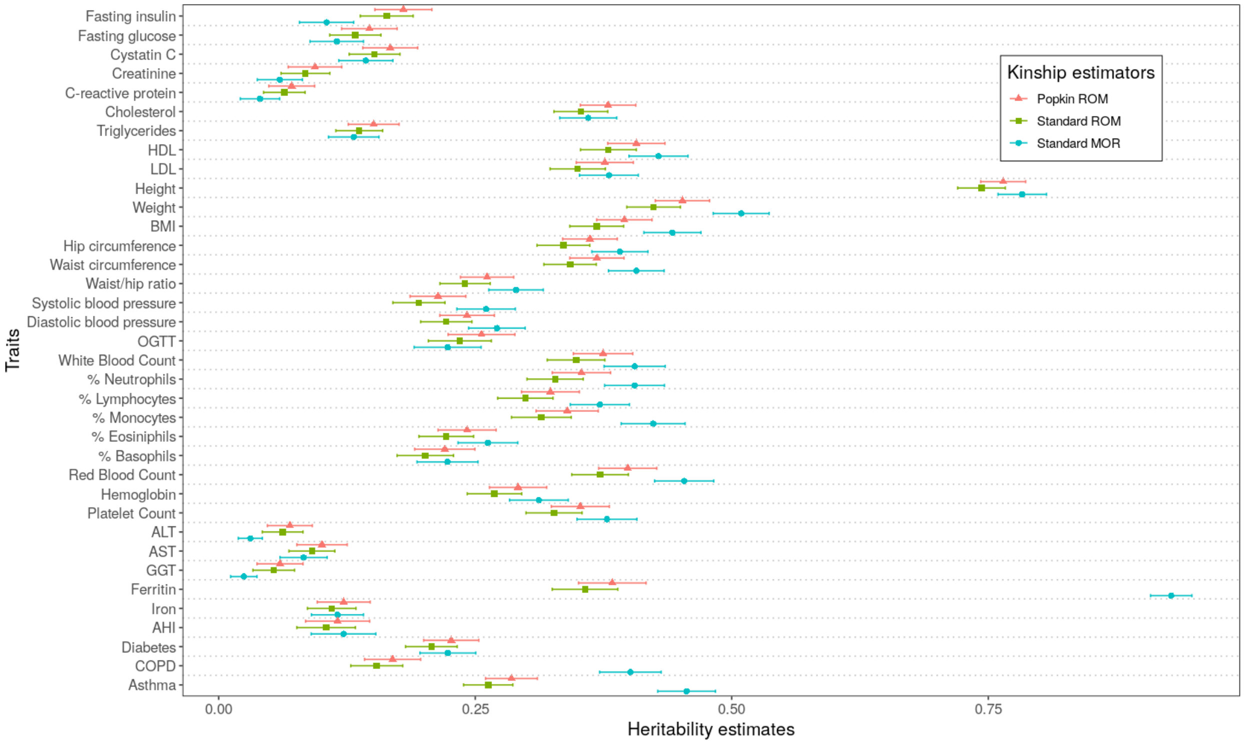

Figure 5: **Heritability estimation by standard kinship MOR estimator and popkin ROM estimator on real datasets-2.** The figure shows results from the Hispanic Community Health Study / Study of Latinos (middle mean kinship).

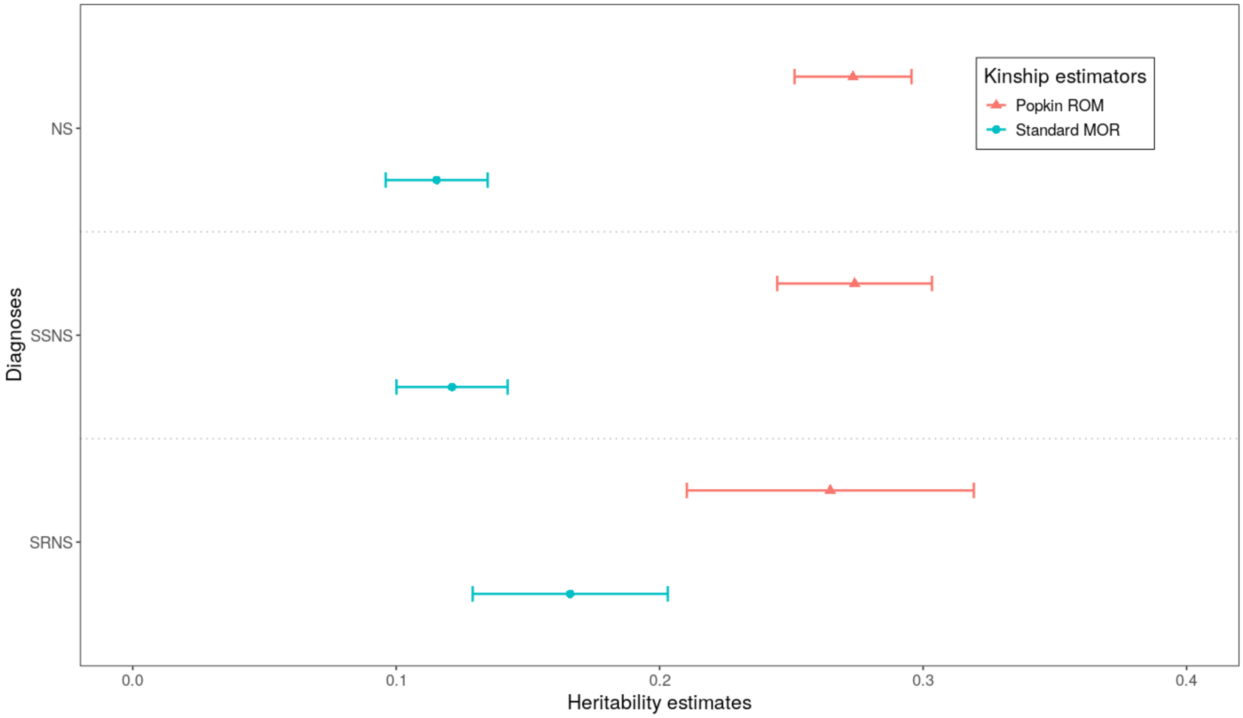

Figure 6: **Heritability estimation by standard kinship MOR estimator and popkin ROM estimator on real datasets-3.** The figure shows results from Nephrotic Syndrome: multiethnic cohort (high mean kinship).

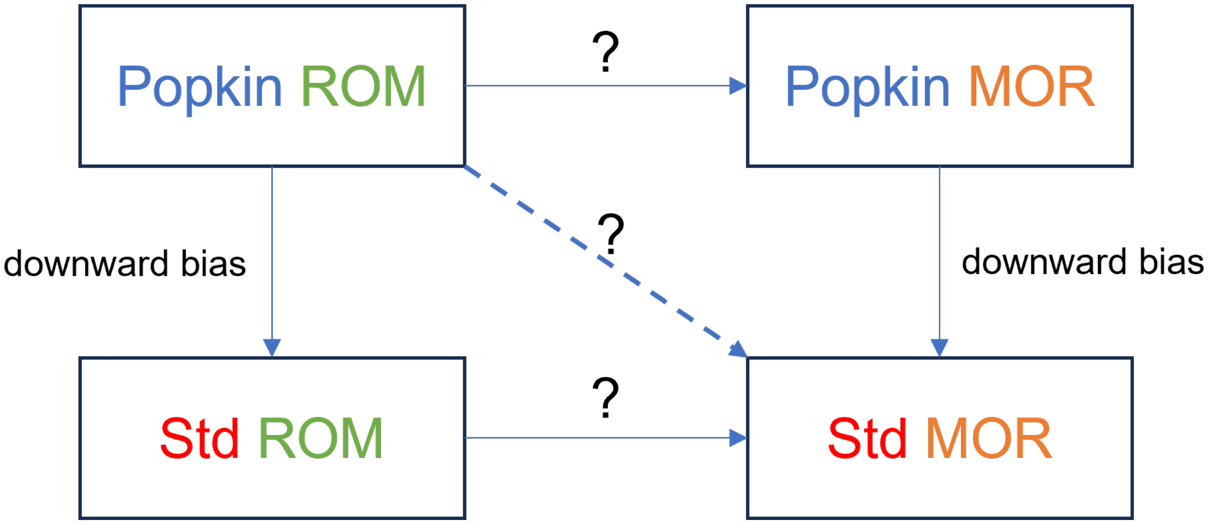

Figure 7: **Relationships among popkin/standard ROM/MOR estimators.**

References

Almasy, L. and J. Blangero (1998). “Multipoint quantitative-trait linkage analysis in general pedigrees”. *Am. J. Hum. Genet.* 62(5), pp. 1198–1211.

Astle, William and David J. Balding (2009). “Population Structure and Cryptic Relatedness in Genetic Association Studies”. *Statist. Sci.* 24(4), pp. 451–471.

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

Bhatia, Gaurav et al. (2013). “Estimating and interpreting FST: the impact of rare variants”. *Genome Res.* 23(9), pp. 1514–1521.

Bulik-Sullivan, Brendan K et al. (2015). “LD Score regression distinguishes confounding from polygenicity in genome-wide association studies”. *Nature genetics* 47(3), pp. 291–295.

Cason, Rachel K et al. (2023). “Genetic risk variants for childhood nephrotic syndrome and corticosteroid response”. *Frontiers in Pediatrics* 11, p. 1248733.

Chang, Christopher C. et al. (2015). “Second-generation PLINK: rising to the challenge of larger and richer datasets”. *GigaScience* 4(1), p. 7.

Choi, Shing Wan, Timothy Shin-Heng Mak, and Paul F O’Reilly (2020). “Tutorial: a guide to performing polygenic risk score analyses”. *Nature protocols* 15(9), pp. 2759–2772.

Fairley, Susan et al. (2020). “The International Genome Sample Resource (IGSR) collection of open human genomic variation resources”. *Nucleic Acids Research* 48(D1), pp. D941–D947.

Falconer, Douglas Scott (1996). *Introduction to quantitative genetics.* Pearson Education India.

Hou, Zhuoran and Alejandro Ochoa (2023). “Genetic association models are robust to common population kinship estimation biases”. *Genetics* 224(1), iyad030.

Jacquard, Albert (1970). *Structures génétiques des populations.* Paris: Masson et Cie.

Loh, Po-Ru et al. (2015). “Efficient Bayesian mixed-model analysis increases association power in large cohorts”. *Nat. Genet.* 47(3), pp. 284–290.

Luo, Yang et al. (2021). “Estimating heritability and its enrichment in tissue-specific gene sets in admixed populations”. *Human molecular genetics* 30(16), pp. 1521–1534.

Lush, Jay L et al. (1949). “Heritability of quantitative characters in farm animals.” *Heritability of quantitative characters in farm animals.*

Malécot, Gustave (1948). *Mathématiques de l’hérédité.* Masson et Cie.

Mitchell, Braxton D et al. (1996). “Genetic and environmental contributions to cardiovascular risk factors in Mexican Americans: the San Antonio Family Heart Study”. *Circulation* 94(9), pp. 2159–2170.

Ochoa, Alejandro and John D. Storey (2021). “Estimating FST and kinship for arbitrary population structures”. *PLoS Genet* 17(1), e1009241.

Price, Alkes L. et al. (2006). “Principal components analysis corrects for stratification in genome-wide association studies”. *Nat. Genet.* 38(8), pp. 904–909.

Rakovski, Cyril S. and Daniel O. Stram (2009). “A kinship-based modification of the armitage trend test to address hidden population structure and small differential genotyping errors”. *PLoS ONE* 4(6), e5825.

Sorlie, Paul D et al. (2010). “Design and implementation of the Hispanic community health study/study of Latinos”. *Annals of epidemiology* 20(8), pp. 629–641.

Speed, Doug and David J. Balding (2015). “Relatedness in the post-genomic era: is it still useful?” *Nat. Rev. Genet.* 16(1), pp. 33–44.

Speed, Doug and David J Balding (2019). “SumHer better estimates the SNP heritability of complex traits from summary statistics”. *Nature genetics* 51(2), pp. 277–284.

Speed, Doug et al. (2012). “Improved heritability estimation from genome-wide SNPs”. *Am. J. Hum. Genet.* 91(6), pp. 1011–1021.

Speed, Doug et al. (2017). “Reevaluation of SNP heritability in complex human traits”. *Nat Genet* 49(7), pp. 986–992.

Sul, Jae Hoon, Lana S. Martin, and Eleazar Eskin (2018). “Population structure in genetic studies: Confounding factors and mixed models”. *PLoS Genet.* 14(12), e1007309.

Thornton, Timothy and Mary Sara McPeek (2010). “ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure”. *Am. J. Hum. Genet.* 86(2), pp. 172–184.

Visscher, Peter M, William G Hill, and Naomi R Wray (2008). “Heritability in the genomics era—concepts and misconceptions”. *Nature reviews genetics* 9(4), pp. 255–266.

Wang, Bowen, Serge Sverdlov, and Elizabeth Thompson (2017). “Efficient Estimation of Realized Kinship from SNP Genotypes”. *Genetics*, genetics.116.197004.

Wright, Sewall (1949). “The Genetical Structure of Populations”. *Annals of Eugenics* 15(1), pp. 323–354.

Yang, Jian et al. (2010). “Common SNPs explain a large proportion of the heritability for human height”. *Nat. Genet.* 42(7), pp. 565–569.

Yang, Jian et al. (2011). “GCTA: a tool for genome-wide complex trait analysis”. *The American Journal of Human Genetics* 88(1), pp. 76–82.

Yao, Yiqi and Alejandro Ochoa (2022). “Limitations of principal components in quantitative genetic association models for human studies”, p. 2022.03.25.485885.

Zhou, Xiang and Matthew Stephens (2012). “Genome-wide efficient mixed-model analysis for association studies”. *Nat. Genet.* 44(7), pp. 821–824.

S1 Supplemental figures

...