

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

Zhuoran Hou^{1,2}, 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. Kinship matrices, also known as Genetic Relatedness Matrices or “GRMs”, are required for heritability estimation with variance components models. However, the most common “standard” kinship estimator employed by GCTA and other approaches, can be severely biased in structured populations. In this study, we characterize heritability estimation biases in GCTA due to kinship estimation biases under population structure. For the standard *ratio-of-means* (ROM) kinship estimator, we derive a closed-form expression for heritability bias given by the mean kinship value and the true heritability. The standard *mean-of-ratios* (MOR) estimator is the most widely used in practice, and exhibits more severe bias than ROM due to upweighting low-frequency variants. Using simulation studies with admixture and family structures, as well as simulated traits from 1000 Genomes genotypes, we find that only Popkin, which is the only unbiased population kinship estimator, produces unbiased heritability estimates in structured settings. Pedigree-only estimates have upward heritability biases when there is population structure. Finally, we analyze three structured datasets with real phenotypes—the San Antonio Family Study, the Hispanic Community Health Study / Study of Latinos, and a multiethnic Nephrotic Syndrome cohort. The standard MOR estimator can produce both downward and upward heritability biases depending on population structure and variant frequency spectrum, compared to the other two estimators. Overall, common kinship estimators result in heritability estimation biases when applied to structured populations, a challenge that Popkin successfully overcomes.

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, which is the reason relatedness plays a large role in heritability estimation (Visscher et al., 2008). In addition, heritability is an important parameter of many trait models, including polygenic risk scores (PRS), whose performance is bounded above by heritability (Choi et al., 2020). There are two primary types of heritability: broad-sense heritability (H^2) and narrow-sense heritability (h^2). Broad-sense heritability includes all genetic variance components—additive (A), dominance (D), and epistatic (I) effects—while narrow-sense heritability focuses specifically on additive genetic variance (Falconer, 1996). Because additive effects are transmitted from parent to offspring in a predictable manner, narrow-sense heritability is the relevant parameter most often estimated in GWAS, PRS modeling, and many other applications. Accurately estimating h^2 is therefore fundamental for both understanding genetic contributions to traits and for the development of predictive models.

Heritability has long been estimated from close relatives, such as twins or siblings (Falconer, 1996). The variance component model of SOLAR, which is based on estimating kinship matrices from pedigrees, enables the use of more complex pedigrees and distant relatives (Almasy and Blangero, 1998a). GCTA extended the last approach to population data, employing population kinship matrices estimated from genetic data only, which was used to demonstrate that SNPs likely explain the majority of missing heritability for height (Yang et al., 2010; Yang et al., 2011). Lastly, there are alternate approaches for estimating heritability from GWAS summary statistics and LD estimates, which enable partitioning heritability within gene sets, such as LD Score Regression (Bulik-Sullivan et al., 2015; Luo et al., 2021) and SumHer (Speed and Balding, 2019), but since they do not use kinship matrices they fall out of the scope of the present work.

Despite widespread use, all heritability estimation methods face important limitations. Traditional twin studies, for example, assume that monozygotic and dizygotic twins share environments to the same extent—a questionable assumption that can inflate heritability estimates by conflating genetic and environmental similarity (Tenesa and Haley, 2013; Charney, 2012). Pedigree-based methods like SOLAR are sensitive to incomplete or biased family structures. GCTA, although groundbreaking, has also sparked debate: it primarily captures the additive effects of common SNPs, potentially missing contributions from rare variants, non-additive effects, or poorly tagged genomic regions (Speed et al., 2012a; Speed et al., 2017a; Zaitlen et al., 2013; Yang et al., 2017). Moreover, population structure and cryptic relatedness can bias GCTA-based estimates if not properly controlled, especially in admixed samples (Price et al., 2010; Yang et al., 2017; Huang et al., 2025). These issues have led to an ongoing reevaluation of how heritability is conceptualized and measured in the genomic era.

Accurate kinship estimation is crucial for heritability estimation based on variance components,

such as GCTA, which are special cases of linear mixed-effects models (LMMs). However, the most common kinship estimator employed by these approaches, which we refer to as the "standard" estimator, can be severely biased in structured populations (Ochoa and Storey, 2021). We previously showed that association tests are invariant to the use of common biased kinship estimators compared to 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. Thus, we hypothesize kinship estimation bias will have a strong effect on heritability estimation accuracy, particularly when there is population structure. However, we do not expect these biases to explain missing heritability in previous studies of relatively homogeneous populations, such as the landmark height paper which only analyzed Australian individuals of European ancestry (Yang et al., 2010).

In addition to concerns about population structure, recent efforts have sought to improve heritability estimation by modeling the effects of rare variants through partitioned genomic relationship matrices (GRMs) or by adjusting assumptions about genetic architecture. Several studies have emphasized that the contribution of a variant to heritability depends not only on its effect size but also on its minor allele frequency (MAF) and linkage disequilibrium (LD) patterns. Our theory suggests that these differences could be in part due to kinship estimation biases specific to rare variants. However, another potentially complementary explanation is that rare variants have a different structure than common variants in humans (Zaidi and Mathieson, 2020). Traditional methods overestimate the contribution of common compared to rare variants, and it has been proposed that heritability is more evenly distributed across the MAF spectrum when accounting for LD (Speed et al., 2017a). Furthermore, low-frequency variants are enriched for functional annotations under negative selection, reinforcing the view that rare variants often have larger per-allele effects (Gazal et al., 2018). Using whole-genome sequencing data, it has been shown that rare variants contribute substantially to the heritability of complex traits—possibly explaining a significant portion of the missing heritability (Wainschtein et al., 2022). More flexible models have been introduced that accommodate diverse genetic architectures, including skewed effect size distributions correlated with allele frequency, but still found that rare variant heritability remains difficult to estimate accurately (Hou et al., 2019). SumHer models effect sizes as a function of both LD and MAF, producing more realistic assumptions about rare variants compared to GCTA (Speed and Balding, 2019). Taken together, these studies highlight the importance of modeling rare variant contributions explicitly, both in terms of effect-size distributions and their relationships to allele frequency and LD, which are essential for accurate heritability estimation in complex trait genetics. However, our work highlights important unsolved challenges estimating variance for rare variants, particularly under population structure, which remain an open problem.

In this study, we characterize the theoretically predicted heritability estimation bias due to kinship bias, following the derivation in our previous work (Hou and Ochoa, 2023) and others (Chen and Storey, 2022), and empirically compare key kinship estimators in various population structure

scenarios. We conduct simulation studies to evaluate heritability estimation accuracy and bias under scenarios such as admixture structure only and admixture combined with family structure, as well as simulated traits drawn from the real 1000 Genomes genotypes. We then apply these kinship estimators to datasets with population structure and real phenotypes, to further characterize the sources and extent of bias in practice: the San Antonio Family Study (Mitchell et al., 1996), the Hispanic Community Health Study/Study of Latinos (Sorlie et al., 2010), and a Nephrotic Syndrome multiethnic cohort. Our results show that the standard kinship estimators—particularly the commonly used mean-of-ratios (MOR) version—introduce systematic and sometimes severe biases in heritability estimation, especially in the presence of population structure and rare variants. On the other hand, pedigree-derived kinship matrices exhibit upward biases when there is population structure. Only the Popkin estimator yields unbiased heritability estimates across all simulation and real data settings. Overall, our findings highlight the importance of using unbiased kinship estimators to obtain reliable heritability estimates in structured populations.

2 Methods

2.1 Genetic and trait models

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. Let φ_{jk} is the kinship coefficient of two individuals j and k , and p_i is the ancestral allele frequency at locus i , which are in terms of an implicit ancestral population (usually the most recent common ancestor) that will remain fixed in this work. Under the kinship model (Malécot, 1948; Wright, 1949; Jacquard, 1970; Astle and Balding, 2009; Ochoa and Storey, 2021) the expectation and covariance of genotypes are given by

$$\mathbb{E} [\mathbf{x}_i] = 2p_i \mathbf{1}_n, \quad \text{Cov} (\mathbf{x}_i) = 4p_i(1-p_i) \boldsymbol{\Phi}, \quad (1)$$

where $\mathbf{x}_i = (x_{ij})$ is the length- n vector of genotypes at locus i , $\boldsymbol{\Phi} = (\varphi_{jk})$ is the $n \times n$ kinship matrix, and $\mathbf{1}_n$ is a length- n vector of ones. The definition of kinship that we follow, which is a probability of identity by descent, is such that the maximum kinship of 1 is achieved for fully inbred individuals (who have only homozygote genotypes) in the case of self-kinship, or fully inbred identical twins for a pair of different individuals. In this definition, the self kinship of an outbred individual is 1/2, the kinship between a parent and their outbred child is 1/4, and so is the expected kinship between outbred siblings. In contrast, GCTA and related models typically define kinship as twice the value that we use, so that self kinship for an outbred individual is 1, and kinship between outbred parent and child or between siblings is 1/2 (Yang et al., 2010; Yang et al., 2011).

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

$$\begin{aligned}\mathbf{y} &= \mathbf{1}_n\alpha + \mathbf{X}'\boldsymbol{\beta} + \boldsymbol{\epsilon}, \\ \boldsymbol{\epsilon} &\sim \text{Normal}(\mathbf{0}, \sigma_e^2 \mathbf{I}_n),\end{aligned}\tag{2}$$

where α is the intercept coefficient, $\boldsymbol{\beta} = (\beta_i)$ is a length- m vector of genetic effect coefficients for each locus i , $\boldsymbol{\epsilon}$ is a length- n vector of non-genetic independent residual effects with zero mean and standard deviation σ_e^2 , which is also called the environmental variance component, \mathbf{I}_n is the $n \times n$ identity matrix, and the prime ('') denotes matrix transposition.

We now derive the variance component model of interest from the previous genotype and trait models. As stated in those models, \mathbf{X} and $\boldsymbol{\epsilon}$ are random, while we treat α and $\boldsymbol{\beta}$ as fixed parameters. Note that the mean trait is given by

$$\mathbb{E}[\mathbf{y}] = \mathbf{1}_n\mu, \quad \mu = \alpha + \sum_{i=1}^m 2p_i\beta_i.\tag{3}$$

Denote the genetic effect by $\mathbf{s} = \mathbf{X}'\boldsymbol{\beta}$. The covariance structure of the genetic effect is also a scaled version of the kinship matrix. In particular, assuming Eq. (1) and independent loci, then

$$\begin{aligned}\text{Cov}(\mathbf{s}) &= \text{Cov}\left(\sum_{i=1}^m \mathbf{x}_i\beta_i\right) \\ &= \sum_{i=1}^m \text{Cov}(\mathbf{x}_i\beta_i) \\ &= \sum_{i=1}^m 4p_i(1-p_i)\beta_i^2\Phi \\ &= \sigma_g^2 2\Phi, \\ \sigma_g^2 &= \sum_{i=1}^m 2p_i(1-p_i)\beta_i^2.\end{aligned}\tag{4}$$

This particular scale for the genetic variance component σ_g^2 , which leaves a factor of two behind, is used traditionally so that σ_g^2 corresponds to the variance of outbred individuals, who have $\varphi_{jj} = 1/2$. Further, as noted earlier, GCTA and other models define their kinship matrices as 2Φ under our notation.

Implicit in the assumed Eq. (1) and the resulting Eq. (4) is that the covariance structure is the same for causal variants as for the rest of the genome (from which the kinship matrix will be estimated, as shown later). However, there is evidence for humans that rare variants have a different structure than common variants (Zaidi and Mathieson, 2020), a potential misspecification to keep in mind as we analyze the real data later.

If we shift the mean of genotypes to the intercept, and assume that \mathbf{s} is well approximated by

a multivariate distribution with the above covariance, then we arrive at the model that GCTA fits:

$$\begin{aligned}\mathbf{y} &= \mathbf{1}_n \alpha + \mathbf{s} + \boldsymbol{\epsilon}, \\ \mathbf{s} &\sim \text{Normal}(\mathbf{0}, \sigma_g^2 2\Phi), \\ \mathbf{s} + \boldsymbol{\epsilon} &\sim \text{Normal}(\mathbf{0}, \sigma_g^2 2\Phi + \sigma_e^2 \mathbf{I}_n).\end{aligned}$$

The narrow-sense heritability h^2 is defined as the proportion of variance that corresponds to the genetic variance component:

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

If we define the trait variance scale as $\sigma^2 = \sigma_g^2 + \sigma_e^2$, which is the trait variance of an outbred individual, then note that $\sigma_g^2 = \sigma^2 h^2$ and $\sigma_e^2 = \sigma^2(1 - h^2)$.

2.2 Statistical problems of rare variant allele frequency estimates

A recurrent theme in this work, surfacing in both the biases of kinship estimators and of trait simulations, is statistical problems due to estimation of allele frequencies under population structure. The standard ancestral allele frequency estimator,

$$\hat{p}_i = \frac{1}{2n} \sum_{j=1}^n x_{ij}, \quad (5)$$

is unbiased ($E[\hat{p}_i] = p_i$) under the kinship model of Eq. (1), and has a variance of (Ochoa and Storey, 2021)

$$\begin{aligned}\text{Var}(\hat{p}_i) &= p_i(1 - p_i)\bar{\varphi}, \\ \bar{\varphi} &= \frac{1}{n^2} \sum_{j=1}^n \sum_{k=1}^n \varphi_{jk}\end{aligned} \quad (6)$$

where $\bar{\varphi}$ is the mean kinship value in the sample, and converges to a non-zero value when there is population structure. For this reason, \hat{p}_i is not a consistent estimator when there is population structure, in other words it does not converge to the true p_i as sample sizes go to infinity, which is an underlying assumption of many previous works in statistical genetics that does not hold for real data with population structure.

This work requires estimates of the Binomial variance factor $p_i(1 - p_i)$. Previous work found that the sample estimator is biased, although it can be readily unbiased with a good estimate of $\bar{\varphi}$ (Ochoa and Storey, 2021):

$$E[\hat{p}_i(1 - \hat{p}_i)] = p_i(1 - p_i)(1 - \bar{\varphi}). \quad (7)$$

However, using a normality approximation, in Appendix A we prove that the variance of this estimator also converges to a non-zero value, so $\hat{p}_i(1 - \hat{p}_i)/(1 - \bar{\varphi})$ is also not consistent:

$$\text{Var}(\hat{p}_i(1 - \hat{p}_i)) = p_i(1 - p_i)\bar{\varphi}((1 - 2p_i)^2 + 2p_i(1 - p_i)\bar{\varphi}).$$

As a direct consequence, the standard estimate of the inverse Binomial variance can be severely biased. Specifically, applying Jensen's inequality to the inverse function with positive arguments, combined with the unbiased form that follows from Eq. (7), we obtain that

$$E \left[\frac{1 - \bar{\varphi}}{\hat{p}_i(1 - \hat{p}_i)} \right] \geq \frac{1}{p_i(1 - p_i)}. \quad (8)$$

Equality occurs when $\hat{p}_i(1 - \hat{p}_i)$ has no variance, which again does not occur under population structure, and conversely, higher variance exacerbates the inequality. Empirically, biases are greater for rare variants, where p_i or $1 - p_i$ are close to zero.

2.3 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. Let $\hat{\Phi}^{\text{name}} = (\hat{\varphi}_{jk}^{\text{name}})$ relate the scalar and matrix formulas of each named kinship estimator.

2.3.1 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., 2017a), including heritability estimation (Yang et al., 2010; Yang et al., 2011; Speed et al., 2012b; Speed and Balding, 2015; Speed et al., 2017b) 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 ROM and MOR versions of the standard kinship estimator are, respectively,

$$\hat{\varphi}_{jk}^{\text{std-ROM}} = \frac{\sum_{i=1}^m (x_{ij} - 2\hat{p}_i)(x_{ik} - 2\hat{p}_i)}{\sum_{i=1}^m 4\hat{p}_i(1 - \hat{p}_i)}, \quad (9)$$

$$\hat{\varphi}_{jk}^{\text{std-MOR}} = \frac{1}{m} \sum_{i=1}^m \frac{(x_{ij} - 2\hat{p}_i)(x_{ik} - 2\hat{p}_i)}{4\hat{p}_i(1 - \hat{p}_i)}. \quad (10)$$

The ROM estimator has a biased limit (Ochoa and Storey, 2021; Hou and Ochoa, 2023):

$$\hat{\Phi}^{\text{std-ROM}} \xrightarrow[m \rightarrow \infty]{\text{a.s.}} \frac{1}{1 - \bar{\varphi}} \mathbf{C} \boldsymbol{\Phi} \mathbf{C}, \quad (11)$$

where $\boldsymbol{\Phi}$ is the true kinship matrix, the scalar mean kinship $\bar{\varphi}$ is as in Eq. (6), and $\mathbf{C} = \mathbf{I}_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}'_n$ is the centering matrix. The MOR estimator does not have a closed-form limit, but in practice it is well approximated by Eq. (11) when rare variants are excluded prior to calculating this estimate, though they can differ greatly otherwise due to the additional biases described in Eq. (8).

2.3.2 Popkin kinship estimator

The popkin (population kinship) estimator (Ochoa and Storey, 2021) is given by

$$\varphi_{jk}^{\text{popkin}} = 1 - \frac{A_{jk}}{\hat{A}_{\min}}, \quad A_{jk} = \frac{1}{m} \sum_{i=1}^m (x_{ij} - 1)(x_{ik} - 1) - 1,$$

where $\hat{A}_{\min} = \min_{j \neq k} A_{jk}$. This estimator of type ROM has an unbiased almost sure limit as the number of loci m goes to infinity,

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

under the assumption that the true minimum kinship is zero. Popkin avoids the biases of the standard estimators because it does not rely on the inconsistent \hat{p}_i estimator.

The mean kinship values $\bar{\varphi}$ reported here are always calculated from the true kinship matrix, or estimated with popkin when the true kinship matrix is unavailable, as is the case for real datasets, since popkin estimates this value without bias. Note both Standard ROM and MOR mean kinship estimates are always exactly zero (algebraically, with no variance), which indirectly confirms that they must be biased.

2.3.3 Software

Standard MOR kinship matrices and all heritability estimates are calculated using GCTA (version 1.93.2beta) (Yang et al., 2011). Popkin kinship estimates are computed using the `popkin` R package, while Standard ROM kinship estimates are calculated using the `popkinsuppl` R package (Ochoa and Storey, 2021). Kinship matrices are calculated from pedigrees using the `kinship2` R package (Sinnwell et al., 2014). GCTA is provided alternative kinship matrices encoded as standard GRM files using the `genio` R package (Yao and Ochoa, 2022), doubling kinship values to match Eq. (4) as needed. Plink (version 2.00a3LM) is used to process genotype data (Chang et al., 2015).

2.4 Heritability estimation bias due to Standard ROM kinship bias

We recently characterized the theoretical relationship between variance component estimates of a biased kinship matrix and its unbiased counterpart (Hou and Ochoa, 2023). In particular, when provided with the Standard ROM kinship matrix of Eq. (11), the LMM fits the genetic variance component with an algebraically biased form compared to the unbiased estimate from Popkin, whereas the environment variance component estimate is the same for both:

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

where $\bar{\varphi}$ is the mean value of the true kinship matrix of Eq. (6). Therefore, since $\hat{\sigma}_e^2 = \left(\frac{1}{\hat{h}^2} - 1\right)\hat{\sigma}_g^2$, Standard ROM heritability estimates will be biased with the following form, where \hat{h}^2 is the unbiased

Popkin estimate:

$$\begin{aligned}
\hat{h}^{2,\text{biased}} &= \frac{\hat{\sigma}_g^{2,\text{biased}}}{\hat{\sigma}_g^{2,\text{biased}} + \hat{\sigma}_e^{2,\text{biased}}} \\
&= \frac{(1 - \bar{\varphi})\hat{\sigma}_g^2}{(1 - \bar{\varphi})\hat{\sigma}_g^2 + \left(\frac{1}{\hat{h}^2} - 1\right)\hat{\sigma}_g^2} \\
&= \hat{h}^2 \frac{1 - \bar{\varphi}}{1 - \bar{\varphi}\hat{h}^2}.
\end{aligned} \tag{12}$$

The above formula not only relates Popkin and Standard ROM estimates in a given dataset, but it also holds in their limit, in which case the Popkin estimate becomes the true heritability. In terms of estimate limits, bias is larger for intermediate true heritability estimates and increases with $\bar{\varphi}$, while bias decreasing to zero when the heritability is close to 0 or 1 (Fig. 1).

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. Each scenario was replicated 50 times, in each case producing a new genotype matrix.

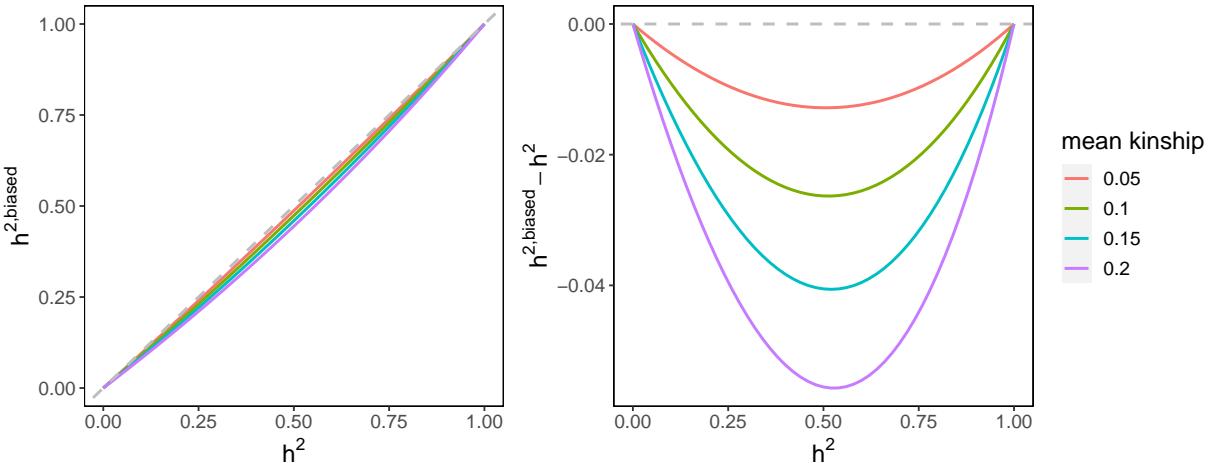


Figure 1: **Relationship between the true heritability and biased Standard ROM estimates.** We illustrate the functional form of Eq. (12) for a range values of the mean kinship parameter $\bar{\varphi}$ observed in real human datasets (Table 1). Note Standard MOR (the most commonly used estimator) has even larger biases empirically. Left: the relationship between biased estimates and the true heritability. Right: the relationship between heritability bias and true heritability.

2.5.1 Admixture simulation for genotype matrices

An admixed family is simulated as before (Yao and Ochoa, 2022; Hou and Ochoa, 2023), with $K = 3$ ancestries and $F_{ST} = 0.3$ for the admixed individuals, which resembles Hispanics and African Americans. Briefly, our admixture model simulates $n = 1000$ individuals with $m = 100,000$ loci. Random ancestral allele frequencies p_i for $i \in \{1, \dots, m\}$, subpopulation allele frequencies $p_i^{S_u}$ for $u \in \{1, \dots, K\}$, and individual-specific allele frequencies π_{ij} and genotypes x_{ij} for $j \in \{1, \dots, n\}$ 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 $p_i(1 - p_i)f_{S_u}$. The admixture proportions q_{ju} and subpopulation inbreeding values f_{S_u} are constructed based on a diffusion model on a 1D geography (Ochoa and Storey, 2021).

For simulations with family structure, 20 generations are generated iteratively, as follows. Individuals in the first generation ($n = 1000$) are drawn from our admixture model, ordered by 1D geography, randomly assigned sex, and treated as locally unrelated. From subsequent generations, individuals are paired iteratively: randomly choosing males from the pool and pairing them with the nearest available female with local kinship $< 1/4^3$ (to preserve the admixture structure) until there are no available males or females. Family sizes are drawn randomly ensuring every family has at least one child. Children are reordered by the average coordinates of their parents, their sex are assigned randomly, and their alleles are drawn from parents independently per locus. The simulation is implemented in the R package `simfam`.

2.5.2 Trait simulation algorithms

We wish to simulate a quantitative trait \mathbf{y} that follows the linear polygenic model of Eq. (2), assuming the genotype matrix \mathbf{X} is given, but the intercept α and the genetic coefficients $\boldsymbol{\beta}$ have not been determined, and they can be random, but they must result in m_1 number of causal loci (with non-zero coefficients), a trait mean of μ , variance scale of σ^2 , and a heritability of h^2 . In this study, we set $m_1 = 500$, $\mu = 0$, $\sigma^2 = 1$, $h^2 \in \{0, 0.1, \dots, 1\}$ for the admixture simulations and $h^2 = 0.8$ for 1000 Genomes. The algorithms require either true ancestral allele frequencies p_i (available for simulations but never for real genotype data) or both its sample estimate \hat{p}_i from Eq. (5) and an unbiased estimate of the mean kinship $\bar{\varphi}$ in order to unbias estimates using Eq. (7).

These algorithms have been used to evaluate GWAS methods (Yao and Ochoa, 2022; Hou and Ochoa, 2023), but here we motivate them for evaluating heritability estimation.

In all cases, independent residual effects ϵ are simulated from Eq. (2) with $\sigma_e^2 = \sigma^2(1 - h^2)$, we randomly select m_1 loci to be causal, and reindex loci to only include causal loci. Afterwards, the following paragraphs describe how to construct β under two evolutionary models, followed by the construction of α . In turn, each of those models has two cases, namely whether the true ancestral allele frequencies p_i are known or not.

In this paper, we treat effect sizes β_i as random variables only in the context of trait simulation (RC and FES), where they are explicitly sampled from a distribution to reflect the assumed genetic architecture. This randomness is necessary for defining and analyzing the properties of our estimators. However, in the rest of the paper we treat the β_i as fixed parameters. This separation maintains both the statistical rigor of our simulation-based proofs and the interpretability expected in applied genetics. In the simulation procedure, we explicitly track both the random variables defined by the model and their realized values in each replicate.

Random Coefficients (RC) model. The initial effect sizes β_{i0} are treated as independent random variables drawn from a standard normal distribution:

$$\beta_{i0} \sim \text{Normal}(0, 1).$$

Let b_{i0} denote the realized value of β_{i0} used in the simulation procedure.

Then, if p_i are known, the theoretical genetic variance component following Eq. (4) is defined as:

$$\sigma_{g0}^2 = E \left[\sum_{i=1}^{m_1} 2p_i(1-p_i)\beta_{i0}^2 \right] = \sum_{i=1}^{m_1} 2p_i(1-p_i),$$

which is a fixed quantity. For a given simulation replicate, we compute the realized variance:

$$s_{g0}^2 = \sum_{i=1}^{m_1} 2p_i(1-p_i)b_{i0}^2.$$

We obtain the desired variance of $\sigma_g^2 = h^2\sigma^2$ by dividing each b_{i0} by s_{g0} (which results in a variance of 1) and then multiply by $h\sigma$. Combining both steps, the final coefficients b_i used in the algorithm and the corresponding random variable β_i are:

$$b_i = b_{i0} \frac{h\sigma}{s_{g0}}, \quad \beta_i = \beta_{i0} \frac{h\sigma}{\sigma_{g0}}.$$

If only \hat{p}_i are available, the plug-in Binomial variance estimator $\hat{p}_i(1-\hat{p}_i)$ is downwardly biased as shown in Eq. (7). Using that result to unbias our estimator, the initial genetic variance component, which is now a function of two random variables \hat{p}_i and β_{i0} , is estimated without bias by $\hat{\sigma}_{g0}^2$, and the realized value of this estimator used in the simulation procedure is denoted \hat{s}_{g0}^2 as:

$$\hat{s}_{g0}^2 = \sum_{i=1}^{m_1} 2\frac{\hat{p}_i^{obs}(1-\hat{p}_i^{obs})}{1-\bar{\varphi}} b_{i0}^2, \quad \hat{\sigma}_{g0}^2 = \sum_{i=1}^{m_1} 2\frac{\hat{p}_i(1-\hat{p}_i)}{1-\bar{\varphi}} \beta_{i0}^2,$$

where we denote the realized value of \hat{p}_i used in simulations as \hat{p}_i^{obs} for clarity and $\bar{\varphi}$ is the mean kinship of Eq. (6). This estimator not only satisfies $E[\hat{\sigma}_{g0}^2] = \sigma_{g0}^2$, but is a consistent estimator of σ_{g0}^2 over our random coefficients as the number of independent causal variants m_1 goes to infinity (Lemma 1 in Appendix B). Thus, we define the final coefficients used in a simulation replicate as \hat{b}_i , corresponding to the random variable $\hat{\beta}_i$ in the model:

$$\hat{b}_i = b_{i0} \frac{h\sigma}{\hat{s}_{g0}}, \quad \hat{\beta}_i = \beta_{i0} \frac{h\sigma}{\hat{\sigma}_{g0}},$$

which results in the desired heritability. We obtain reasonable results in practice for large numbers of causal loci ($m_1 = 500$ in this study) which are not enriched for rare variants, following Theorem 1 in Appendix B.

Fixed Effect Sizes (FES) model. In this model, the effect size of locus i , defined as $2p_i(1 - p_i)\beta_i^2$, has the same value for every causal locus, so we desire each to equal $\sigma_g^2/m_1 = h^2\sigma^2/m_1$. If p_i are known, we simply solve for the desired coefficients:

$$\beta_i = \pm \frac{h\sigma}{\sqrt{2p_i(1 - p_i)m_1}},$$

where the signs are chosen randomly with equal probability.

If only \hat{p}_i are available, we again unbias the variance estimate following Eq. (7), which results in

$$\hat{\beta}_i = \pm \frac{h\sigma\sqrt{1 - \bar{\varphi}}}{\sqrt{2\hat{p}_i(1 - \hat{p}_i)m_1}}.$$

However, unlike RC, the FES estimates are more likely to have biases, meaning that $E[\hat{\beta}_i^2] \geq \beta_i^2$, where β_i^2 is fixed and non-random in the FES model. This bias may cause a substantial misspecification in heritability estimation, because of the key result in Eq. (8), which follows since the single-locus estimator $\hat{p}_i(1 - \hat{p}_i)/(1 - \bar{\varphi})$ is not consistent as the number of individuals goes to infinity. In particular, this simulation tends to behave poorly if rare variants are causal. As a consequence, the data generated for this paper in this case only used an alternative estimator: let \hat{p}_i^{MLE} be the maximum likelihood estimate of p_i obtained from the model

$$\hat{p}_i \sim \text{Beta}\left(p_i \left(\frac{1}{\bar{\varphi}} - 1\right), (1 - p_i) \left(\frac{1}{\bar{\varphi}} - 1\right)\right),$$

which is the Balding-Nichols model for \hat{p}_i that matches its known moments (Eq. (6)): $E[\hat{p}_i] = p_i$ and $\text{Var}(\hat{p}_i) = p_i(1 - p_i)\bar{\varphi}$. In particular, \hat{p}_i^{MLE} is estimated from \hat{p}_i as the only data point, and using the known $\bar{\varphi}$ estimate. While this strategy reduces bias somewhat, it does not result in unbiased estimates as shown in the results.

Construction of intercept. When p_i are known, we obtain the desired trait mean μ by solving for the intercept in Eq. (3), and we use a to denote the realized value of α used in the simulation:

$$a = \mu - 2 \sum_{i=1}^{m_1} p_i b_i,$$

where b_i are computed from the RC or FES procedure above.

If only \hat{p}_i are available, we construct the intercept coefficient using

$$a = \mu - 2\hat{p}^{obs} \sum_{i=1}^{m_1} b_i, \quad \hat{p}^{obs} = \frac{1}{m_1} \sum_{i=1}^{m_1} \hat{p}_i^{obs}.$$

Note that \hat{p}^{obs} is computed among causal loci only. This works well in practice since in both models $E[\beta_i] = 0$.

We avoid the naive construction $a = \mu - 2 \sum_{i=1}^{m_1} \hat{p}_i^{obs} b_i$, since it is equivalent to centering genotypes at each locus in the model:

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

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

$$\text{Cov}(\mathbf{x}_i - 2\hat{p}_i \mathbf{1}_n) = 4p_i(1-p_i)\mathbf{C}\Phi\mathbf{C},$$

which resembles the standard kinship estimator bias in Eq. (11). These undesirable distortions propagate to the trait, whereas the intercept we constructed does not have these distortions. Our theory proves that such distortions by themselves do not influence heritability estimation using an LMM, since the model's intercept compensates for this centering (Hou and Ochoa, 2023). Nevertheless, we preserve our trait intercept construction in case future applications are sensitive to the centering effect.

2.6 Real data analysis

We utilize the high-coverage NYGC version of the 1000 Genomes Project (Fairley et al., 2020), which is publicly available at ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20190425_NYGC_GATK/. 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. We simulate traits with causal variants that satisfy MAF thresholds in $\{0, 0.01, 0.05\}$, and separately, estimate the kinship matrix using genotypes filtered by MAF thresholds in $\{0, 0.01, 0.05\}$ as well. Mean kinship was estimated by popkin for each MAF and used to simulate traits with matching MAF filter.

The 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). We used the Phase Ia data available on dbGaP (accession phs000810.v2.p2), which genotyped individuals at 2.5 million SNPs from the Illumina SOL HCHS Custom 15041502 B3 array (core set of SNPs from the Illumina HumanOmni2.5-8 array, with the addition of 110k custom SNPs). We filter genotypes using $\text{MAF} \geq 0.01$ and $\text{HWE} \leq 1e-10$. 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. Again, we log-transformed all traits except height, % Neutrophils, % Lymphocytes, % Monocytes, % Eosinophils, % Basophils, and Red Blood Count to improve the model fits.

The San Antonio Mexican American Family Study (SAMAFS) is a 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 obtained the exome chip data from dbGaP (accession phs000847.v2.p1). We exclude all unplaced and non-autosomal variants, and further filter using $\text{MAF} \geq 0.01$ and Hardy-Weinberg Equilibrium p-value (HWE) $\leq 1\text{e-}10$. 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. Most continuous traits are skewed, so we log-transformed all traits, except height, to improve the model's fit.

The Nephrotic Syndrome (NS) multiethnic cohort is a multi-ancestry study exploring the etiology of nephrotic syndrome. We analyzed the imputed version of the genotypes filtered using minor allele count ≥ 20 , as described elsewhere. The final dataset consists of $m = 16,605,628$ loci and $n = 1,981$ individuals (the 1000 Genomes controls were excluded). We adjust for sex in the heritability estimate. Liability-scale heritability estimates were calculated using the formula of Lee et al. (2011) assuming a prevalence for NS of $K = 1.69\text{e-}4$ (Gbadegezin et al., 2015; Jia et al., 2020; Barry et al., 2023), and that SSNS (steroid sensitive NS) and SRNS (steroid resistant NS) are 80% and 20% of NS cases, respectively (Gbadegezin et al., 2015; Jia et al., 2018).

3 Results

In order to evaluate the effect of kinship estimation bias on heritability estimation, we consider simulations of genotypes and traits, real genotypes with simulated traits, and lastly real genotypes and traits. Our theory identifies two parameters as key determinants of heritability estimation bias using standard kinship estimators: the mean kinship value $\bar{\varphi}$, and the proportion of rare variants (in this work defined as the proportion of loci with $\text{MAF} \leq 0.05$, following Biddanda et al. (2020)). We focused on datasets with higher values of $\bar{\varphi}$, which reflect more population structure, since standard kinship estimators are more biased in those cases. On the other hand, the proportion of rare variants in the data, from which kinship estimates are obtained, is highly variable and depends on genotyping platform and quality control filters used, so we consider datasets that span this range (Table 1).

3.1 Evaluations based on admixture and family simulations

We begin by considering simulated genotypes and phenotypes where the true heritability is known. The genotype simulations are based on an admixture model, with or without family structure, with

parameters chosen previously to match those of real multiethnic datasets (Yao and Ochoa, 2022). The mean kinship parameters of these two simulations are $\bar{\varphi} = 0.150$ and 0.163 (first replicate, varies slightly per replicate), respectively, which are only slightly larger than those of the real datasets we analyze in this work (Table 1).

First we consider the admixture simulation without family structure. The true kinship matrix of this simulation, as well as estimates for all methods from the first replicate are visualized in Fig. S1. As expected from our theoretical results, only the Popkin kinship estimator produces unbiased heritability estimates, whose distribution closely resembles estimates obtained using the true kinship matrix of the simulation (Fig. 2A and Fig. S2A using traits simulated from the FES and RC models, respectively). The Standard ROM kinship estimator results in a small but noticeable downward bias, in accordance to our theoretical predictions (Fig. 1). Surprisingly, the Standard MOR, which is the most commonly used kinship estimator in practice, but whose bias does not have a closed form, results in considerably more downwardly biased heritability estimates than the ROM counterpart. Furthermore, downward biases are greater for higher true heritability values. To further illustrate the pattern of these biases, we plotted the bias trends for each kinship estimators (Fig. 2B and Fig. S2B).

Next, we consider the simulations with admixture followed by 20 generations of family structures (kinship matrices in Fig. S3). Heritability estimation biases are similar in magnitude in this scenario, but estimation variance is much reduced, apparently a consequence of the presence of closely related individuals. Importantly, the observed trends remain consistent with the previous scenario, namely that Popkin yields unbiased estimates in agreement with the true kinship matrix, while Standard ROM has a consistent downward estimation bias and Standard MOR has an even more severe bias that is higher for larger values of the true heritability (Fig. 2C-D and Fig. S2C-D). Additionally, here we tested using the kinship matrix derived from the true pedigree, which models the relatedness due to family structure but ignores the population structure due to shared ancestry that is also present. Unlike other cases, we find that these pedigree-based estimates are upwardly biased, and also have

Table 1: Proportions of rare variants and mean kinship values in the simulated and real datasets.

Dataset	QC Filter	Proportion of loci with $MAF \leq 0.05$	Mean kinship
Admix Sim.	$MAC > 0$	0.194	0.150
Admix+Fam Sim.	$MAC > 0$	0.180	0.163
1000 Genomes	$MAC > 0$	0.930	0.133
1000 Genomes	$MAF \geq 0.01$	0.438	0.134
1000 Genomes	$MAF \geq 0.05$	0.000	0.100
HCHS/SOL	$MAF \geq 0.01$	0.256	0.108
SAMAFLS	$MAF \geq 0.01$	0.281	0.063
NS	$MAC \geq 20$	0.564	0.124

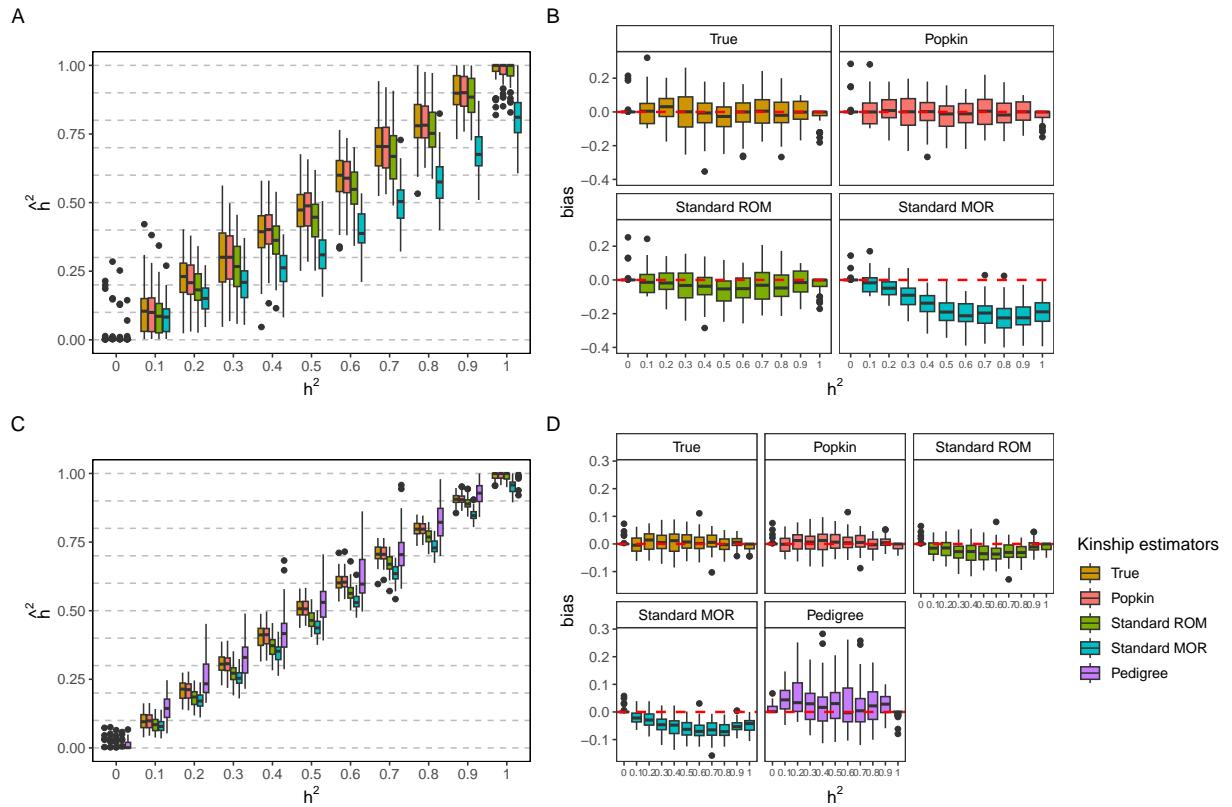


Figure 2: Heritability estimation using GCTA with various kinship estimators evaluated using simulated genotypes and traits. The RC trait model was used to simulate traits in this figure. True heritability values h^2 (x-axis) are compared to their estimates \hat{h}^2 or their bias $\hat{h}^2 - h^2$ (y-axis). **A-B.** Admixture simulation. **C-D.** Admixture plus 20 generations of family structure simulation. Every true heritability simulation was replicated 50 times.

a much greater variance than the rest of the estimates.

3.2 Evaluations based on 1000 Genomes genotypes and simulated traits

The previous simulations differ from real data in their relative dearth of rare variants ($\text{MAF} \leq 0.05$). In our simulations, ancestral allele frequencies are drawn uniformly (since the maximum MAF is $1/2$, the expected proportion of rare variants is $0.05 \times 2 = 0.10$), but after the admixture and family structure simulation variants tend to become rarer, resulting in a proportion of rare variants of 0.18-0.19 (Table 1). However, in WGS datasets like 1000 Genomes, rare variants are the majority (0.93 when there are no MAF filters; Table 1). In order to characterize the effects of rare variants, we used 1000 Genomes genotypes with simulated traits, in this case with a fixed heritability of $h^2 = 0.8$. 1000 Genomes is a multiethnic cohort which includes African, European, South Asian, East Asian, and Hispanic individuals (kinship matrices in Fig. S4), with a high mean kinship estimate of $\bar{\varphi} = 0.133$ without MAF filter ($\bar{\varphi} = 0.134$ if only including $\text{MAF} \geq 0.01$, and $\bar{\varphi} = 0.100$ for $\text{MAF} \geq 0.05$). Notably, we apply MAF filters separately to causal variants, which determine the trait’s genetic architecture, and to the variants used for kinship estimation, which affect the kinship matrices and consequently the bias in heritability estimation.

To evaluate how trait architecture influences heritability estimation bias, we simulated traits under two distinct models that differ in how they assign effects to rare variants. These models yield distinct behaviors: RC results align with previous simulations confirming Popkin’s unbiasedness, while FES reveals a novel bias in Standard MOR, though this simulation strategy has limitations on real genotypes.

The RC (random coefficients) model assigns causal effect sizes independently of allele frequency, resulting in smaller effect sizes—defined as the product of the coefficient and genotype variance—for rare variants. Importantly, RC is the only model that guarantees correctly specified heritabilities when true allele frequencies are unknown, as is typical with real genotype data. In this setting, Popkin again yields unbiased heritability estimates, with estimates closely overlapping the true value of $h^2 = 0.8$ in most cases. A small downward bias may occur if only common variants are causal but rare variants are included in the kinship estimate (Fig. S5). The Standard ROM estimator has a small downward bias in all cases, and always has smaller estimates than Popkin, as predicted by Eq. (12). In contrast, the Standard MOR estimator exhibits its strongest downward bias when rare variants are included in the kinship estimation, regardless of whether the causal variants are rare or common. However, when estimated using only common variants, Standard MOR produces heritability estimates that are nearly unbiased. These results reinforce our earlier findings for common variant simulations, while further highlighting the central role of rare variants in modulating the downward bias of the Standard MOR estimator.

Next, we turn to traits simulated with the FES (fixed effects sizes) model, which assigns causal coefficients that are roughly inversely proportional to the square root of the minor allele frequency

(see Methods), thus upweighing rare variants. When causal variants are more common ($\text{MAF} \geq 0.01$) we saw results as before, namely that Popkin gives relatively unbiased estimates, while Standard MOR is downwardly biased unless only common variants are used to estimate kinship. Unfortunately, this simulation model results in a misspecified heritability when using real genotypes and rare causal variants, since their inverse genotype variance of individual variants cannot currently be estimated without bias (Eq. (8)). Thus, when traits include rare causal variants, heritability estimates were systematically lower than the desired $h^2 = 0.8$ for all kinship estimators (first column of Fig. 3). Our theory suggests that these trait $\text{MAF} \geq 0$ results are a bias in how traits are simulated, although at present we cannot disentangle it from estimation biases. Nevertheless, keeping this heritability misspecification in mind, these results yield a new pattern: Standard MOR can yield estimates that exceed those of Popkin and Standard ROM. Although not consistently, we see higher estimates for Standard MOR when the trait has rare causal variants: either no MAF filter for trait and no MAF filter for kinship estimator, no MAF filter for trait and $\text{MAF} \geq 0.05$ for kinship estimator, and $\text{MAF} \geq 0.01$ for trait and $\text{MAF} \geq 0.05$ for kinship estimator. Thus, Standard MOR can result in both downward and upward heritability biases depending on the rare variant architecture of the trait and the inclusion of rare variants in the kinship estimate.

3.3 Analysis of real genotype and phenotype datasets

Having established that only Popkin results in unbiased heritability estimates, and having characterized the biases of the Standard ROM and MOR estimators, we now demonstrate how these biases manifest in real genotype and phenotype data. Recall that the Standard MOR kinship estimator is the most commonly used, being the default for GCTA and related methods. We analyzed three real datasets with complex population structures and otherwise complementary features: the Hispanic Community Health Study/Study of Latinos (HCHS/SOL), the San Antonio Mexican American Family Study (SAMAFS), and the Nephrotic Syndrome (NS) multiethnic cohort.

HCHS/SOL is a population study with strong population structure due to the admixture structure of the Hispanic individuals sampled from various nationalities in this study ($\bar{\varphi} = 0.108$; Fig. S6). Throughout the 33 traits that we analyzed, Popkin always estimates larger heritabilities than Standard ROM (Fig. 4), as predicted by our theoretical results. These differences are large relative to the standard errors of these estimates, where often the Popkin estimate is about a standard error larger than the Standard ROM estimate. In contrast, Standard MOR estimates are often the highest among the three estimators, sometimes exceeding Popkin by up to one standard error, although for some traits, they are instead comparable to or smaller than Popkin estimates. The proportion of rare variants, which influences the kinship estimates and the bias of Standard MOR in particular, is the lowest among the real datasets (0.26), but it is slightly higher than the simulated data (Table 1). Therefore, we do not expect the downward biases due to rare variants in the Standard MOR kinship estimate. Instead, the upward bias observed for many traits is likely driven by the presence of rare

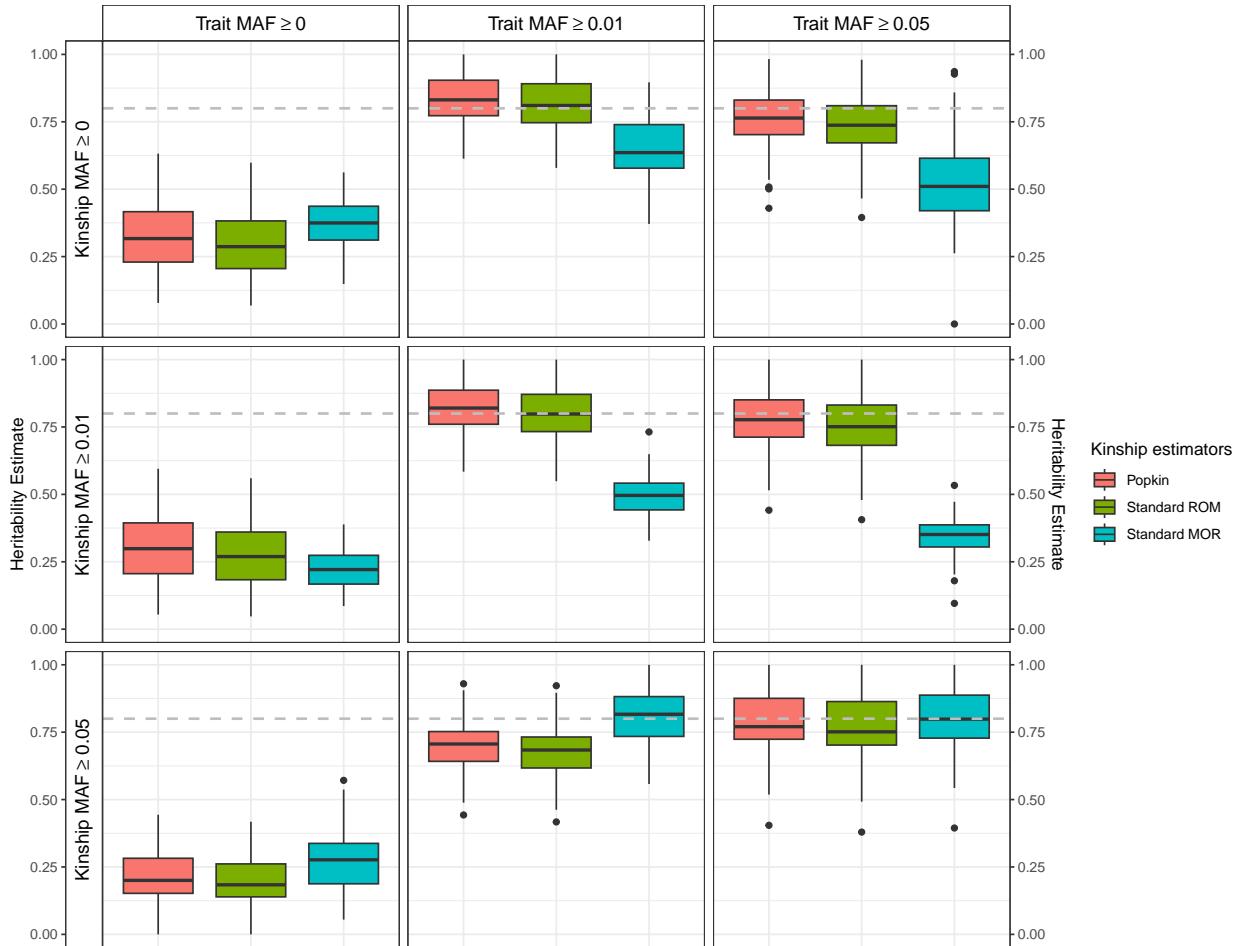


Figure 3: Testing effect of rare variants using 1000 Genomes genotypes and simulated traits. Trait simulations under the FES model, which upweights rare variants and can produce upward bias in Standard MOR estimates. Heritability is misspecified when applied to real genotypes with rare causal variants (affecting the first column only; see Methods). The true heritability is $h^2 = 0.8$ in all panels. Each trait was simulated 50 times. For simulated traits, MAF thresholds shown are applied before causal variants are selected, thus influencing the architecture of the trait. For kinship estimates, MAF thresholds are applied to genotypes before kinship matrices are estimated from these genotypes, thus influencing only kinship estimation.

causal variants, which are underrepresented in the genotype data and thus not accounted for in the kinship matrix. Accordingly, traits with larger MOR estimates relative to ROM may be those with greater rare variant contributions to heritability. Interestingly, the estimated heritability of height is high for all estimators, just under the $h^2 = 0.8$ estimated from classic sibling studies, which is achieved here only after conditioning for sex and age as fixed effects.

The SAMAFS data is considerably less structured, potentially since it is a more homogeneous set of Hispanics sampled from a single city and belonging to only 20 families ($\bar{\varphi} = 0.0634$; Fig. S7). Although we see the same patterns as in HCHS/SOL, namely that Popkin yields larger estimates than Standard ROM, differences are much smaller here relative to their standard errors (Fig. 5). Interestingly, unlike HCHS/SOL, here Standard MOR and ROM produce very similar estimates,

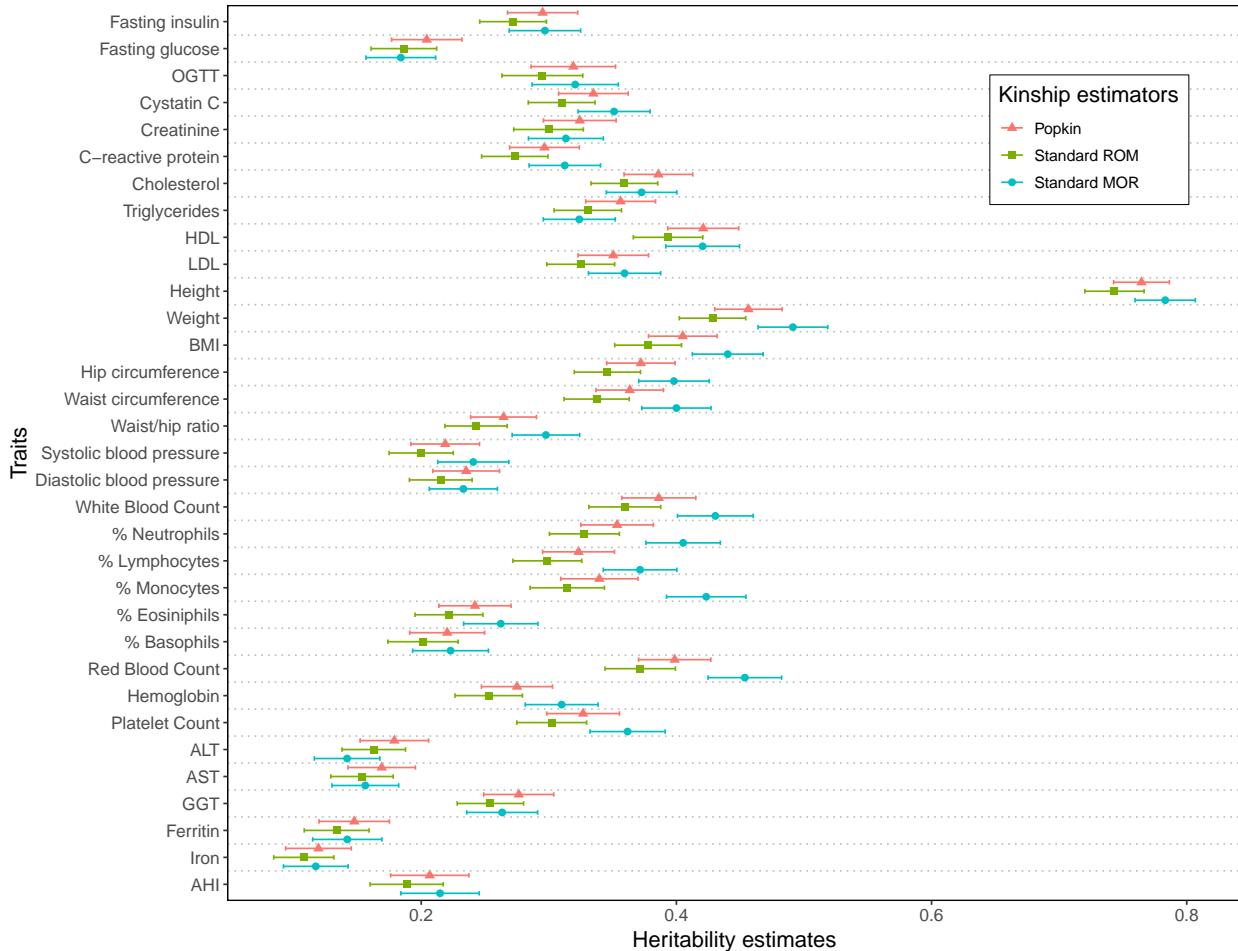


Figure 4: Heritability estimates on Hispanic Community Health Study/Study of Latinos (HCHS/SOL). The figure compares heritability estimates with GCTA using Popkin, Standard ROM, and Standard MOR kinship estimators.

which could again be partly due to a low proportion of rare variants used to estimate kinship (0.28; Table 1), although upward biases due to causal rare variants were not observed either. Furthermore, SAMAFS has a well-documented pedigree structure, which we use to estimate kinship to mimic family analyses, although it does not model the population structure due to admixture expected for Hispanics. Pedigree-based heritability estimates were generally higher than those of other methods, in some cases by more than one standard error (e.g., Adiponectin), though for one trait, Fasting glucose, the pedigree estimate was the lowest. Since pedigree-based heritability estimates were upwardly biased and highly variable in our simulations (Fig. 2), we conclude that these estimates are likely also upwardly biased. We again see high height heritability estimates, also obtained by conditioning for sex and age, which is less surprising for a family study.

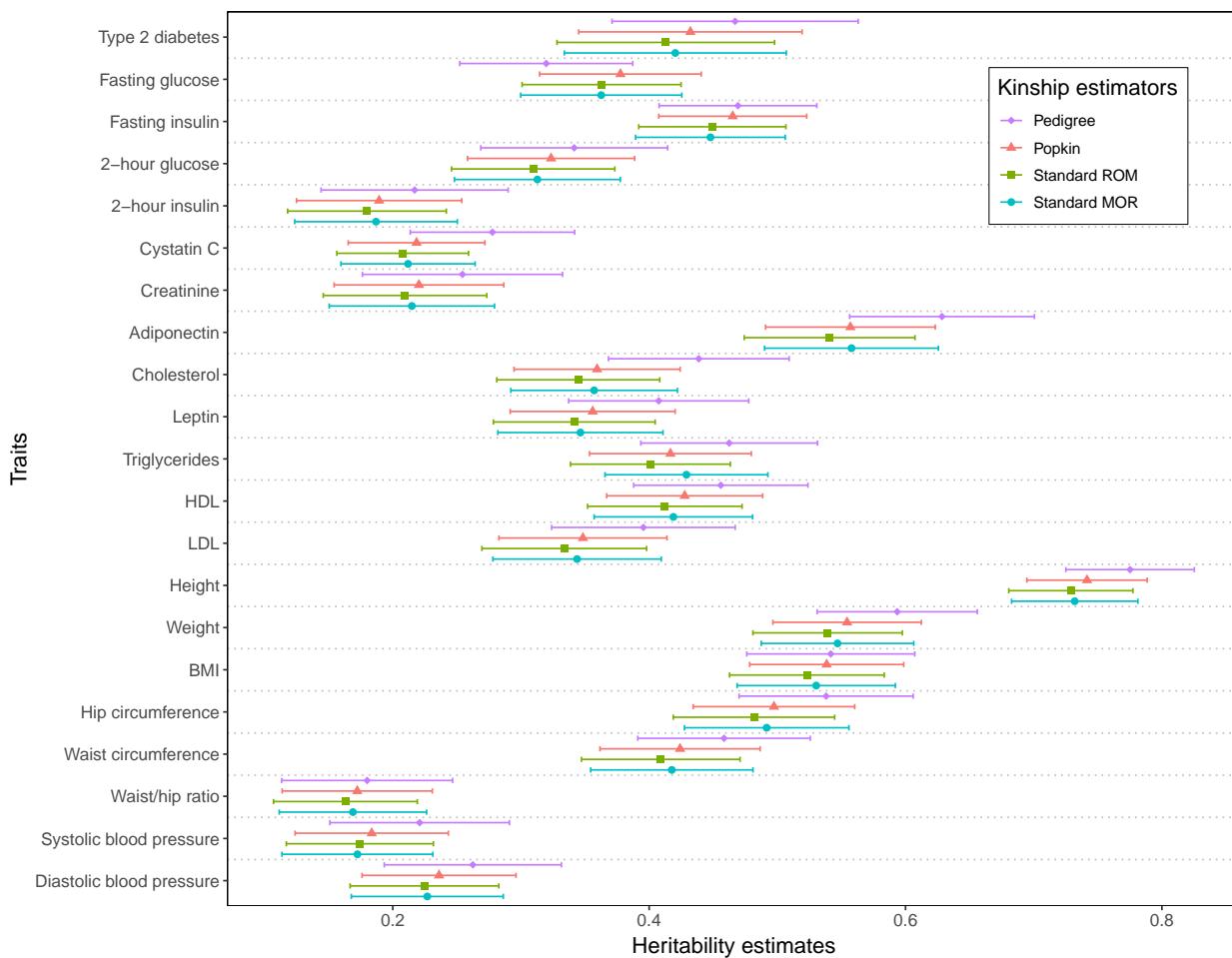


Figure 5: Heritability estimates on San Antonio Mexican American Family Study: Type 2 Diabetes (SAMAFS). Heritability estimates are calculated using GCTA and the Pedigree, Popkin, Standard ROM, and Standard MOR kinship estimates.

Since HCHS/SOL and SAMAFS are studies of Hispanics that overlap on 16 traits, we compared their heritability estimates. Despite the potential for heritability differences due to differences in environment or genetics, we find high Pearson correlations of ≈ 0.8 between the estimates of each method in both datasets (Fig. S8). We also observe slightly higher heritability estimates in SAMAFS compared to HCHS/SOL. However, for this small number of traits there are no clear differences in consistency between Popkin, Standard ROM and MOR estimates, whose regression lines all closely follow the $y = x$ line.

The NS dataset is a multiethnic dataset including large proportions of African, European, South Asian, and admixed ancestry individuals (Fig. S9). Its mean kinship value of $\bar{\varphi} = 0.124$ is the largest analyzed with real phenotypes, and just below that of 1000 Genomes. This dataset has three binary traits: each of NS and its subtypes SSNS (steroid sensitive NS) and SRNS (steroid resistant NS) contrasts individuals with those diseases to control samples. As before, Popkin estimates are larger than those of Standard ROM, although the differences are negligible. However, here Standard MOR yields considerably smaller heritability estimates than Popkin, a pattern that was not observed in the other datasets (Fig. 6). The proportion of rare variants is second highest here (0.56; Table 1), which likely explains the severe downward biases observed for Standard MOR. The visualized estimates are in the liability scale. We noticed that Popkin estimates in the observed scale were the maximum possible (0.999) for NS and SSNS (Table S1), yet liability scale estimates are less than 0.3, so we suggest interpreting these estimates with caution.

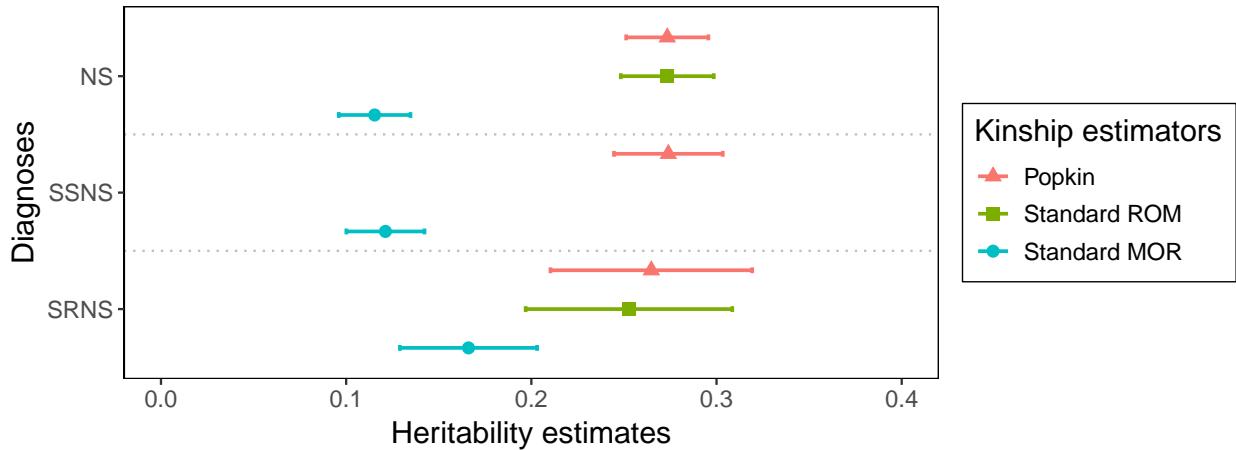


Figure 6: Heritability estimates on Nephrotic Syndrome (NS) dataset. The figure shows heritability estimates in the liability scale using Popkin ROM, Standard ROM, and Standard MOR kinship estimators on the NS multiethnic cohort. The binary traits are NS (NS vs control), SSNS (steroid sensitive NS subset vs control), and SRNS (steroid resistant NS subset vs control). Note the Standard ROM result for SSNS did not converge. For observed scale estimates, see Table S1.

4 Discussion

Previous research has shown that commonly used kinship estimators are biased when there is population structure (Ochoa and Storey, 2021). These kinship estimators are widely used to conduct association studies, directly in linear mixed-effects models or indirectly via PCA, models that are remarkably robust to these kinship estimation biases (Hou and Ochoa, 2023). In this study, we characterized the effect of kinship estimation bias on heritability estimation using variance components models, finding that this bias carries over to heritability estimation, in agreement with recent work (Chen and Storey, 2022). Our findings are summarized in Fig. 7. In the presence of population structure, the Standard ROM kinship estimator consistently underestimates heritability, with a closed-form bias that depends on the mean kinship value. The more widely used Standard MOR estimator lacks a closed-form limit and exhibits even more severe biases in practice, particularly influenced by the presence and effect sizes of rare variants. Pedigree-based kinship estimates tend to produce upward biases and higher variance under population structure. In contrast, the Popkin estimator consistently yields unbiased heritability estimates across all scenarios.

Rare variants can affect heritability estimation in two ways: by introducing biases in kinship estimation, and by having large effects as causal variants. Our evaluations demonstrate that the Standard MOR estimator has stronger biases than the Standard ROM estimator, but the direction of the bias depends on the properties of the rare variant effects. In particular, when rare variants are not causal or when they have small effect sizes, Standard MOR is more downwardly biased than Standard ROM, which is itself always downwardly biased. In contrast, when there is a large

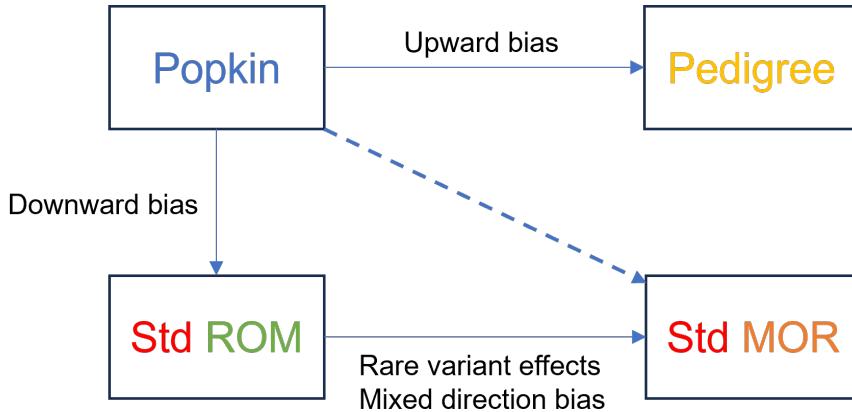


Figure 7: **Relationships between kinship estimators in terms of their effects on heritability estimation.** Under population structure, Popkin remains unbiased, Standard ROM shows consistent downward bias, and pedigree-based estimates are upwardly biased. Standard MOR can exhibit either downward or upward bias, depending on the proportion of rare variants in kinship estimation or among causal variants.

proportion of rare causal variants and they have large effect sizes, the Standard MOR has an upward bias. Our hypothesis is that this behavior is due to Standard MOR upweighing rare variants compared to the other two estimators. In HCHS/SOL, many traits appear to have an architecture where rare variants have significant effects, since Popkin's heritability estimates are lower than those of Standard MOR. However, it remains an open challenge to quantitatively determine the direction and magnitude of the heritability estimation bias due to using the Standard MOR kinship estimator.

Although we drew firm conclusions about the effects of rare variants on heritability estimation across kinship estimators, our work has limitations in how rare variants and their effect sizes are simulated. Our admixture and family simulations begin from pre-existing ancestral variation and do not introduce new mutations, so they primarily generate common variants. To simulate traits influenced by more realistic rare variants, we utilized the genotypes from the 1000 Genomes Project. When rare variants were causal under the FES model, all kinship estimators underestimated heritability (Fig. 3), whereas this bias was not present under the RC model (Fig. S5). This discrepancy reflects a limitation of the FES trait simulation algorithm when applied to real genotypes, for which true ancestral allele frequencies are unknown, leading to upwardly biased estimates of the inverse of genotype variance and misspecified heritability. Overall, estimation of heritability due to rare variants remains a theoretically underdeveloped area, necessitating further research.

Our model has another limitation: it assumes that rare and common variants have the same covariance structure, although there is evidence to the contrary (Zaidi and Mathieson, 2020). In our 1000 Genomes simulation, we applied different MAF thresholds to causal variants and to kinship estimation, allowing mismatches in terms of rare variant content between the estimated kinship matrix and the true trait covariance. Popkin is clearly unbiased when causal variants are drawn from the same set used to estimate kinship (diagonal cases in Fig. S5), while we observed small biases for mismatched causal variants and kinship matrices (off-diagonal cases in the figure). Thus, our data supports the hypothesis that rare and common variants have somewhat different structures, although the resulting bias due to ignoring their distinction was relatively small compared to the biases resulting from using the more common Standard MOR estimator.

Pedigrees are available for the San Antonio Family Study (SAMAFS), allowing us to estimate heritability using the pedigree in a manner that more closely resembles classic family studies (Almasy and Blangero, 1998b) as well as more recent approaches from large EHR or insurance claim data (Polubriaginof et al., 2018; Wang et al., 2017b). Notably, SAMAFS was originally analyzed using the pedigree to partition phenotypic variance more precisely than our present analysis (Mitchell et al., 1996). However, our simulation results indicate that these heritability estimates are upwardly biased when there is population structure, since the kinship estimate is missing the additional relatedness due to shared ancestry. In practice, incomplete pedigrees also result in unmodeled cryptic relatedness. Further, studies based on close relatives may overestimate heritability due to epistasis effects (Hemani et al., 2013; Young and Durbin, 2014). Supporting this, heritability estimates from close relatives are often inflated due to shared environmental effects, further highlighting the

limitations of pedigree-based methods (Zaitlen et al., 2013).

In the Nephrotic Syndrome (NS) multiethnic cohort, traits are binary (case vs. control). In our study, although the liability-scale heritability estimates ranged from 0.1 to 0.3 after transformation, the observed heritability estimates for NS and SSNS using the popkin estimator were 0.99 (Table S1), which occurred because σ_e^2 converged to nearly zero. Liability models rely on untestable assumptions (e.g., multivariate normality of latent liability) that, when violated, can produce severely biased estimates (Benck and Morris, 2013). Our results extend these concerns, suggesting that even when liability-scale estimates appear plausible, they may mask extreme variance components. In addition, it is known that NS, and SSNS in particular, have a single large effect locus in the HLA region, which affects the assumptions of heritability estimation using variance components (Gbadegesin et al., 2015; Adeyemo et al., 2018; Debiec et al., 2018; Jia et al., 2018; Dufek et al., 2019; Jia et al., 2020; Barry et al., 2023). Nevertheless, this data also serves as another illustration of the heritability biases present in the Standard estimators compared to Popkin.

Overall, our study highlights the importance of selecting appropriate kinship estimators for heritability analysis, particularly in structured populations and in the presence of rare variants. These findings provide key insights into the biases inherent in existing kinship estimators and underscore the need for future research to refine estimation methods for more accurate and reliable heritability inference across diverse genetic datasets.

Acknowledgments

This work was funded in part by the Duke University School of Medicine Whitehead Scholars Program, a gift from the Whitehead Charitable Foundation. Thanks to Rasheed Gbadegesin for sharing his NS data. Thanks to Tiffany Tu for sharing processed data and derivatives for the NS data, and a trait transformation analysis for SAMAFS and HCHS/SOL datasets. Thanks to Grace Rhodes for developing the maximum likelihood estimator of allele frequency used for fixed effect size simulation for real genotype data.

The 1000 Genomes data were generated at the New York Genome Center with funds provided by NHGRI Grant 3UM1HG008901-03S1.

The Hispanic Community Health Study/Study of Latinos is supported by contracts from the National Heart, Lung, and Blood Institute (NHLBI) to the University of North Carolina at Chapel Hill, Chapel Hill, NC (N01-HC65233), University of Miami, Miami, FL (N01-HC65234), Albert Einstein College of Medicine, Bronx NY (N01-HC65235), University of Illinois, Chicago IL (N01-HC65236), and San Diego State University, San Diego CA (N01-HC65237). The following Institutes/Centers/Offices contribute to the HCHS/SOL through a transfer of funds to the NHLBI: National Center on Minority Health and Health Disparities, the National Institute of Deafness and Other Communications Disorders, the National Institute of Dental and Craniofacial Research, the National Institute of Diabetes and Digestive and Kidney Diseases, The National Institute of Neu-

rological Disorders and Stroke, and the Office of Dietary Supplements. The authors thank the staff and participants of the HCHS/SOL study for their important contributions.

The research reported in this article was supported by National Institutes of Health grants. The genetic and phenotypic data were provided by the San Antonio Family Heart Study (SAFHS) investigators and supported by the National Heart, Lung, and Blood Institute (NHLBI) [P01 HL045222] and the San Antonio Family Diabetes/Gallbladder Study (SAFDGS) investigators and supported by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) [R01 DK047482, R01 DK053889]. The phenotypic data were also provided by the Veterans Administration Genetic Epidemiology Study (VAGES) investigators and supported by the Health Services Research and Development, U.S., Department of Veteran Affairs and the Family Investigation Acknowledgement Statement : of Nephropathy and Diabetes (FIND) - San Antonio (FIND-SA) Component and its extension called the Extended FIND (E-FIND) [U01 DK57295] investigators and supported by the NIDDK. The SAFHS gene expression assays were supported by a donation from the Azar and Shepperd families. The exome sequencing, exome chip genotypic, and whole genome sequencing data were provided by the T2D-GENES Consortium grants U01 DK085524, U01 DK085584, U01 DK085501, U01 DK085526, and U01 DK085545 and investigators supported by the NIDDK. This manuscript was not prepared in collaboration with investigators of the T2D-GENES SAMAFS/Consortium and does not necessarily reflect the opinions or views of the members of the T2D-GENES SAMAFS/Consortium, or the NIDDK

Competing interests

The authors declare no competing interests.

Data and code availability

The data and code generated during this study are available on GitHub at <https://github.com/OchoaLab/bias-herit-paper>

References

- Adeyemo, Adebowale et al. (2018). "HLA-DQA1 and APOL1 as risk loci for childhood-onset steroid-sensitive and steroid-resistant nephrotic syndrome". *American Journal of Kidney Diseases* 71(3), pp. 399–406.
- Almasy, L. and J. Blangero (1998a). "Multipoint quantitative-trait linkage analysis in general pedigrees". *Am. J. Hum. Genet.* 62(5), pp. 1198–1211. DOI: 10.1086/301844.
- Almasy, Laura and John Blangero (1998b). "Multipoint quantitative-trait linkage analysis in general pedigrees". *The American Journal of Human Genetics* 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. DOI: 10.1214/09-STS307.
- 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. DOI: <https://doi.org/10.1007/BF01441146>.
- Barry, Alexandra et al. (2023). "Multi-population genome-wide association study implicates immune and non-immune factors in pediatric steroid-sensitive nephrotic syndrome". *Nature communications* 14(1), p. 2481.
- Benchek, Penny H and Nathan J Morris (2013). "How meaningful are heritability estimates of liability?" *Human genetics* 132, pp. 1351–1360.
- Bhatia, Gaurav et al. (2013). "Estimating and interpreting FST: the impact of rare variants". *Genome Res.* 23(9), pp. 1514–1521. DOI: 10.1101/gr.154831.113.
- Biddanda, Arjun, Daniel P Rice, and John Novembre (2020). "A variant-centric perspective on geographic patterns of human allele frequency variation". *Elife* 9, e60107.
- 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.
- Chang, Christopher C. et al. (2015). "Second-generation PLINK: rising to the challenge of larger and richer datasets". *GigaScience* 4(1), p. 7. DOI: 10.1186/s13742-015-0047-8.
- Charney, Evan (2012). "Behavior genetics and postgenomics". *Behavioral and brain sciences* 35(5), pp. 331–358.
- Chen, Danfeng and John D. Storey (2022). "How Kinship Estimation Bias Propagates to Heritability". *The 72nd Annual Meeting of The American Society of Human Genetics*.
- 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.
- Debiec, Hanna et al. (2018). "Transethnic, genome-wide analysis reveals immune-related risk alleles and phenotypic correlates in pediatric steroid-sensitive nephrotic syndrome". *Journal of the American Society of Nephrology* 29(7), pp. 2000–2013.

- Dufek, Stephanie et al. (2019). "Genetic identification of two novel loci associated with steroid-sensitive nephrotic syndrome". *Journal of the American Society of Nephrology* 30(8), pp. 1375–1384.
- 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. DOI: 10.1093/nar/gkz836.
- Falconer, Douglas Scott (1996). *Introduction to quantitative genetics*. Pearson Education India.
- Gazal, Steven et al. (2018). "Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations". *Nature genetics* 50(11), pp. 1600–1607.
- Gbadegesin, Rasheed A et al. (2015). "HLA-DQA1 and PLCG2 are candidate risk loci for childhood-onset steroid-sensitive nephrotic syndrome". *Journal of the American Society of Nephrology* 26(7), pp. 1701–1710.
- Hemani, Gibran, Sara Knott, and Chris Haley (2013). "An evolutionary perspective on epistasis and the missing heritability". *PLoS genetics* 9(2), e1003295.
- Hou, Kangcheng et al. (2019). "Accurate estimation of SNP-heritability from biobank-scale data irrespective of genetic architecture". *Nature genetics* 51(8), pp. 1244–1251.
- Hou, Zhuoran and Alejandro Ochoa (2023). "Genetic association models are robust to common population kinship estimation biases". *Genetics* 224(1), iyad030.
- Huang, Jinguo et al. (2025). "Interpreting SNP heritability in admixed populations". *bioRxiv*. DOI: 10.1101/2023.08.04.551959.
- Jacquard, Albert (1970). *Structures génétiques des populations*. Paris: Masson et Cie.
- Jia, Xiaoyuan et al. (2018). "Strong association of the HLA-DR/DQ locus with childhood steroid-sensitive nephrotic syndrome in the Japanese population". *Journal of the American Society of Nephrology* 29(8), pp. 2189–2199.
- Jia, Xiaoyuan et al. (2020). "Common risk variants in NPHS1 and TNFSF15 are associated with childhood steroid-sensitive nephrotic syndrome". *Kidney international* 98(5), pp. 1308–1322.
- Lee, Sang Hong et al. (2011). "Estimating missing heritability for disease from genome-wide association studies". *Am. J. Hum. Genet.* 88(3), pp. 294–305. DOI: 10.1016/j.ajhg.2011.02.002.
- Loh, Po-Ru et al. (2015). "Efficient Bayesian mixed-model analysis increases association power in large cohorts". *Nat. Genet.* 47(3), pp. 284–290. DOI: 10.1038/ng.3190.
- 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. DOI: 10.1371/journal.pgen.1009241.
- Polubriaginof, Fernanda CG et al. (2018). "Disease heritability inferred from familial relationships reported in medical records". *Cell* 173(7), pp. 1692–1704.
- Price, Alkes L. et al. (2006). "Principal components analysis corrects for stratification in genome-wide association studies". *Nat. Genet.* 38(8), pp. 904–909. DOI: 10.1038/ng1847.
- Price, Alkes L et al. (2010). "New approaches to population stratification in genome-wide association studies". *Nature reviews genetics* 11(7), pp. 459–463.
- 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. DOI: 10.1371/journal.pone.0005825.
- Sinnwell, Jason P., Terry M. Therneau, and Daniel J. Schaid (29, 2014). "The kinship2 R Package for Pedigree Data". *Human Heredity* 78(2), pp. 91–93. DOI: 10.1159/000363105.
- 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. DOI: 10.1038/nrg3821.
- 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. (2012a). "Improved heritability estimation from genome-wide SNPs". *The American Journal of Human Genetics* 91(6), pp. 1011–1021.
- Speed, Doug et al. (2012b). "Improved heritability estimation from genome-wide SNPs". *Am. J. Hum. Genet.* 91(6), pp. 1011–1021. DOI: 10.1016/j.ajhg.2012.10.010.
- Speed, Doug et al. (2017a). "Reevaluation of SNP heritability in complex human traits". *Nature genetics* 49(7), pp. 986–992.
- Speed, Doug et al. (2017b). "Reevaluation of SNP heritability in complex human traits". *Nat Genet* 49(7), pp. 986–992. DOI: 10.1038/ng.3865.
- 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. DOI: 10.1371/journal.pgen.1007309.
- Tenesa, Albert and Chris S Haley (2013). "The heritability of human disease: estimation, uses and abuses". *Nature Reviews Genetics* 14(2), pp. 139–149.
- 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. DOI: 10.1016/j.ajhg.2010.01.001.

- 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.
- Wainschtein, Pierrick et al. (2022). “Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data”. *Nature genetics* 54(3), pp. 263–273.
- Wang, Bowen, Serge Sverdlov, and Elizabeth Thompson (2017a). “Efficient Estimation of Realized Kinship from SNP Genotypes”. *Genetics, genetics*.116.197004. DOI: 10.1534/genetics.116.197004.
- Wang, Kanix et al. (2017b). “Classification of common human diseases derived from shared genetic and environmental determinants”. *Nature genetics* 49(9), pp. 1319–1325.
- Wright, Sewall (1949). “The Genetical Structure of Populations”. *Annals of Eugenics* 15(1), pp. 323–354. DOI: 10.1111/j.1469-1809.1949.tb02451.x.
- Yang, Jian et al. (2010). “Common SNPs explain a large proportion of the heritability for human height”. *Nat. Genet.* 42(7), pp. 565–569. DOI: 10.1038/ng.608.
- 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.
- Yang, Jian et al. (2017). “Concepts, estimation and interpretation of SNP-based heritability”. *Nature genetics* 49(9), pp. 1304–1310.
- Yao, Yiqi and Alejandro Ochoa (2022). “Limitations of principal components in quantitative genetic association models for human studies”, p. 2022.03.25.485885. DOI: 10.1101/2022.03.25.485885.
- Young, Alexander I and Richard Durbin (2014). “Estimation of epistatic variance components and heritability in founder populations and crosses”. *Genetics* 198(4), pp. 1405–1416.
- Zaidi, Arslan A and Iain Mathieson (2020). “Demographic history mediates the effect of stratification on polygenic scores”. *eLife* 9. Ed. by George H Perry, Michael C Turchin, and Alicia R Martin, e61548. DOI: 10.7554/eLife.61548.
- Zaitlen, Noah et al. (2013). “Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits”. *PLoS genetics* 9(5), e1003520.
- Zhou, Xiang and Matthew Stephens (2012). “Genome-wide efficient mixed-model analysis for association studies”. *Nat. Genet.* 44(7), pp. 821–824. DOI: 10.1038/ng.2310.

Appendices

A Variance of the plug-in Binomial variance estimator $\hat{p}_i(1 - \hat{p}_i)$

To explain the inconsistency of the FES variance estimator, as well as the Standard MOR kinship estimator, we calculate the variance of $\hat{p}_i(1 - \hat{p}_i)$ shown in the main Methods, where $\hat{p}_i = \frac{1}{2n} \mathbf{1}'_n \mathbf{x}_i$ is the empirical allele frequency estimated from genotypes \mathbf{x}_i .

We begin by rewriting the quantity of interest as a bilinear form:

$$\hat{p}_i(1 - \hat{p}_i) = \mathbf{x}'_i \mathbf{A} \mathbf{y}_i, \quad \mathbf{A} = \frac{\mathbf{1}'_n \mathbf{1}_n'}{4n^2}, \quad \mathbf{y}_i = 2\mathbf{1}_n - \mathbf{x}_i.$$

This allows us to use a known identity to complete the variance calculation, which assumes multi-variate normal distributions for \mathbf{x}_i and \mathbf{y}_i and a symmetric \mathbf{A} , namely

$$\text{Var}(\mathbf{x}'_i \mathbf{A} \mathbf{y}_i) = \boldsymbol{\mu}'_x \mathbf{A} \boldsymbol{\Sigma}_y \mathbf{A} \boldsymbol{\mu}_x + \boldsymbol{\mu}'_y \mathbf{A} \boldsymbol{\Sigma}_x \mathbf{A} \boldsymbol{\mu}_y + 2\boldsymbol{\mu}'_x \mathbf{A} \boldsymbol{\Sigma}_{yx} \mathbf{A} \boldsymbol{\mu}_y + \text{Tr}(\mathbf{A} \boldsymbol{\Sigma}_x \mathbf{A} \boldsymbol{\Sigma}_y) + \text{Tr}(\mathbf{A} \boldsymbol{\Sigma}_{xy} \mathbf{A} \boldsymbol{\Sigma}_{xy}),$$

where $\boldsymbol{\mu}_x, \boldsymbol{\mu}_y$ are expectation vectors, and $\boldsymbol{\Sigma}_x, \boldsymbol{\Sigma}_y, \boldsymbol{\Sigma}_{xy}$ are covariance and cross-covariance matrices, respectively. Thus, we assume a normal version of the kinship model of Eq. (1), namely

$$\mathbf{x}_i \sim \mathcal{N}(2p_i \mathbf{1}_n, 4p_i(1 - p_i)\boldsymbol{\Phi}),$$

which is expected to result in a more accurate calculation for p_i near 0.5 (high MAF). It follows directly from these assumptions that \mathbf{y}_i is also normal and has moments

$$\begin{aligned} \boldsymbol{\mu}_y &= \text{E}[\mathbf{y}_i] = 2(1 - p_i)\mathbf{1}_n, \\ \boldsymbol{\Sigma}_y &= \text{Cov}(\mathbf{y}_i) = 4p_i(1 - p_i)\boldsymbol{\Phi}, \\ \boldsymbol{\Sigma}_{xy} &= \text{Cov}(\mathbf{x}_i, \mathbf{y}_i) = -4p_i(1 - p_i)\boldsymbol{\Phi}. \end{aligned}$$

Substituting in, gathering like terms and simplifying, we obtain that:

$$\text{Var}(\hat{p}_i(1 - \hat{p}_i)) = 16p_i(1 - p_i) \left((1 - 2p_i)^2 \mathbf{1}'_n \mathbf{A} \boldsymbol{\Phi} \mathbf{A} \mathbf{1}_n + 2p_i(1 - p_i) \text{Tr}(\mathbf{A} \boldsymbol{\Phi} \mathbf{A} \boldsymbol{\Phi}) \right).$$

Since $\mathbf{A} = \frac{\mathbf{1}'_n \mathbf{1}_n'}{4n^2}$ and $\frac{\mathbf{1}'_n \boldsymbol{\Phi} \mathbf{1}_n}{n^2} = \bar{\varphi}$, and using the cyclic property of traces, the following terms simplify:

$$\begin{aligned} \mathbf{1}'_n \mathbf{A} \boldsymbol{\Phi} \mathbf{A} \mathbf{1}_n &= \frac{1}{(4n^2)^2} \cdot \mathbf{1}'_n \mathbf{1}_n \cdot \mathbf{1}'_n \boldsymbol{\Phi} \mathbf{1}_n \cdot \mathbf{1}'_n \mathbf{1}_n = \frac{n \cdot (\mathbf{1}'_n \boldsymbol{\Phi} \mathbf{1}_n) \cdot n}{16n^4} = \frac{\bar{\varphi}}{16}, \\ \text{Tr}(\mathbf{A} \boldsymbol{\Phi} \mathbf{A} \boldsymbol{\Phi}) &= \frac{1}{(4n^2)^2} \text{Tr}(\mathbf{1}_n \mathbf{1}'_n \boldsymbol{\Phi} \mathbf{1}_n \mathbf{1}'_n \boldsymbol{\Phi}) = \frac{1}{16n^4} \mathbf{1}'_n \boldsymbol{\Phi} \mathbf{1}_n \mathbf{1}'_n \boldsymbol{\Phi} \mathbf{1}_n = \frac{\bar{\varphi}^2}{16}, \end{aligned}$$

Thus, the final variance equation is:

$$\text{Var}(\hat{p}_i(1 - \hat{p}_i)) = p_i(1 - p_i)\bar{\varphi} \left((1 - 2p_i)^2 + 2p_i(1 - p_i)\bar{\varphi} \right).$$

Note that even though $\bar{\varphi}$ is defined as a sample parameter (for finite n), its value will converge to the population value as n grows, and this value will be non-zero when there is population structure.

Note that

$$\text{Var}(\hat{p}_i(1 - \hat{p}_i)) \leq p_i(1 - p_i)\bar{\varphi}, \quad (13)$$

since the second factor $(1 - 2p_i)^2 + 2p_i(1 - p_i)\bar{\varphi}$ is a quadratic on p_i with values of 1 at both edge points ($p_i = 0$ and $p_i = 1$) and a minimum at $p_i = 1/2$ that evaluates to $\bar{\varphi}/2$, so the whole second factor is non-negative and bounded above by 1. Thus, the inequality is tighter near the edges, and looser in the middle (near $p_i = 1/2$).

B Consistency of the RC initial variance estimator $\hat{\sigma}_{g0}^2$ and the final variance $\hat{\sigma}_g^2$

Lemma 1. Let $\hat{\sigma}_{g0}^2 := \sum_{i=1}^{m_1} \frac{2\hat{p}_i(1-\hat{p}_i)}{1-\bar{\varphi}} \cdot \beta_{i0}^2$, where $\beta_{i0} \sim \mathcal{N}(0, 1)$ i.i.d., and \hat{p}_i are allele frequency estimates satisfying:

$$\begin{aligned} \mathbb{E}[\hat{p}_i(1 - \hat{p}_i)] &= p_i(1 - p_i)(1 - \bar{\varphi}), \\ \text{Var}(\hat{p}_i(1 - \hat{p}_i)) &\leq p_i(1 - p_i)\bar{\varphi}, \end{aligned}$$

with fixed $p_i \in [\delta, 1 - \delta]$ for some $\delta > 0$, and each term in the sum is independent across i . Define the target initial genetic variance component as

$$\sigma_{g0}^2 := \sum_{i=1}^{m_1} 2p_i(1 - p_i).$$

Then, as $m_1 \rightarrow \infty$,

$$\frac{\hat{\sigma}_{g0}^2}{\sigma_{g0}^2} \xrightarrow{P} 1.$$

That is, $\hat{\sigma}_{g0}^2$ is a consistent estimator of σ_{g0}^2 .

Proof. Define $A_i := \frac{2\hat{p}_i(1-\hat{p}_i)}{1-\bar{\varphi}}$, so that $\hat{\sigma}_{g0}^2 = \sum_{i=1}^{m_1} A_i \cdot \beta_{i0}^2$. Since $\beta_{i0} \sim \mathcal{N}(0, 1)$ and are independent of A_i , we have:

$$\begin{aligned} \mathbb{E}[A_i \cdot \beta_{i0}^2] &= 2p_i(1 - p_i), \\ \text{Var}(A_i \cdot \beta_{i0}^2) &= 2\mathbb{E}[A_i]^2 + 3\text{Var}(A_i), \end{aligned}$$

because $\mathbb{E}[\beta_{i0}^2] = 1$ and $\text{Var}(\beta_{i0}^2) = 2$. Then $\mathbb{E}[\hat{\sigma}_{g0}^2] = \sigma_{g0}^2$.

Therefore:

$$\text{Var}\left(\frac{\hat{\sigma}_{g0}^2}{\sigma_{g0}^2}\right) = \frac{1}{\sigma_{g0}^4} \sum_{i=1}^{m_1} (2\mathbb{E}[A_i]^2 + 3\text{Var}(A_i)).$$

It follows from our assumptions that:

$$\mathbb{E}[A_i]^2 = \left(\frac{2p_i(1-p_i)(1-\bar{\varphi})}{1-\bar{\varphi}} \right)^2 = 4p_i^2(1-p_i)^2 \leq \frac{1}{4},$$

$$\text{Var}(A_i) = \frac{4 \text{Var}(\hat{p}_i(1-\hat{p}_i))}{(1-\bar{\varphi})^2} \leq C_1 \cdot p_i(1-p_i),$$

for constants C_1 depending on $\bar{\varphi}$.

Thus, for some constant C_2 :

$$\sum_{i=1}^{m_1} (2\mathbb{E}[A_i]^2 + 3\text{Var}(A_i)) \leq C_2 \sum_{i=1}^{m_1} p_i(1-p_i) \leq \frac{C_2}{4} \cdot m_1,$$

since $p_i(1-p_i) \leq 1/4$ for all i , the right-hand side grows at most linearly with m_1 , and hence the numerator of the variance expression increases at most proportionally to the number of causal variants m_1 .

Under the assumption that all $p_i \in [\delta, 1-\delta]$, each term $p_i(1-p_i) \geq \delta(1-\delta)$, so the initial variance satisfies

$$\sigma_{g0}^2 = \sum_{i=1}^{m_1} 2p_i(1-p_i) \geq c \cdot m_1 \quad \text{for some constant } c > 0.$$

Therefore, as $m_1 \rightarrow \infty$:

$$\text{Var}\left(\frac{\hat{\sigma}_{g0}^2}{\sigma_{g0}^2}\right) \leq \frac{C_2 \cdot m_1}{4\sigma_{g0}^4} \leq \frac{C_2 \cdot m_1}{4c^2 \cdot m_1^2} = \frac{C_2}{4c^2 \cdot m_1} \rightarrow 0.$$

Therefore, by Chebyshev's inequality,

$$\frac{\hat{\sigma}_{g0}^2}{\sigma_{g0}^2} \xrightarrow{P} 1.$$

□

Remark. When causal variants have small minor allele frequencies (i.e., small p_i), the terms $p_i(1-p_i)$ become small, reducing the denominator σ_{g0}^2 . This leads to a looser (i.e., larger) upper bound in the variance of $\hat{\sigma}_{g0}^2/\sigma_{g0}^2$, and thus slower convergence in probability.

Theorem 1 (Consistency of Final Genetic Variance). *Let $\hat{\beta}_i := \beta_{i0} \cdot \frac{h\sigma}{\sqrt{\hat{\sigma}_{g0}^2}}$ be the final standardized effect sizes constructed from i.i.d. draws $\beta_{i0} \sim \mathcal{N}(0, 1)$, and define the final genetic variance as*

$$\hat{\sigma}_g^2 := \sum_{i=1}^{m_1} 2p_i(1-p_i)\hat{\beta}_i^2.$$

Let the target variance be $\sigma_g^2 := h^2\sigma^2$. Then, under the assumptions of Lemma 1,

$$\hat{\sigma}_g^2 \xrightarrow{P} \sigma_g^2 \quad \text{as } m_1 \rightarrow \infty.$$

Proof. By Lemma 1, we have

$$\hat{\sigma}_{g0}^2 \xrightarrow{P} \sigma_{g0}^2.$$

Then,

$$\frac{h^2\sigma^2}{\hat{\sigma}_{g0}^2} \xrightarrow{P} \frac{h^2\sigma^2}{\sigma_{g0}^2},$$

by the continuous mapping theorem.

Substituting into the definition of $\hat{\sigma}_g^2$, we have:

$$\hat{\sigma}_g^2 = \sum_{i=1}^{m_1} 2p_i(1-p_i)\hat{\beta}_i^2 = \frac{h^2\sigma^2}{\hat{\sigma}_{g0}^2} \sum_{i=1}^{m_1} 2p_i(1-p_i)\beta_{i0}^2.$$

The sum $\sum_i 2p_i(1-p_i)\beta_{i0}^2$ is a consistent estimator of σ_{g0}^2 by construction, so

$$\hat{\sigma}_g^2 \xrightarrow{P} \frac{h^2\sigma^2}{\sigma_{g0}^2} \cdot \sigma_{g0}^2 = h^2\sigma^2 = \sigma_g^2.$$

Hence, $\hat{\sigma}_g^2 \xrightarrow{P} \sigma_g^2$, as claimed. □

S1 Supplemental Tables and Figures

Table S1: **Heritability Estimates for the NS multiethnic cohort.**

Diagnoses	Kinship Estimators	Observed Heritability (CI)	Liability-Scaled Heritability (CI)
NS	Popkin	0.999 (0.919–1.081)	0.273 (0.251–0.296)
NS	Standard ROM	0.999 (0.908–1.092)	0.273 (0.248–0.299)
NS	Standard MOR	0.422 (0.351–0.493)	0.115 (0.096–0.135)
SSNS	Popkin	0.999 (0.893–1.107)	0.274 (0.245–0.303)
SSNS*	Standard ROM	-	-
SSNS	Standard MOR	0.442 (0.365–0.519)	0.121 (0.100–0.142)
SRNS	Popkin	0.616 (0.490–0.743)	0.265 (0.210–0.319)
SRNS	Standard ROM	0.588 (0.458–0.718)	0.253 (0.197–0.308)
SRNS	Standard MOR	0.387 (0.300–0.473)	0.166 (0.129–0.203)

*GCTA algorithm does not converge.

NS: NS vs control; SSNS: steroid sensitive NS subset vs control; SRNS: steroid resistant NS subset vs control.

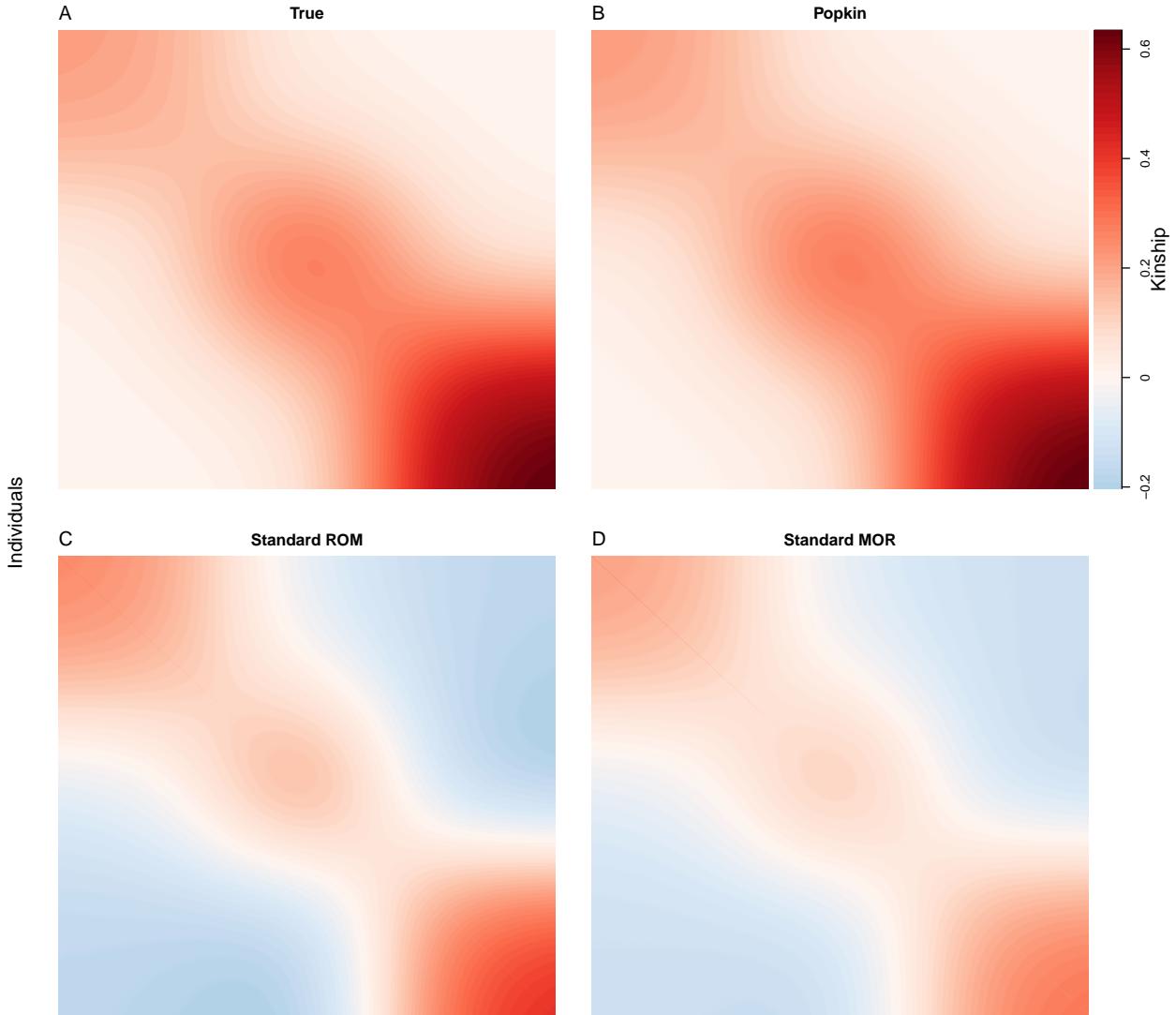


Figure S1: Kinship matrix estimates for the simulated admixture dataset. Estimates are calculated on the genotype matrix of the first replicate of this simulation. In each panel, individuals are placed along both x and y axes, and the kinship value φ_{jk} of a pair of individuals j and k is visualized as color, where values near zero appear white, large positive values are darker red, and large negative values are darker blue. The diagonal displays inbreeding values $f_j = 2\varphi_{jj} - 1$ rather than self-kinship values φ_{jj} because the inbreeding value of an individual equals the kinship of the parents, so it is on the same scale as kinship whereas self-kinship is not. **A.** True kinship matrix of the simulation, calculated from the admixture model parameters as in Ochoa and Storey (2021). **B.** Popkin estimate, which is unbiased. **C.** Standard ROM (ratio-of-means) estimate, which has a bias described in the Methods. **D.** Standard MOR (mean-of-ratios) estimate, which has a similar bias to Standard ROM but has additional biases that do not have a closed form; the difference is driven by rare variants, which are upweighted in the MOR version.

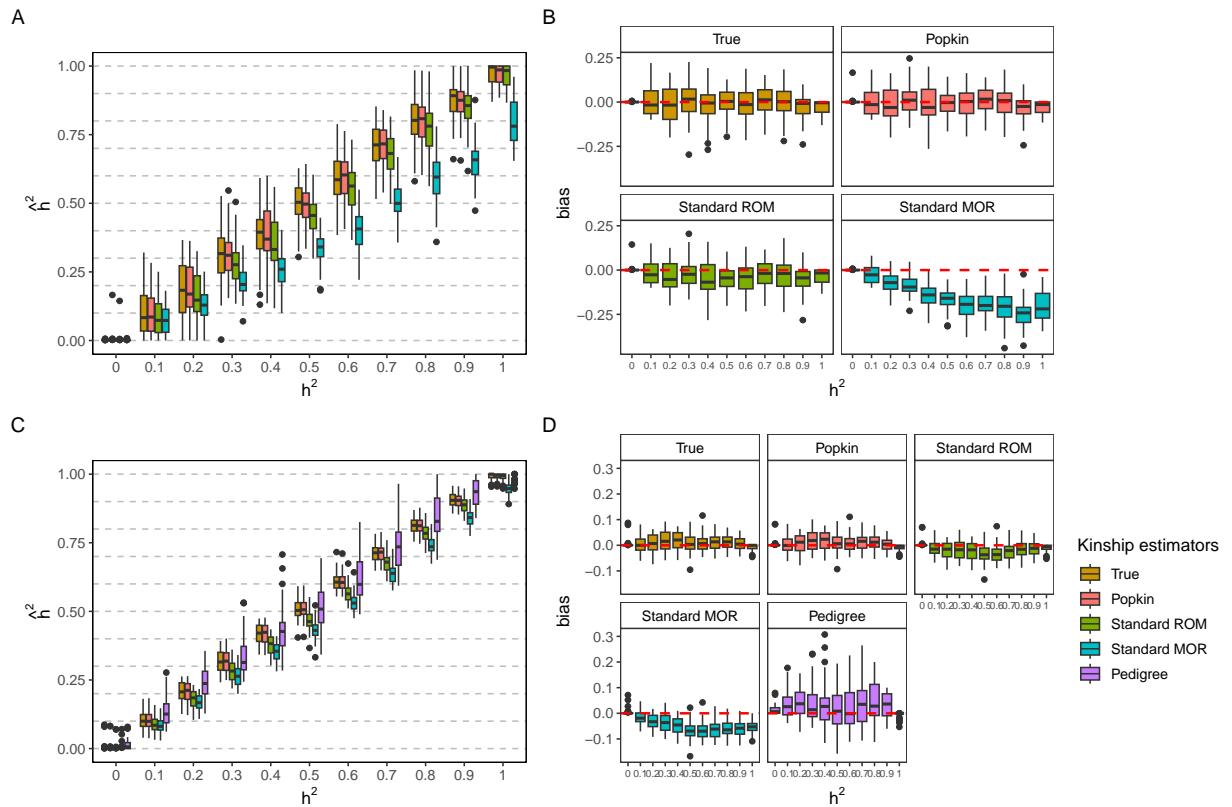


Figure S2: **Heritability estimation using GCTA with various kinship estimators evaluated using simulated genotypes and traits.** The FES trait model was used to simulate traits in this figure. **A-B.** Admixture simulation. **C-D.** Admixture plus 20 generations of family structure simulation. See Fig. 2 for more information.

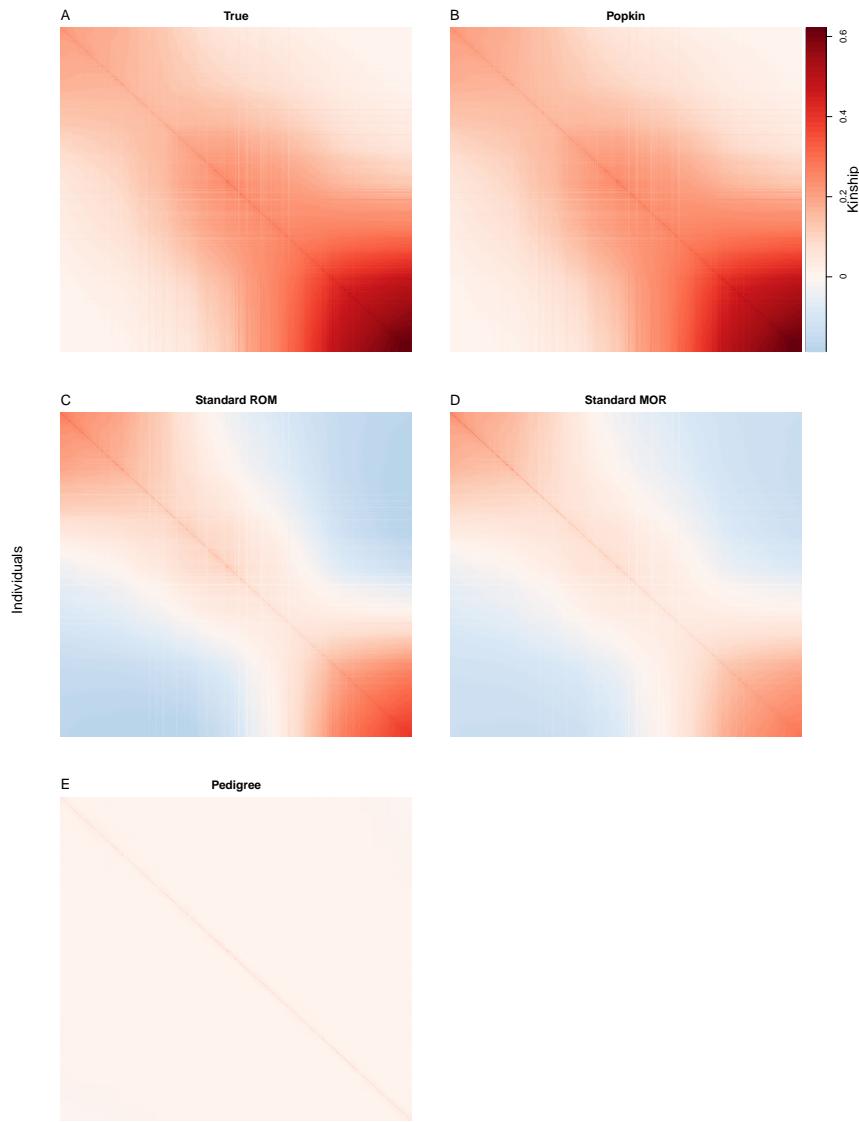


Figure S3: Kinship matrix estimates for the simulated admixture plus family structure dataset. Estimates are calculated on the genotype matrix of the first replicate of this simulation. The simulation reorders individuals so that families appear closer together along the diagonal (notice darker reds corresponding to higher kinship values). Darker or brighter lines in these figures correspond to individuals with more different ancestry than their neighbors on average (including family members such as spouses) which occur with some frequency in this simulation. See Fig. S1 for more details, including descriptions of panels, except: **E.** The kinship matrix calculated from the true pedigree of the simulation using standard methods, which erroneously treat founders as unrelated even though in this simulation they are admixed so their relatedness is as in Fig. S1A.

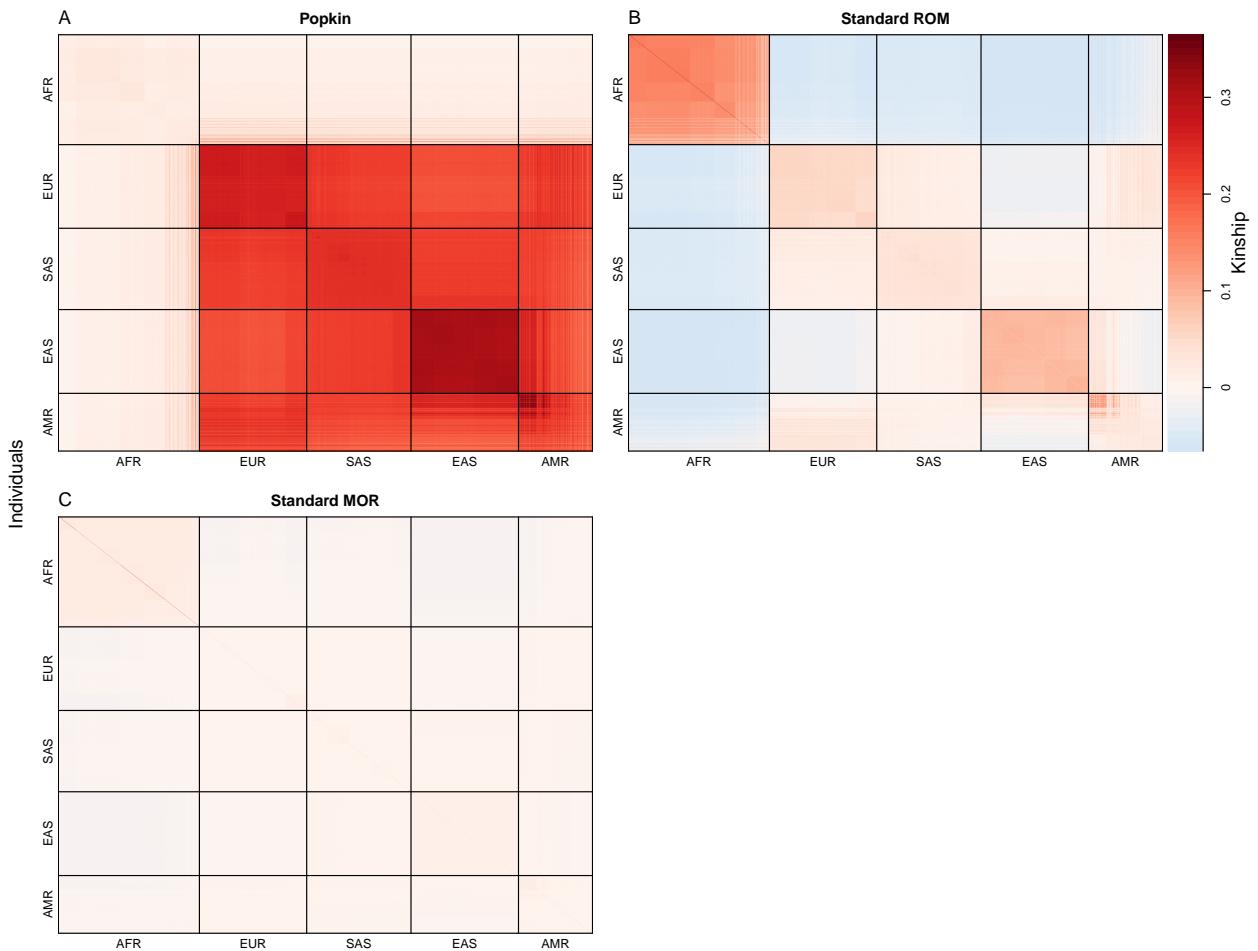


Figure S4: Kinship matrix estimates for the 1000 Genome dataset. AFR = African, EUR = European, SAS = South Asian, EAS = East Asian, AMR = Admixed Americans (Hispanics). See Fig. S1 for more details.

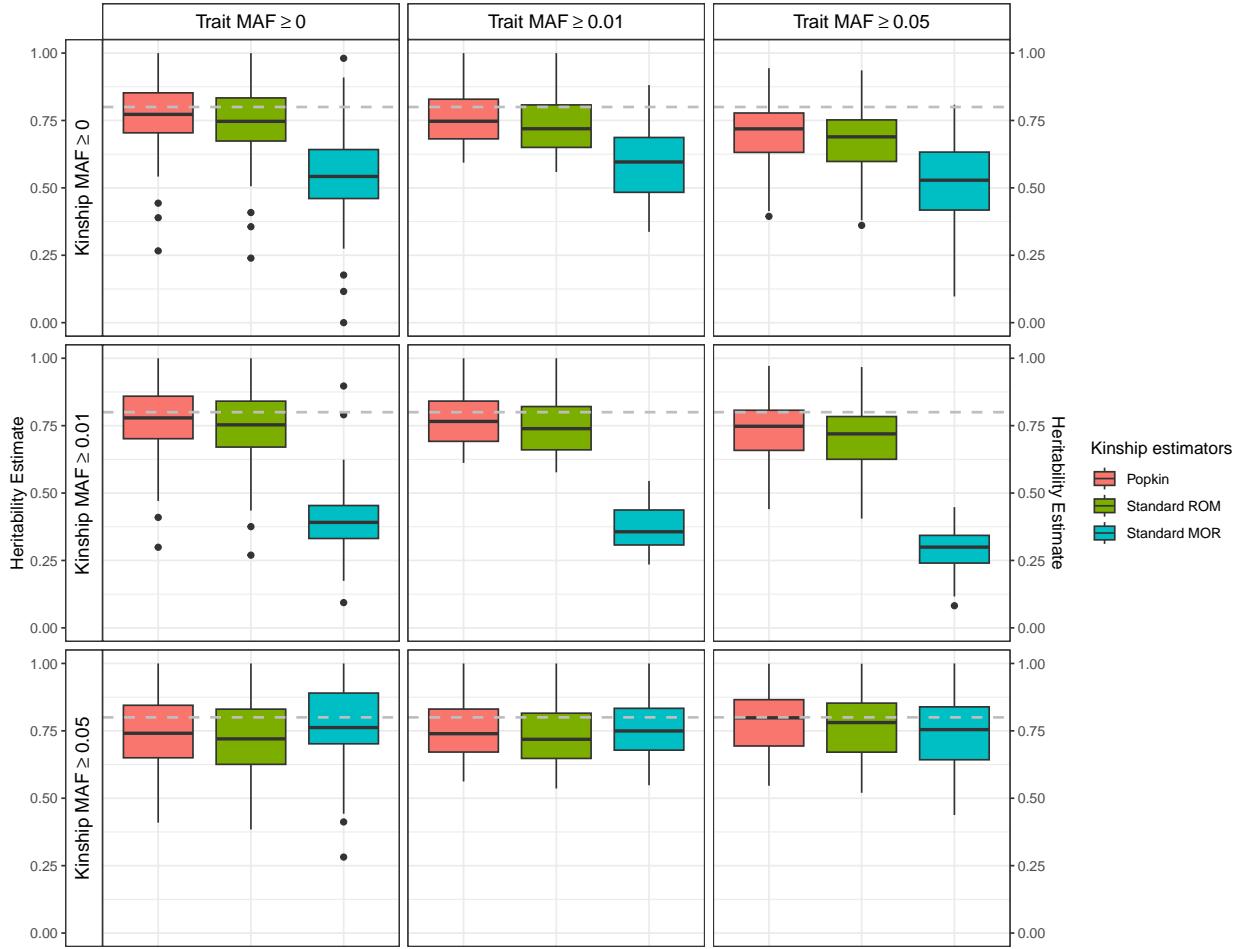


Figure S5: Testing effect of rare variants using 1000 Genomes genotypes and simulated traits. Trait simulations under the RC trait model, which assigns smaller effects to causal variants and always has a correctly specified heritability. The true heritability is $h^2 = 0.8$ in all panels. Each trait was simulated 50 times. For simulated traits, MAF thresholds shown are applied before causal variants are selected, thus influencing the architecture of the trait. For kinship estimates, MAF thresholds are applied to genotypes before kinship matrices are estimated from these genotypes, thus influencing only kinship estimation.

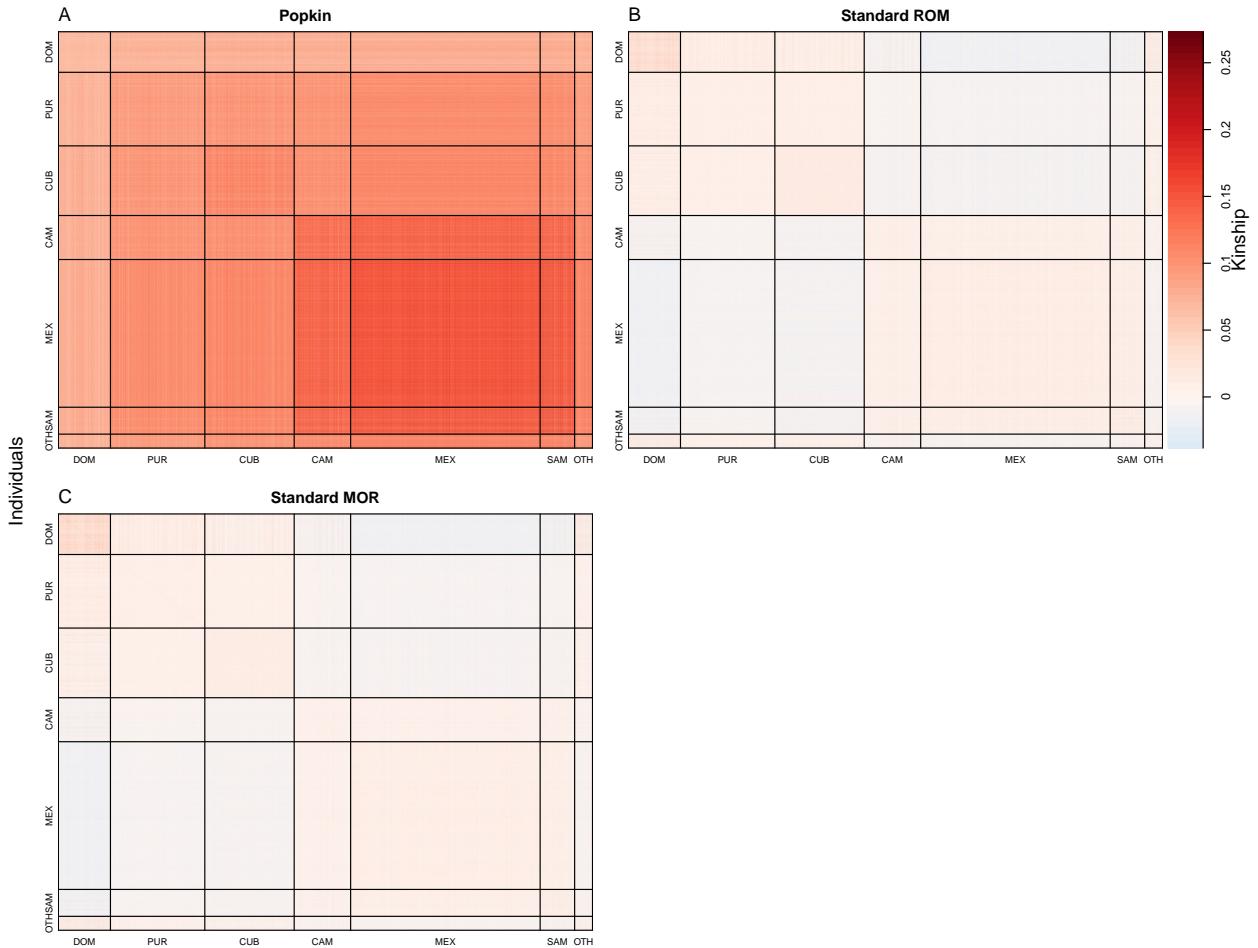


Figure S6: Kinship matrix estimates for the Hispanic Community Health Study / Study of Latinos dataset. DOM = Dominican, PUR = Puerto-Rican, CUB = Cuban, CAM = Central American, MEX = Mexican, SAM = South American, OTH = Other. See Fig. S1 for more details.

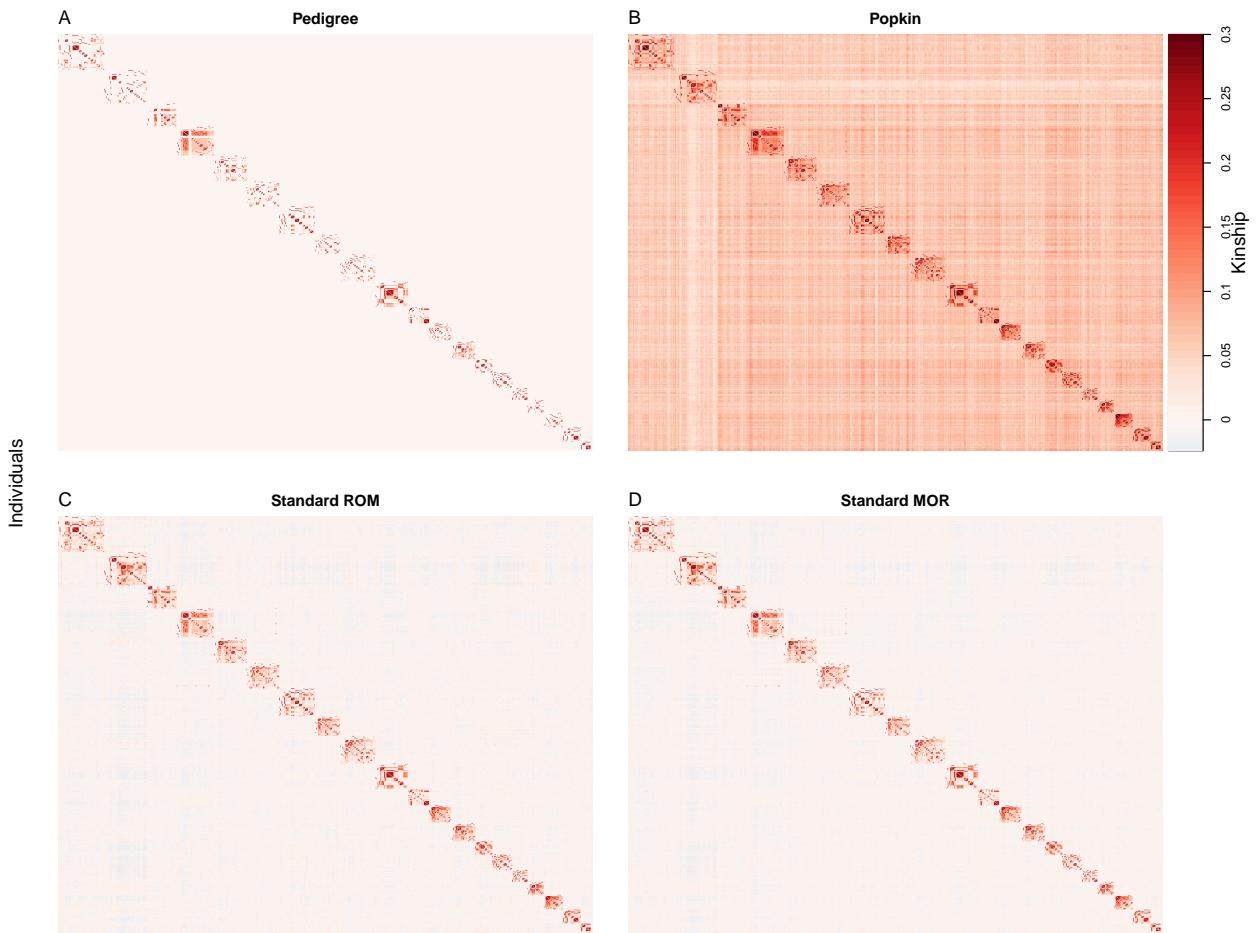


Figure S7: **Kinship matrix estimates for the San Antonio Family Study dataset.** See Fig. S1 for more details.

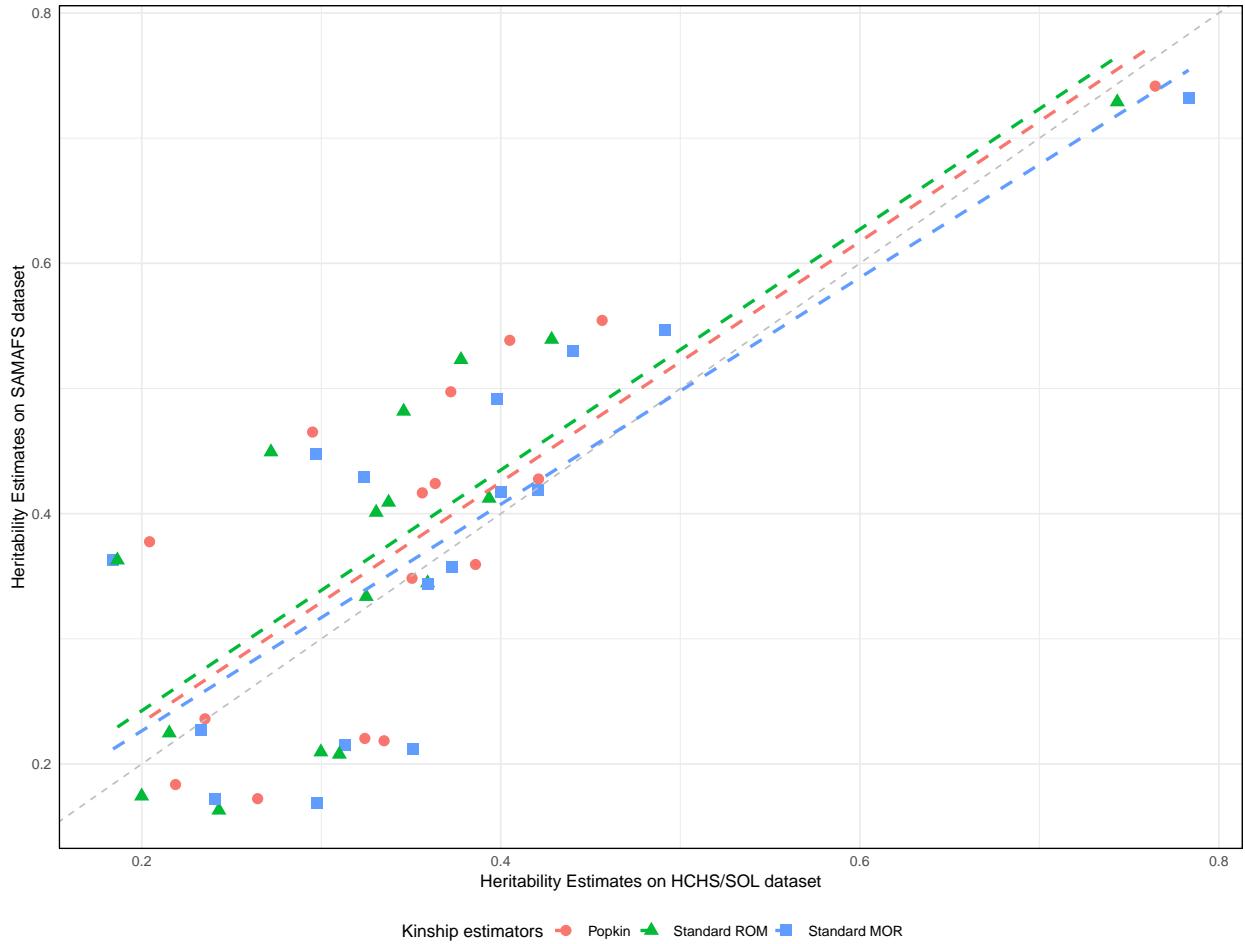


Figure S8: **Comparison of Heritability Estimates Between HCHS/SOL and SAMAFS datasets.** 16 overlapping traits are included in the analysis. Pearson correlation coefficients for each kinship estimator are: 0.799 for Popkin, 0.802 for Standard ROM, and 0.794 for Standard MOR.

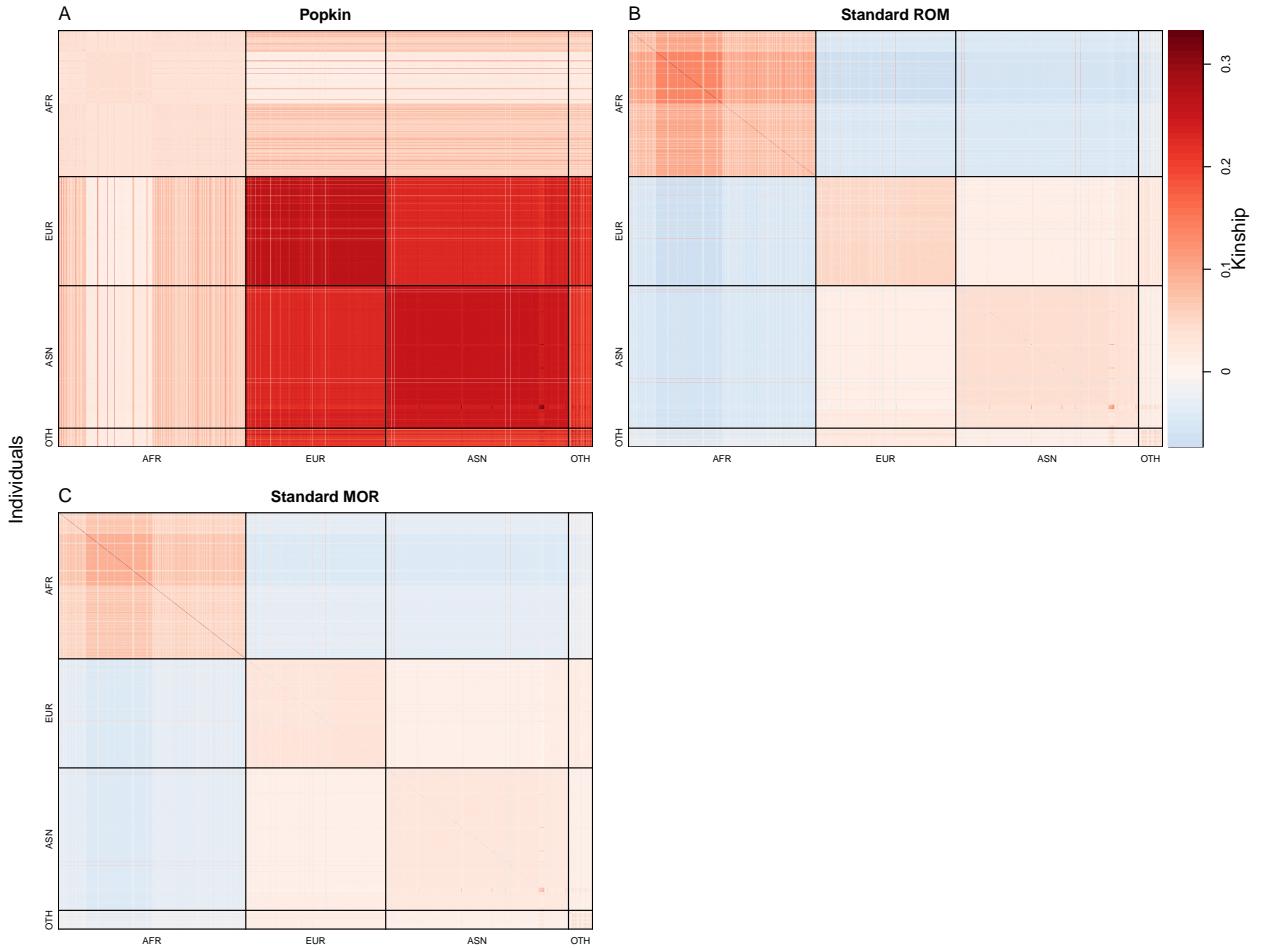


Figure S9: Kinship matrix estimates for the Nephrotic Syndrome multiethnic cohort dataset. AFR = African, EUR = European, ASN = Asian, OTH = Other. See Fig. S1 for more details.