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

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

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

Abstract

Heritability is a fundamental parameter of diseases and other traits, quantifying the contribution of genetics to that trait. 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 upweighing 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). 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. 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, 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, the Hispanic Community Health Study/Study of Latinos, 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

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

where $\mathbf{x}_i = (x_{ij})$ is the length-n vector of genotypes at locus i, $\mathbf{\Phi} = (\varphi_{jk})$ 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,

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

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

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

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

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

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 $2\boldsymbol{\Phi}$ under our notation. 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:

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

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,

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

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

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

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):

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

However, using a normality approximation, in Section S1 we prove that the variance of this estimator also converges to a non-zero value, so it is also not consistent:

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

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

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

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 $\hat{\mathbf{\Phi}}^{\text{name}} = (\hat{\varphi}_{jk}^{\text{name}})$ 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,

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

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

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

$$\hat{\mathbf{\Phi}}^{\text{std-ROM}} \xrightarrow[m\to\infty]{\text{a.s.}} \frac{1}{1-\bar{\varphi}}\mathbf{C}\mathbf{\Phi}\mathbf{C}, \tag{11}$$

where $\mathbf{\Phi}$ 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

$$\hat{\varphi}_{jk}^{\text{popkin}} = 1 - \frac{A_{jk}}{\hat{A}_{\min}}, \qquad A_{jk} = \frac{1}{m}\sum_{i=1}^{m}(x_{ij}-1)(x_{ik}-1) - 1, \tag{12}$$

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,

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

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). 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:

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

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.

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 $F_{\mathrm{ST}} = 0.3$ for the admixed individuals, which more closely 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 $i \in \{1, ..., m\}$, subpopulation allele frequencies $p_i^{S_u}$ for $u \in \{1, ..., K\}$, and individual-specific allele frequencies π_{ij} and genotypes x_{ij} for $j \in \{1, ..., n\}$ are drawn from this hierarchical model:

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

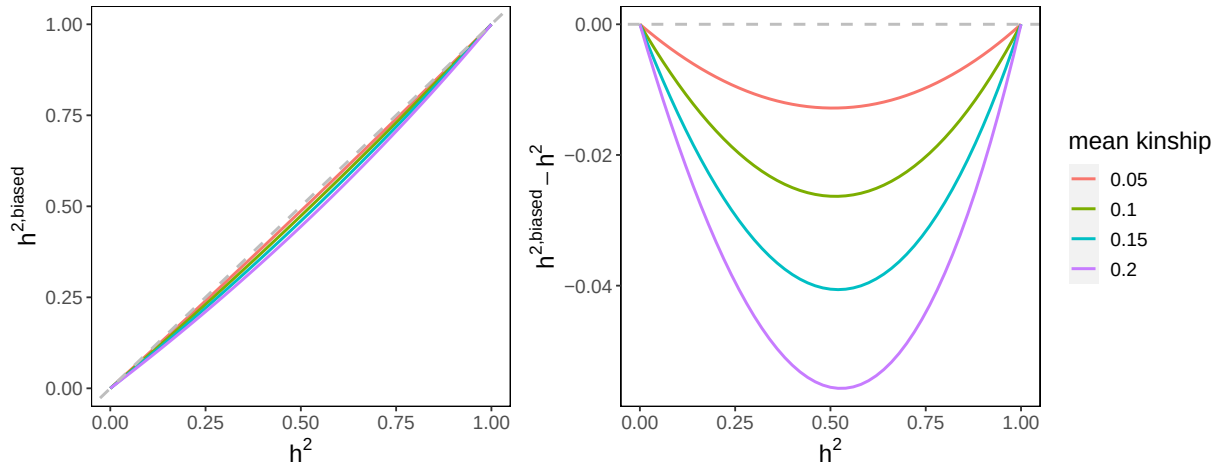

Figure 1: **Relationship between the true heritability and biased estimates.** Left: the relationship between biased estimates and the true heritability. Right: the relationship between hertability bias and true heritability. The range of mean kinship values matches what we observed in real data, as shown later.

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 $m_1 = 500$, $\mu = 0$, $\sigma^2 = 1$, $h^2 \in \{0, 0.1, \ldots, 1\}$ 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 $\boldsymbol{\epsilon}$ 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 $\boldsymbol{\beta}$ 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, throughout the rest of the paper, including genetic and trait models and empirical analyses, 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:

$$\sigma^2_{g0} = \text{E}\left[\sum_{i=1}^{m_1} 2p_i(1-p_i)\beta^2_{i0}\right] = \sum_{i=1}^{m_1} 2p_i(1-p_i),$$

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

$$s^2_{g0} = \sum_{i=1}^{m_1} 2p_i(1-p_i)b^2_{i0}.$$

We obtain the desired variance of $\sigma^2_g = h^2\sigma^2$ 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 $\hat{\sigma}^2_{g0}$, and the realized value of this estimator used in the simulation procedure is denoted $\hat{s}^2_{g0}$ as:

$$\hat{s}^2_{g0} = \sum_{i=1}^{m_1} 2\frac{\hat{p}^{obs}_i(1-\hat{p}^{obs}_i)}{1-\bar{\varphi}}b^2_{i0}, \quad \hat{\sigma}^2_{g0} = \sum_{i=1}^{m_1} 2\frac{\hat{p}_i(1-\hat{p}_i)}{1-\bar{\varphi}}\beta^2_{i0}, \tag{13}$$

where we denote the realized value of $\hat{p}_i$ used in simulations as $\hat{p}^{obs}_i$ for clarity and $\bar{\varphi}$ is the mean kinship of Eq. (6). This estimator not only satisfies $\text{E}[\hat{\sigma}^2_{g0}] = \sigma^2_{g0}$, but is a consistent estimator of σ^2_{g0} over our random coefficients as the number of independent causal variants m_1 goes to infinity (Lemma 1 in Section S2). 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:

$$\hat{b}_i = b_{i0}\frac{h\sigma}{\hat{s}_{g0}}, \quad \hat{\beta}_i = \beta_{i0}\frac{h\sigma}{\hat{\sigma}_{g0}}, \tag{14}$$

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 Section S2.

Fixed Effect Sizes (FES) model. In this model, the effect size of locus i, defined as $2p_i(1-p_i)\beta^2_i$, has the same value for every causal locus, so we desire each to equal $\sigma^2_g/m_1 = h^2\sigma^2/m_1$. 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 $\mathrm{E}[\hat{\beta}_i^2] \geq \beta_i^2$, 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.

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

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

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

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):

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

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.

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 2.5M 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 MAF $\geq$ 0.01 and 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, % Eosiniphils, % Basophils, and Red Blood Count to improve the model fits.

San Antonio Family Study (SAMAFS) is a complex pedigree-based study designed to identify low frequency or rare variants influencing susceptibility to T2D, conducted in 20 Mexican American T2D-enriched pedigrees from San Antonio, Texas (Mitchell et al., 1996). We obtained the exome chip data from dbGaP (accession phs000847.v2.p1). We exclude all unplaced and non-autosomal variants, and further filter using MAF $\geq$ 0.01 and Hardy-Weinberg Equilibrium p-value (HWE) $\leq$ 1e-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.

Nephrotic Syndrome (NS) multiethnic cohort is a multi-ancestry study exploring the etiology of nephrotic syndrome. We imputed the genotypes of study cases and controls and filtered genotypes using minor allele count $\geq$ 20. The final dataset consists of $m = 16,605,628$ loci and $n = 1,981$ individuals. We adjust for sex in the heritability estimate.

3 Results

3.1 Evaluations based on admixture and family simulations

First, we considered a scenario with population structure due to admixture but no family structure, where the simulated mean kinship value is comparable to that of real multiethnic human datasets (as shown below), at $\bar{\varphi} = 0.150$. 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 estimator produces unbiased heritability estimates, whose distribution closely resembles estimates obtained when we use the true kinship matrix of the simulation in the variance components model, whereas all other kinship estimators lead to downwardly biased estimates (Fig. 2A and Fig. S2A using traits simulated from the FES and RC models, respectively). To further illustrate the pattern of these biases, we plotted the bias trends for different kinship estimators (Fig. 2B and Fig. S2B), which closely align with our theoretical predictions (Fig. 1). Notably, our theory applies to the Standard ROM estimator only, which is the only one with a closed form biased limit for its kinship estimates and the resulting heritability estimates. In these simulations, the Standard ROM estimator has a small but measurable bias in our simulations. However, the Standard MOR estimator, which is the most frequently used in the literature and the default for GCTA and other methods, is also by far the most biased, with biases that are greater as the true heritability value increases.

Next, we extended our simulations to incorporate both population and family structures, which due to the addition of 20 generations post admixture while preserving the overall population structure, with a mean kinship value of $\bar{\varphi} = 0.163$ (kinship matrices in Fig. S3). Thus, heritability estimation biases are similar in magnitude in this scenario, although estimation variance is much reduced, apparently a direct 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 sever bias that is higher for larger values of the true heritability (Fig. 2C-D and Fig. S2C-D). Additionally, since we simulated a pedigree to generate the family structure, we tested estimation of heritability using the kinship matrix derived from the pedigree only, which models the relatedness due to family structure only, ignoring the population structure due to shared ancestry that is also present in this data. Unlike other cases, we find that these pedigree-based estimates are upwardly biased, and also have a much greater variance than the rest of the estimates (Fig. 2C-D and Fig. S2C-D).

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, defined here as those with MAF ≤ 0.05 for simplicity, as others have done recently (Biddanda et al., 2020). In our

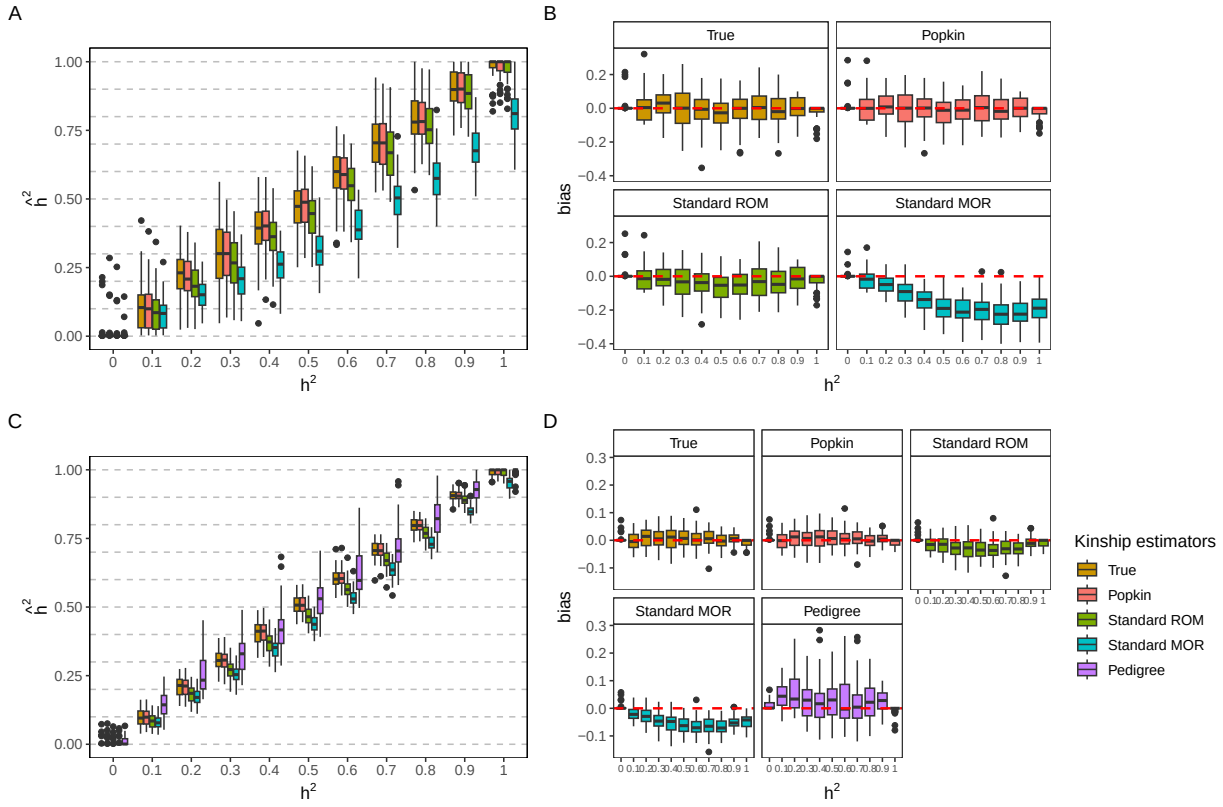

Figure 2: **Heritability estimation simulation by GCTA with various kinship matrices based on the RC trait model.** The upper and lower rows show simulation results for admixture structure only and admixture plus family structure, respectively. Bias is defined as $\hat{h}^2 - h^2$.

simulations, ancestral allele frequencies are drawn uniformly, which after the admixture and family structure results in a proportion of rare variants of 0.18-0.19 (Table 1). (Note that under perfect uniformity, since MAF is folded with range between 0 and 0.5, the proportion of rare variants is $0.05 \times 2 = 0.10$, which is considerably less than our simulations). On the other hand, rare variants are the majority of variants in WGS datasets like 1000 Genomes (0.93 proportion when there are no MAF filters; Table 1). In order to characterize the effects of rare variants in our analysis, we carried out additional simulations in which the real 1000 Genomes Project genotypes are used along with simulated traits, in this case with a fixed heritability to $h^2 = 0.8$. 1000 Genomes is a multiethnic cohort which includes African, European, South Asian, East Asian, and Hispanic individuals (Fig. S4), which results in a high mean kinship estimate of $\bar{\varphi} = 0.133$ without MAF filter ($\bar{\varphi} = 0.134$ if only including MAF ≥ 0.01, and $\bar{\varphi} = 0.100$ for MAF ≥ 0.05). We removed rare variants with a given MAF threshold that differs for causal variants (which influence the trait directly) and separately for inclusion in the kinship estimator (which affects kinship estimation bias and as a consequence heritability estimation with the variance components model). In this case we find different behaviors for the two trait models we considered, which differ precisely in how they assign coefficients to rare variants.

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

Dataset	QC Filter	Proportion of loci with MAF ≤ 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 ≥ 0.01	0.438	0.134
1000 Genomes	MAF ≥ 0.05	0.000	0.100
HCHS/SOL	MAF ≥ 0.01	0.256	0.108
SAMAFS	MAF ≥ 0.01	0.281	0.063
NS	MAC ≥ 20	0.564	0.124

We first considered traits simulated with the RC model, which assigns causal coefficients independently of allele frequency, resulting in smaller effect sizes (product of coefficient and genotype variance) for rare variants; it is also the only model one guaranteed to result in correctly specified heritabilities for simulated traits when true allele frequencies are unknown, as is the case for real human datasets. We confirm again that Popkin yields estimates that are consistent with being unbiased in most of these cases (overlap with the true value of $h^2 = 0.8$), although possible small downward biases are observed in the case in which causal variants have MAF ≥ 0.05 but variants with MAF < 0.05 are included in the kinship estimate (Fig. S5). The Standard ROM estimator has a small downward bias in all these cases, and it always results in smaller estimates than Popkin. Lastly, the Standard MOR estimator has its most extreme underestimates of heritability whenever

its kinship estimate includes variants with MAF < 0.05, regardless of the presence or absence of rare variants among causal loci, whereas its estimates become practically unbiased and resemble those of the other two estimators if only variants with MAF ≥ 0.05 are included when calculating the kinship estimate. Thus, in these cases, our results are consistent with our previous simulations for common variants, particularly with rare variants being a large source of bias that affects the Standard MOR estimator.

Next, we turn to traits simulated with the FES model, which assign causal coefficients that are roughly inverse proportional to allele frequency (see Methods), thus upweighing rare variants, but which as a consequence does not always result in correctly specified heritabilities, since the genotype variance of rare variants is challenging to estimate. Thus, when traits were simulated using all loci, including rare variants, heritability estimates were systematically lower than the desired value of $h^2 = 0.8$ across all kinship estimators (Fig. 3). Our work suggests that these are not real biases of these heritability estimators, but rather a bias in how traits are simulated, which is a known problem that is the target of ongoing work. As expected, for traits with causal MAF ≥ 0.01 we tended to see results as before, consistent with RC traits and with our common variant genotype simulations, namely that Popkin gives relatively unbiased estimates, and Standard ROM gives similar or smaller estimates than Popkin. However, here we see a new pattern that had not been observed before, namely that 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 MAF ≥ 0.05 for kinship estimator, and MAF ≥ 0.01 for trait and MAF ≥ 0.05 for kinship estimator.

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 kinship MOR estimator is the most commonly used estimator in population heritability analyses with GCTA by default 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 a large degree of population structure, resulting in a mean kinship of $\bar{\varphi} = 0.108$, owing to the admixture structure of the Hispanic individuals sampled from various nationalities in this study (Fig. S6). Throughout the 33 traits that we analyzed, we see the consistent pattern that Popkin estimates larger heritabilities than Standard ROM (Fig. 4), as predicted by our theoretical results. Though these differences appear small overall, they are

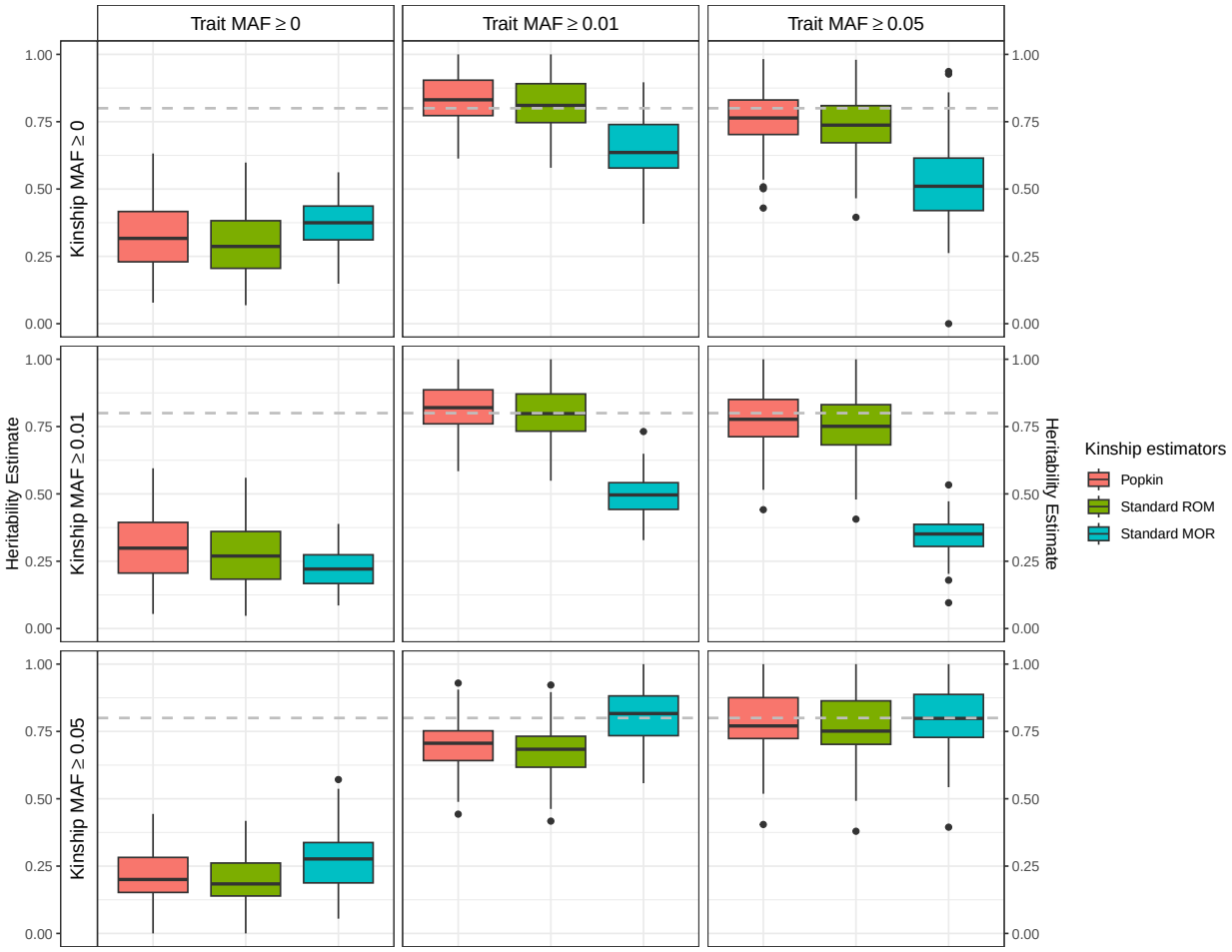

Figure 3: **Heritability estimation evaluation based on the 1000 Genomes Project testing effect of rare variants.** Traits were simulated using the FES model. For simulated traits, MAF thresholds shown are applied before causal variants are selected, thus influencing the architecture of the trait. For kinship estimates, MAF thresholds are applied to genotypes before kinship matrices are estimated from these genotypes, thus influencing only kinship estimation.

large relative to the standard errors of these estimates, such that in most of these cases the Popkin estimate is about a standard error larger than the expected distribution of the Standard ROM estimates. In contrast, Standard MOR estimates, which are the most common estimates, are often the largest of the three estimates, again with values up to a standard error larger than Popkin's estimates, although some traits have Standard MOR estimates that are instead smaller or similar to the 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). Thus, we do not expect the severe downward biases that Standard MOR experiences due to inclusion of rare variants in the kinship estimate, while a large proportion of rare variants among causal loci (which are likely not genotyped) could explain the large upward bias of Standard MOR in this data. If so, traits with larger Standard MOR estimates (relative to Standard ROM) may have architectures driven by more causal rare variants, whereas traits with similar estimates may have a larger common variant component. It is also worth noting that in this data the estimated heritability of height is high for all estimators, just under the estimate of $h^2 = 0.8$ 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, with a mean kinship of $\bar{\varphi} = 0.0634$, potentially owing to a more homogeneous subset of Hispanics, all arising from a single city, and belonging to only 20 families (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 the standard errors of these estimates (Fig. 5). Interestingly, unlike HCHS/SOL, here Standard MOR produces estimates are very similar to Standard ROM estimates, which could again be partly due to a low proportion of rare variants used to estimate the kinship matrices (0.28; Table 1), although there do not seem to be upward biases due to causal rare variants either. Furthermore, SAMAFS has a well-documented pedigree structure, which we can use as an alternate estimate of kinship that may be more accurate for close relatives, although it has the disadvantage of not modeling ancestry differences expected for admixed populations such as Hispanics. We find that heritability estimates based on the pedigree kinship tended to be larger, sometimes considerably so relative to the standard errors (for example, Adiponectin), whereas in only one case it results in the smallest estimate (Fasting glucose). Since pedigree-based heritability estimates were shown to be upwardly biased and highly variable in our simulations when there is also population structure (Fig. 2), we conclude that these estimates are likely also generally upwardly biased. Here we again see high heritability estimates for height, which were also obtained by conditioning for sex and age, although the conclusion is less surprising given that this is a family study containing numerous siblings.

Since HCHS/SOL and SAMAFS are both studies of Hispanic populations that overlap on 16 comparable traits, we compared the heritability estimates that we obtained with all of the kinship estimators used in both studies. Despite the potential for differences in heritability due to differences in environment or genetics, we find high Pearson correlations of ≈ 0.8 between the estimates of

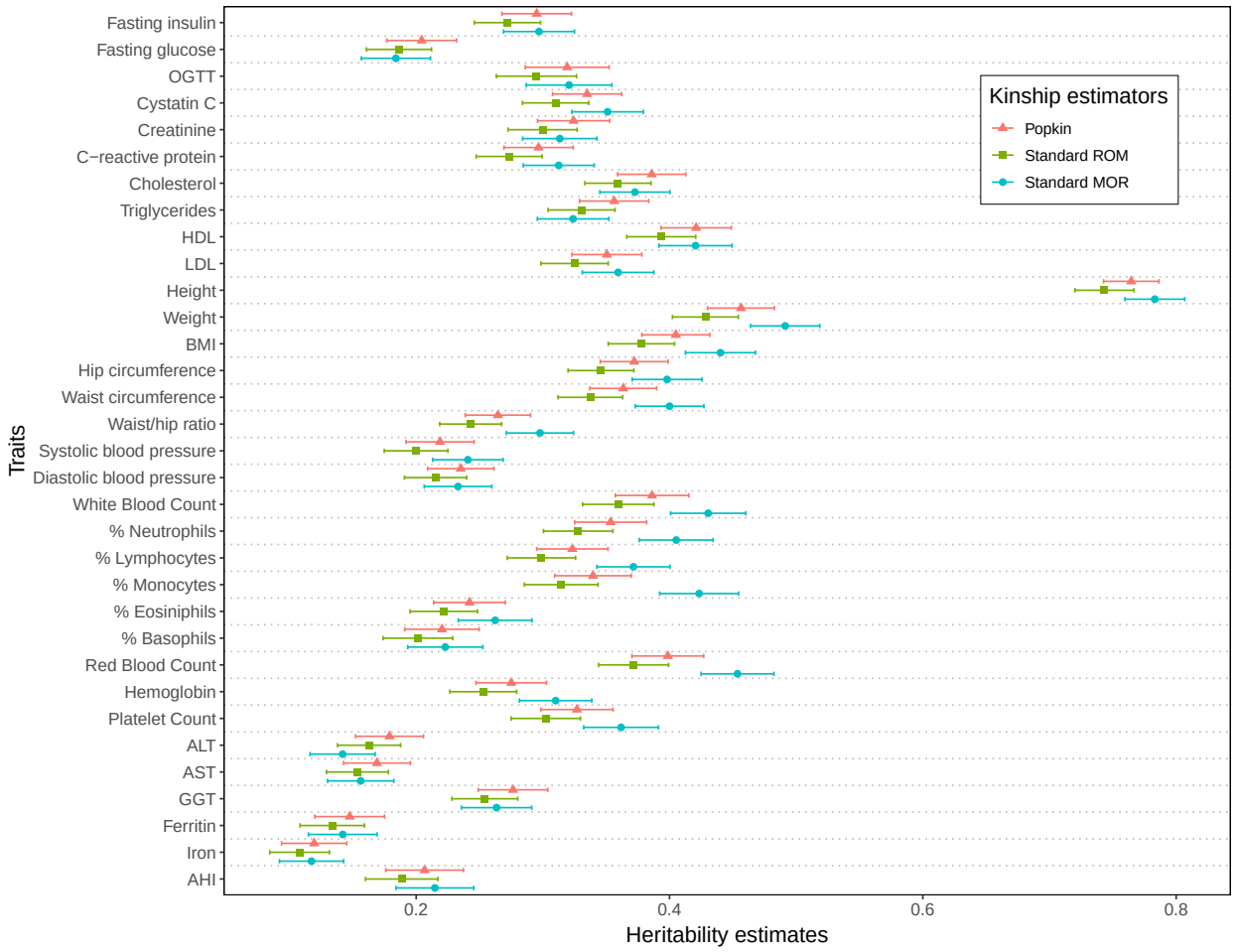

Figure 4: **Heritability estimates on HCHS/SOL dataset.** The figure shows heritability estimates using Popkin, Standard ROM, and Standard MOR kinship estimators on the Hispanic Community Health Study / Study of Latinos.

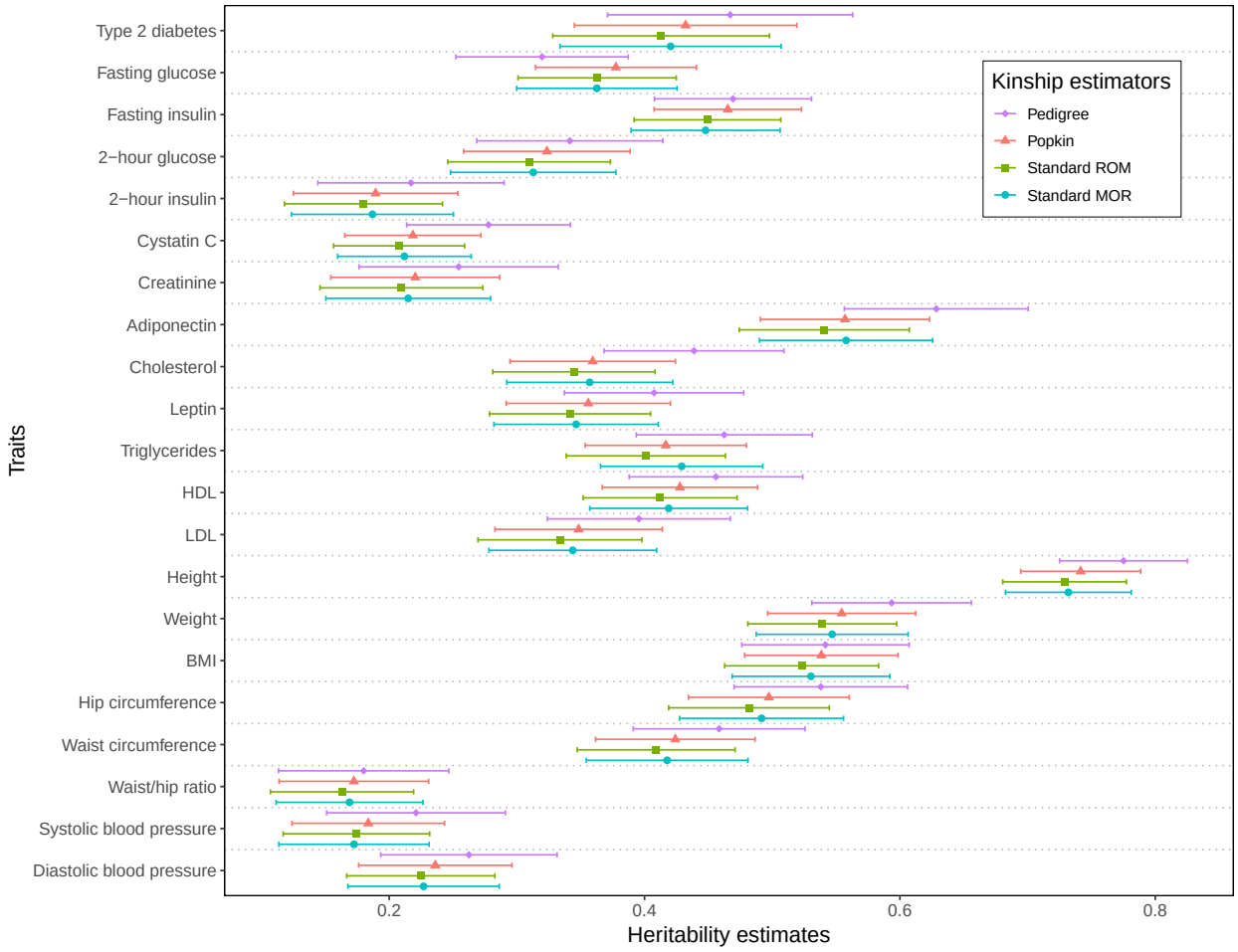

Figure 5: **Heritability estimates on SAMAFS dataset.** Heritability estimates are calculated using the Pedigree, Popkin, Standard ROM, and Standard MOR kinship estimates on the San Antonio Mexican American Family Study: Type 2 Diabetes.

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 we are unable to identify substantial differences in consistency between Popkin, Standard ROM and MOR estimates, in the sense that all of their regression lines closely follow the $y = x$ line.

The NS dataset is a multiethnic dataset including large proportions of African, European, South Asian, and admixed individuals (Fig. S9). Its mean kinship value of $\bar{\varphi} = 0.124$ is one of the largest analyzed that includes real phenotypes, and which is nearly as large as that of 1000 Genomes. This dataset has three binary traits: each of the diseases NS and its subtypes SSNS (steroid sensitive) and SRNS (steroid resistant) 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, in this data Standard MOR consistently results in considerably smaller heritability estimates than Popkin (Fig. 6). The proportion of rare variants in this data is second highest here (0.56; Table 1), which likely explains the severe downward biases observed for Standard MOR. The visualized estimates were transformed to be in the liability scaled; for heritability estimates in the observed scale, see Table S1.

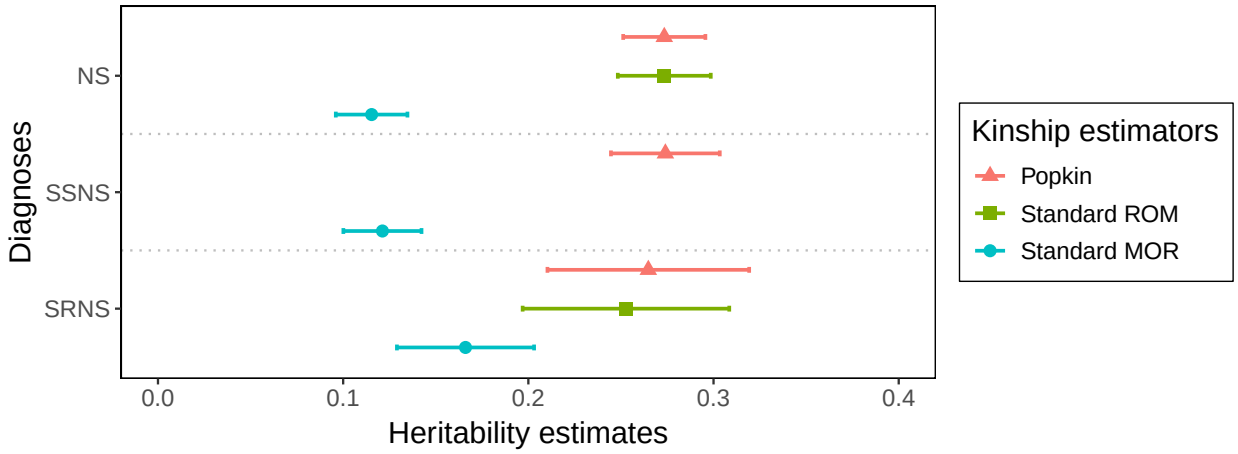

Figure 6: **Heritability estimates on NS dataset.** The figure shows heritability estimates in the liability scale using Popkin ROM, Standard ROM, and Standard MOR kinship estimators on the Nephrotic Syndrome (NS) multiethnic cohort. The binary traits are NS (NS vs control), SSNS (steroid sensitive NS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

4 Discussion

Previous research has shown that commonly used kinship estimators are biased (Ochoa and Storey, 2021). However, although these kinship estimators are widely used to conduct association studies, either directly in linear mixed-effects models or indirectly in the PCA regression approach, these biases do not affect association test statistics (Hou and Ochoa, 2023). In this study, we investigated the impact of kinship estimation bias on heritability estimation using variance components models, which are another common application of biased kinship estimators, and found that this bias propagates to heritability estimation, in agreement with recent work (Chen and Storey, 2022). Specifically, we observed that the Standard ROM kinship estimator systematically underestimates heritability, while the bias introduced by the much more common MOR estimator can be even more severe and depends on the presence of rare variants.

Based on our simulation and real data analysis, the bias between the Standard ROM and Popkin (which is also of type ROM) estimators is evident, with the magnitude of the differences depending on the mean kinship values. However, it remains challenging to quantitatively determine both the direction and the extent of bias between the Standard ROM and MOR versions. The formula of the MOR estimator in Eq. (10) assigns greater weight to rare variants, so we hypothesize that rare variants are responsible for the observed biases specific to MOR estimators. Our simulation studies (Figs. 2, 3 and S5) confirm that the proportion and effect size distributions of rare variants influence the extent of bias in the MOR estimator. Thus, the relationship of heritability using different kinship estimators can be characterized in Fig. 7.

Rare variants can affect heritability estimation in two ways: by introducing biases in kinship estimation, as well as by having large effects as causal variants. All of our evaluations demonstrate that the Standard MOR estimator has stronger biases than the Standard ROM estimator, but the

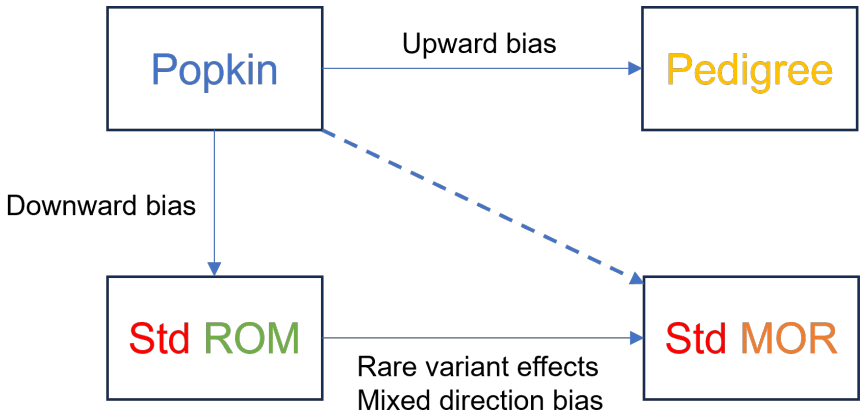

Figure 7: **Relationships between kinship estimators in terms of their effects on heritability estimation.**

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 compared to Popkin (which is unbiased). In contrast, when a large proportion of causal variants are rare and they have large effect sizes, this appears to lead the Standard MOR to have an upward bias relative to Standard ROM and often Popkin too. Our hypothesis is that this behavior is due to Standard MOR upweighing rare variants compared to the other two estimators. In practical applications, as demonstrated by our real data analysis of HCHS/SOL, many traits appear to have an architecture where rare variants have significant effects, which explains in those cases why popkin's heritability estimates are lower than those of the standard MOR 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. Because our admixture and family simulations begin from pre-existing ancestral variation, they primarily generate common variants and do not introduce new mutations, limiting the realism of rare variant modeling. Under these conditions, only the popkin estimator yielded unbiased heritability estimates; standard estimators showed consistent downward bias, more severe for MOR than ROM, as shown in the second and third columns of Figs. 3 and S5. To simulate traits influenced by more realistic rare variants, we utilized genotype data from the 1000 Genomes Project. When rare variants were set as 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 trait simulation algorithm: with real genotypes, true ancestral allele frequencies are unknown, leading to upwardly biased estimates of the inverse of genotype variance and misspecified heritability under the FES model. Overall, the heritability of traits affected by rare variants remains a theoretically underdeveloped area, necessitating further research. These challenges are consistent with prior work showing that standard and LD-weighted kinship matrices can misestimate heritability when rare variants are involved. Alternative modeling strategies have been proposed to better account for rare variant effects, reinforcing the need for continued methodological development in this area (Lee et al., 2013; Speed et al., 2017a).

Pedigree information was available for the San Antonio Family Study (SAMAFS), allowing us to estimate kinship from this pedigree and obtaining heritability estimates, in a manner that more closely resembles classic family (including twin and sibling) studies (Almasy and Blangero, 1998b) as well as more recent pedigree-based approaches from large EHR data or insurance claim database (Polubriaginof et al., 2018; Wang et al., 2017b). Notably, SAMAFS was originally analyzed using pedigree-derived kinship matrices and variance component models to partition phenotypic variance (Mitchell et al., 1996). However, our results indicate that heritability estimated using the pedigree-based kinship matrix is upwardly biased compared to estimates obtained with the popkin ROM estimator. This bias may arise because pedigree-derived kinship matrices do not account for ancestry differences or due to cryptic relatedness (because the pedigree is incomplete between

families), leading to inflated heritability estimates. Some research has suggested that studies based on close relatives tend to estimate a higher 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, the traits are binary (case vs. control). Heritability was initially estimated directly (observed heritability) and subsequently transformed into liability-scale heritability. In our study, although the liability-scaled heritability estimates ranged from 0.1 to 0.3 after transformation, the observed heritability estimates for NS vs. control and SSNS vs. control using the popkin estimator approached 0.99 (Table S1). This is because σ_e^2 converged to nearly zero after several iterations in the GCTA software. This discrepancy highlights fundamental limitations in liability-based heritability for binary traits. Liability models rely on untestable assumptions (e.g., multivariate normality of latent liability) that, when violated, can produce severely biased estimates (Benchek and Morris, 2013). Our results extend these concerns, suggesting that even when liability-scale estimates appear plausible, they may mask extreme biases in variance component partitioning. Improved algorithms and transformation frameworks are urgently needed to address these pitfalls, especially in admixed cohorts where cryptic relatedness or non-normal liability distributions may further distort estimates. 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's.

In this study, we empirically and theoretically examined the impact of kinship estimation bias on heritability estimation, expanding on prior research that identified biases in commonly used kinship estimators that are overcome by popkin. While these biases do not affect association test statistics, our findings show that they propagate to heritability estimates. Specifically, Popkin results in unbiased estimates, the Standard ROM kinship estimator consistently underestimates heritability, whereas bias in MOR estimators depends on the presence of rare variants. Through simulations and real data analyses, we observed that the extent of Standard MOR estimator bias is influenced by mean kinship values, but rare variants playing a critical role in shaping bias patterns as well. Overall, our study highlights the importance of selecting appropriate kinship estimators in heritability analyses, particularly in structured populations and rare variant studies. 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 and Tiffany Tu for sharing processed data and derivatives for NS, SAMAFS, and HCHS/SOL datasets.

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 Neurological 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.

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.

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.

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.

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.

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.

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.

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.

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, S Hong et al. (2013). "Estimation of SNP heritability from dense genotype data". *The American Journal of Human Genetics* 93(6), pp. 1151–1155.

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

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

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

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

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

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

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.

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.

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.

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

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

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

Speed, Doug et al. (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.

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.

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.

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.

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.

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.

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

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

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.

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.

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.

S1 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 multivariate normal distributions for $\mathbf{x}_i$ and $\mathbf{y}_i$ and a symmetric $\mathbf{A}$, namely

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

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

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

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

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

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:

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

Thus, the final variance equation is:

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

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

$$\mathrm{Var}(\hat{p}_i(1-\hat{p}_i)) \leq p_i(1-p_i)\bar{\varphi}, \tag{S1}$$

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 $p_i = 1/2$).

S2 Consistency of the RC initial variance estimator $\hat{\sigma}^2_{g0}$ and the final variance $\hat{\sigma}^2_g$

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

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

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*

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

Then, as $m_1 \to \infty$,

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

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

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

$$\mathrm{E}[A_i \cdot \beta^2_{i0}] = 2p_i(1-p_i),$$
$$\mathrm{Var}(A_i \cdot \beta^2_{i0}) = 2\,\mathrm{E}[A_i]^2 + 3\,\mathrm{Var}(A_i),$$

because $\mathrm{E}[\beta^2_{i0}] = 1$ and $\mathrm{Var}(\beta^2_{i0}) = 2$. Then $\mathrm{E}[\hat{\sigma}^2_{g0}] = \sigma^2_{g0}$.

Therefore:

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

It follows from our assumptions that:

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

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

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

Thus, for some constant C_2:

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

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 $m_1 \to \infty$:

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

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* $\hat{\beta}_i := \beta_{i0} \cdot \dfrac{h\sigma}{\sqrt{\hat{\sigma}_{g0}^2}}$ *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,*

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

Proof. By Lemma 1, we have

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

Then,

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

by the continuous mapping theorem.

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

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

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

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

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

S3 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.

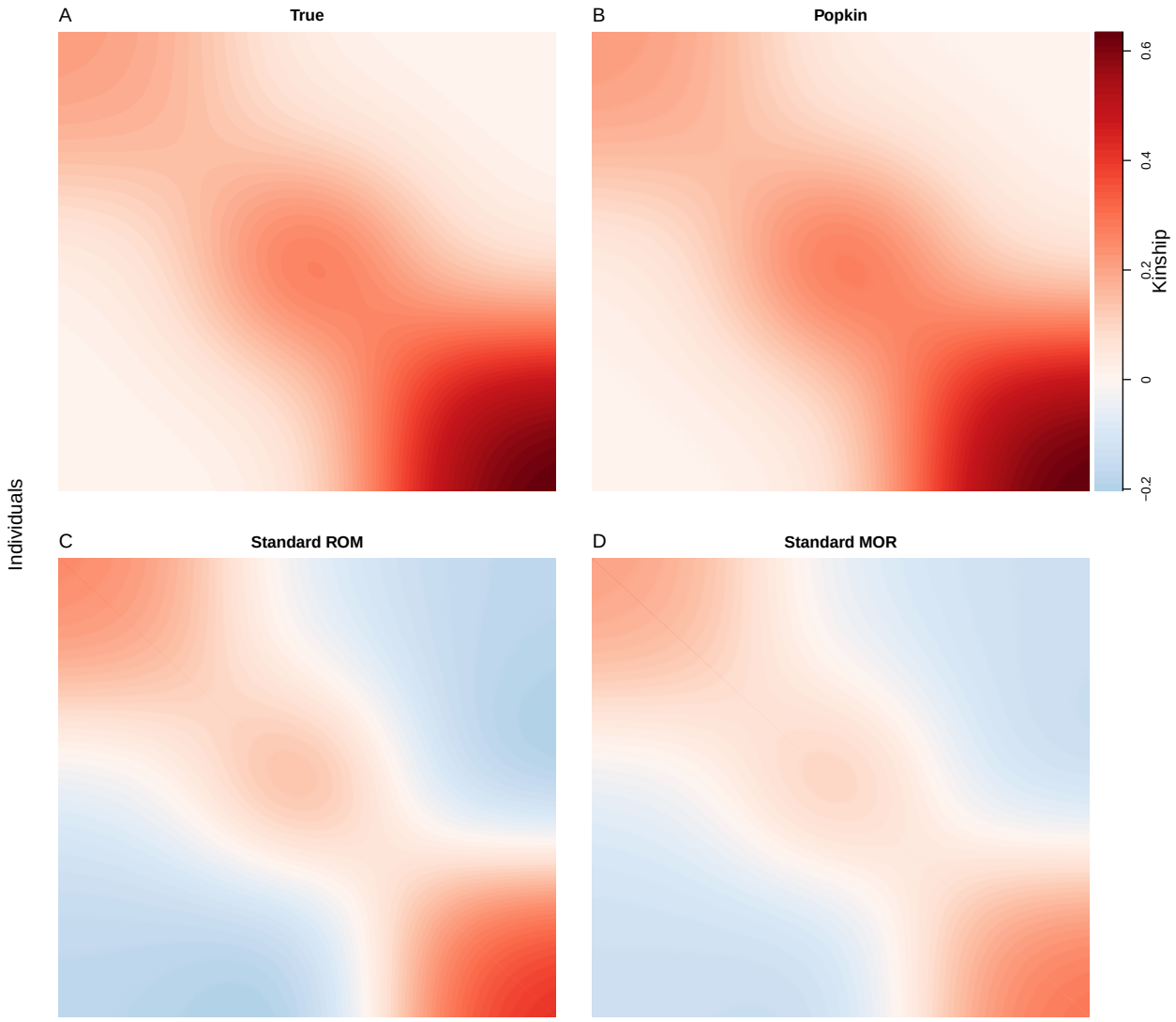

Figure S1: **Kinships heatmap for the heritability estimation simulation admixture structure only.** Estimates are calculated on the genotype matrix of the first replicate of this simulation. In each panel, individuals are placed along both x and y axes, and the kinship value φ_{jk} of a pair of individuals 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 upweighed in the MOR version.

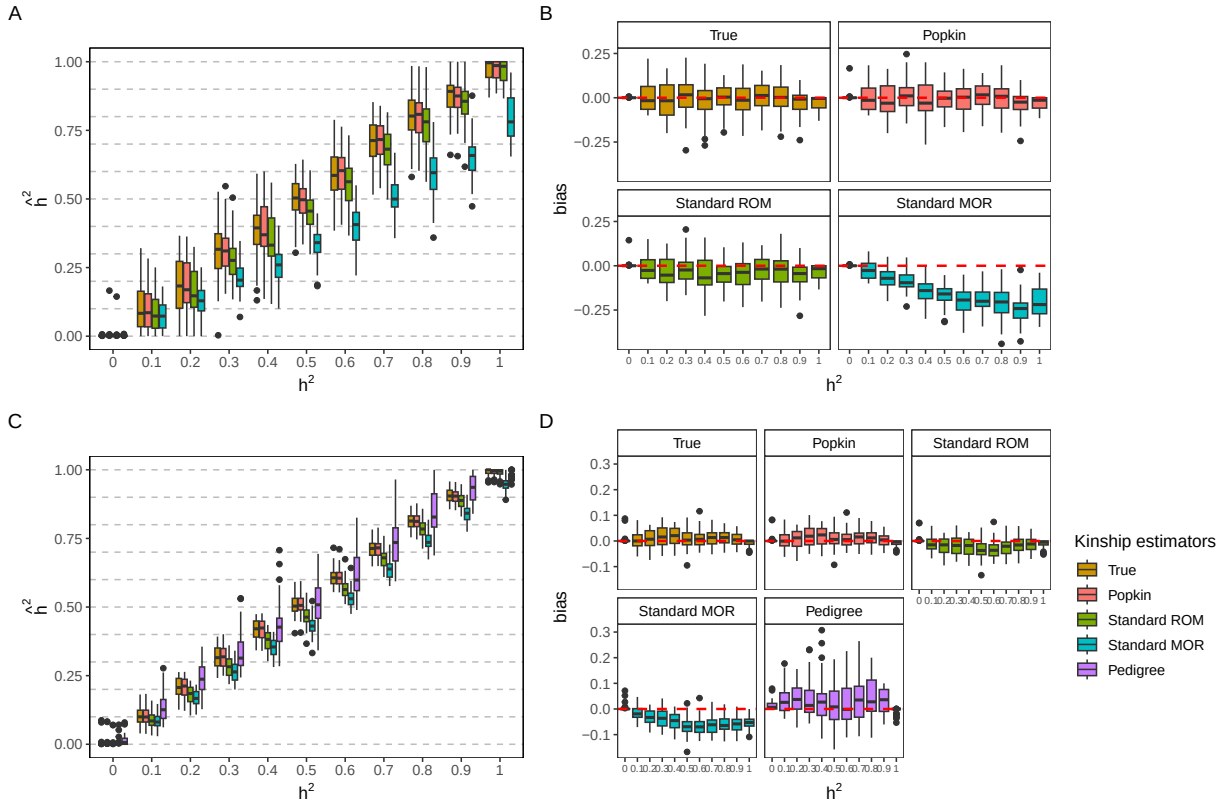

Figure S2: **Heritability estimation simulation by GCTA with various kinship matrices based on the FES trait model.** The upper and lower rows show simulation results for admixture structure only and admixture plus family structure, respectively.

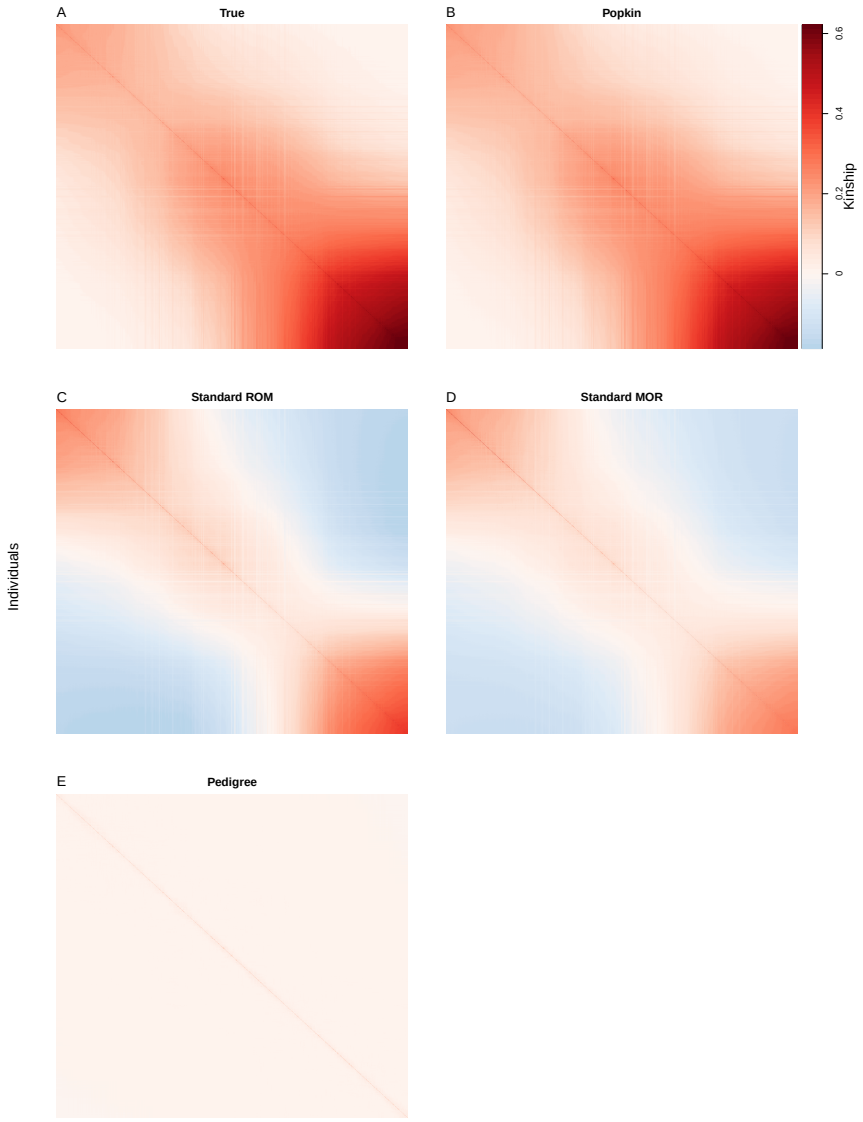

Figure S3: **Kinships heatmap for the heritability estimation simulation admixture structure plus family structure.** Estimates are calculated on the genotype matrix of the first replicate of this simulation. The simulation reorders individuals so that families appear closer together along the diagonal (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.

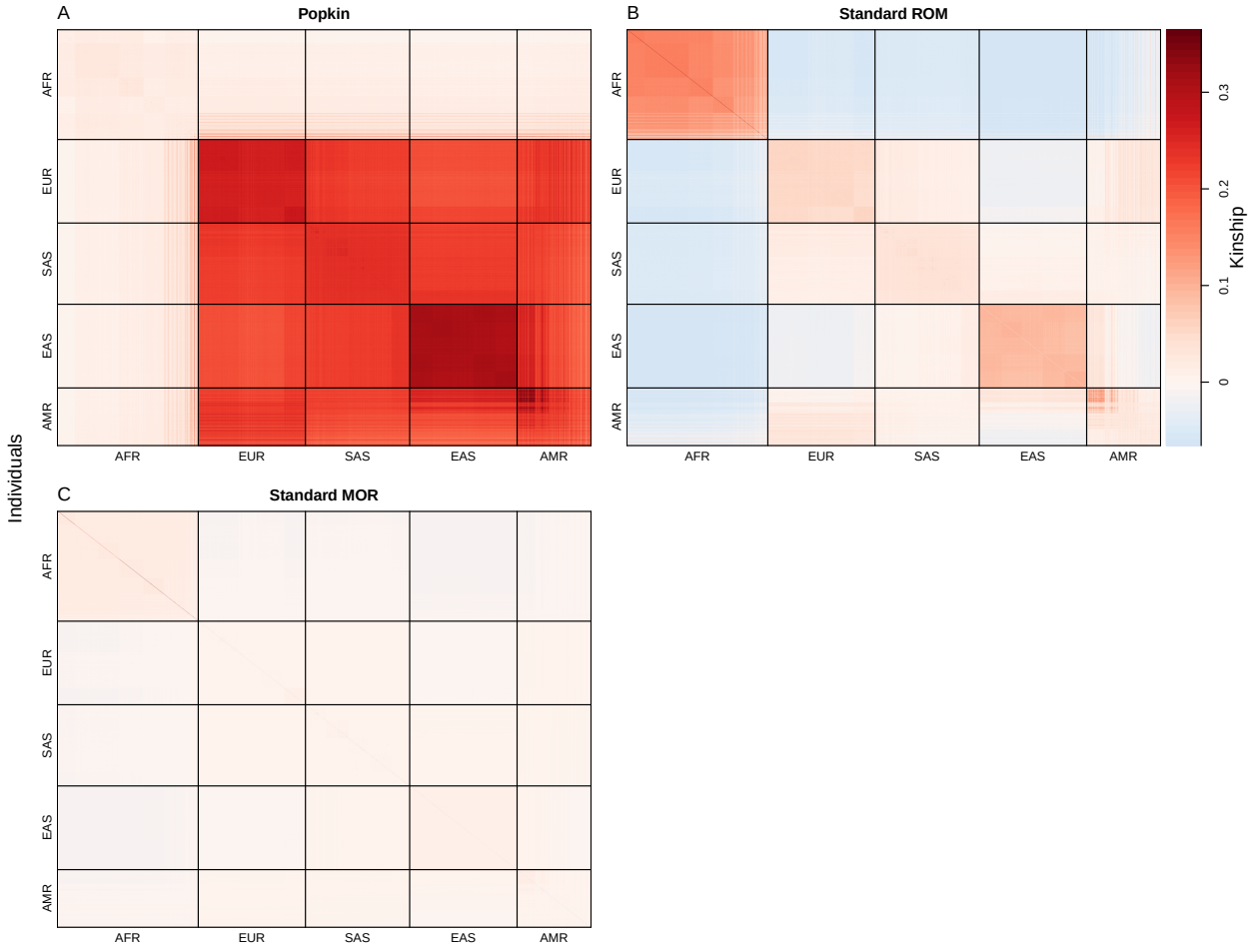

Figure S4: **Kinships heatmap for the 1000 Genome dataset.** AFR = African, EUR = European, SAS = South Asian, EAS = East Asian, AMR = Admixed Americans (Hispanics). See Fig. S1 for more details.

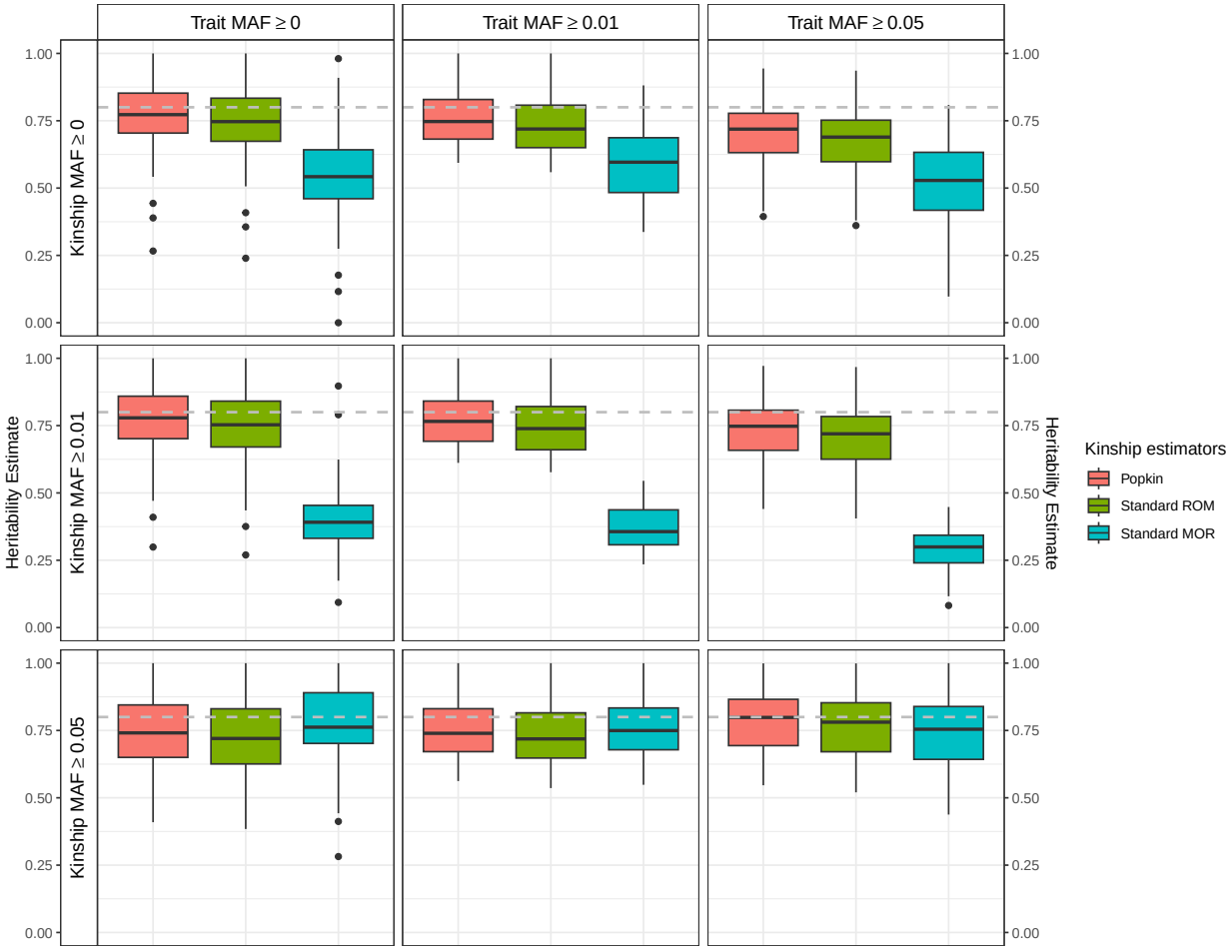

Figure S5: **Heritability estimation simulation based on 1000 Genomes Project with various kinship matrices using different rare variants filters based on the RC model.**

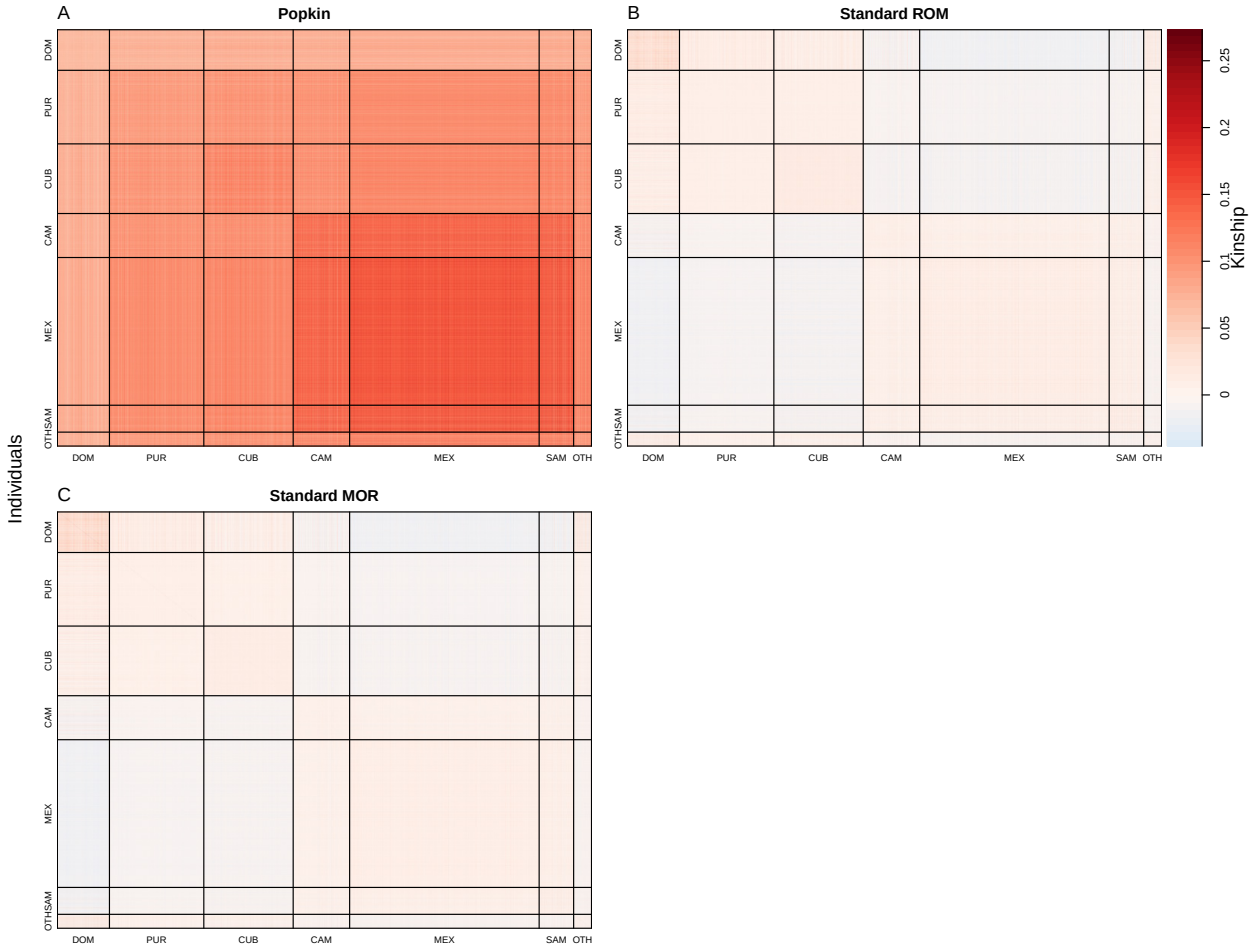

Figure S6: **Kinships heatmap for the Hispanic Community Health Study / Study of Latinos dataset.** DOM = Dominican, PUR = Puerto-Rican, CUB = Cuban, CAM = Central American, MEX = Mexican, SAM = South American, OTH = Other. See Fig. S1 for more details.

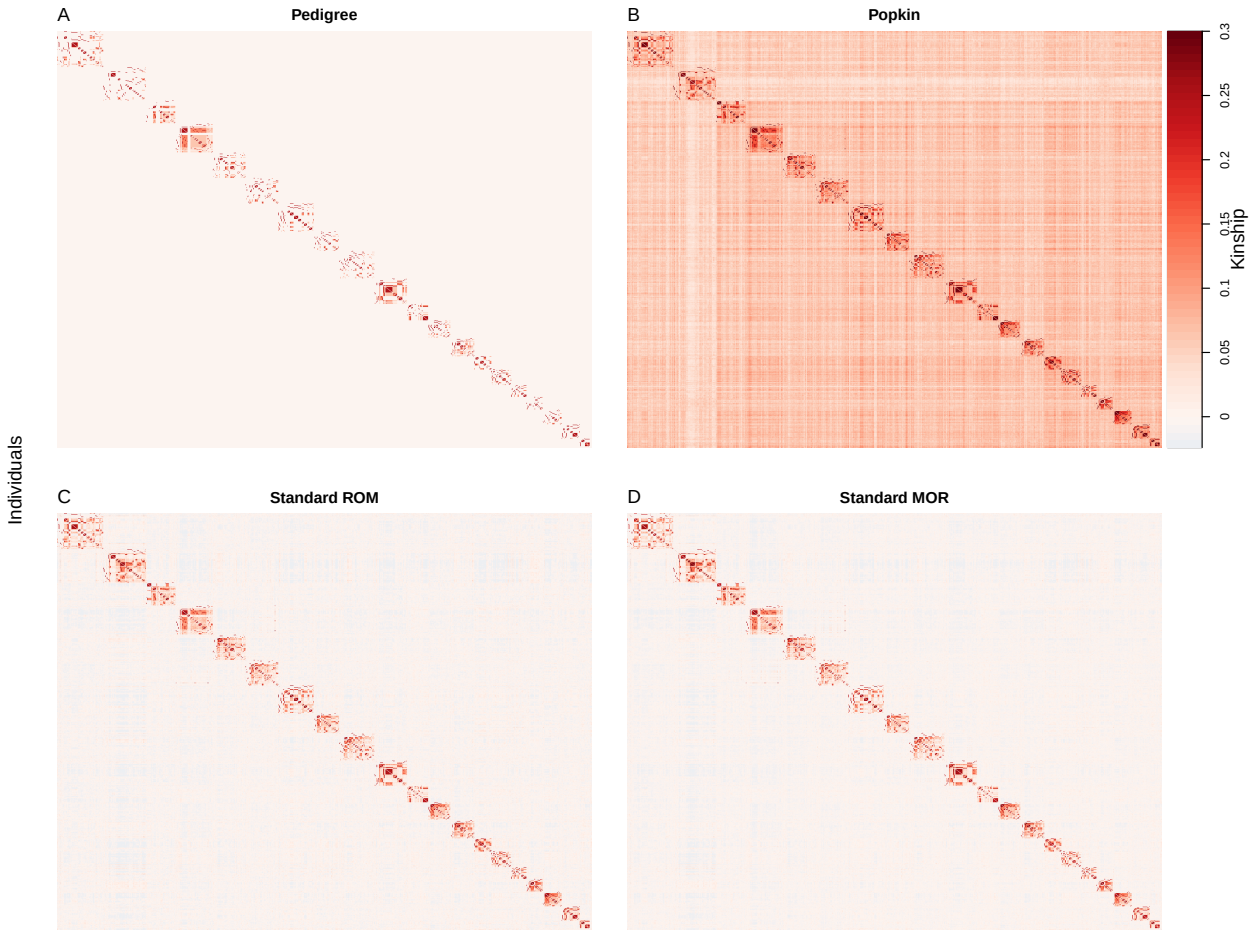

Figure S7: **Kinships heatmap for the San Antonio Family Study dataset.** See Fig. S1 for more details.

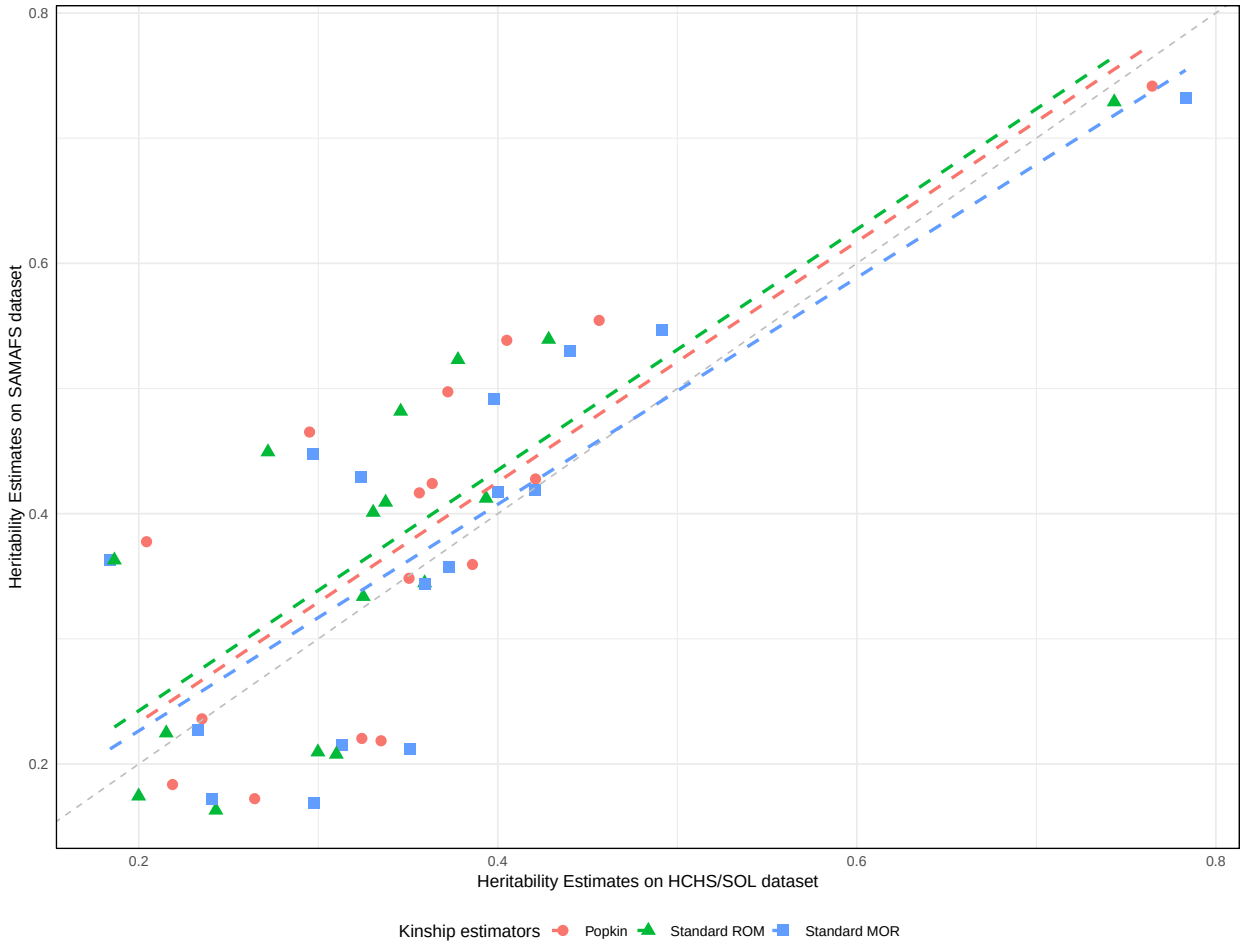

Figure S8: **Comparison of Heritability Estimates Between HCHS and SAMAFS datasets.** 16 overlapping traits are included in the analysis. Pearson correlation coefficients for each kinship estimator are: 0.799 for Popkin, 0.802 for Standard ROM, and 0.794 for Standard MOR.

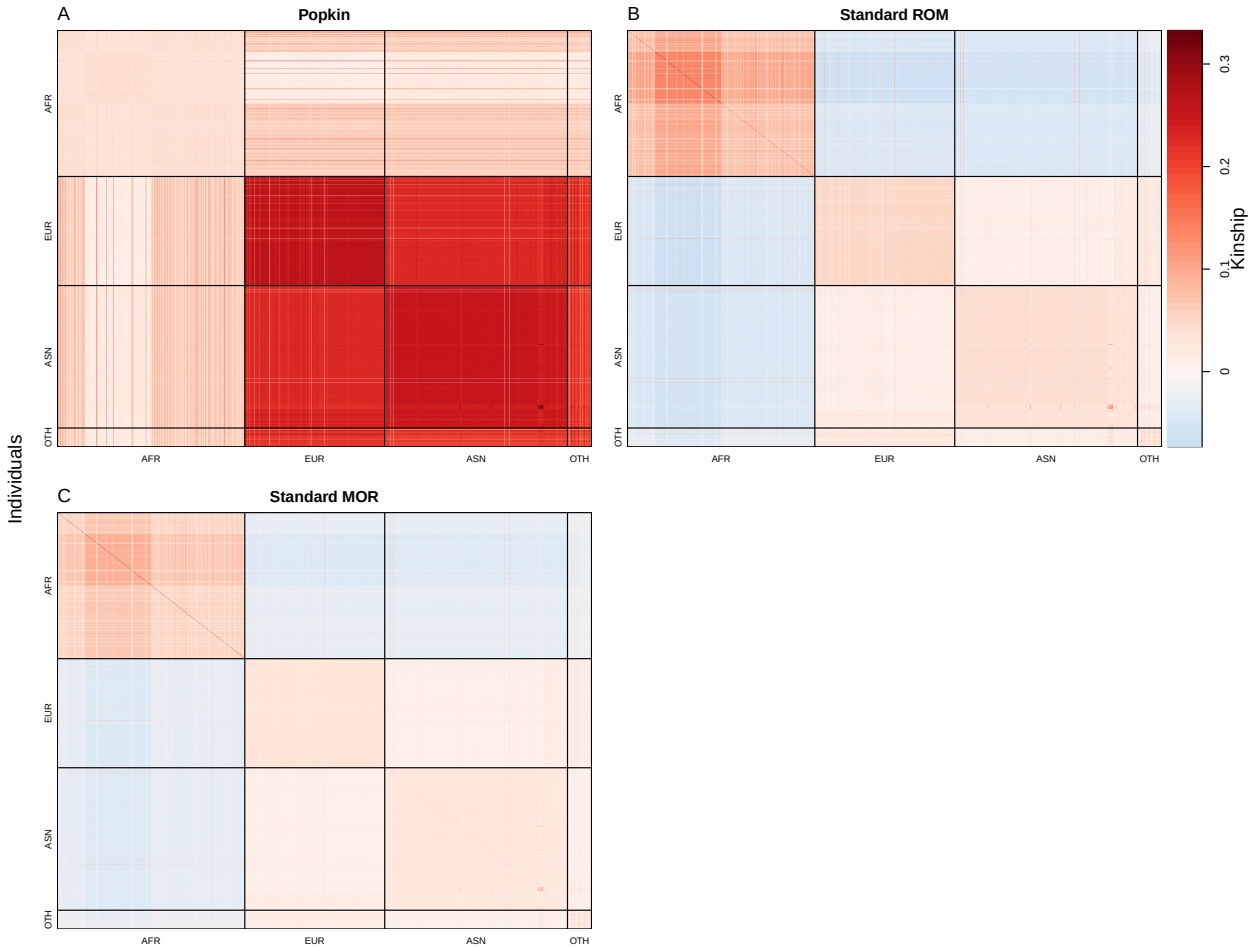

Figure S9: **Kinships heatmap for the Nephrotic Syndrome multiethnic cohort dataset.** AFR = African, EUR = European, ASN = Asian, OTH = Other. See Fig. S1 for more details.