Genetic association models are robust to common population kinship estimation biases

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

Common genetic association models for structured populations, including Principal Component Analysis (PCA) and Linear Mixed-effects Models (LMM), model the correlation structure between individuals using population kinship matrices, also known as Genetic Relatedness Matrices or "GRMs". However, the most common kinship estimators can have severe biases that were only recently determined. Here we characterize the effect of these kinship biases on genetic association. We employ a large simulated admixed family and genotypes from the 1000 Genomes Project, both with simulated traits, to evaluate key kinship estimators. Remarkably, we find practically invariant association statistics for kinship matrices of different bias types (matching all other features). We then prove using statistical theory and linear algebra that LMM association tests are invariant to these kinship biases, and PCA approximately so. Our proof shows that the intercept and relatedness effect coefficients compensate for the kinship bias, an argument that extends to generalized linear models. As a corollary, association testing is also invariant to changing the reference ancestral population of the kinship matrix. Lastly, we observed that all kinship estimators, except for popkin ROM, can give improper non-positive semidefinite matrices, which can be problematic although some LMMs handle them surprisingly well, and condition numbers can be used to choose kinship estimators. Overall, we find that existing association studies are robust to kinship estimation bias, and our calculations may help improve association methods by taking advantage of this unexpected robustness, as well as help determine the effects of kinship bias in related problems.

Abbreviations: PCA: principal component analysis; PCs: principal components; LMM: linear mixed-effects model; MOR: mean of ratios; ROM: ratio of means; WG: Weir-Goudet (kinship estimator); MRCA: Most Recent Common Ancestor; SRMSD_p: p-value Signed Root Mean Square Deviation; AUC_{PR}: Area Under the Precision Recall Curve; GCTA: Genome-wide Complex Trait Analysis (software); PSD: positive semidefinite.

1 Introduction

The goal of genetic association is to detect loci that are related to a specific trait, either causally or by proximity to causal loci. When applied to structured populations with admixed individuals, multiethnic cohorts, or close relatives, controlling for relatedness is crucial to avoid spurious associations and loss of power (Devlin and Roeder, 1999; Voight and Pritchard, 2005; Astle and Balding, 2009; Yao and Ochoa, 2022). The most popular association models for structured populations are Linear Mixed-effects Models (LMM) and Principal Component Analysis (PCA), which are closely related except LMM is capable of modeling high-dimensional structures whereas PCA is strictly a low-dimensional model (Astle and Balding, 2009; Hoffman, 2013; Yao and Ochoa, 2022).

Various association models, including both PCA and LMM, parametrize relatedness using kinship matrices, also known as Genetic Relatedness Matrices or "GRMs". Kinship coefficients are well suited for this task since they model the covariance structure of genotypes (Malécot, 1948; Jacquard, 1970). Kinship is often encountered in family studies, where they reflect recent relatedness and can be calculated from pedigrees (Wright, 1922; Emik and Terrill, 1949; García-Cortés, 2015). However, as kinship is defined as a probability of identity by descent, it may also capture ancient population relatedness (Malécot, 1948; Astle and Balding, 2009), and common non-parametric kinship estimators from genotypes indeed include population structure in their estimates (Ochoa and Storey, 2021). In LMMs, the kinship matrix is an explicit parameter determining the random effect covariance structure (Xie et al., 1998; Yu et al., 2006; Aulchenko et al., 2007; Astle and Balding, 2009; Kang et al., 2008; Kang et al., 2010; Zhou and Stephens, 2012; Yang et al., 2014; Loh et al., 2015; Sul et al., 2018). In PCA, the principal components (PCs) are in practice the eigenvectors of an empirical genetic covariance matrix that is equivalent to the most common kinship estimator (Price

et al., 2006; Astle and Balding, 2009; Hoffman, 2013; Yao and Ochoa, 2022).

Although several kinship estimators have been used with LMMs in the past, work from the last 15 years has converged on what we call the "standard" kinship estimator, which is the same estimator used in PCA and other related models (Price et al., 2006; Astle and Balding, 2009; Rakovski and Stram, 2009; Thornton and McPeek, 2010; Yang et al., 2010; Yang et al., 2011; Zhou and Stephens, 2012; Speed et al., 2012; Yang et al., 2014; Speed and Balding, 2015; Loh et al., 2015; Wang et al., 2017; Sul et al., 2018). The impetus of our work is the recent characterization of a complex bias for this standard estimator, which varies for every pair of individuals (Weir and Goudet, 2017; Ochoa and Storey, 2021). These recent works also produced two new kinship estimators, which we are interested in characterizing in the context of association. The Weir-Goudet (WG) estimator constitutes a key improvement in that it has a uniformly downward bias (Weir and Goudet, 2017; Ochoa and Storey, 2021). Lastly, the popkin estimator is the only unbiased estimator under arbitrary relatedness (Ochoa and Storey, 2021). To the best of our knowledge, the new WG and popkin estimators have not been used in association studies before, but represent potential improvements over the use of the standard estimator for association.

One potential confounder when comparing the above kinship estimators is that the standard estimator upweighs rare variants in a formulation previously called "mean-of-ratios" (MOR), whereas WG and popkin do not, instead following a "ratio-of-means" (ROM) estimation strategy (Bhatia et al., 2013; Ochoa and Storey, 2021). Recent work also formulated a ROM version of the standard estimator, which has a more predictable bias than the widely used MOR version (Ochoa and Storey, 2021). Following a locus weight formulation that allows the standard estimator to weigh loci in both ways (Wang et al., 2017), here we generalize the popkin and WG estimators to have both MOR and ROM versions, to test estimators without confounding by locus weighing strategy.

In this work, we originally hypothesized that kinship estimation bias would affect association testing. We perform evaluations using an admixed family simulation (Yao and Ochoa, 2022) as well as real genotypes from the 1000 Genomes project (Consortium, 2010; 1000 Genomes Project Consortium et al., 2012; Fairley et al., 2020), in both cases with simulated traits, to characterize type I error control and power using robust statistics. Surprisingly, we find that both LMM and PCA

association statistics are largely invariant to kinship estimation bias. We theoretically characterize the conditions under which these kinship biases result in invariant association statistics, which encompass changing ancestral population in the kinship matrix too. As we discover that most kinship estimates are non-positive semidefinite (non-PSD), breaking a key modeling assumption, we perform additional empirical validations and discover that some LMMs can handle these improper covariance matrices surprisingly well. Overall, we find that long-used association approaches are unaffected by the most common kinship estimation biases, and develop theory that may help improve association and related approaches such as heritability estimation.

2 Methods

2.1 Genetic model

The following genetic model justifies the use of kinship matrices in association studies, and is the basis of all kinship estimation bias calculations that our theoretical work depends upon.

Suppose there are m biallelic loci and n diploid individuals. The genotype $x_{ij} \in \{0, 1, 2\}$ at a locus i of individual j is encoded as the number of reference alleles, for a preselected but otherwise arbitrary reference allele per locus. Genotypes are treated as random variables structured according to relatedness. If T is the ancestral population on which allele frequencies are conditioned, φ_{jk}^T is the kinship coefficient of two individuals j and k, and p_i^T is the ancestral allele frequency at locus i, then under the kinship model (Malécot, 1948; Wright, 1949; Jacquard, 1970; Astle and Balding, 2009; Ochoa and Storey, 2021) the expectation and covariance are given by

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

where $\mathbf{x}_i = (x_{ij})$ is the length-n column vector of genotypes at locus i, $\mathbf{\Phi}^T = (\varphi_{jk}^T)$ is the $n \times n$ kinship matrix, and $\mathbf{1}$ is a length-n column vector of ones. Both $\mathbf{\Phi}^T$ and p_i^T are parameters that depend on the choice of ancestral population, for which the Most Recent Common Ancestor (MRCA) population is the most sensible choice (Ochoa and Storey, 2021). However, one of the results of this work is proof that the choice of ancestral population does not affect association testing.

2.2 Kinship estimators

Each subsection below corresponds to a kinship estimator bias type: Popkin is unbiased, while Standard and WG have different bias functions (defined shortly). Each estimator bias type has two locus weight types called *ratio-of-means* (ROM) and *mean-of-ratios* (MOR), a terminology that follows previous convention for these and related estimators (Bhatia et al., 2013; Ochoa and Storey, 2021). Only ROM estimators have closed-form limits. Below $\hat{p}_i^T = \frac{1}{2n}\mathbf{x}_i^\intercal \mathbf{1}$ is the standard ancestral allele frequency estimator, where the $\intercal$ superscript denotes matrix transposition (do not confuse with ancestral population superscript T), and $\hat{\mathbf{\Phi}}^{T,\text{name}} = (\hat{\varphi}_{jk}^{T,\text{name}})$ relates the scalar and matrix formulas of each named kinship estimator.

2.2.1 Popkin estimator

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

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

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

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

under the assumption that the true minimum kinship is zero. The MOR version, introduced here, upweighs rare variants by using $w_i = \left(\hat{p}_i^T\left(1-\hat{p}_i^T\right)\right)^{-1}$; although it has no closed-form limit, it is approximately unbiased as well (Appendix A) and it is connected to the most common estimator, Standard MOR (Appendix B). The use of locus weights here is inspired by previous calculations relating the standard ROM and MOR estimators (Wang et al., 2017).

2.2.2 Standard estimator

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

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

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

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

$$\hat{\boldsymbol{\Phi}}^{T,\text{std-ROM}} \xrightarrow[m\to\infty]{\text{a.s.}} F^{\text{std}}\left(\boldsymbol{\Phi}^T\right) = \frac{1}{1-\bar{\varphi}^T}\left(\boldsymbol{\Phi}^T + \bar{\varphi}^T\mathbf{J} - \boldsymbol{\varphi}^T\mathbf{1}^\intercal - \mathbf{1}\left(\boldsymbol{\varphi}^T\right)^\intercal\right), \tag{4}$$

where $\mathbf{J} = \mathbf{1}\mathbf{1}^\intercal$ is the $n \times n$ matrix of ones, $\boldsymbol{\varphi}^T = \frac{1}{n}\boldsymbol{\Phi}^T\mathbf{1}$ is a length-n vector of per-row mean kinship values, and $\bar{\varphi}^T = \frac{1}{n^2}\mathbf{1}^\intercal\boldsymbol{\Phi}^T\mathbf{1}$ is the scalar overall mean kinship. The MOR estimator does not have closed-form limit, but it is well approximated by Eq. (4) in practice, especially when loci with small minor allele frequencies are excluded prior to calculating this estimate. In Appendix B we prove that, when there are no missing genotypes, the two standard estimators are functions of the corresponding popkin estimators, given by the bias function F^{std}:

$$\hat{\boldsymbol{\Phi}}^{T,\text{std-ROM}} = F^{\text{std}}\left(\hat{\boldsymbol{\Phi}}^{T,\text{popkin-ROM}}\right),$$
$$\hat{\boldsymbol{\Phi}}^{T,\text{std-MOR}} = F^{\text{std}}\left(\hat{\boldsymbol{\Phi}}^{T,\text{popkin-MOR}}\right).$$

2.2.3 Weir-Goudet estimator

The ROM version of the Weir-Goudet (WG) kinship estimator is given by (Weir and Goudet, 2017; Ochoa and Storey, 2021)

$$\hat{\varphi}_{jk}^{T,\text{WG-ROM}} = 1 - \frac{A_{jk}}{\hat{A}_{\text{avg}}}, \qquad \hat{A}_{\text{avg}} = \frac{2}{n(n-1)}\sum_{j=2}^{n}\sum_{k=1}^{j-1} A_{jk}, \tag{5}$$

where A_{jk} is as in Eq. (1). Its biased limit is also a function of the true kinship matrix:

$$\hat{\mathbf{\Phi}}^{T,\text{WG-ROM}} \xrightarrow[m\to\infty]{\text{a.s.}} F^{\text{WG}}\left(\mathbf{\Phi}^T\right) = \frac{1}{1-\tilde{\varphi}^T}\left(\mathbf{\Phi}^T - \tilde{\varphi}^T\mathbf{J}\right), \tag{6}$$

where $\tilde{\varphi}^T$ is the mean kinship excluding the matrix diagonal:

$$\tilde{\varphi}^T = \frac{2}{n(n-1)}\sum_{j=2}^{n}\sum_{k=1}^{j-1}\varphi_{jk}^T. \tag{7}$$

In Appendix C we prove that

$$0 \le \tilde{\varphi}^T \le \bar{\varphi}^T \le 1,$$

and equalities are achieved if and only if all kinship values are equal. Since the WG-ROM estimator closely resembles the popkin estimator in Eq. (1), it is clear that they are related by the bias function F^{WG}, while WG-MOR is introduced here and defined by the below formula:

$$\hat{\mathbf{\Phi}}^{T,\text{WG-ROM}} = F^{\text{WG}}\left(\hat{\mathbf{\Phi}}^{T,\text{popkin-ROM}}\right),$$
$$\hat{\mathbf{\Phi}}^{T,\text{WG-MOR}} = F^{\text{WG}}\left(\hat{\mathbf{\Phi}}^{T,\text{popkin-MOR}}\right).$$

2.3 Association models

LMM and PCA are closely related association models (Astle and Balding, 2009; Hoffman, 2013; Yao and Ochoa, 2022):

$$\text{LMM:}\quad \mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta_i + \mathbf{s} + \boldsymbol{\epsilon}, \tag{8}$$

$$\mathbf{s} \sim \text{Normal}\left(\mathbf{0}, 2\sigma^2\mathbf{\Phi}^T\right), \tag{9}$$

$$\text{PCA:}\quad \mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta_i + \mathbf{U}_r\boldsymbol{\gamma}_r + \boldsymbol{\epsilon}, \tag{10}$$

$$\mathbf{\Phi}^T = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\intercal, \tag{11}$$

where $\mathbf{y}$ is a length-n vector of continuous trait values, α is the intercept coefficient, β_i is the genetic effect (association) coefficient of locus i, $\mathbf{s}$ is the (genetic) random effect, σ^2 is the random effect

variance factor, $\mathbf{U}_r$ is the $n \times r$ matrix of top-r eigenvectors (PCs) of $\mathbf{\Phi}^T$, $\boldsymbol{\gamma}_r$ is a length-r vector of coefficients for each eigenvector, $\boldsymbol{\epsilon} \sim \text{Normal}(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I})$ are random independent residuals, and $\mathbf{I}$ is the $n \times n$ identity matrix. Furthermore, Eq. (11) is the complete eigendecomposition of $\mathbf{\Phi}^T$, where $\mathbf{U}$ is the $n \times n$ matrix of eigenvectors, and $\mathbf{\Lambda}$ is the $n \times n$ diagonal matrix of eigenvalues. As $\mathbf{s}$ and $\mathbf{U}_r$ play analogous roles in modeling the effect of relatedness in LMM and PCA, respectively, we refer to them jointly as relatedness effects, and σ^2 and $\boldsymbol{\gamma}_r$ as their coefficients.

2.4 Simulations

2.4.1 Admixed family genotype simulation

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

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

where this Beta is the Balding-Nichols distribution (Balding and Nichols, 1995) with mean p_i^T and variance $p_i^T \left(1 - p_i^T\right) f_{S_u}^T$. This is implemented in the R package `bnpsd`.

We also include family structure in the simulation. 20 generations are generated iteratively. Individuals in the first generation ($n = 1000$) are ordered by 1D geography, randomly assigned sex, and treated as locally unrelated. From the next generation, 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`.

2.4.2 Trait simulation algorithm

Given an $m \times n$ genotype matrix $\mathbf{X} = (\mathbf{x}_i^\intercal)$, traits are simulated from

$$\mathbf{y} = \mathbf{1}\alpha + \mathbf{X}^\intercal\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \text{Normal}(\mathbf{0}, (1-h^2)\mathbf{I}).$$

Given a desired number of causal loci $m_1 = n/10$ and heritability $h^2 = 0.8$, the goal is to choose causal coefficients $\boldsymbol{\beta}$ and the intercept α that result in zero mean and the desired trait heritability. Here, we use the "fixed effect sizes" trait simulation model described in (Yao and Ochoa, 2022). Briefly, first m_1 causal loci are randomly selected, and for these steps only $\mathbf{X}$ is subset to these loci and reindexed. For known p_i^T, causal coefficients are constructed as:

$$\beta_i = \sqrt{\frac{h^2}{2m_1 v_i^T}},$$

where $v_i^T = p_i^T\left(1-p_i^T\right)$; for unknown p_i^T (real genotypes), the unbiased estimate $\hat{v}_i^T = \hat{p}_i^T\left(1-\hat{p}_i^T\right)/\left(1-\bar{\varphi}^T\right)$ is used, where $\bar{\varphi}^T$ is the mean kinship estimated from `popkin`. Coefficients are made negative randomly with probability 0.5. For known p_i^T, we obtain the desired zero trait mean with $\alpha = -2\left(\mathbf{p}^T\right)^\intercal\boldsymbol{\beta}$, where here $\mathbf{p}^T = (p_i^T)$ contains causal loci only. When p_i^T are unknown, to avoid covariance distortions, the intercept coefficient is constructed as

$$\alpha = -2\hat{\bar{p}}^T\mathbf{1}_{m_1}^\intercal\boldsymbol{\beta}, \qquad \hat{\bar{p}}^T = \frac{1}{m_1}\mathbf{1}_{m_1}^\intercal\hat{\mathbf{p}}^T,$$

where $\mathbf{1}_{m_1}$ is a length-m_1 column vector of ones.

2.5 Real genotype data processing

To evaluate different kinship estimators on a real dataset, we use the high-coverage NYGC version of the 1000 Genomes Project (Fairley et al., 2020), which is processed as before (Yao and Ochoa, 2022). Briefly, using `plink2` (Chang et al., 2015) we keep only autosomal biallelic SNP loci with filter "PASS", pruned for linkage disequilibrium with parameters "`--indep-pairwise 1000kb 0.3`" to remove loci that have a greater than 0.3 squared correlation coefficient with other loci within 1000kb, and lastly remove loci with minor allele frequencies < 0.01. The resulting data have $m = 1,111,266$ loci and $n = 2,504$ individuals. Traits are simulated for this dataset with $m_1 = n/10 = 250$ causal loci.

2.6 Evaluation of performance

AUC_{PR} and SRMSD_p are used to evaluate approaches as before (Yao and Ochoa, 2022). Briefly, SRMSD_p (Signed Root Mean Square Deviation) measures the difference between the observed null p-value quantiles and the expected uniform quantiles:

$$\text{SRMSD}_p = \text{sgn}(u_{\text{median}} - p_{\text{median}}) \sqrt{\frac{1}{m_0} \sum_{i=1}^{m_0} \left(u_i - p_{(i)}\right)^2},$$

where $m_0 = m - m_1$ is the number of null (non-causal) loci, i indexes null loci only, $p_{(i)}$ is the ith ordered null p-value, $u_i = (i-0.5)/m_0$ is its expectation, p_{median} is the median observed null p-value, $u_{\text{median}} = \frac{1}{2}$ is its expectation, and sgn is the sign function (1 if $u_{\text{median}} \geq p_{\text{median}}$, -1 otherwise). $\text{SRMSD}_p = 0$ corresponds to calibrated p-values, $\text{SRMSD}_p > 0$ indicate anti-conservative p-values, and $\text{SRMSD}_p < 0$ are conservative p-values.

AUC_{PR} (Area Under the Precision and Recall Curve) is a binary classification measure that reflects calibrated power (Yao and Ochoa, 2022), which is calculated from the total numbers of true

positives (TP), false positives (FP) and false negatives (FN) at some threshold or parameter t:

$$\text{Precision}(t) = \frac{\text{TP}(t)}{\text{TP}(t) + \text{FP}(t)},$$
$$\text{Recall}(t) = \frac{\text{TP}(t)}{\text{TP}(t) + \text{FN}(t)},$$

followed by calculating the area under the curve traced as t varies recall from zero to one. Higher AUC_{PR} is better, with best performance at $\text{AUC}_{\text{PR}} = 1$ for a perfect classifier, while worst performance at $\text{AUC}_{\text{PR}} = \frac{m_1}{m}$ (overall proportion of causal loci) is for random classifiers.

2.7 Software

Popkin kinship estimates are calculated with the `popkin` R package. Standard MOR kinship estimates are calculated with GCTA (version 1.93.2beta). All other kinship estimators and limits are calculated using the `popkinsuppl` R package. PCs are calculated with the `eigen` function of R.

GCTA is used to run all LMM associations (Yang et al., 2011; Yang et al., 2014). We pass $2\mathbf{\Phi}^T$ for all kinship matrices tested (the same scale as its own kinship estimate). PCA association is performed with `plink2` (Chang et al., 2015). We use $r = K - 1 = 2$ PCs for the admixed family simulations, and $r = 10$ PCs for 1000 Genomes.

3 Results

3.1 Empirical analysis using admixed family simulation

To quantify the effect of kinship estimation bias, we simulate genotypes and traits, and calculate association p-values using a factorial design that tests all kinship matrix (3 bias types, times two locus weight types and one limit) and association model (PCA and LMM) combinations. We simulate an admixed population with $K = 3$ ancestries, who serve as founders for a 20-generation random pedigree. This high-dimensional admixed family scenario yields a large difference in performance between PCA and LMM (Yao and Ochoa, 2022).

Kinship estimates and limits for this simulation are shown in Fig. 1. The true kinship matrix

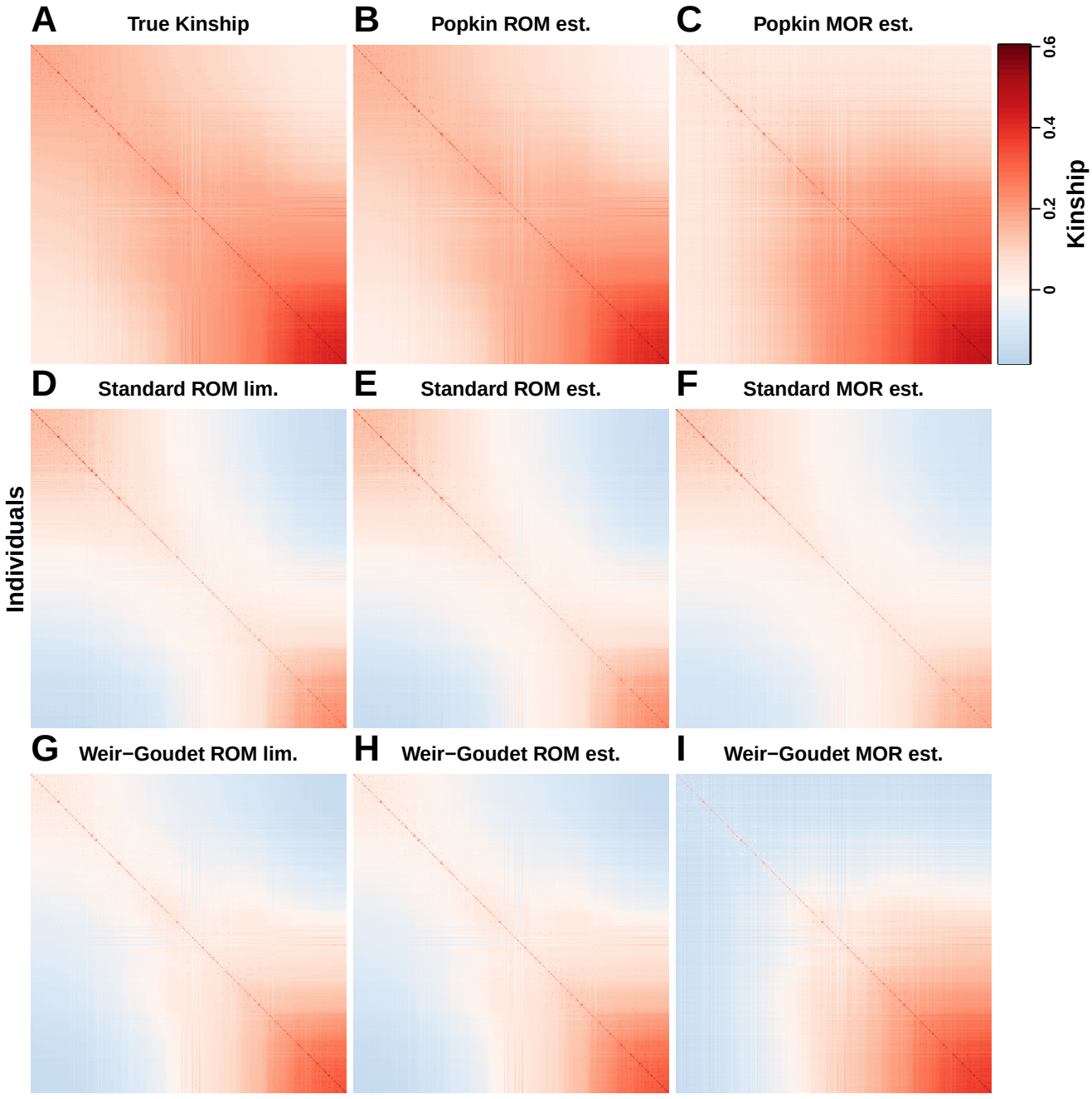

Figure 1: **Kinship estimates and limits on the admixed family simulation.** Each panel shows a kinship matrix as a heatmap, with each of the $n = 1000$ individuals along both x and y axes, color represents kinship: positive values in red, negative in blue. Diagonal contains inbreeding values. Each estimator bias type (Popkin, Standard, and Weir-Goudet; rows) has three matrices (columns): two locus weight types (ROM (ratio of means) and MOR (mean of ratios)) and limit of ROM.

shows the family relatedness as high values concentrated near the diagonal and the ancestry-driven population structure as the broad patterns off-diagonal. Only Popkin ROM is unbiased, while popkin MOR has a slight upward bias that varies across the matrix (Fig. S1A). In contrast, the Standard and Weir-Goudet (WG) estimates have large downward biases overall, resulting in abundant negative values; Standard biases vary for every pair of individuals, while WG has a uniform bias.

We perform LMM and PCA association tests to determine how kinship biases affect association performance. Surprisingly, we find that kinship bias type does not have a discernible effect on association performance, as summarized by $\mathrm{AUC_{PR}}$ (a robust proxy for power; Fig. 2) and SRMSD_p (measures null statistic calibration; Fig. S2). The largest differences in performance are explained

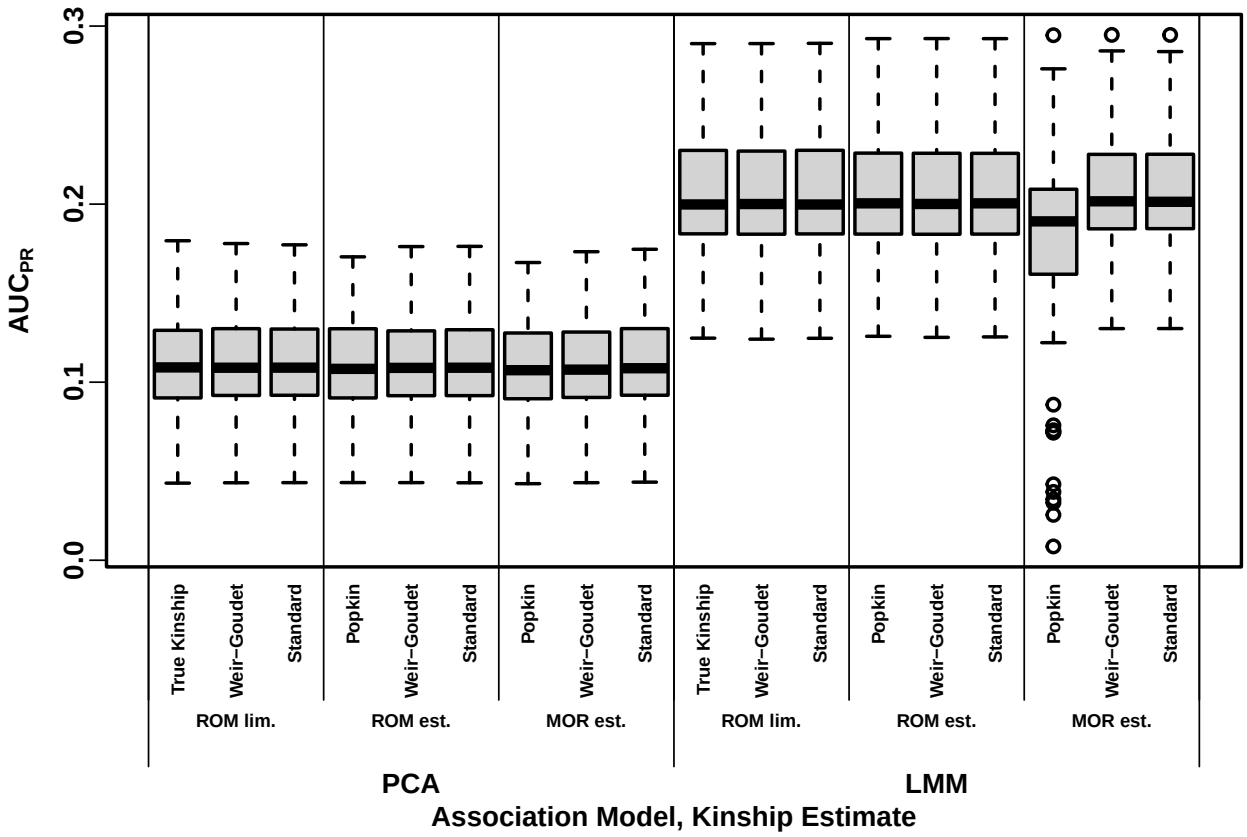

Figure 2: **Distributions of Area Under the Precision-Recall Curve ($\mathbf{AUC_{PR}}$) on the admixed family simulation.** Higher $\mathrm{AUC_{PR}}$ is better performance. Results for 100 replicates (each a random genotype matrix and trait vector). Approaches cluster primarily by association model (LMM or PCA), and vary little across bias types.

by the association model (LMM vs PCA), as expected due to our use of a family simulation where PCA performs poorly. Within association models, there are no clear differences between the performance of any of the kinship matrices, in fact many appear to have identical distributions (both statistics), the only clear exception being LMM popkin MOR, which has a few outlier replicates where performance is exceedingly poor.

To better characterize the nearly identical performance distribution just observed, we next mea-

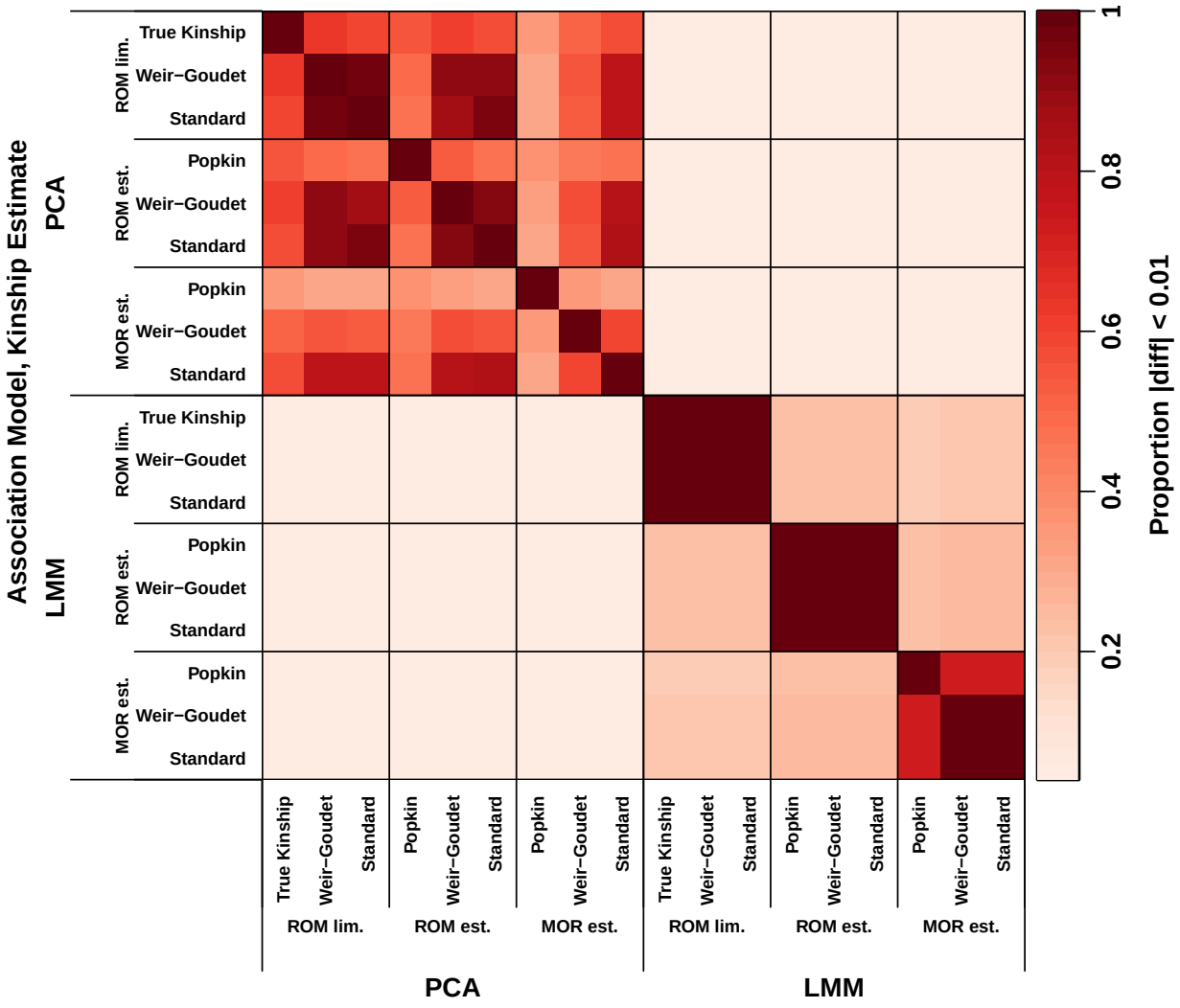

Figure 3: **Approximate agreement between p-values on the admixed family simulation.** Calculated agreement (absolute difference under 0.01) averaged over loci (color) of association p-values between association models (LMM vs PCA) and kinship matrices (x and y axes). All 100 replicates are used. Different bias types (matched for association model and locus weight type) have large proportions of nearly identical p-values.

sure the agreement between individual association p-values. We calculate the proportion of loci between two methods with p-values within 0.01 of each other, which is an approximate measure of agreement, and find a remarkably high agreement between estimators of different bias types after matching association model and locus weight type or limit (Fig. 3). This is in contrast to low agreement between PCA and LMM statistics, and between LMM statistics with different locus weight types or limits. Minimum agreements are higher across PCA methods, though here the true kinship or popkin estimates disagree more from Standard and WG matrices. Overall, kinship matrices with different bias types (otherwise matched) result in nearly identical association statistics.

3.2 Empirical analysis using 1000 Genomes

Now we repeat our analysis using the real genotypes of 1000 Genomes. Kinship estimates are shown in Fig. 4 (note real data have no true kinship or estimator limits). Popkin ROM estimates display an approximate nested block structure that arises from the tree relationships between subpopulations (Fig. 4A; trees were explicitly fit to this data in previous work (Yao and Ochoa, 2022)). However, popkin MOR estimates do not follow the nested blocks tree structure, since kinship between African and non-African populations is higher than kinship within African populations (Fig. 4B). Standard estimates have values are closer to zero, and a different bias for each pair of individuals, resulting in higher relative kinship for African compared to non-African populations (Fig. 4C-D). Lastly, WG estimates are uniformly smaller than popkin's and attain large negative values (Fig. 4E-F).

Our association test conclusion are similar to our simulation study: AUC_{PR} and SRMSD_p distributions are nearly identical for estimators of different bias types but same locus weight type (ROM or MOR) and association model. However, unlike the simulation, here the MOR estimates noticeably outperform ROM estimates (LMM only), in terms of both AUC_{PR} (Fig. 5) and SRMSD_p (Fig. S3). P-values are again nearly identical at a large proportion of loci between approaches with matched association model and locus weight type (MOR or ROM), regardless of bias type (Fig. S4).

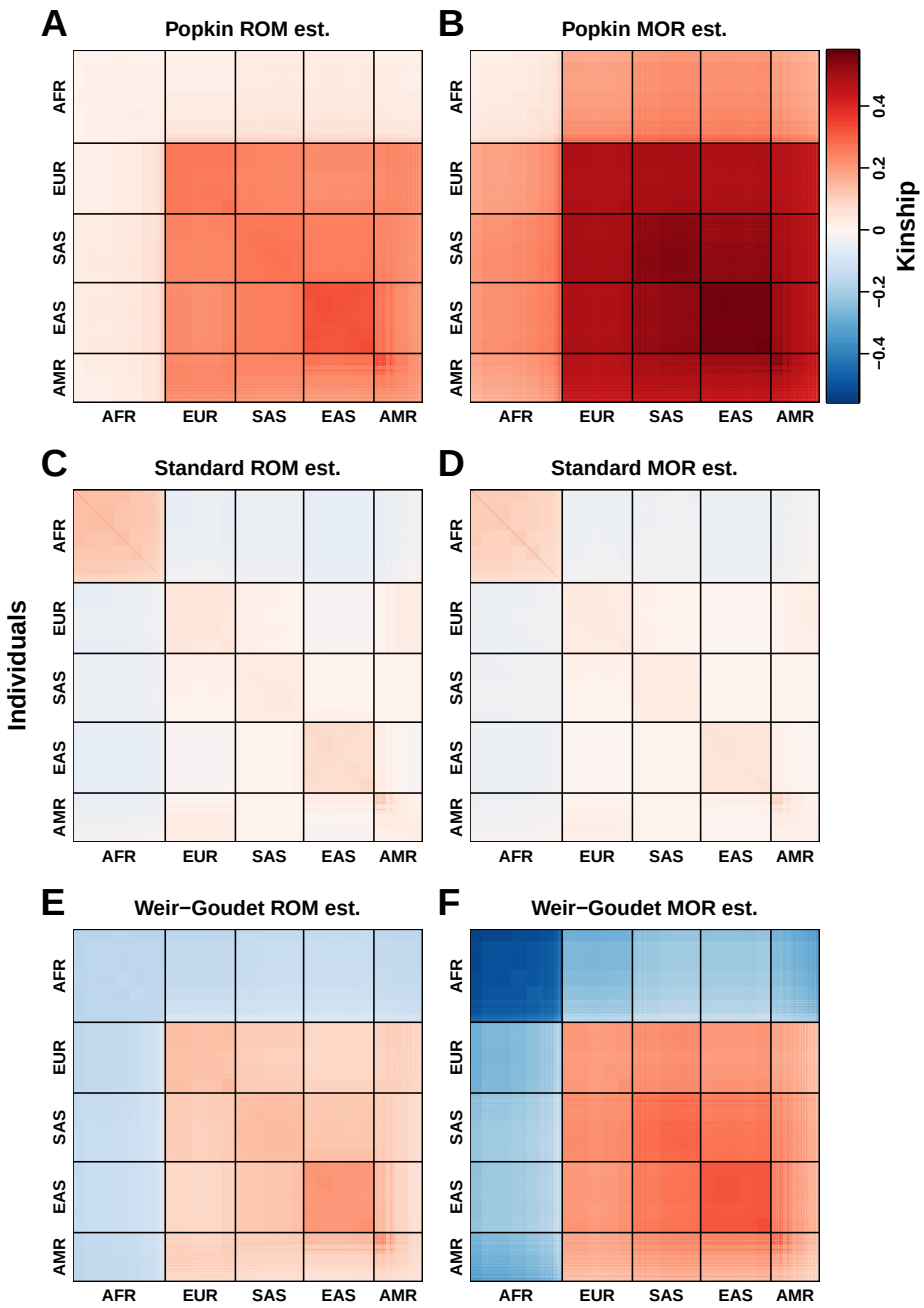

Figure 4: **Kinship estimates on 1000 Genomes.** Each panel represents a kinship matrix as a heatmap, as in Fig. 1. Superpopulation codes: AFR = African, EUR = European, SAS = South Asian, EAS = East Asian, AMR = Admixed Americans (Hispanics). Each estimator bias type (Popkin, Standard, and Weir-Goudet; rows) has two locus weight types (columns): ROM (ratio of means) and MOR (mean of ratios). In this visualization the upper range of all panels is capped to the 99 percentile of the diagonal (population inbreeding values) of the popkin MOR estimates.

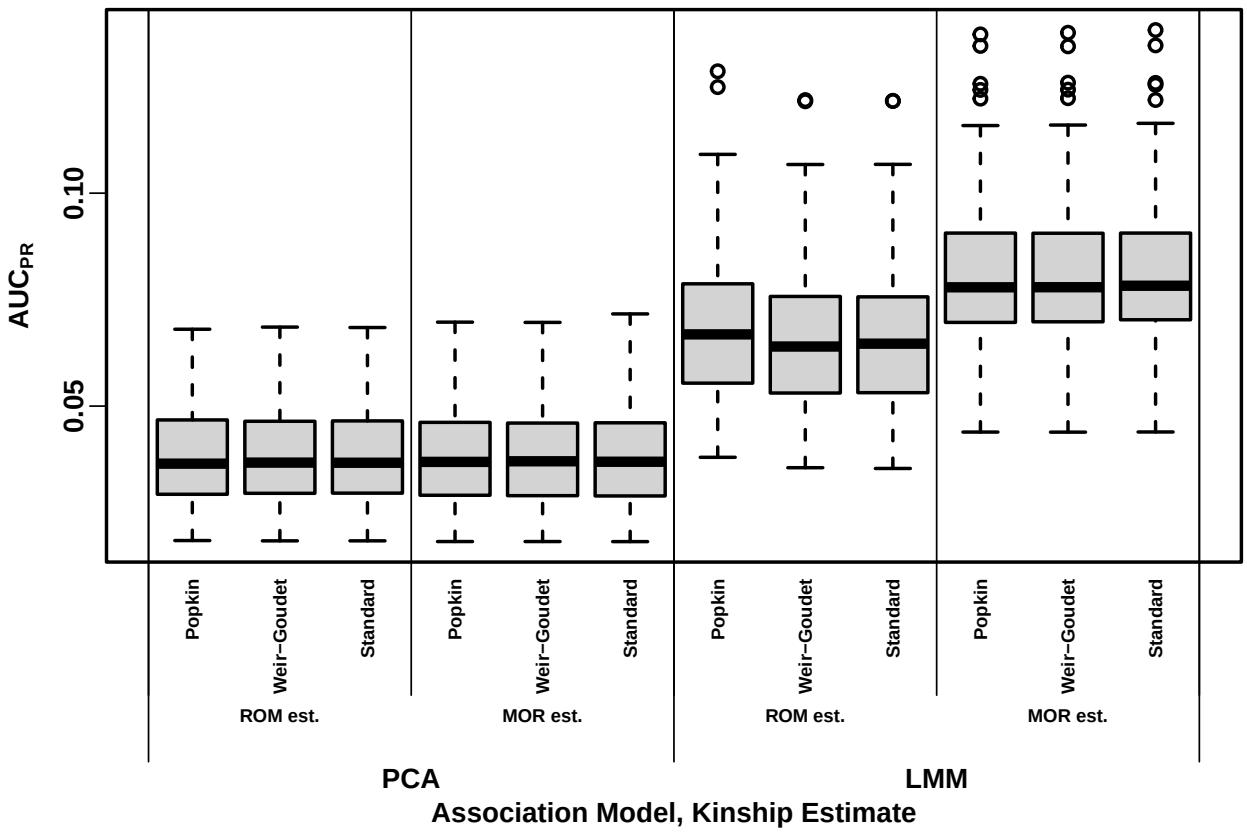

Figure 5: **Distributions of Area Under the Precision-Recall Curve ($\mathbf{AUC_{PR}}$) on 1000 Genomes.** Higher $\mathrm{AUC_{PR}}$ is better performance. Results based on 100 simulated trait replicates (real genotype matrix is fixed). Approaches cluster primarily by association model (LMM or PCA) and locus weight type (ROM or MOR), and do not depend much at all on the bias type.

3.3 Proof of association invariability to common kinship biases

Our empirical observations suggest that replacing a kinship matrix with either the Standard or WG-biased version does not alter association statistics (with exceptions we attribute to numerical precision artifacts); here prove a more general version of these facts mathematically. Our constructive proof shows that only a regression model with relatedness effects as covariates and an intercept is required, whose coefficients adapt to the bias, and no other coefficients change. This is fortunate, as the intercept and relatedness effect coefficients are nuisance parameters that usually go unreported, while the focal genetic association coefficient and its p-value are unchanged by these biases.

The most general form we identified of the bias function, mapping a kinship matrix to its bias-transformed version, and for which association invariability holds, is

$$\mathbf{\Phi}^{T\prime} = F\left(\mathbf{\Phi}^{T}\right) = \frac{1}{c}\mathbf{B}\mathbf{\Phi}^{T}\mathbf{B}^{\intercal}, \qquad \mathbf{B} = \mathbf{I} - \mathbf{1}\mathbf{b}^{\intercal}, \tag{12}$$

where c is any positive scalar and $\mathbf{b}$ is any length-n vector. The key property that the linear operator $\mathbf{B}$ must satisfy is that it shifts the input vector by the same scalar across its values, or

$$\mathbf{B}\mathbf{s} = \mathbf{s} - \mathbf{1}\eta, \tag{13}$$

where $\mathbf{s}$ is any vector and the scalar $\eta = \mathbf{b}^{\intercal}\mathbf{s}$ is a function of the input vector. $\mathbf{B}$ in Eq. (12) is the only form that results in Eq. (13).

The Standard bias function $F = F^{\text{std}}$ of Eq. (4) can be written as Eq. (12) with $c = 1 - \bar{\varphi}^{T}$ and $\mathbf{b} = \frac{1}{n}\mathbf{1}$, in which case $\mathbf{B}$ equals the centering matrix. Further, the generalized Standard estimator studied in Ochoa and Storey (2021) has $\mathbf{b}$ be a vector of individual weights that sum to one: $\mathbf{b}^{\intercal}\mathbf{1} = 1$. These $\mathbf{B}$ and $\mathbf{\Phi}^{T\prime}$ are singular transformations (they are not invertible and have a zero eigenvalue), since $\mathbf{B}\mathbf{1} = \mathbf{0}$ and $\mathbf{B}^{\intercal}\mathbf{b} = \mathbf{0}$.

The WG bias function $F = F^{\mathrm{WG}}$ of Eq. (6) can be written as Eq. (12) with $c = 1 - \tilde{\varphi}^T$ and

$$\mathbf{b} = q\frac{\left(\mathbf{\Phi}^T\right)^{-1}\mathbf{1}}{\mathbf{1}^\intercal\left(\mathbf{\Phi}^T\right)^{-1}\mathbf{1}}, \tag{14}$$

$$q = 1 \pm \sqrt{1 - \tilde{\varphi}^T\left(\mathbf{1}^\intercal\left(\mathbf{\Phi}^T\right)^{-1}\mathbf{1}\right)}. \tag{15}$$

The derivation of this factorization is given in Appendix D. The determinant of the quadratic solution q would be non-negative if $\tilde{\varphi}^T$ satisfied $\tilde{\varphi}^T \leq 1/\left(\mathbf{1}^\intercal\left(\mathbf{\Phi}^T\right)^{-1}\mathbf{1}\right)$. However, the actual $\tilde{\varphi}^T$ does not satisfy this inequality in any of our empirical cases, and in fact $1/\left(\mathbf{1}^\intercal\left(\mathbf{\Phi}^T\right)^{-1}\mathbf{1}\right) \leq \bar{\varphi}^T$ holds (proven in Appendix E; although $\tilde{\varphi}^T \leq \bar{\varphi}^T$ (Appendix C), in practice those two are very close while $1/\left(\mathbf{1}^\intercal\left(\mathbf{\Phi}^T\right)^{-1}\mathbf{1}\right)$ is much smaller than both), so b above is complex. This is a consequence of WG estimates being non-PSD, which we elaborate in the following sections. Nevertheless, PCA as well as the GCTA algorithms work for non-PSD matrices without invoking complex numbers (following sections and Appendix F).

3.3.1 Proof for LMM case

Consider a random effect $\mathbf{s}$ drawn using $\mathbf{\Phi}^T$, as given in Eq. (9). Using the affine transformation property of Multivariate Normal distributions (which holds even if $\mathbf{B}$ below is singular) and Eq. (12), then

$$\mathbf{s}' = \mathbf{B}\mathbf{s} \sim \mathrm{Normal}\left(\mathbf{0}, 2\sigma^{2\prime}\mathbf{\Phi}^{T\prime}\right),$$

$$\sigma^{2\prime} = c\sigma^2. \tag{16}$$

(This $\mathbf{s}'$ has a degenerate distribution for Standard bias, since $\mathbf{\Phi}^{T\prime}$ is singular, but $\mathbf{s}' + \boldsymbol{\epsilon}$ is usually non-degenerate, since its covariance $\mathbf{V}' = 2\sigma^{2\prime}\mathbf{\Phi}^{T\prime} + \sigma_\epsilon^2\mathbf{I}$ is invertible as long as $\sigma_\epsilon^2 \neq 0$.) Replacing $\mathbf{B}\mathbf{s}$ with the shift form in Eq. (13) shows that $\mathbf{s}' = \mathbf{s} - \mathbf{1}\eta$ are equal in distribution. Therefore, the random effect $\mathbf{s}'$ of the biased kinship matrix differs from the random effect $\mathbf{s}$ of the original kinship only by $\mathbf{1}\eta$, a difference compensated for by adjusting the intercept coefficient in Eq. (8):

$$\alpha' = \alpha + \eta. \tag{17}$$

No other regression coefficients, or the total residuals, change when $\mathbf{\Phi}^T$ is replaced with $\mathbf{\Phi}^{T\prime}$, including the association coefficient β_i that is the focus of the test.

The above results require PSD kinship matrices $\mathbf{\Phi}^{T\prime}$, which covariance matrices must be, and which are characterized by non-negative eigenvalues and determinants. Nevertheless, for the non-PSD WG bias (has a negative eigenvalue) combined with the generalized least squares association algorithm, which is used by GCTA and other LMMs (Kang et al., 2008; Kang et al., 2010; Yang et al., 2014), we find a stronger result consistent with Eq. (17), namely that $\alpha' = \alpha$, or in other words, $\eta = 0$ (Appendix F).

The LMM association p-value does not change in several common tests, including the F-test, since it only depends on the residuals and these do not change, as well as the likelihood ratio test, because although covariance determinants change, they cancel out in the ratio. The Wald test used by GCTA (Yang et al., 2014) is also invariant to these kinship biases given our empirical results in Figs. 3 and S4 and proven explicitly for WG bias in Appendix F. Lastly, we confirmed empirically that the Score test for the GCTA model is also invariant to these kinship biases (not shown). These arguments hold whether variance components are fit with maximum likelihood or restricted maximum likelihood (Kang et al., 2008; Kang et al., 2010; Yang et al., 2014), since multiplying the estimated genetic variance component σ^2 by c and adjusting the intercept compensates for the bias regardless of how $\sigma^2, \sigma^2_\epsilon$ are estimated.

3.3.2 Proof for PCA case

We present a proof for the PCA case that relies on an approximation that holds well in practice. Based on the PCA model of Eqs. (10) and (11), let $\mathbf{U}_r$ be the top eigenvectors of $\mathbf{\Phi}^T$, and $\mathbf{U}'_r$ those of $\mathbf{\Phi}^{T\prime}$. They key approximation is that

$$\mathbf{U}'_r \approx \mathbf{B}\mathbf{U}_r, \tag{18}$$

which is not strictly equal (since $\mathbf{B}\mathbf{U}_r$ is not generally orthogonal, as eigenvectors must be), but we have found it to be a good approximation in practice. In this case the eigenvector coefficients need not change, $\boldsymbol{\gamma}'_r = \boldsymbol{\gamma}_r$, since the difference in scale of the kinship matrices (c in Eq. (12)) is absorbed

by the eigenvalues not present in this model. Applying the shift of Eq. (13) shows that

$$\mathbf{U}_r'\boldsymbol{\gamma}_r' = \mathbf{B}\mathbf{U}_r\boldsymbol{\gamma}_r = \mathbf{U}_r\boldsymbol{\gamma}_r - \mathbf{1}\eta,$$

where $\eta = \mathbf{b}^\intercal\mathbf{U}_r\boldsymbol{\gamma}_r$ is a scalar. Therefore, the relatedness effects again differ only by $\mathbf{1}\eta$, which is compensated for by adjusting the intercept using Eq. (17), so the association coefficient β_i and the residuals are the same in both cases. This proof works if there are small numbers of zero or negative eigenvalues in $\boldsymbol{\Phi}^{T\prime}$ (non-PSD cases), as those rank last and are simply ignored. The observations from LMMs, that p-values are invariant to bias types, also hold for PCA.

We visualize the top PCs of our datasets in Fig. 6 to assess the validity of Eq. (18). The approximation is equivalent to each biased PC (Standard or Weir-Goudet) being shifted from the unbiased PC (Popkin), as described in Eq. (13). Fig. 6 indeed shows that PC1 is shifted by noticeable amounts in each of these cases, while PC2 is less shifted. However, a rotation of the PCs is also noticeable, particularly in the simulated data, and other large differences between MOR estimators, as expected since we know the approximation cannot be exact. Also, PCs can change order upon bias transformation, which we notice in the admixed family simulation, where PC2 and PC3 from popkin (and true kinship) actually correspond to PC1 and PC2, respectively, in both Standard and WG, and are plotted as such. No PC reordering occurs in 1000 Genomes. Overall, while the approximation of Eq. (18) can be weakened to merely require that the biased PCs plus intercept span the same subspace of the unbiased PCs plus intercept, the approximate PC shifts better explain intuitively why the result for LMM is also observed for PCA association.

3.4 Proof of association invariability to change in ancestral population

The kinship matrices we used so far have values that depend on the choice of ancestral population T. Here we consider the effect on association of changing ancestral population, and prove that it is also compensated for by the relatedness and intercept coefficients.

Start from a kinship matrix $\boldsymbol{\Phi}^S$ in terms of ancestral population S, and let T is a population ancestral to S. If the inbreeding coefficient of S when T is the reference ancestral population is f_S^T,

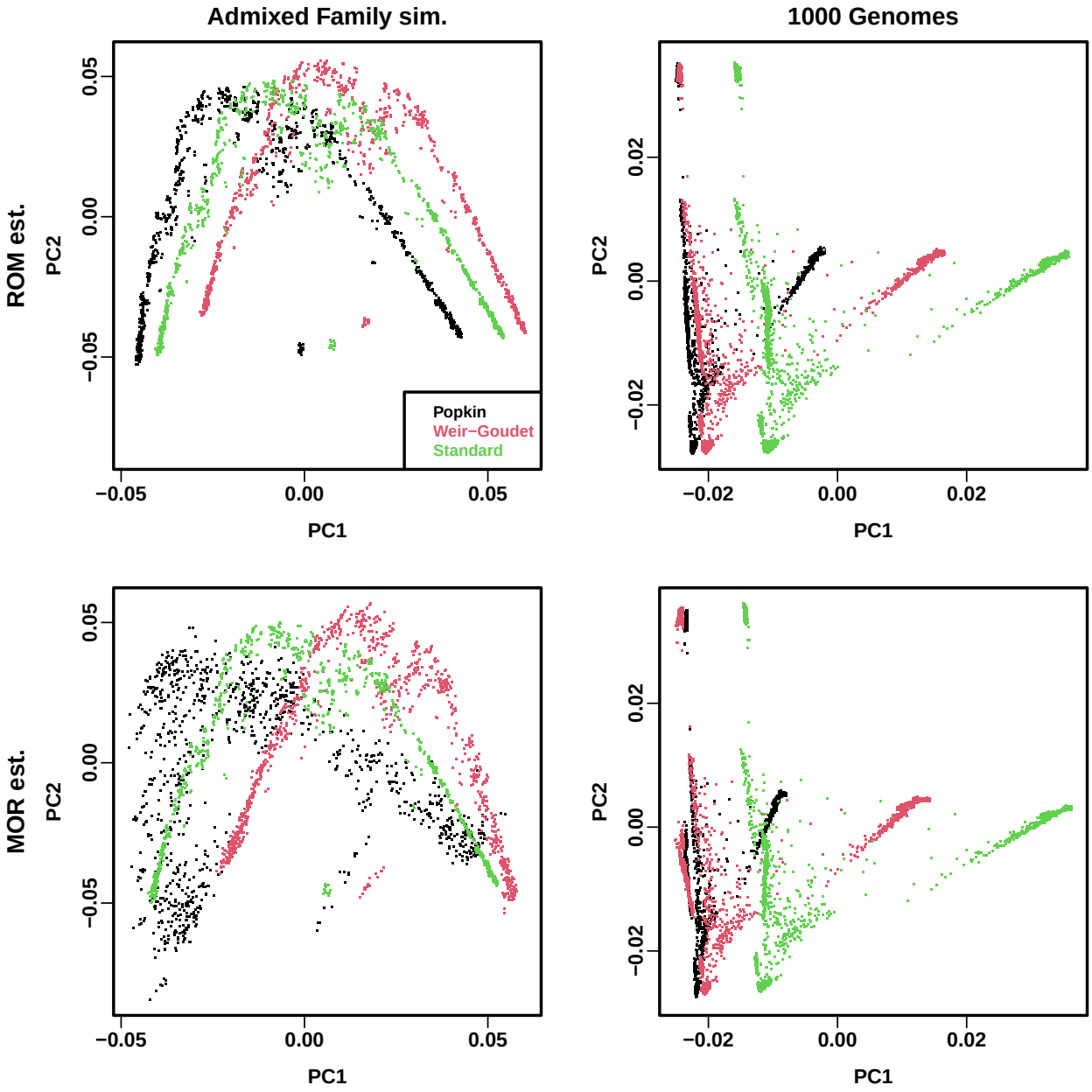

Figure 6: **Visualization of PC shift due to kinship biases.** Each panel shows three estimates (bias types): Popkin, Standard, and Weir-Goudet. ROM estimates are in first row, MOR in second row. (In admixed family, ROM limits are very similar to ROM estimates (not shown).) Columns show estimates from each dataset: admixed family simulation (first replicate) and 1000 Genomes. For popkin (both ROM and MOR estimates) in admixed family only, PC1 and PC2 are replaced with PC2 and PC3 (see text).

then the kinship matrix $\mathbf{\Phi}^T$ in terms of T is given by (Ochoa and Storey, 2021)

$$\left(\mathbf{J} - \mathbf{\Phi}^T\right) = \left(\mathbf{J} - \mathbf{\Phi}^S\right)\left(1 - f_S^T\right).$$

Solving for $\mathbf{\Phi}^T$ and simplifying results in

$$\mathbf{\Phi}^T = \left(1 - f_S^T\right)\mathbf{\Phi}^S + f_S^T\mathbf{J}.$$

This resembles WG bias but in reverse: whereas WG reduces and rescales kinship by $\tilde{\varphi}^T$, changing to a more ancestral population rescales and increases kinship by f_S^T. Indeed, excluding $f_S^T = 1$, this transformation can be written as Eq. (12) with $c = \left(1 - f_S^T\right)^{-1}$ and

$$\mathbf{b} = q\frac{\left(\mathbf{\Phi}^S\right)^{-1}\mathbf{1}}{\mathbf{1}^\intercal\left(\mathbf{\Phi}^S\right)^{-1}\mathbf{1}},$$
$$q = 1 \pm \sqrt{1 + \frac{f_S^T}{1 - f_S^T}\left(\mathbf{1}^\intercal\left(\mathbf{\Phi}^S\right)^{-1}\mathbf{1}\right)}.$$

The determinant of q is strictly positive, since $\mathbf{1}^\intercal\left(\mathbf{\Phi}^S\right)^{-1}\mathbf{1} > 0$ (since $\mathbf{\Phi}^S$ is positive definite, its inverse is too) and $0 \leq f_S^T < 1$. Thus, our previous results apply: ancestor change is compensated for by the relatedness and intercept coefficients, so the association statistics are invariant to this transformation.

3.5 Characterization of non-PSD and singular kinship and trait covariance estimators

While attempting to validate and characterize the earlier factorization of the WG bias function (Eqs. (12) to (15)), we discovered that it does not produce PSD matrices, which covariance matrices are required to be. To characterize this problem more broadly, we calculate the eigenvalues of all kinship matrices $\mathbf{\Phi}^T$ and trait covariance matrices $\mathbf{V} = 2\sigma^2\mathbf{\Phi}^T + \sigma_\epsilon^2\mathbf{I}$, the latter used by LMMs and which we calculate using GCTA's estimates of σ^2 and σ_ϵ^2.

We find that all WG matrices have very large negative minimum eigenvalues, and popkin MOR

estimates also have smaller negative minimum eigenvalues (Fig. S5). Moreover, besides all WG matrices and most popkin MOR estimates, Standard matrices are also often non-PSD but only in 1000 Genomes (Fig. 7), which has missing genotypes (the admixed family simulation does not have missing genotypes). Each of these non-PSD matrices only has one negative eigenvalue. Notably, all popkin ROM estimates are PSD in every evaluation, including under missingness in 1000 Genomes.

In order to quantify matrix singularity, as well as numerical accuracy problems caused by multiplying by inverses of nearly-singular matrices, we calculate condition numbers, which equal the maximum absolute eigenvalue divided by the minimum absolute eigenvalue of our covariance matrices. As expected, we see that Standard kinship matrices are singular on our admixed family simulation (which lacks missingness), as reflected by extremely high condition numbers, but their trait covariances have small condition numbers (Fig. S6). No other matrices are singular, but popkin

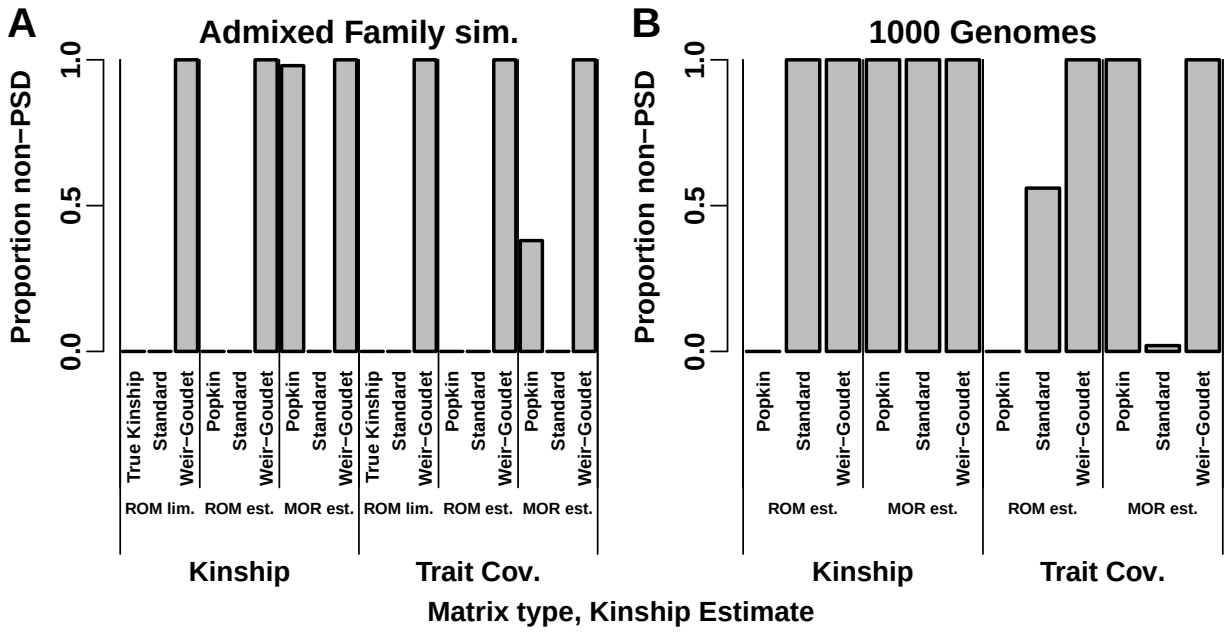

Figure 7: **Proportion of kinship and trait covariance (V) matrices that are not positive semidefinite (PSD).** A matrix is non-PSD if it has negative eigenvalues (below -10^{-7} to allow for limited machine precision). Proportion is calculated over 100 replicates (1000 Genomes kinship has one value since genotypes are fixed, but **V** varies per replicate). **A.** In admixed family simulation, which does not have missing genotypes, all WG matrices and most popkin MOR estimates are non-PSD. All non-PSD kinship matrices results in non-PSD **V** except some popkin ROM estimates yield PSD **V**. **B.** In 1000 Genomes, which has missingness, all kinship estimates are non-PSD except popkin ROM. Of the non-PSD kinship matrices, only some Standard estimates yield PSD **V**.

MOR estimates in the admixed family simulation have relatively high condition numbers for both kinship and trait covariance.

Consider the theoretical connection between the eigenvalues of $\mathbf{\Phi}^T$ and those of $\mathbf{V}$. The eigen-decomposition trick widely used to fit variance components in LMMs (Kang et al., 2008; Lippert et al., 2011; Svishcheva et al., 2012; Zhou and Stephens, 2012; Sul et al., 2018) yields

$$\mathbf{V} = \mathbf{U}\left(2\sigma^2\mathbf{\Lambda} + \sigma_\epsilon^2\mathbf{I}\right)\mathbf{U}^\intercal,$$

where $\mathbf{U}$ and $\mathbf{\Lambda}$ are the eigenvectors and eigenvalues of $\mathbf{\Phi}^T$, respectively (Eq. (11)), so the eigenvectors of $\mathbf{V}$ are also $\mathbf{U}$ and its eigenvalues are $2\sigma^2\mathbf{\Lambda} + \sigma_\epsilon^2\mathbf{I}$. Therefore, since $\sigma^2, \sigma_\epsilon^2 \geq 0$, then if $\mathbf{\Phi}^T$ is positive definite (all of its eigenvalues are positive) then so is $\mathbf{V}$, and the condition number of $\mathbf{V}$ is always smaller (better) or equal than that of $\mathbf{\Phi}^T$. A negative kinship eigenvalue λ_k may become positive for $\mathbf{V}$ only if $\lambda_k > -\sigma_\epsilon^2/(2\sigma^2) = -(1-h^2)/(2h^2)$, so very large negative λ_k values as observed for WG do not become positive in $\mathbf{V}$, in fact they can become more negative (Fig. S5). $\mathbf{V}$ is always invertible and well-conditioned even when $\mathbf{\Phi}^T$ is singular PSD (has zero eigenvalues), as the Standard estimator is under no missingness, since a kinship zero eigenvalue becomes σ_ϵ^2 for $\mathbf{V}$. Conversely, the above equation explains why some non-PSD kinship matrices are particularly problematic: negative eigenvectors near $-\sigma_\epsilon^2/(2\sigma^2)$ can result in ill-conditioned $\mathbf{V}$. We see that popkin MOR estimates are non-PSD (Fig. S5) in such a way that some of their $\mathbf{V}$ are ill-conditioned (Fig. S6), and this explains its poorer performance in the admixed family evaluations (Figs. 2 and S2), as shown in the next subsection.

3.6 Further empirical validation of theoretical predictions

Seeing that WG is always non-PSD, and to query other instances where predictions are not fully met, here we analyze estimation accuracy of various parameters to better understand theoretically and empirically how broken assumptions affect them. With PCA, no deviations from expectation of AUC_{PR} and SRMSD_p are observed for WG (Figs. 2 and 5, Figs. S2 and S3), which makes sense since PCA simply ignores eigenvectors with negative or zero eigenvalues. Therefore, our analysis focuses on LMM, where deviations are observed and clarification regarding WG is needed.

LMMs such as GCTA perform association testing in two steps. First is the restricted maximum likelihood step used to fit variance components. Although the eigendecomposition approaches (Kang et al., 2008; Lippert et al., 2011; Svishcheva et al., 2012; Zhou and Stephens, 2012; Sul et al., 2018) require positive definite $\mathbf{V}$ (lest the determinant of $\mathbf{V}$ be negative), surprisingly the GCTA average information algorithm only requires in practice that $\mathbf{V}$ be invertible (Yang et al., 2011). Thus, the relationship between WG, Standard, and True or Popkin variance components are largely as expected from our theoretical prediction $\sigma^{2\prime} = c\sigma^2$ in Eq. (16), with the exception of popkin ROM on 1000 Genomes only, whose genetic variance estimates are slightly smaller than expected (Fig. S7).

Next we determine the effect of WG bias on coefficient estimates. In this second step of LMM association testing, once $\mathbf{V}$ is determined, GCTA and other LMMs use generalized least squares to estimate fixed effects coefficients (Kang et al., 2008; Kang et al., 2010; Yang et al., 2014). Using the first replicate of the admixed family simulation and the true kinship matrix and the Standard and WG limits only, we recalculate the genetic effect β_i and intercept coefficients α in R for all loci, and confirm that we recover the GCTA estimates for β_i to the given precision. We then compare intercept coefficients, which are not given by GCTA, and confirm our theoretical prediction (Appendix F) that they are identical whether the True or WG ROM limit kinship matrices are used (the mean absolute difference is below 10^{-7}). In contrast, intercepts fit using the Standard ROM limit kinship matrix are different than those of the true kinship (not shown), which agrees with our theoretical prediction that the intercept varies to compensate for the kinship matrix bias ($\alpha' = \alpha + \eta$ in Eq. (17)).

Lastly, we explain the largest deviations from our predictions of the performance metrics $\mathrm{AUC}_{\mathrm{PR}}$ and SRMSD_p. We find that the small performance errors of popkin ROM in 1000 Genomes (Figs. S3 and 5) are driven by errors in genetic variance component estimation σ^2 (Fig. S8). However, the larger performance errors of popkin MOR in the admixed family simulation (Figs. 2 and S2) are instead explained by the condition number of $\mathbf{V}$ (Fig. S9). This result makes sense since the condition number by definition quantifies regression coefficient estimation accuracy.

4 Discussion

Previous research showed that commonly used kinship estimators are biased, and that these biases can be large (Ochoa and Storey (2021); Fig. 1). Our initial hypothesis was that these kinship biases would affect association testing, but surprisingly found that association is unaffected. We then proved theoretically that it is the intercept and relatedness effect (random effect or PCs) coefficients that compensate for the bias, and result in identical association coefficients and significance statistics.

Kinship estimates depend on the choice of ancestral population, which conditions the distributions of allele frequencies and genotypes, but the effect of this choice of association testing was not only unknown but completely disregarded. A corollary of our theoretical results is that changes of ancestral population, which behave algebraically like kinship bias, are also compensated for by the relatedness and intercept coefficients, so association testing is also invariant to the choice of ancestral population. Thus, although a choice of ancestral population is always being made when estimating kinship, this choice is fortunately inconsequential to association testing, as it ought to be since relatedness is being conditioned upon in these tests.

Given that kinship bias type is not important for association studies, we are free to choose a kinship estimator based on other properties. Ideally, kinship matrices result in well conditioned trait covariance matrices, since that has the largest effect in numerical accuracy and power in LMMs. Well-conditioned association is guaranteed for PSD kinship matrices, and popkin ROM is the only estimator that produces PSD matrices consistently across our evaluations (Fig. 7). Popkin ROM is also the only unbiased kinship estimator (Ochoa and Storey, 2021). We observed that Standard kinship estimates are also not PSD when genotypes are missing, a well understood phenomenon for related sample covariance estimators outside genetics (Jurczak and Rohde, 2017). Fortunately, non-PSD kinship estimators often perform well for association. Nevertheless, in our admixed family simulation we did see the other popkin estimator (the MOR version) perform particularly poorly due to being non-PSD, which in combination with the heritability parameters of this simulation results in ill-conditioned association tests and substantial loss of accuracy and power (Figs. 2, S2 and S9). Theory predicts that the same can happen with any non-PSD estimator, depending on

unknowns such as the heritability and the value of the negative eigenvalues of the kinship estimator, so it is risky to use MOR estimators (all of which are non-PSD in 1000 Genomes), as well as the WG estimator generally (which is non-PSD in all replicates of all of our evaluations). We also observe smaller numerical inaccuracies for popkin ROM, the estimator we recommend, in 1000 Genomes only, although the result is mixed: performance is slightly better (Fig. 5) although null p-value calibration is slightly worse (Fig. S3). The cause is variance components are poorly estimated (Fig. S7), but we did not find a more fundamental explanation. Overall, our assessment suggests that the popkin ROM estimator is the safest choice due to its guarantee of well-conditioned associations that other estimators cannot make.

Despite being non-PSD, we observe better performance for MOR versus ROM estimators in LMM association of 1000 Genomes (Fig. 5). Perhaps this is expected because we simulated larger coefficients for rare variants, while MOR estimators upweigh rare variants. This effect is not observed in the admixed family simulation, where MOR and ROM versions give similar kinship estimates (Fig. 1) and performed similarly (Fig. 2), compared to 1000 Genomes where kinship estimates are also strikingly different (Fig. 4). However, only popkin ROM is unbiased (Fig. 1B, Fig. S1). One potential explanation is that our kinship model assumes that all variants existed in the MRCA population, whereas rare variants in human data are known to be more recent mutations, and thus their effective kinship matrix is different than that of ancestral variants. Therefore, despite its biases, the popkin MOR estimator may better capture the covariance of rare variants and thus model them better in association tests, particularly in LMMs where the effect is most pronounced. Future work should focus on better approaches for upweighing rare variants or otherwise estimating their covariance structure while resulting in positive definite kinship estimates.

Our conclusions that common kinship biases do not affect association studies extend to variations of the Standard kinship estimator that weigh loci according to linkage disequilibrium (Speed et al., 2017; Wang et al., 2017), which also have the Standard bias type since this bias is present in each locus (Ochoa and Storey, 2021). As shown in our theoretical results, another form of the Standard kinship estimator that weighs individuals to estimate ancestral allele frequencies $\hat{p}_i^T$, including the best unbiased linear estimator in Appendix E (Astle and Balding, 2009; Thornton and McPeek,

2010), is also subject to the same conclusions.

In this study, we show empirically and theoretically that association tests are invariant to the use of common kinship estimators that are biased versus a more recent unbiased estimator. The underpinnings of our proof show that the same result holds for association with generalized linear models, since the intercept and relatedness effects interact in the same way as for linear models (the link function goes around the trait only); these models include case/control models such as logistic PCA and LMM. However, heritability estimation requires unbiased estimates of the random effect coefficient (σ^2), so it is biased when the standard kinship estimator is used, as it is using GCTA (Yang et al., 2011; Yang et al., 2014). Nevertheless, heritability estimation is a complex problem and a complete analysis is beyond the scope of this work. Overall, we have described an unexpected robustness of association studies, and our theoretical understanding of this result may help guide future improvements for association and other related models.

Declaration of interests

The authors declare no competing interests.

Acknowledgments

This work was funded in part by the Duke University School of Medicine Whitehead Scholars Program, a gift from the Whitehead Charitable Foundation. The 1000 Genomes data were generated at the New York Genome Center with funds provided by NHGRI Grant 3UM1HG008901-03S1.

Web resources

plink2, `https://www.cog-genomics.org/plink/2.0/`

GCTA, `https://yanglab.westlake.edu.cn/software/gcta/`

bnpsd, `https://cran.r-project.org/package=bnpsd`

simfam, `https://cran.r-project.org/package=simfam`

simtrait, `https://cran.r-project.org/package=simtrait`

popkin, https://cran.r-project.org/package=popkin

popkinsuppl, https://github.com/OchoaLab/popkinsuppl

Data and code availability

The data and code generated during this study are available on GitHub at https://github.com/OchoaLab/bias-assoc-paper. The high-coverage version of the 1000 Genomes Project was downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20190425_NYGC_GATK/.

References

1000 Genomes Project Consortium et al. (2012). "An integrated map of genetic variation from 1,092 human genomes". *Nature* 491(7422), pp. 56–65. DOI: 10.1038/nature11632.

Altschul, Stephen F., Raymond J. Carroll, and David J. Lipman (1989). "Weights for data related by a tree". *Journal of Molecular Biology* 207(4), pp. 647–653. DOI: 10.1016/0022-2836(89)90234-9.

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.

Aulchenko, Yurii S., Dirk-Jan de Koning, and Chris Haley (2007). "Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis". *Genetics* 177(1), pp. 577–585. DOI: 10.1534/genetics.107.075614.

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.

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.

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.

Consortium, The 1000 Genomes Project (2010). "A map of human genome variation from population-scale sequencing". *Nature* 467(7319), pp. 1061–1073. DOI: 10.1038/nature09534.

Devlin, B. and Kathryn Roeder (1999). "Genomic Control for Association Studies". *Biometrics* 55(4), pp. 997–1004. DOI: 10.1111/j.0006-341X.1999.00997.x.

Emik, L. Otis and Clair E. Terrill (1949). "Systematic procedures for calculating inbreeding coefficients". *J Hered* 40(2), pp. 51–55. DOI: 10.1093/oxfordjournals.jhered.a105986.

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.

García-Cortés, Luis Alberto (2015). "A novel recursive algorithm for the calculation of the detailed identity coefficients". *Genetics Selection Evolution* 47(1), p. 33. DOI: 10.1186/s12711-015-0108-6.

Hoffman, Gabriel E. (2013). "Correcting for population structure and kinship using the linear mixed model: theory and extensions". *PLoS ONE* 8(10), e75707. DOI: 10.1371/journal.pone.0075707.

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

Jurczak, Kamil and Angelika Rohde (2017). "Spectral analysis of high-dimensional sample covariance matrices with missing observations". *Bernoulli* 23(4A), pp. 2466–2532. DOI: 10.3150/16-BEJ815.

Kang, Hyun Min et al. (2008). "Efficient control of population structure in model organism association mapping". *Genetics* 178(3), pp. 1709–1723. DOI: 10.1534/genetics.107.080101.

Kang, Hyun Min et al. (2010). "Variance component model to account for sample structure in genome-wide association studies". *Nat. Genet.* 42(4), pp. 348–354. DOI: 10.1038/ng.548.

Lippert, Christoph et al. (2011). "FaST linear mixed models for genome-wide association studies". *Nat. Methods* 8(10), pp. 833–835. DOI: 10.1038/nmeth.1681.

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.

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

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.

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.

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.

Sherman, Jack and Winifred J. Morrison (1950). "Adjustment of an Inverse Matrix Corresponding to a Change in One Element of a Given Matrix". *The Annals of Mathematical Statistics* 21(1), pp. 124–127. DOI: 10.1214/aoms/1177729893.

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 et al. (2012). "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. (2017). "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.

Svishcheva, Gulnara R. et al. (2012). "Rapid variance components–based method for whole-genome association analysis". *Nat Genet* 44(10), pp. 1166–1170. DOI: 10.1038/ng.2410.

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.

Voight, Benjamin F. and Jonathan K. Pritchard (2005). "Confounding from Cryptic Relatedness in Case-Control Association Studies". *PLOS Genetics* 1(3), e32. DOI: 10.1371/journal.pgen.0010032.

Wang, Bowen, Serge Sverdlov, and Elizabeth Thompson (2017). "Efficient Estimation of Realized Kinship from SNP Genotypes". *Genetics*, genetics.116.197004. DOI: 10.1534/genetics.116.197004.

Weir, Bruce S. and Jérôme Goudet (2017). "A Unified Characterization of Population Structure and Relatedness". *Genetics* 206(4), pp. 2085–2103. DOI: 10.1534/genetics.116.198424.

Wright, Sewall (1922). "Coefficients of Inbreeding and Relationship". *The American Naturalist* 56(645), pp. 330–338.

— (1949). "The Genetical Structure of Populations". *Annals of Eugenics* 15(1), pp. 323–354. DOI: 10.1111/j.1469-1809.1949.tb02451.x.

Xie, C., D. D. Gessler, and S. Xu (1998). "Combining different line crosses for mapping quantitative trait loci using the identical by descent-based variance component method". *Genetics* 149(2), pp. 1139–1146.

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". *Am. J. Hum. Genet.* 88(1), pp. 76–82. DOI: 10.1016/j.ajhg.2010.11.011.

Yang, Jian et al. (2014). "Advantages and pitfalls in the application of mixed-model association methods". *Nat Genet* 46(2), pp. 100–106. DOI: 10.1038/ng.2876.

Yao, Yiqi and Alejandro Ochoa (2022). *Limitations of principal components in quantitative genetic association models for human studies*. Tech. rep. bioRxiv, p. 2022.03.25.485885. DOI: 10.1101/2022.03.25.485885.

Yu, Jianming et al. (2006). "A unified mixed-model method for association mapping that accounts for multiple levels of relatedness". *Nat. Genet.* 38(2), pp. 203–208. DOI: 10.1038/ng1702.

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 Justification for popkin generalizations

The popkin estimator in Eq. (1) has been generalized in this work to include locus weights w_i. The original ROM formulation had $w_i = 1$ for all loci i (Ochoa and Storey, 2021). Recalling from that original work that

$$\mathrm{E}\left[(x_{ij}-1)(x_{ik}-1)-1|T\right] = 4p_i^T\left(1-p_i^T\right)\left(\varphi_{jk}^T-1\right),$$

then for fixed w_i we get

$$\begin{aligned}\mathrm{E}\left[A_{jk}|T\right] &= v_m^T\left(\varphi_{jk}^T-1\right),\\ v_m^T &= \frac{4}{m}\sum_{i=1}^{m} w_i p_i^T\left(1-p_i^T\right).\end{aligned}$$

Therefore, as before all the unknowns p_i^T and now also the known weights w_i collapse into a single parameter v_m^T, which is estimated under the assumption that the minimum kinship is zero, giving $\hat{A}_{\min} = -v_m^T$, so that

$$\hat{\varphi}_{jk}^{T,\text{popkin-ROM}} = 1 - \frac{A_{jk}}{\hat{A}_{\min}} \xrightarrow[m\to\infty]{\text{a.s.}} \varphi_{jk}^T$$

as desired.

The MOR case of $w_i = \left(\hat{p}_i^T\left(1-\hat{p}_i^T\right)\right)^{-1}$ does not fit the previous case because this w_i is a random variable (it is a function of the genotypes). The term of interest $w_i((x_{ij}-1)(x_{ik}-1)-1)$ is a ratio of random variables whose expectation does not have a closed form. In this case, we rely

on the first-order approximation to this expectation, namely

$$\begin{aligned}
\mathrm{E}\left[\frac{(x_{ij}-1)(x_{ik}-1)-1}{\hat{p}_i^T\left(1-\hat{p}_i^T\right)}\middle| T\right] &\approx \frac{\mathrm{E}\left[(x_{ij}-1)(x_{ik}-1)-1|T\right]}{\mathrm{E}\left[\hat{p}_i^T\left(1-\hat{p}_i^T\right)|T\right]} \\
&= \frac{4p_i^T\left(1-p_i^T\right)\left(\varphi_{jk}^T-1\right)}{p_i^T\left(1-p_i^T\right)\left(1-\bar{\varphi}^T\right)} \\
&= \frac{4\left(\varphi_{jk}^T-1\right)}{1-\bar{\varphi}^T},
\end{aligned}$$

where the expectation of $\hat{p}_i^T\left(1-\hat{p}_i^T\right)$ was calculated previously (Ochoa and Storey, 2021). In this case the expectation of A_{jk}, summing across loci, is also approximated by

$$\mathrm{E}\left[A_{jk}|T\right] \approx \frac{4\left(\varphi_{jk}^T-1\right)}{1-\bar{\varphi}^T}.$$

The same strategy as before applies to estimate the unknown factor $4/\left(1-\bar{\varphi}^T\right)$, namely that if the minimum kinship is zero then $\hat{A}_{\min} \approx -4/\left(1-\bar{\varphi}^T\right)$, resulting in

$$\hat{\varphi}_{jk}^{T,\text{popkin-MOR}} = 1 - \frac{A_{jk}}{\hat{A}_{\min}} \approx \varphi_{jk}^T.$$

B Connection between popkin and standard kinship estimator

Since the connection we discovered holds when data are complete, but not under missingness, to determine necessary conditions we introduce more complete forms of the estimators that handle missingness. Popkin (with locus weights) has the following parts updated:

$$\begin{aligned}
A_{ijk} &= I_{ij}I_{ik}((x_{ij}-1)(x_{ik}-1)-1), \\
A_{jk} &= \frac{1}{m_{jk}}\sum_{i=1}^{m} w_i A_{ijk}, \\
m_{jk} &= \sum_{i=1}^{m} I_{ij}I_{ik},
\end{aligned}$$

where $I_{ij} = 1$ if x_{ij} is not missing, 0 otherwise (this way missing x_{ij} can have any value and not contribute to the estimator). Only loci with both genotypes (x_{ij} and x_{ik}) non-missing are included in the above average, and m_{jk} counts the total number of such loci. The ancestral allele frequency estimator with missingness is

$$\hat{p}_i^T = \frac{1}{2n_i}\sum_{j=1}^{n} I_{ij}x_{ij},$$
$$n_i = \sum_{j=1}^{n} I_{ij},$$

which averages over individuals rather than loci, so its denominator is the number of non-missing individuals at this locus. Let us compute some averages of the popkin estimator. Since the result we want holds at every locus separately, let us formulate the averages of interest at locus i only:

$$\bar{A}_{ij} = \frac{1}{n}\sum_{k=1}^{n} A_{ijk} = I_{ij}\frac{n_i}{n}\left(\left(x_{ij}-1\right)\left(2\hat{p}_i^T-1\right)-1\right),$$
$$\bar{A}_i = \frac{1}{n}\sum_{k=1}^{n} \bar{A}_{ij} = -\left(\frac{n_i}{n}\right)^2 4\hat{p}_i^T\left(1-\hat{p}_i^T\right).$$

Therefore, the combination of interest is:

$$\begin{aligned} A_{ijk} + \bar{A}_i - \bar{A}_{ij} - \bar{A}_{ik} = {} & I_{ij}I_{ik}\left(x_{ij}-2\hat{p}_i^T\right)\left(x_{ik}-2\hat{p}_i^T\right) \\ & + \frac{n_i}{n}\left(I_{ij}-\frac{n_i}{n}\right)4\hat{p}_i^T + \left(\left(\frac{n_i}{n}\right)^2 - I_{ij}I_{ik}\right)4\left(\hat{p}_i^T\right)^2 \\ & + I_{ij}\left(I_{ik}-\frac{n_i}{n}\right)x_{ij}\left(2\hat{p}_i^T-1\right) + I_{ik}\left(I_{ij}-\frac{n_i}{n}\right)x_{ik}\left(2\hat{p}_i^T-1\right). \end{aligned}$$

For the above to equal $I_{ij}I_{ik}\left(x_{ij}-2\hat{p}_i^T\right)\left(x_{ik}-2\hat{p}_i^T\right)$, which is the first term above, the rest of the terms must vanish for arbitrary values of $\hat{p}_i^T$, x_{ij}, and x_{ik}. Since $n_i > 0$ (there is at least one non-missing individual at every locus), the term $\frac{n_i}{n}(I_{ij}-\frac{n_i}{n})4\hat{p}_i^T$ vanishes if and only if $I_{ij} = \frac{n_i}{n}$, and since $I_{jk} = 0$ does not solve this equation (because $n_i > 0$), then $I_{jk} = 1$, which requires $n_i = n$, so no individuals can have missing data at this locus (the rest of the terms vanish when this is so).

Thus,

$$A_{ijk} + \bar{A}_i - \bar{A}_{ij} - \bar{A}_{ik} = I_{ij} I_{ik} \left(x_{ij} - 2\hat{p}_i^T\right)\left(x_{ik} - 2\hat{p}_i^T\right)$$

if and only if there is no missing data at locus i. The other desired result of

$$\bar{A}_i = -4\hat{p}_i^T\left(1 - \hat{p}_i^T\right)$$

also requires $n_i = n$.

Assuming now no missingness, transforming the popkin estimates using the Standard bias function of Eq. (4) gives

$$\begin{aligned}
\frac{\hat{\varphi}_{jk}^{T,\text{popkin}} + \bar{\hat{\varphi}}^{T,\text{popkin}} - \bar{\hat{\varphi}}_j^{T,\text{popkin}} - \bar{\hat{\varphi}}_k^{T,\text{popkin}}}{1 - \bar{\hat{\varphi}}^{T,\text{popkin}}} &= \frac{A_{jk} + \bar{A} - \bar{A}_j - \bar{A}_k}{-\bar{A}} \\
&= \frac{\sum_{i=1}^m w_i (A_{ijk} + \bar{A}_i - \bar{A}_{ij} - \bar{A}_{ik})}{-\sum_{i=1}^m w_i \bar{A}_i} \\
&= \frac{\sum_{i=1}^m w_i \left(x_{ij} - 2\hat{p}_i^T\right)\left(x_{ik} - 2\hat{p}_i^T\right)}{\sum_{i=1}^m w_i 4\hat{p}_i^T\left(1 - \hat{p}_i^T\right)}.
\end{aligned}$$

Therefore, if popkin ROM is input ($w_i = 1$), this transformation yields Standard ROM. On the other hand, if popkin MOR is used ($w_i^{-1} = \hat{p}_i^T\left(1 - \hat{p}_i^T\right)$), the transformation yields Standard MOR.

C Mean kinship inequalities

Denote the mean of the diagonal kinship terms as $\bar{\delta}^T = \frac{1}{n}\sum_{j=1}^n \varphi_{jj}^T$. Here we prove that

$$0 \le \tilde{\varphi}^T \le \bar{\varphi}^T \le \bar{\delta}^T \le 1,$$

with each of $\tilde{\varphi}^T = \bar{\varphi}^T$ and $\bar{\varphi}^T = \bar{\delta}^T$ if and only if all kinship values are equal.

The inequalities $0 \le \bar{\varphi}^T \le \bar{\delta}^T \le 1$ follow directly from previous work, applied to a kinship matrix rather than a coancestry matrix as done originally, as the proof required solely a covariance matrix with values between 0 and 1 (Ochoa and Storey, 2021). $\tilde{\varphi}^T$ is defined in Eq. (7). $0 \le \tilde{\varphi}^T$

follows since every kinship value is non-negative. $\bar{\varphi}^T$ and $\tilde{\varphi}^T$ are related by

$$\bar{\varphi}^T = \frac{\tilde{\varphi}^T(n-1) + \bar{\delta}^T}{n}. \tag{19}$$

Applying $\bar{\varphi}^T \leq \bar{\delta}^T$ to Eq. (19) and simplifying yields $\tilde{\varphi}^T \leq \bar{\delta}^T$. Lastly, since $\bar{\varphi}^T - \tilde{\varphi}^T = \left(\bar{\delta}^T - \tilde{\varphi}^T\right)/n$ (from rearranging Eq. (19)), it also follows that $\tilde{\varphi}^T \leq \bar{\varphi}^T$, as desired. Furthermore, $\tilde{\varphi}^T = \bar{\varphi}^T$ holds if and only if all $\varphi^T_{jk} = \bar{\delta}^T$, since that is necessary and sufficient for $\bar{\varphi}^T = \bar{\delta}^T$.

D Derivation of WG bias factorization

Here we rewrite the WG bias function of Eq. (6) as a factorization of the form of Eq. (12). It is easy to see that $c = 1 - \tilde{\varphi}^T$. Expanding Eq. (12) gives

$$\begin{aligned}\mathbf{B}\boldsymbol{\Phi}^T\mathbf{B}^\intercal &= \left(\mathbf{I} - \mathbf{1}\mathbf{b}^\intercal\right)\boldsymbol{\Phi}^T\left(\mathbf{I} - \mathbf{b}\mathbf{1}^\intercal\right) \\ &= \boldsymbol{\Phi}^T - \mathbf{1}\left(\boldsymbol{\Phi}^T\mathbf{b}\right)^\intercal - \left(\boldsymbol{\Phi}^T\mathbf{b}\right)\mathbf{1}^\intercal + \mathbf{J}(\mathbf{b}^\intercal\boldsymbol{\Phi}^T\mathbf{b}),\end{aligned}$$

where $\mathbf{b}^\intercal\boldsymbol{\Phi}^T\mathbf{b}$ is a scalar and $\boldsymbol{\Phi}^T\mathbf{b}$ a vector. Equating the above to Eq. (6) and rearranging, we obtain

$$\mathbf{J}\left(\tilde{\varphi}^T + \left(\mathbf{b}^\intercal\boldsymbol{\Phi}^T\mathbf{b}\right)\right) = \mathbf{1}\left(\boldsymbol{\Phi}^T\mathbf{b}\right)^\intercal + \left(\boldsymbol{\Phi}^T\mathbf{b}\right)\mathbf{1}^\intercal.$$

Since $\tilde{\varphi}^T + \left(\mathbf{b}^\intercal\boldsymbol{\Phi}^T\mathbf{b}\right)$ is a scalar and $\mathbf{J} = \mathbf{1}\mathbf{1}^\intercal$, we can see that the solution requires the right side to also be a constant matrix, which is only achieved if $\boldsymbol{\Phi}^T\mathbf{b} \propto \mathbf{1}$. We choose the scaling factor for the last $\mathbf{1}$ to be $q\left(\mathbf{1}^\intercal\left(\boldsymbol{\Phi}^T\right)^{-1}\mathbf{1}\right)^{-1}$ as this simplifies notation later, and solving for $\mathbf{b}$ results in Eq. (14). To solve for q, we replace $\mathbf{b}$ from Eq. (14) into the above equation, which after rearranging results in

$$q^2 - 2q + \tilde{\varphi}^T\left(\mathbf{1}^\intercal\left(\boldsymbol{\Phi}^T\right)^{-1}\mathbf{1}\right) = 0.$$

The solution to the above quadratic equation is given by Eq. (15), as desired.

E Minimum weighted mean kinship

Consider the weighted mean kinship value $\mathbf{w}^\intercal \mathbf{\Phi}^T \mathbf{w}$, where $\mathbf{w}$ are weights that sum to one ($\mathbf{w}^\intercal \mathbf{1} = 1$). The ordinary mean kinship $\bar{\varphi}^T$ is the special case with $\mathbf{w} = \frac{1}{n}\mathbf{1}$. The weights that minimize the weighted mean kinship are the solution of the Lagrangian multiplier problem

$$G = \mathbf{w}^\intercal \mathbf{\Phi}^T \mathbf{w} + \lambda(\mathbf{w}^\intercal \mathbf{1} - 1).$$

The derivatives are the constraint and $\frac{dG}{d\mathbf{w}} = 2\mathbf{\Phi}^T\mathbf{w} + \lambda\mathbf{1} = \mathbf{0}$. The optimal weights thus satisfy $\mathbf{w} = \frac{-\lambda}{2}\left(\mathbf{\Phi}^T\right)^{-1}\mathbf{1}$. Multiplying by $\mathbf{1}^\intercal$, since $\mathbf{1}^\intercal\mathbf{w} = 1$, allows us to solve for $\lambda^{-1} = -\frac{1}{2}\mathbf{1}^\intercal\left(\mathbf{\Phi}^T\right)^{-1}\mathbf{1}$. Thus, the optimal weights are

$$\mathbf{w} = \frac{\left(\mathbf{\Phi}^T\right)^{-1}\mathbf{1}}{\mathbf{1}^\intercal\left(\mathbf{\Phi}^T\right)^{-1}\mathbf{1}},$$

a solution that recurs in related settings, and applied to genotypes as $\hat{p}_i^T = \mathbf{w}^\intercal\mathbf{x}_i/2$ yields the best linear unbiased estimator of p_i^T (Altschul et al., 1989; Astle and Balding, 2009; Thornton and McPeek, 2010). Therefore, the minimum weighted mean kinship is, and satisfies,

$$\mathbf{w}^\intercal\mathbf{\Phi}^T\mathbf{w} = \frac{1}{\mathbf{1}^\intercal\left(\mathbf{\Phi}^T\right)^{-1}\mathbf{1}} \leq \bar{\varphi}^T \approx \tilde{\varphi}^T.$$

F Proof that WG bias results in zero intercept shift under LMM generalized least squares estimation

For this section suppose that variance components have been estimated, so $\mathbf{V} = 2\sigma^2\mathbf{\Phi}^T + \sigma_\epsilon^2\mathbf{I}$ is given, assume it is invertible, and rewrite the LMM as

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}_V, \qquad \boldsymbol{\epsilon}_V \sim \text{Normal}\left(\mathbf{0}, \mathbf{V}\right),$$

where the design matrix $\mathbf{Z} = (\mathbf{1}, \mathbf{x}_i, ...)$ contains the intercept, genotype and now additional covariates, and $\boldsymbol{\beta} = (\alpha, \beta_i, ...)$ are their coefficients. The generalized least squares coefficients estimate,

used by GCTA and other LMMs, is

$$\hat{\beta} = \left(\mathbf{Z}^\intercal \mathbf{V}^{-1} \mathbf{Z}\right)^{-1} \mathbf{Z}^\intercal \mathbf{V}^{-1} \mathbf{y}.$$

Now suppose $\mathbf{V}$ corresponds to some kinship matrix $\mathbf{\Phi}^T$ while $\mathbf{V}'$ corresponds to $\mathbf{\Phi}^{T\prime} = F^{\mathrm{WG}}(\mathbf{\Phi}^T)$, and $\mathbf{V}'$ is also invertible. Our strategy involves repeated application of the Sherman-Morrison formula for calculating inverses of matrices after a rank-1 update, which for a symmetric update of a matrix $\mathbf{A}$ with a vector $\mathbf{z}$ and a scalar b takes the form (Sherman and Morrison, 1950)

$$\left(\mathbf{A} + b\mathbf{z}\mathbf{z}^\intercal\right)^{-1} = \mathbf{A}^{-1} - \frac{b}{1 + b\left(\mathbf{z}^\intercal \mathbf{A}^{-1} \mathbf{z}\right)} \left(\mathbf{A}^{-1}\mathbf{z}\right)\left(\mathbf{A}^{-1}\mathbf{z}\right)^\intercal.$$

Since $F^{\mathrm{WG}}(\mathbf{\Phi}^T)$ is a rank-1 update of $\mathbf{\Phi}^T$ by Eq. (6), then $\mathbf{V}'$ is also a rank-1 update of $\mathbf{V}$:

$$\begin{aligned}
\mathbf{V}' &= 2\sigma^{2\prime}\mathbf{\Phi}^{T\prime} + \sigma_\epsilon^2 \mathbf{I} \\
&= 2\sigma^2\left(\mathbf{\Phi}^T - \tilde{\varphi}^T \mathbf{1}\mathbf{1}^\intercal\right) + \sigma_\epsilon^2 \mathbf{I} \\
&= \mathbf{V} - d\mathbf{1}\mathbf{1}^\intercal,
\end{aligned}$$

where $d = 2\sigma^2\tilde{\varphi}^T$ and we used $\sigma^{2\prime} = \left(1 - \tilde{\varphi}^T\right)\sigma^2$. Therefore,

$$\left(\mathbf{V}'\right)^{-1} = \mathbf{V}^{-1} + e\mathbf{V}^{-1}\mathbf{1}\left(\mathbf{V}^{-1}\mathbf{1}\right)^\intercal,$$

where $e = d/\left(1 - d\left(\mathbf{1}^\intercal \mathbf{V}^{-1}\mathbf{1}\right)\right)$. Therefore the following remains a rank-1 update,

$$\mathbf{Z}^\intercal\left(\mathbf{V}'\right)^{-1}\mathbf{Z} = \mathbf{Z}^\intercal\mathbf{V}^{-1}\mathbf{Z} + e\mathbf{u}\mathbf{u}^\intercal,$$

where $\mathbf{u} = \mathbf{Z}^\intercal\mathbf{V}^{-1}\mathbf{1}$ is a column vector the length of the number of covariates (including intercept and genotype). Therefore,

$$\left(\mathbf{Z}^\intercal\left(\mathbf{V}'\right)^{-1}\mathbf{Z}\right)^{-1} = \left(\mathbf{Z}^\intercal\mathbf{V}^{-1}\mathbf{Z}\right)^{-1} - g\mathbf{v}\mathbf{v}^\intercal,$$

where $\mathbf{v} = \left(\mathbf{Z}^\intercal \mathbf{V}^{-1} \mathbf{Z}\right)^{-1} \mathbf{u}$ and $g = e/(1 + e(\mathbf{u}^\intercal \mathbf{v}))$. Noting that $\mathbf{Z}^\intercal \mathbf{V}^{-1} \mathbf{1}$ is the first column of $\mathbf{Z}^\intercal \mathbf{V}^{-1} \mathbf{Z}$, then $\mathbf{v}$ is the first column of the identity matrix:

$$\mathbf{v} = \left(\mathbf{Z}^\intercal \mathbf{V}^{-1} \mathbf{Z}\right)^{-1} \mathbf{Z}^\intercal \mathbf{V}^{-1} \mathbf{1} = \begin{pmatrix} 1 \\ \mathbf{0} \end{pmatrix},$$

where $\mathbf{0}$ is a vector the length of the number of covariates minus one (exclude the intercept). As a consequence, $\mathbf{Z}\mathbf{v} = \mathbf{1}$, so $\mathbf{u}^\intercal \mathbf{v} = \mathbf{1}^\intercal \mathbf{V}^{-1} \mathbf{1}$ and

$$\begin{aligned} g &= \frac{e}{1 + e(\mathbf{1}^\intercal \mathbf{V}^{-1} \mathbf{1})} \\ &= \frac{\frac{d}{1 - d(\mathbf{1}^\intercal \mathbf{V}^{-1} \mathbf{1})}}{1 + \frac{d}{1 - d(\mathbf{1}^\intercal \mathbf{V}^{-1} \mathbf{1})} (\mathbf{1}^\intercal \mathbf{V}^{-1} \mathbf{1})} \\ &= d. \end{aligned}$$

The final step yields the coefficient estimates as a rank-1 update:

$$\begin{aligned} \hat{\boldsymbol{\beta}}' &= \left(\mathbf{Z}^\intercal \left(\mathbf{V}'\right)^{-1} \mathbf{Z}\right)^{-1} \mathbf{Z}^\intercal \left(\mathbf{V}'\right)^{-1} \mathbf{y} \\ &= \left(\left(\mathbf{Z}^\intercal \mathbf{V}^{-1} \mathbf{Z}\right)^{-1} - d\mathbf{v}\mathbf{v}^\intercal\right) \mathbf{Z}^\intercal \left(\mathbf{V}^{-1} + e\mathbf{V}^{-1}\mathbf{1} \left(\mathbf{V}^{-1}\mathbf{1}\right)^\intercal\right) \mathbf{y} \\ &= \hat{\boldsymbol{\beta}} + e\mathbf{v} \left(\mathbf{1}^\intercal \mathbf{V}^{-1} \mathbf{y}\right) - d\mathbf{v} \left(\mathbf{1}^\intercal \mathbf{V}^{-1} \mathbf{y}\right) - de\mathbf{v} \left(\mathbf{1}^\intercal \mathbf{V}^{-1} \mathbf{1}\right) \left(\mathbf{1}^\intercal \mathbf{V}^{-1} \mathbf{y}\right) \\ &= \hat{\boldsymbol{\beta}} + \mathbf{v} \left(\mathbf{1}^\intercal \mathbf{V}^{-1} \mathbf{y}\right) \left(e - d - de \left(\mathbf{1}^\intercal \mathbf{V}^{-1} \mathbf{1}\right)\right). \end{aligned}$$

The last factor above vanishes:

$$\begin{aligned} e - d - de \left(\mathbf{1}^\intercal \mathbf{V}^{-1} \mathbf{1}\right) &= \frac{d}{1 - d \left(\mathbf{1}^\intercal \mathbf{V}^{-1} \mathbf{1}\right)} - d - d\frac{d}{1 - d \left(\mathbf{1}^\intercal \mathbf{V}^{-1} \mathbf{1}\right)} \left(\mathbf{1}^\intercal \mathbf{V}^{-1} \mathbf{1}\right) \\ &= 0. \end{aligned}$$

Therefore, $\hat{\boldsymbol{\beta}}' = \hat{\boldsymbol{\beta}}$, which shows that all fixed effect coefficients, including the intercept, are invariant to using a WG-biased kinship matrix instead of the unbiased one when the coefficients are estimated with generalized least squares.

Furthermore, since the diagonal values of $\left(\mathbf{Z}^\intercal \left(\mathbf{V}'\right)^{-1} \mathbf{Z}\right)^{-1}$, which correspond to $\mathrm{Var}(\hat{\beta}'_k)$ for each k, are the same as those of $\left(\mathbf{Z}^\intercal \mathbf{V}^{-1} \mathbf{Z}\right)^{-1}$ except for the first one corresponding to the intercept, then the Wald test statistic of the kth covariate coefficients, given by $\hat{\beta}_k^2 / \mathrm{Var}(\hat{\beta}_k)$, and their p-values, are also the same for $k \neq 1$ for WG bias as for the unbiased kinship matrix.

Supplemental figures

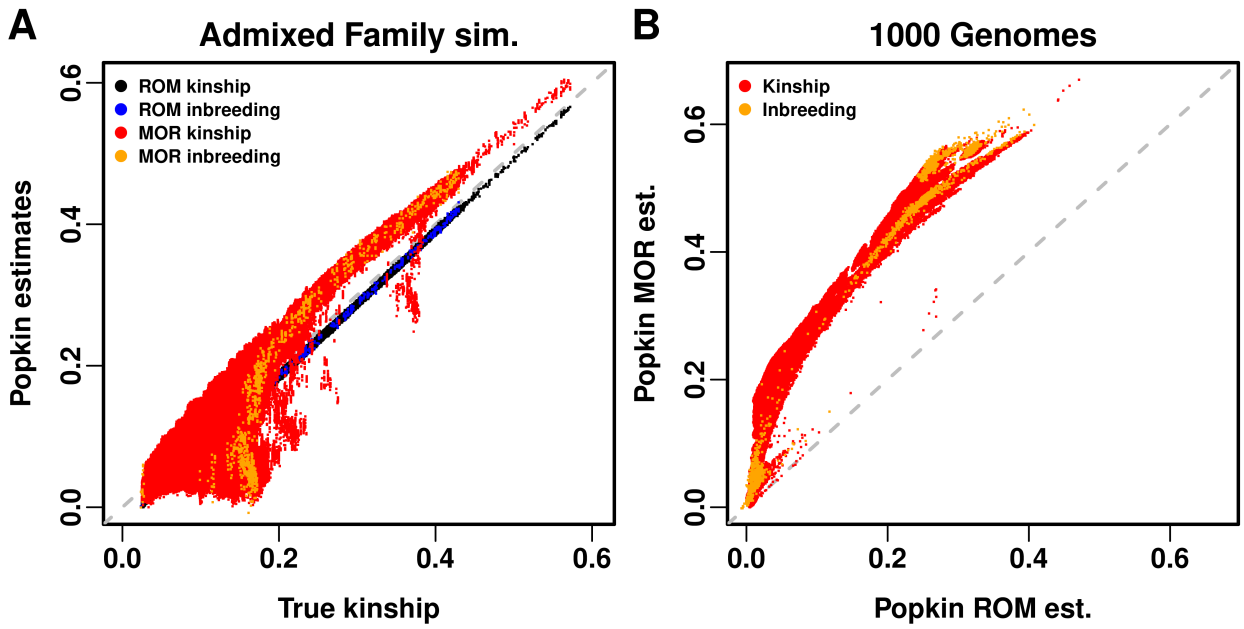

Figure S1: **Comparison of popkin ROM and MOR estimates.** Kinship (off-diagonal of matrix) and inbreeding (transformed diagonal) are plotted in different colors, which shows that their biases (if any) overlap. **A.** In admixed family simulation, both estimates are compared against true kinship. Popkin ROM has a negligible bias, due to the minimum true kinship of the simulation being slightly larger than zero. Popkin MOR has considerable biases, tending to be upward though not always. **B.** In 1000 Genomes, since true kinship is unknown, popkin ROM takes its place. Popkin MOR biases take on a similar shape as panel A.

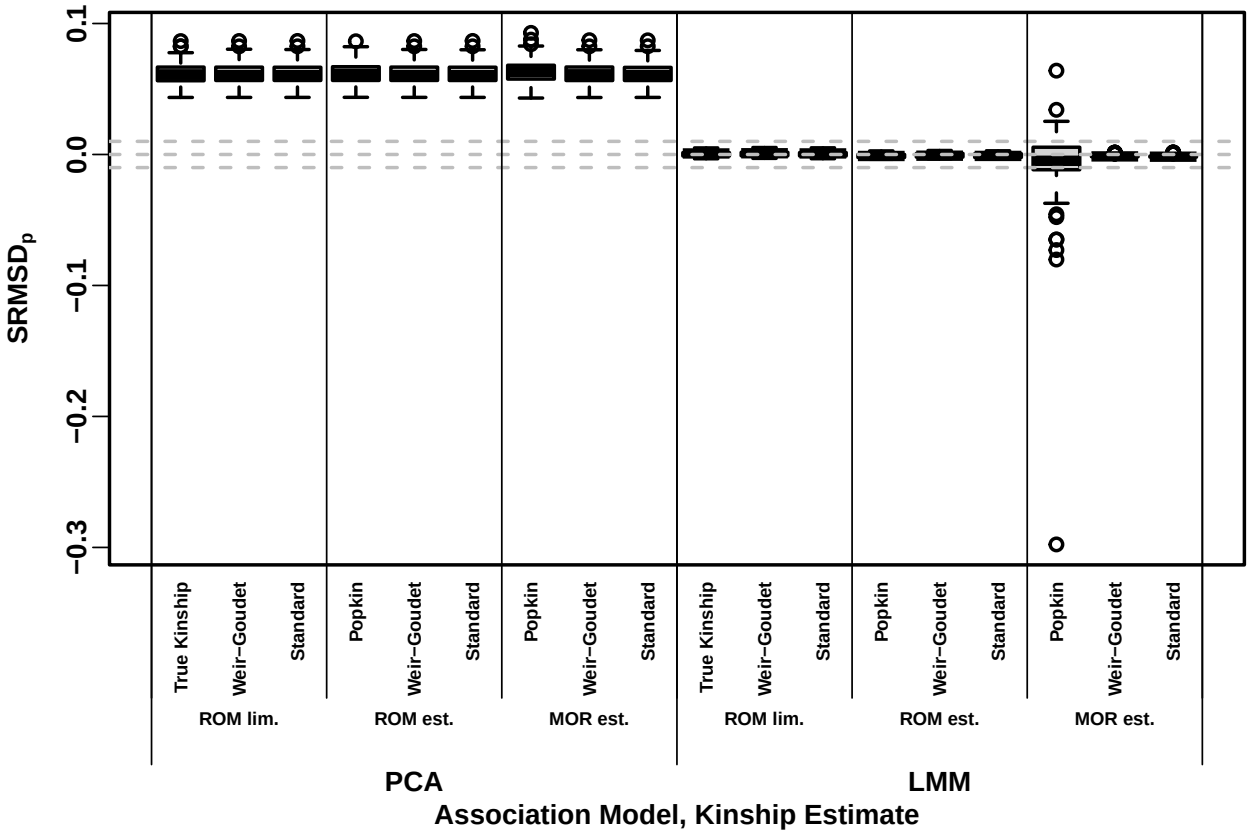

Figure S2: **Signed Root Mean Square Deviation of null p-values ($\mathbf{SRMSD}_p$) on the admixed family simulation.** Same methods and simulation as Fig. 2, see that for more information. $|\text{SRMSD}_p| < 0.01$ (area between gray dashed lines) is considered calibrated. All PCA runs are miscalibrated by similar amounts, whereas most LMM runs are calibrated with few exceptions.

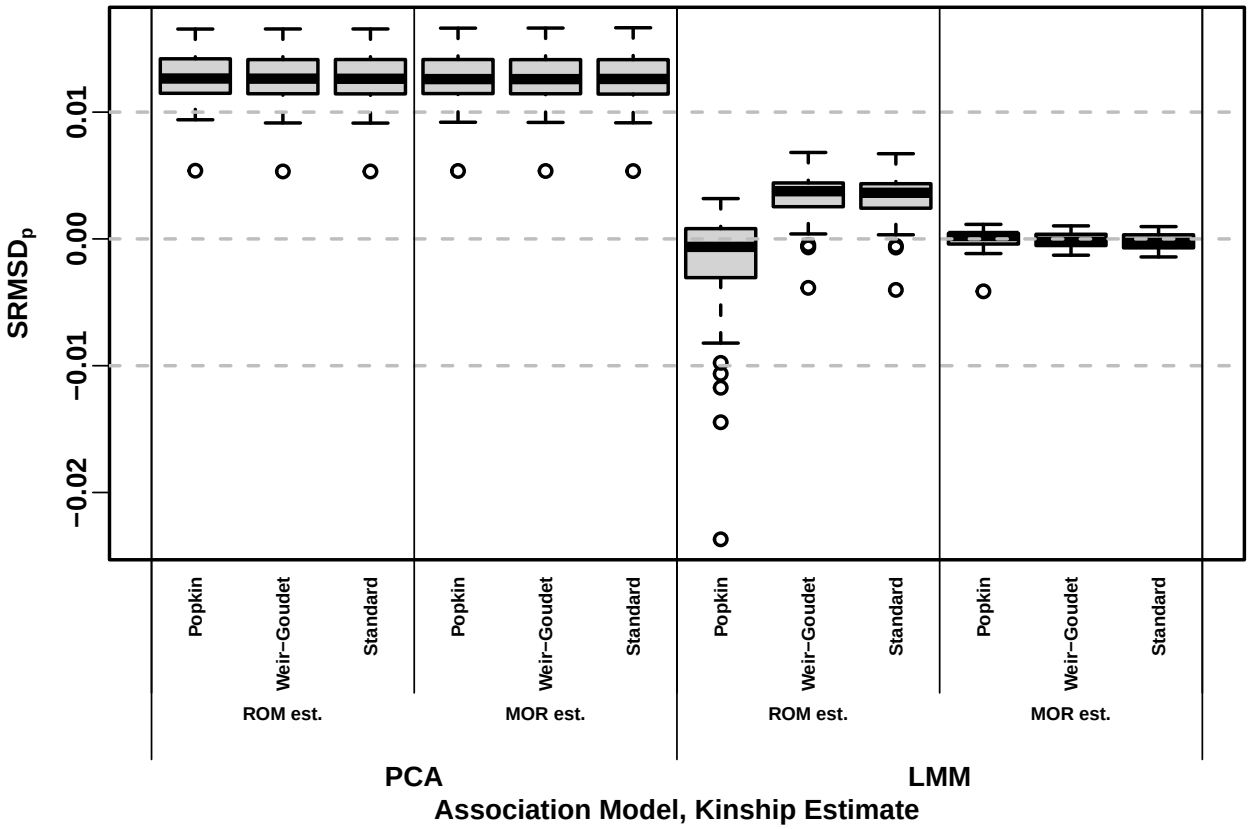

Figure S3: **Signed Root Mean Square Deviation of null p-values (SRMSD$_p$) on 1000 Genomes.** Same methods and simulation as Fig. 5, and y-axis statistic and conclusions of Fig. S2, see those for more information.

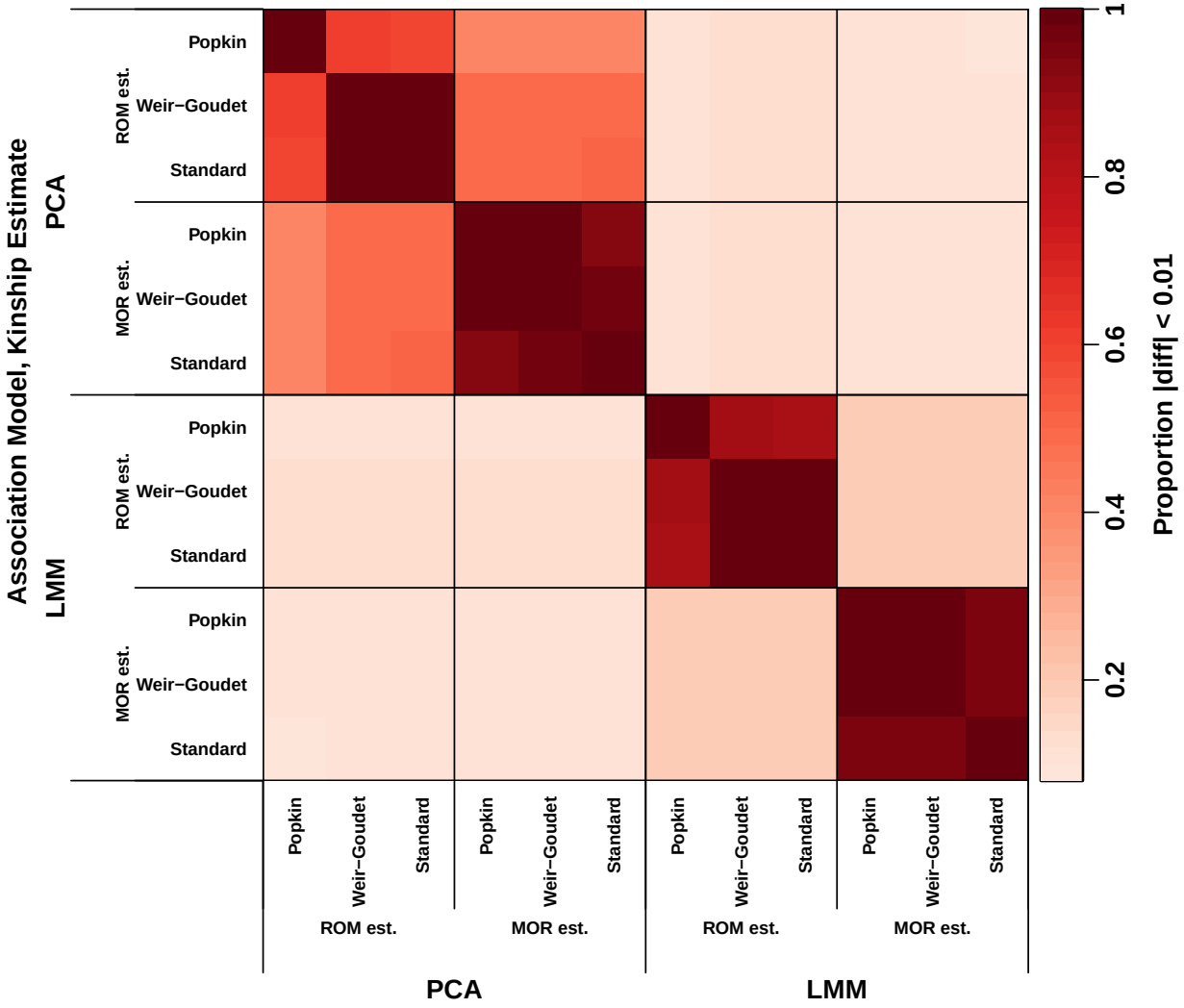

Figure S4: **Approximate agreement between p-values on 1000 Genomes.** See Fig. 3 for more details.

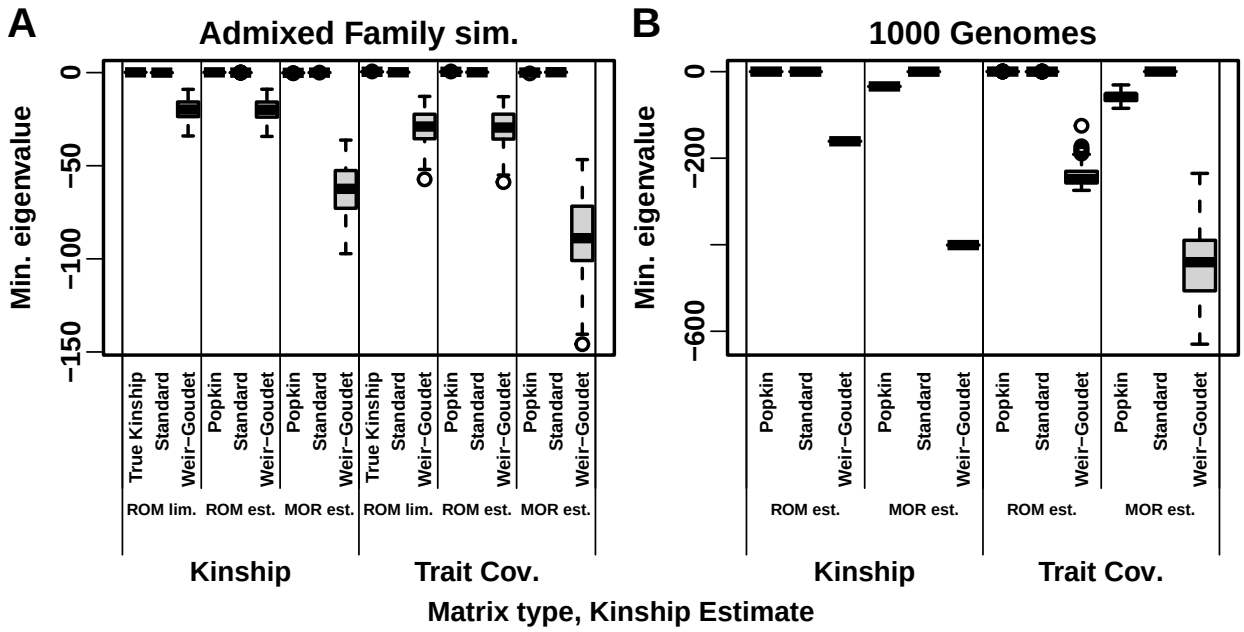

Figure S5: **Minimum eigenvalue of kinship and trait covariance (V) matrices.** Each distribution is over 100 replicates (1000 Genomes kinship has one value since genotypes are fixed, but **V** varies per replicate). All WG matrices has very large negative eigenvalues, and Popkin MOR has negative eigenvalues as well; in these cases **V** always has negative eigenvalues too.

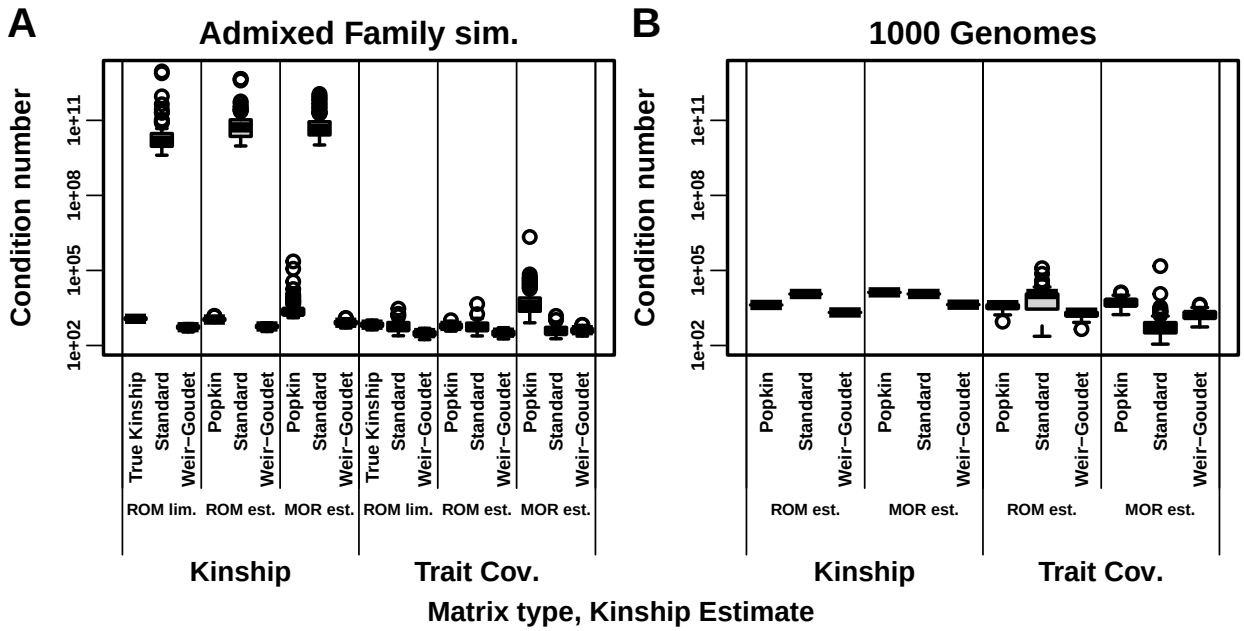

Figure S6: **Condition numbers of kinship and trait covariance (V) matrices.** Larger condition numbers reflect ill-conditioned problems such as near singularity. Each distribution is over 100 replicates (1000 Genomes kinship has one value since genotypes are fixed, but **V** varies per replicate).

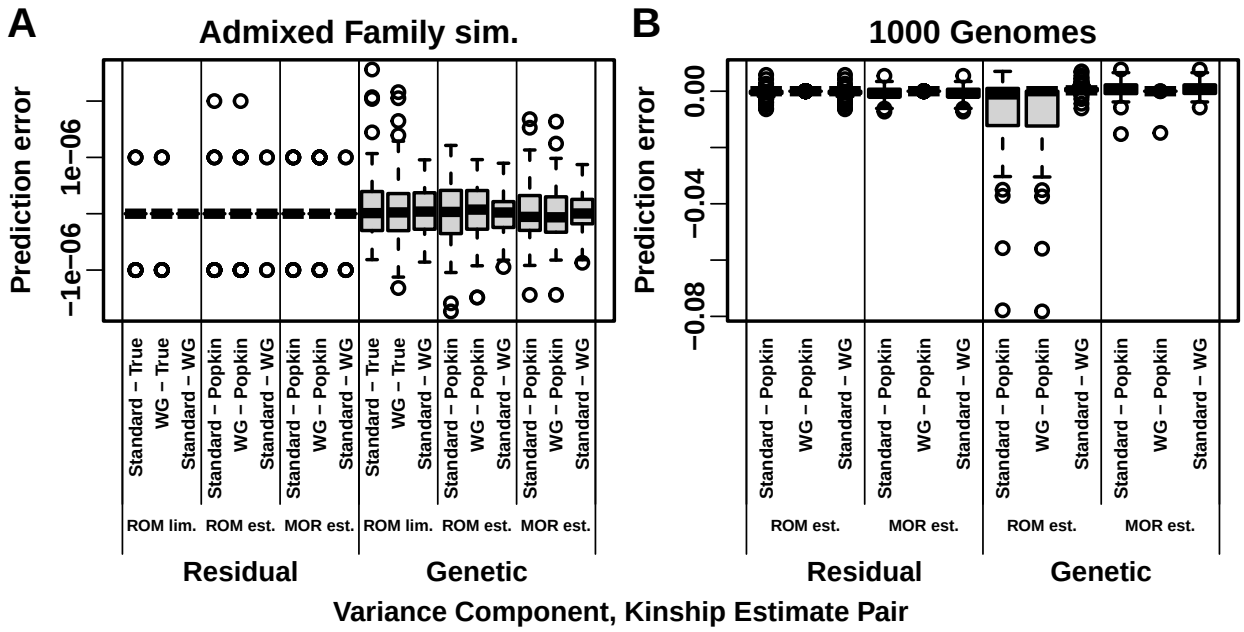

Figure S7: **Variance component prediction errors across evaluations.** Here we test the predictions that $\sigma^{2\prime}_\epsilon = \sigma^2_\epsilon$ and $\sigma^{2\prime} = c\sigma^2$ in Eq. (16). For Residual, prediction error (y-axis) is $\sigma^{2\prime}_\epsilon - \sigma^2_\epsilon$ between pairs of estimates as listed. For Genetic, prediction error is $\sigma^{2\prime} - c\sigma^2$: The biased-unbiased pairs use $c = 1 - \bar{\varphi}^T$ for Standard, $c = 1 - \tilde{\varphi}^T$ for WG, $\sigma^{2\prime}$ is their estimate and σ^2 is True or Popkin; The Standard-WG pair uses σ^2 for WG and $c = \left(1 - \bar{\varphi}^T\right) / \left(1 - \tilde{\varphi}^T\right)$. Each distribution is over the 100 replicates of each simulation. **A.** In admixed family simulation, all errors are zero within machine precision. Excess perfect zero residual prediction errors are due to limited precision of GCTA outputs. **B.** In 1000 Genomes, popkin ROM estimates has large errors compared to Standard and WG.

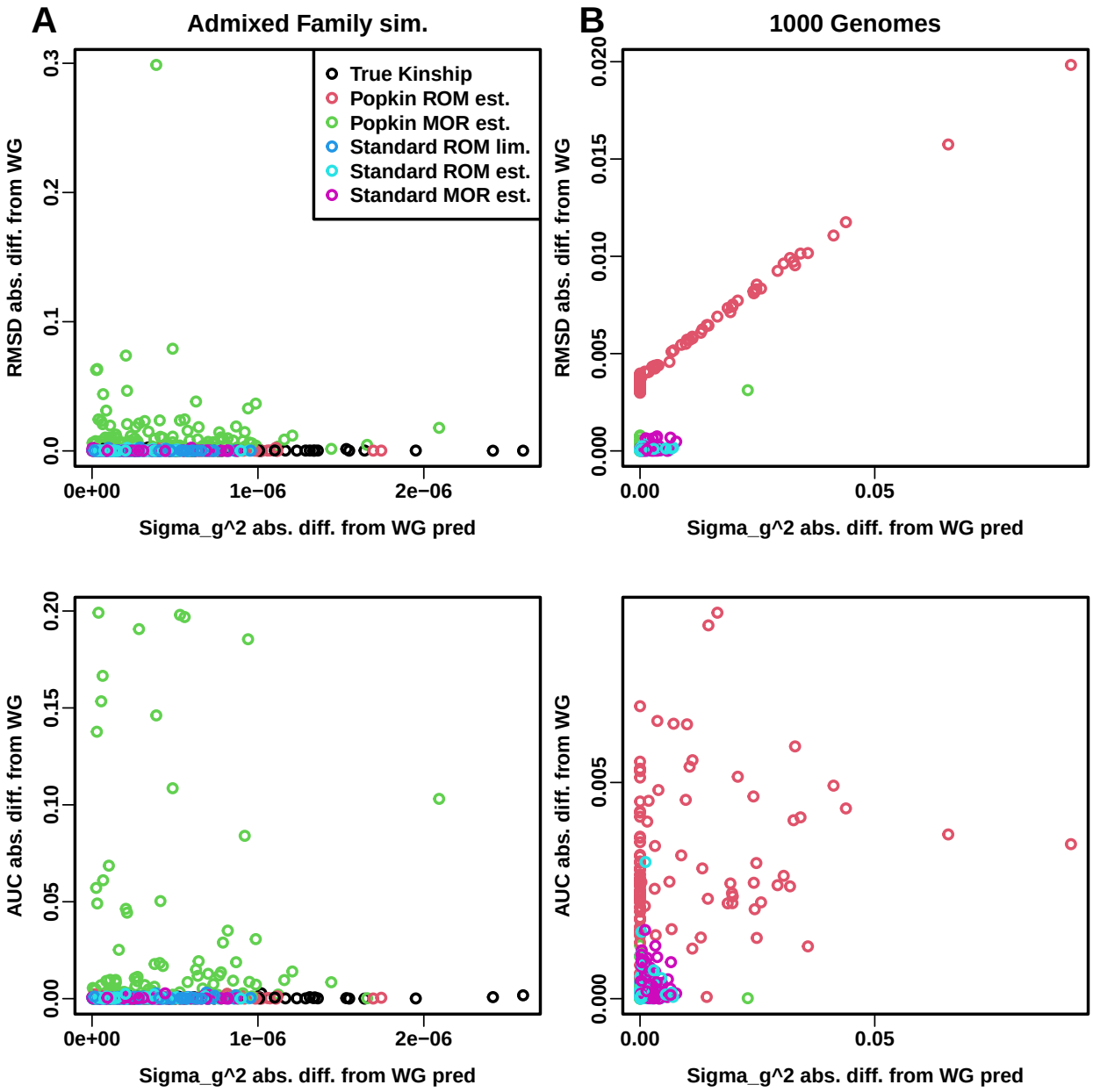

Figure S8: **AUC_{PR} and SRMSD_p prediction errors explained by variance component errors.** Genetic variance component (σ^2) absolute error is calculated with the formulas in Fig. S7 using WG as reference since its $\mathbf{V}$ had the lowest condition numbers (Fig. S6). AUC_{PR} and SRMSD_p are expected to be the same between WG, Standard, and True or Popkin (within each locus weight type). **A.** Large errors in the admixed family simulation are not explained by high σ^2 error. **B.** Smaller popkin ROM prediction errors in 1000 Genomes are explained by high σ^2 error.

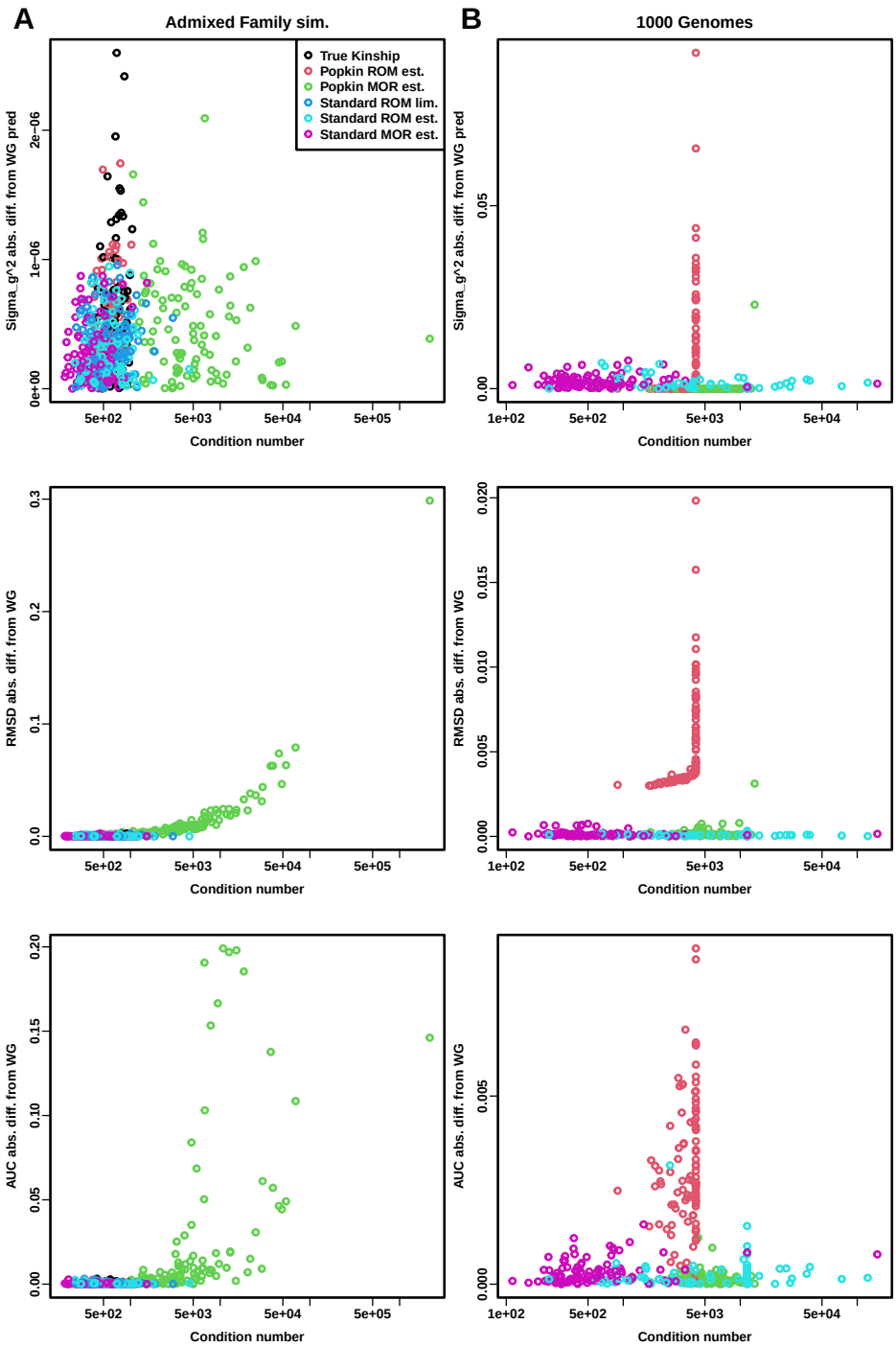

Figure S9: **$\mathrm{AUC_{PR}}$ and SRMSD_p prediction errors explained by the condition number of V.** $\mathrm{AUC_{PR}}$ and SRMSD_p are expected to be the same between WG, Standard, and True or Popkin (within each locus weight type). WG was used as reference since its **V** had the lowest condition numbers (Fig. S6). **A.** The large popkin MOR prediction errors ($\mathrm{AUC_{PR}}, \mathrm{SRMSD}_p$, but not σ^2) in the admixed family simulation are explained by the condition number of **V**. **B.** Smaller errors in 1000 Genomes are not explained by the condition number of **V**.