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 studies 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 characterized. 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 a variety of kinship matrices (every bias type has two locus weight types, and their theoretical limits for the simulation). Remarkably, we find nearly equal association statistics and performance for kinship matrices of different bias types (when all other features are matched). These empirical observations lead us to hypothesize that these association tests are invariant to these kinship biases, which using linear algebra we prove holds exactly for LMM and approximately for PCA. Our constructive proof shows that the intercept and relatedness (PCs in PCA, random effect in LMM) coefficients compensate for the kinship bias, so the result extends to generalized linear models as long as those coefficients are present and are nuisance parameters. Overall, we find that existing association studies are robust to kinship estimation bias, and our theoretical results may help improve association methods by taking advantage of this unexpected robustness, as well as help determine the effects of kinship bias in other settings.

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). This recent work 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 upweights 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 generalized the popkin and WG estimators to have both MOR and ROM versions as well, in order to test for the effect of estimator bias 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 in order to characterize type I error control and power using robust statistics. Surprisingly, we found that both LMM and PCA association are robust to kinship estimation bias to an extent that most association statistics are invariant to these biases. Lastly, we theoretically characterize the conditions under which these kinship biases result in invariant association statistics. Overall, we found that long-used association approaches are robust to the most common kinship estimation biases, and developed theoretical results 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. These genotypes can be treated as random variables structured according to relatedness. If φ_{jk} is the kinship coefficient of two individuals j and k, and p_i is the ancestral allele frequency at locus i, then under the kinship model (Ochoa and Storey, 2021) the expectation and covariance are given by

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

where $\mathbf{x}_i = (x_{ij})$ is the length-n column vector of genotypes at locus i, $\mathbf{\Phi} = (\varphi_{jk})$ is the $n \times n$ kinship matrix, $\mathbf{1}$ is a length-n column vector of ones, and the $^\intercal$ superscript denotes matrix transposition. Both $\mathbf{\Phi}$ and p_i 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). In this work, to simplify notation, we omit cumbersome notation that marks this dependence of parameters on the choice of ancestral population, nor do we explicitly condition on the ancestral population (it is done implicitly) when calculating expectations and covariances as done in previous work.

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 versions, some introduced in this work, 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

$\hat{p}_i = \frac{1}{2n}\mathbf{x}_i^\intercal \mathbf{1}$ is the standard ancestral allele frequency estimator, and $\hat{\mathbf{\Phi}}^{\text{name}} = (\hat{\varphi}_{jk}^{\text{name}})$ 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

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

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,

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

under the assumption that the true minimum kinship is zero. The MOR version, introduced here, upweighs rare variants by using $w_i = (\hat{p}_i(1-\hat{p}_i))^{-1}$; 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,

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

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

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

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

where $\mathbf{J} = \mathbf{1}\mathbf{1}^\intercal$ is the $n \times n$ matrix of ones, $\boldsymbol{\varphi} = \frac{1}{n}\mathbf{\Phi}\mathbf{1}$ is a length-n vector of per-row mean kinship values, and $\bar{\varphi} = \frac{1}{n^2}\mathbf{1}^\intercal\mathbf{\Phi}\mathbf{1}$ 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 the two standard estimators are functions of the corresponding popkin estimators, given by the bias function F^{std}:

$$\hat{\mathbf{\Phi}}^{\text{std-ROM}} = F^{\text{std}}(\hat{\mathbf{\Phi}}^{\text{popkin-ROM}}),$$
$$\hat{\mathbf{\Phi}}^{\text{std-MOR}} = F^{\text{std}}(\hat{\mathbf{\Phi}}^{\text{popkin-MOR}}).$$

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)

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

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

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

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

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

In Appendix C we prove that $F^{\mathrm{WG}}(\mathbf{\Phi})$ is positive definite if $\mathbf{\Phi}$ also is. In Appendix D we prove that

$$0 \leq \tilde{\varphi} \leq \bar{\varphi} \leq \bar{d} \leq 1,$$

where $\bar{d} = \frac{1}{n}\sum_{j=1}^{n} \varphi_{jj}$, 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 follows more straightforwardly that they are related by the bias function F^{WG}, while WG-MOR is introduced here and defined by the below formula:

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

2.3 Association models

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

$$\text{LMM:} \quad \mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta_i + \mathbf{s} + \boldsymbol{\epsilon}, \tag{8}$$
$$\mathbf{s} \sim \text{Normal}\left(\mathbf{0}, \sigma^2\mathbf{\Phi}\right), \tag{9}$$
$$\text{PCA:} \quad \mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta_i + \mathbf{U}_d\boldsymbol{\gamma}_d + \boldsymbol{\epsilon}, \tag{10}$$
$$\mathbf{\Phi} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\intercal}, \tag{11}$$

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 component factor, $\mathbf{U}_d$ is the $n \times d$ matrix of top-d eigenvectors of $\mathbf{\Phi}$ (often refered to as "principal components" in genetics), $\boldsymbol{\gamma}_d$ is a length-d 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 $\mathbf{\Phi}$, 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 $\mathbf{U}_d$ play analogous roles in modeling the effect of relatedness in LMM and PCA, respectively, we refer to them jointly as "relatedness"

effects, and σ and $\boldsymbol{\gamma}_d$ as their respective coefficients.

2.4 Simulations

2.4.1 Admixed family genotype simulation

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

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

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

We also include family structure in the simulation. 20 generations are generated iteratively. To preserve admixture structure mentioned above, individuals in the first generation ($n = 1000$) are ordered by 1D geography, locally unrelated and randomly assigned sex. 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 $< 1/4^3$ until 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 $\mathbf{X}$ is subset to these loci and reindexed. For known p_i, causal coefficients are constructed as:

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

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

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

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 were processed as before (Yao and Ochoa, 2022). Briefly, using `plink2` (Chang et al., 2015) we kept only autosomal biallelic SNP loci with filter "PASS", LD-pruned with parameters "`--indep-pairwise 1000kb 0.3`" to remove loci that have a greater than 0.3 correlation coefficient with other loci within 1000kb, and lastly remove loci

with MAF < 0.01. The resulting data has $m = 1,111,266$ loci and $n = 2,504$ individuals. Traits were 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) is used to measure the difference between the observed null p-value quantiles and the expected uniform quantiles (p-values of continuous test statistics follow a uniform distribution under the null):

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

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

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

3 Results

3.1 Empirical analysis using admixed family simulation

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

Kinship estimates and available limits on this simulation are shown in Fig. 1. The true kinship matrix 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; for Standard these biases vary for every pair of individuals whereas for WG they are uniform.

We performed LMM and PCA association tests to determine how kinship biases affect association performance. Surprisingly, we found that kinship bias type does not have a discernible effect on association performance, as summarized by $\mathrm{AUC}_{\mathrm{PR}}$ (a robust proxy for power; Fig. 2) and SRMSD_p

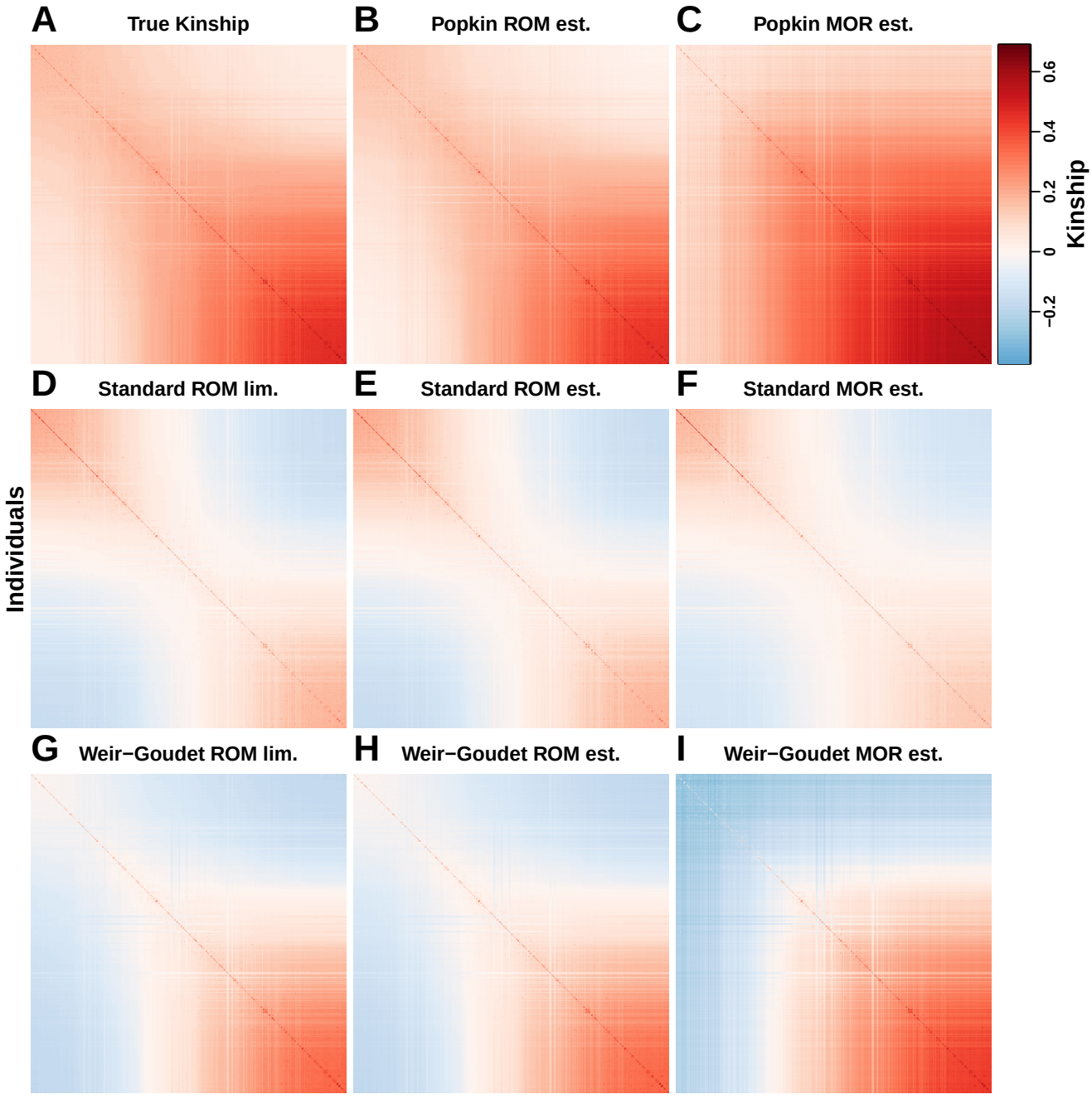

Figure 1: **Kinship estimates and limits on the admixed family simulation.** Each panel shows a kinship matrix as a heatmap, with each of the $n = 1000$ individuals along both x and y axes, color represents kinship: positive estimates in red, negative in blue. Diagonal contains inbreeding estimates.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

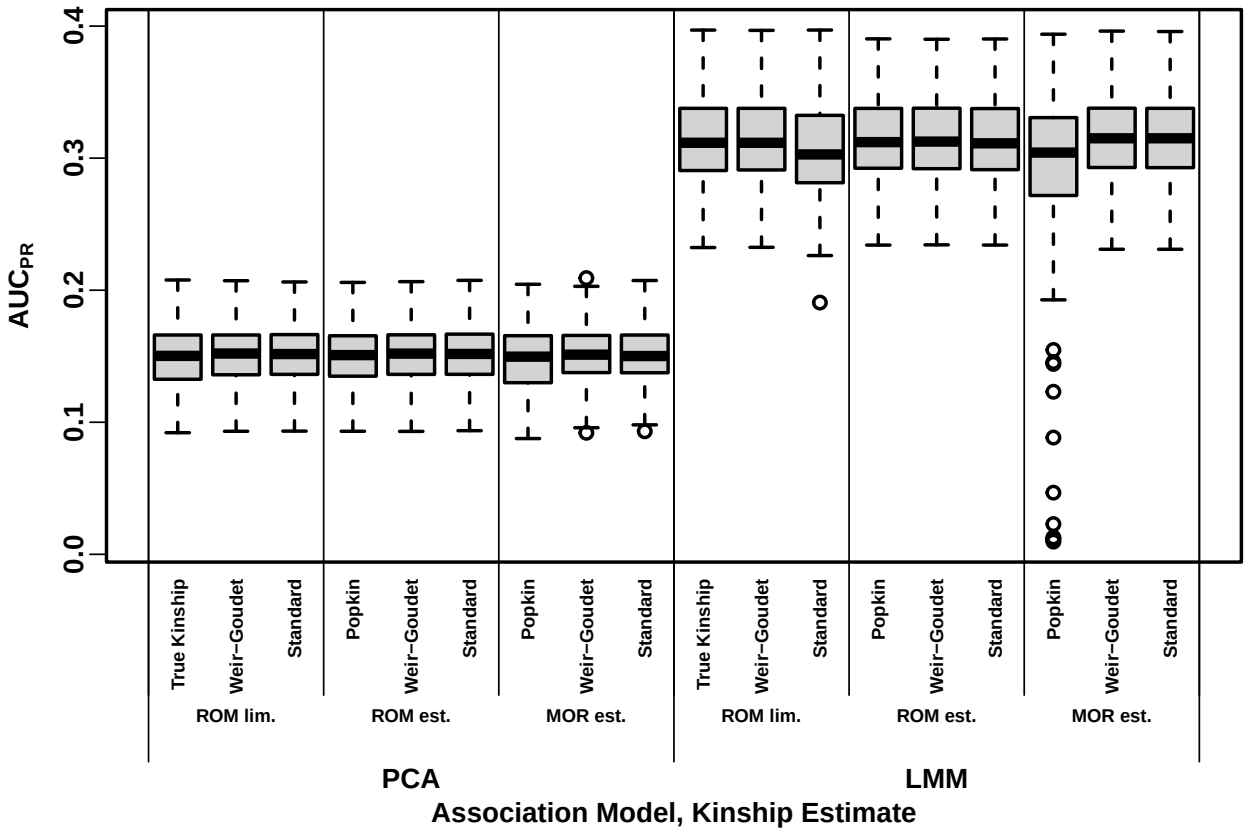

Figure 2: **Distributions of Area Under the Precision-Recall Curve ($\mathbf{AUC_{PR}}$) on the admixed family simulation.** Higher AUC_{PR} is better performance. Results for 100 replicates (each a random genotype matrix and trait vector). Approaches cluster primarily by association model (LMM or PCA), and do not depend much at all on the bias type.

(measure null statistic calibration; Fig. S2). The largest differences in performance are explained by the association model used (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 was exceedingly poor.

To better characterize the nearly-identical performance distribution just observed, we next measured the agreement between individual association p-values. We calculated the proportion of loci between two methods with p-values within 0.01 of each other, which is an approximate measure of agreement, and found a remarkably high agreement between estimators of diferent bias types after matching association model and locus-weight version or limit (Fig. 3). This is in contrast to the low amounts of agreement across PCA and LMM statistics, and even across LMM statistics with different locus-weight or between those and the ROM limits. Minimum agreements tended to be higher across PCA methods, though here use of the true kinship or either popkin estimates resulted in more disagreements than between Standard and WG estimates or limits. Overall, sets of matched kinship matrices except for different bias types 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 has 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: $\mathrm{AUC}_{\mathrm{PR}}$ and SRMSD_p

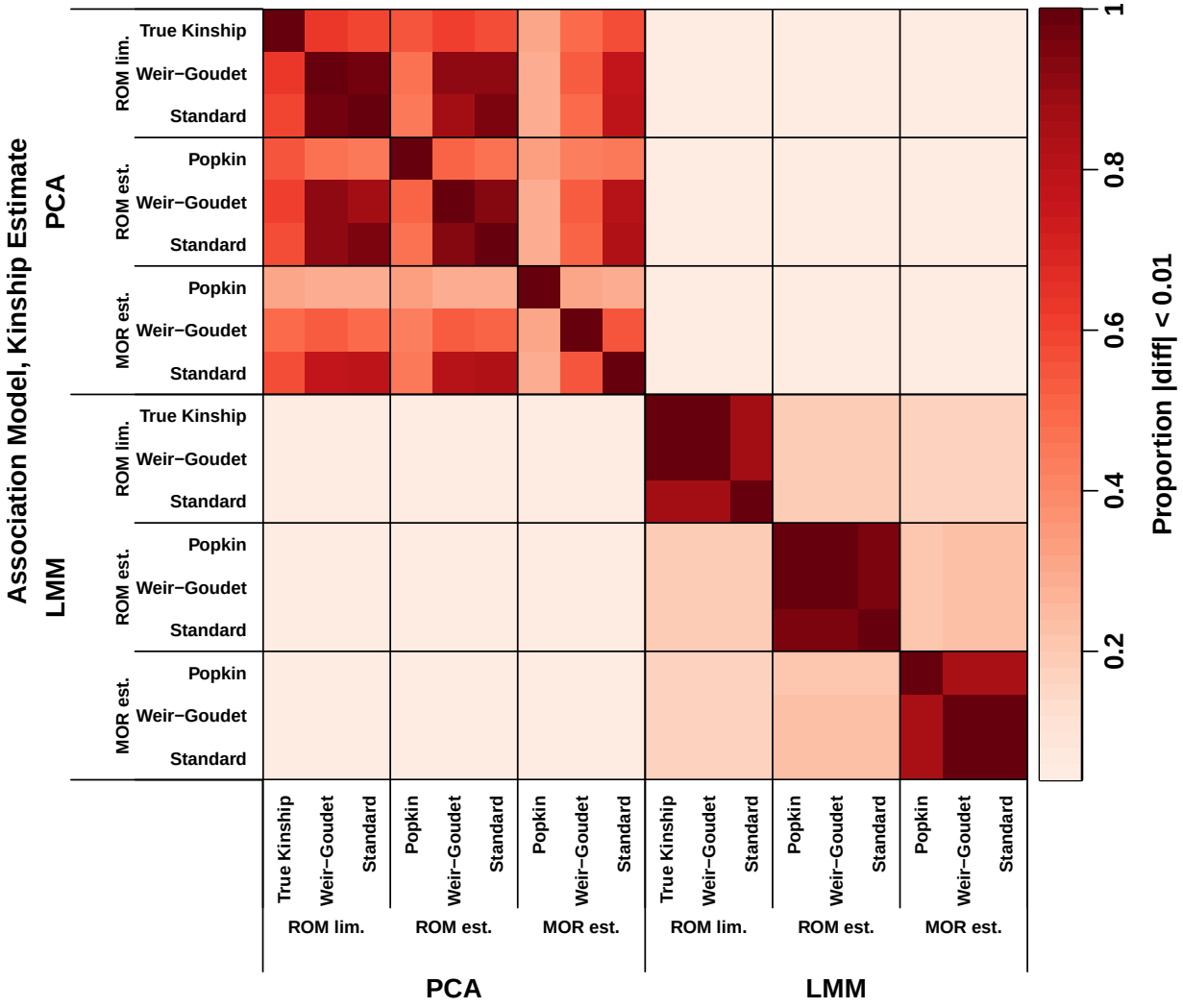

Figure 3: **Approximate agreement between p-values on the admixed family simulation.** The association p-value vector (one value per tested locus) produced by each combination of association model (LMM vs PCA) and kinship matrix (x and y axes) was used to compute proportions of loci with absolute differences under 0.01 (color). All 100 replicates were used. Methods of all bias types (matched for association model and locus weight type) have large proportions of nearly identical p-values.

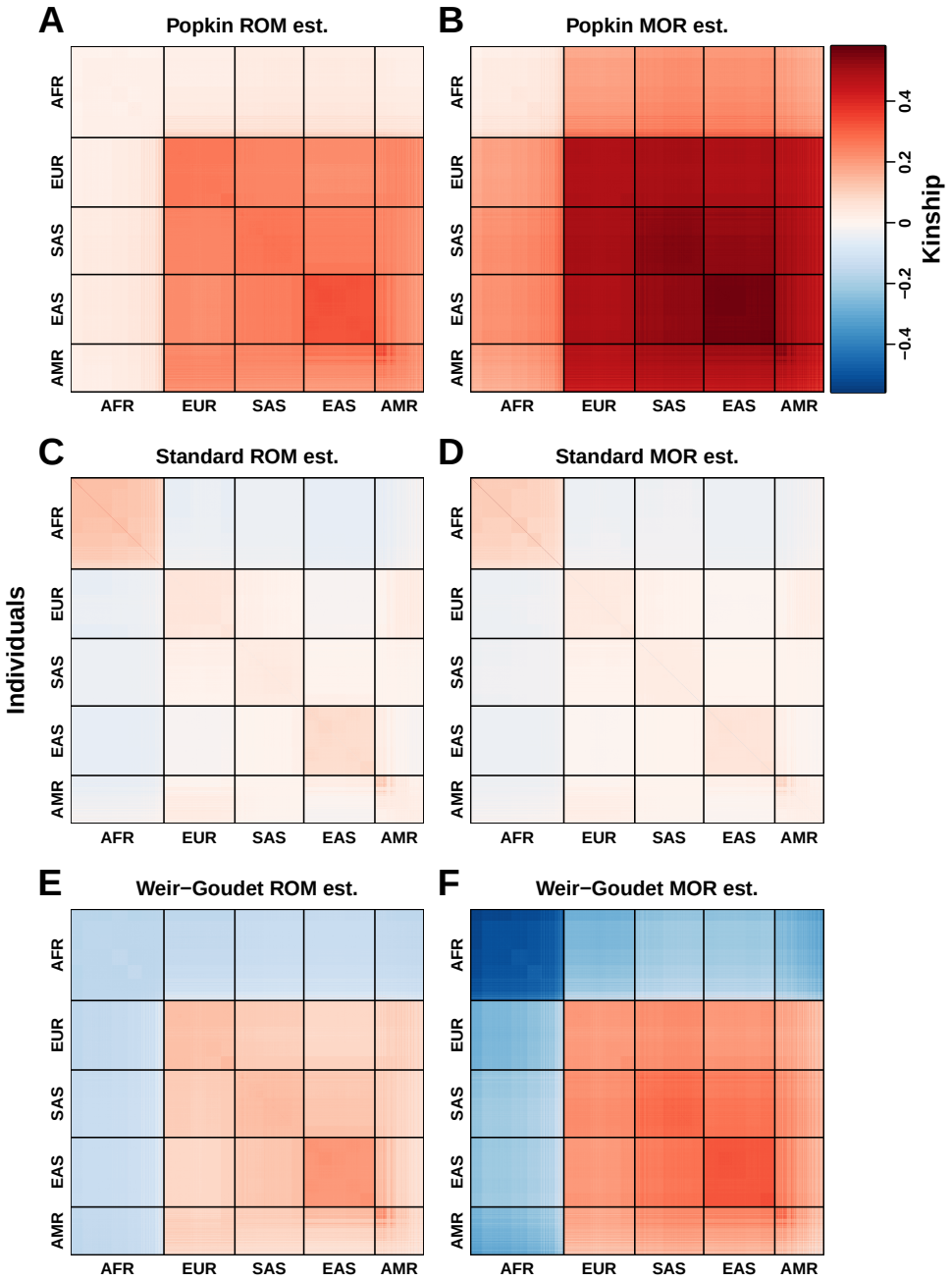

Figure 4: **Kinship estimates on 1000 Genomes.** Each panel represents a kinship matrix as a heatmap, as in Fig. 1. Superpopulation codes: AFR = African, EUR = European, SAS = South Asian, EAS = East Asian, AMR = Admixed Americans (Hispanics). Each estimator bias type (Popkin, Standard, and Weir-Goudet;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

distributions are nearly identical for estimators of different bias types but same locus-weight version (ROM or MOR) and association model (Figs. S3 and 5). However, unlike the simulation, here the MOR estimates greatly outperform ROM estimates (LMM only), on terms of both $\mathrm{AUC_{PR}}$ and SRMSD_p. P-values are again nearly identical at a large proportion of loci between approaches with matched association model and locus-weight version (MOR or ROM), regardless of bias type (Fig. S4).

3.3 Proof of association invariability to common kinship biases

Our empirical observations suggested that replacing a kinship matrix with either the Standard- or WG-biased version does not alter association statistics; here prove a general version of these

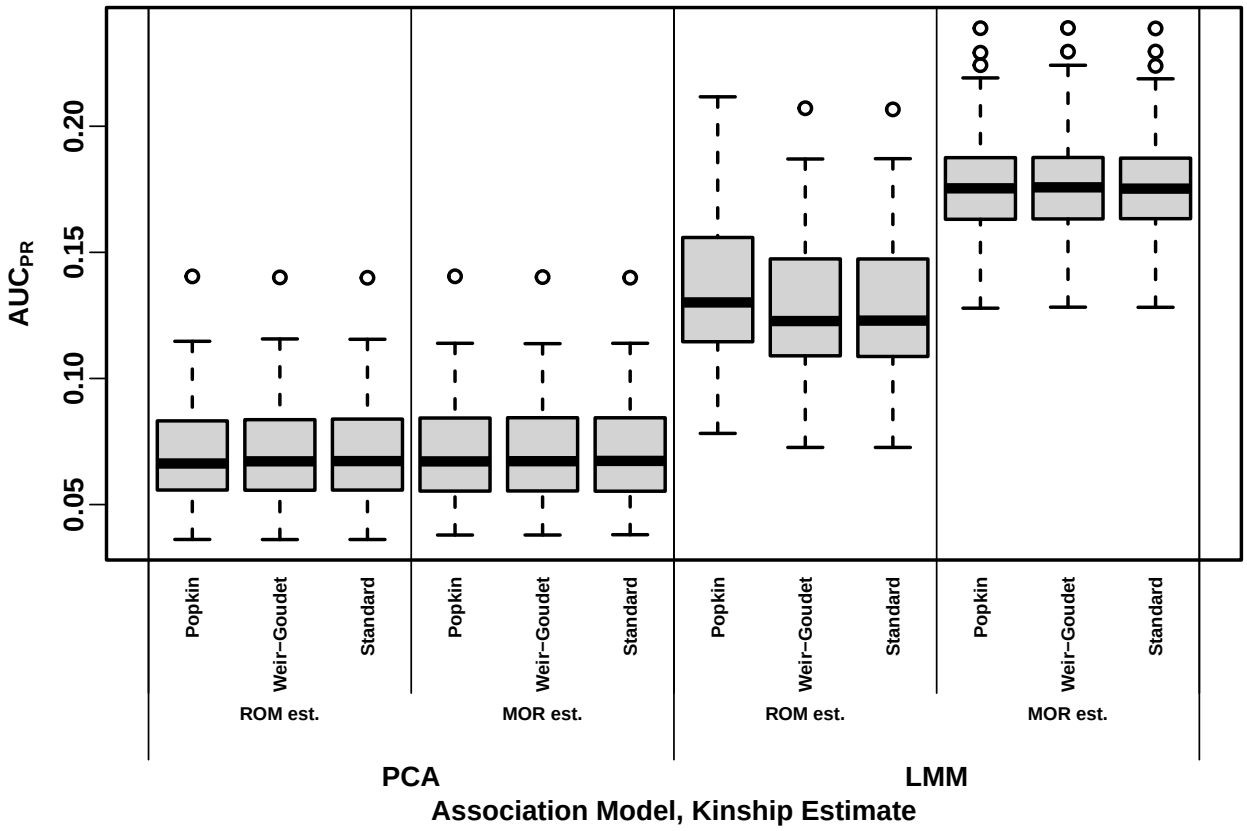

Figure 5: **Distributions of Area Under the Precision-Recall Curve ($\mathbf{AUC_{PR}}$) on 1000 Genomes.** Higher $\mathrm{AUC_{PR}}$ is better performance. Results based on 100 simulated trait replicates (real genotype matrix is fixed). Approaches cluster primarily by association model (LMM or PCA)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

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 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

$$\boldsymbol{\Phi}' = F(\boldsymbol{\Phi}) = \frac{1}{c}\mathbf{B}\boldsymbol{\Phi}\mathbf{B}^\intercal, \qquad \mathbf{B} = \mathbf{I} - \mathbf{1}\mathbf{b}^\intercal, \tag{12}$$

where c is any positive scalar and $\mathbf{b}$ is any length-n vector.

The Standard bias function $F = F^{\text{std}}$ of Eq. (4) can be written as Eq. (12) with $c = 1 - \bar{\varphi}$ and $\mathbf{b} = \frac{1}{n}\mathbf{1}$, in which case $\mathbf{B}$ equals the centering matrix. Further, the generalized standard kinship estimator studied in Ochoa and Storey (2021) instead has $\mathbf{b}$ be the vector of individual weights used in the estimator, whose elements must sums to one, so $\mathbf{b}^\intercal\mathbf{1} = 1$. In all these cases $\mathbf{B}$ and $\boldsymbol{\Phi}'$ are singular transformations, since $\mathbf{B1} = \mathbf{0}$ and $\mathbf{B}^\intercal\mathbf{b} = \mathbf{0}$.

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

$$\mathbf{b} = b\frac{\boldsymbol{\Phi}^{-1}\mathbf{1}}{\mathbf{1}^\intercal\boldsymbol{\Phi}^{-1}\mathbf{1}},$$
$$b = 1 \pm \sqrt{1 - (\mathbf{1}^\intercal\boldsymbol{\Phi}^{-1}\mathbf{1})\tilde{\varphi}}.$$

The determinant of b is non-negative, since $\tilde{\varphi} \leq 1/(\mathbf{1}^\intercal\boldsymbol{\Phi}^{-1}\mathbf{1})$ [TODO: prove it!], so both values of b above are real and valid, the positive root being positive while the negative one is negative. This $\mathbf{b}$ is not a weight vector, since $\mathbf{b}^\intercal\mathbf{1} = b \neq 1$ [TODO: unless there's a degenerate case?]. Further, we prove that WG-biased kinship matrices are invertible if the original was invertible in Appendix C.

3.3.1 Proof for LMM case

Consider a random effect $\mathbf{s}$ drawn using $\boldsymbol{\Phi}$, as given in Eq. (9). Using the affine transformation property of Multivariate Normal distributions (which despite its name also holds for singular linear

transformations) and Eq. (12), it follows that

$$\mathbf{s}' = \mathbf{B}\mathbf{s} \sim \text{Normal}\left(\mathbf{0}, (\sigma')^2\mathbf{\Phi}'\right),$$

where $(\sigma')^2 = c\sigma^2$. (This $\mathbf{s}'$ has a degenerate distribution for Standard bias, since $\mathbf{\Phi}'$ is singular, but this is not problematic as long as $\mathbf{s}' + \boldsymbol{\epsilon}$ is non-degenerate, whose total covariance $(\sigma')^2\mathbf{\Phi}' + \sigma_\epsilon^2\mathbf{I}$ is invertible as long as $\sigma_\epsilon^2 \neq 0$.) The key property that $\mathbf{B}$ must satisfy for association invariability is that $\mathbf{B}\mathbf{s} = \mathbf{s} - \mathbf{1}s$, where $s = \mathbf{b}^\intercal\mathbf{s}$ is a scalar, so

$$\mathbf{s}' = \mathbf{s} - \mathbf{1}s$$

are equal in distribution. Therefore, after matching the variance components σ^2 and $(\sigma')^2$ as above, the random effect $\mathbf{s}'$ of the biased kinship matrix differs from the random effect $\mathbf{s}$ of the original kinship only by $\mathbf{1}s$, a difference compensated for by adjusting the intercept coefficient in Eq. (8): $\alpha' = \alpha + \mathbf{1}s$. No other regression coefficients, or the total residuals, change when $\mathbf{\Phi}$ is replaced with $\mathbf{\Phi}'$, including the association coefficient β_i that is the focus of the test.

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 the covariance matrices do change, their determinants cancel out in the ratio. [TODO: What about Score, Wald tests? I think there can be changes in one of those cases!] The argument holds whether the model is fit with maximum likelihood (ML) or restricted maximum likelihood (REML) (Kang et al., 2008), since the difference only affects how σ^2 is estimated, and in both cases the adjusted estimate, which is given by the original σ^2 multiplied by c, results in an identical likelihood so no other estimates are affected.

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 $\mathbf{U}_d$ be the top eigenvectors of $\mathbf{\Phi}$, and $\mathbf{U}'_d$ those

of $\mathbf{\Phi}'$. They key approximation is that

$$\mathbf{U}'_d \approx \mathbf{B}\mathbf{U}_d,$$

which is not strictly equal (since $\mathbf{BU}$ 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, $\boldsymbol{\gamma}'_d = \boldsymbol{\gamma}_d$, since the difference in scale of the kinship matrices (c in Eq. (12)) is absorbed by the eigenvalues, which are not used in the association model. As before, note that

$$\mathbf{U}'_d\boldsymbol{\gamma}'_d = \mathbf{B}\mathbf{U}_d\boldsymbol{\gamma}_d = \mathbf{U}_d\boldsymbol{\gamma}_d - \mathbf{1}s,$$

where $s = \mathbf{b}^\intercal\mathbf{U}_d\boldsymbol{\gamma}_d$ is a scalar. Therefore, the relatedness effects again differ only by $\mathbf{1}s$, which is compensated for by adjusting the intercept accordingly, so the association coefficient β_i and the residuals are the same in both cases. The observations from LMMs, for how p-values change depending on the type of test used, also hold for PCA.

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). We initiated the present work under the hypothesis that these kinship biases would affect association testing, but surprisingly find that association is unaffected by these kinship biases. We then prove theoretically that it is the intercept and population structure (random effect or PCs) coefficients that compensate for the bias, and result in identical genetic effect coefficients and significance statistics.

Given that kinship bias type is not important for association studies, we are free to choose a kinship estimator based on other properties. The biased standard kinship matrix may be more desirable than the popkin estimator based on the numerical stability we observed in our simulations. In particular, while theory shows that the solutions should be the same for all estimators of the same type, we find that popkin's statistics disagree more often from the standard and WG estimators, namely LMM association with popkin MOR (admixed family simulation, Fig. 2, Fig. S2) and popkin ROM (1000 Genomes, Fig. S3). The standard kinship matrix is orthogonal to the intercept, because

of the centering operation applied to obtain it in our theoretical results, whereas the popkin and true kinship matrix are not orthogonal to the intercept. Thus, PCA regression with the eigenvectors of the standard kinship matrix is more numerically stable (because more covariates are linearly independent) than the popkin counterpart. We believe that the observed popkin disagreemnts in LMMs are due to poor convergence of that algorithm in those cases.

We also found that all MOR estimators perform better in the LMM association (and overall) compared to the ROM versions in the 1000 Genomes evaluation. Perhaps this is expected because out trait simulation follows the "fixed effect sizes" model, in which rare variants have larger coefficients, and the MOR estimators also weigh rare variants more highly in estimating kinship coefficients. This effect was not observed in the admixed family simulation, where MOR and ROM versions gave similar kinship estimates and performed similarly, compared to the real data evaluation, where kinship estimates were also strikingly different. However, only the popkin ROM estimator is unbiased (Fig. 1B, Fig. S1), so it is unclear why the biased popkin MOR estimator performs better in this setting. One potential explanation is that our kinship model assumes that all variants were preexisting in the MRCA population, whereas rare variants in human data are known to be very recent mutations, and thus their effective kinship matrix is different than that of ancestral variants. Therefore, despite its biases, it is possible that the popkin MOR estimator is more accurately capturing the kinship matrix of these rare variants and thus modeling them better in association tests, particularly in LMMs where the effect is most pronounced.

Our conclusions extend to variants of the standard kinship estimator that weigh loci according to linkage disequilibrium (Speed et al., 2017; Wang et al., 2017), which have the same bias form since this bias is present in each individual locus (Ochoa and Storey, 2021). As shown in our proof, the more general form of the standard kinship estimator that weighs individuals to estimate ancestral allele frequencies $\hat{p}_i$ is also subject to the same conclusions. Such weighted $\hat{p}_i$ estimates include the best unbiased linear estimator (Astle and Balding, 2009; Thornton and McPeek, 2010).

In this study, we show empirically and theoretically that association tests are invariant to the use of common kinship estimators that are biased as well as a more recent unbiased estimator. The theoretical underpinnings of our proof show that the same is expected of any generalized linear model

with the same setup, namely intercept and population structure with coefficients that are nuisance variables, which includes case/control models as well as the quantitative trait model we explicitly studied here. However, heritability estimation requires unbiased estimates of the random effect coefficient (σ^2), so our results prove that it will be 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 its full study 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.

Astle, William and David J. Balding (2009). "Population Structure and Cryptic Relatedness in Genetic Association Studies". *Statist. Sci.* 24(4). Mathematical Reviews number (MathSciNet): MR2779337, pp. 451–471.

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.

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

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

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

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

Devlin, B. and Kathryn Roeder (1999). "Genomic Control for Association Studies". *Biometrics* 55(4), pp. 997–1004.

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

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.

García-Cortés, Luis Alberto (2015). "A novel recursive algorithm for the calculation of the detailed identity coefficients". *Genetics Selection Evolution* 47(1), p. 33.

Hefferon, Jim (2020). *Linear Algebra*. 4th. Leanpub.

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

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

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

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.

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

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.

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

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

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

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

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

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

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

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

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

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

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.

Yang, Jian et al. (2011). "GCTA: a tool for genome-wide complex trait analysis". *Am. J. Hum. Genet.* 88(1), pp. 76–82.

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

Yao, Yiqi and Alejandro Ochoa (2022). *Limitations of principal components in quantitative genetic association models for human studies*. Tech. rep. Section: New Results Type: article. bioRxiv, p. 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.

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

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 formulation had $w_i = 1$ for all loci i (Ochoa and Storey, 2021). Recalling from that original work that

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

then for fixed w_i we get

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

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

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

as desired.

The MOR case of $w_i = (\hat{p}_i(1 - \hat{p}_i))^{-1}$ 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

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

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

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

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

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

B Connection between popkin and standard kinship estimator

Since the conection we discovered holds when data is complete but not under missingness, to determine necessary conditions, here we introduce more complete forms of the estimators that handle missingness. The generalized popkin estimator (including both ROM and MOR special cases) is

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

where $I_{ij} = 1$ if x_{ij} is not missing, 0 otherwise (this way missing x_{ij} can be treated as having any finite value and not contribute to the estimator). Note that only loci where both genotypes (x_{ij} and x_{ik}) are 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

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

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 generalized popkin estimator. Since the result we want holds at every locus separately, let us formulate the averages of interest at locus i only:

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

Therefore, the combination of interest is:

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

To arrive at the desired result of $I_{ij}I_{ik}(x_{ij} - 2\hat{p}_i)(x_{ik} - 2\hat{p}_i)$, which is the first term above, it is necessary for the rest of the terms to vanish for arbitrary values of $\hat{p}_i$, x_{ij}, and x_{ik}. Since $n_i > 0$ (there is at least one non-missing individual at every locus), the term $\frac{n_i}{n}(I_{ij} - \frac{n_i}{n})4\hat{p}_i$ vanishes if and only if $I_{ij} = \frac{n_i}{n}$, and since $I_{jk} = 0$ does not solve this equation (because $n_i > 0$) the only other case

is $I_{jk} = 1$, which requires $n_i = n$, so no individuals can have missing data at this locus. Thus,

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

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

$$\bar{A}_i = -4\hat{p}_i(1 - \hat{p}_i)$$

also requires $n_i = n$.

Assuming now no missingness, transforming the popkin estimates as desired gives

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

Therefore, if the ROM version of popkin is input ($w_i = 1$), this transformation yields the ROM version of the standard kinship estimator. On the other hand, if the MOR version of popkin is used ($w_i^{-1} = \hat{p}_i(1 - \hat{p}_i)$), the transformation yields the MOR version of the standard kinship estimator.

C Proof that WG limits are positive definite

Starting from a positive-definite kinship matrix $\mathbf{\Phi}$, we prove that WG-bias transformed matrix $\mathbf{\Phi'} = F^{\text{WG}}(\mathbf{\Phi})$ is also positive definite (and therefore invertible), except in one degenerate case. Recall from Eq. (6) that $\mathbf{\Phi'} = \frac{1}{1-\tilde{\varphi}}(\mathbf{\Phi} - \tilde{\varphi}\boldsymbol{J})$. We shall not consider $\mathbf{\Phi} = \mathbf{J}$ as a valid kinship matrix, which therefore ensures that $\tilde{\varphi} < 1$ as there is at least one kinship value with $\varphi_{jk} < 1$. Now consider two linear subspaces of $\mathbb{R}^n$, S_1 spanned by $\mathbf{1}$ and S_2 its complement (orthogonal to $\mathbf{1}$), and prove that $\mathbf{\Phi'}$ is positive definite in both subspaces, therefore it is positive-definite in the direct sum of the subspaces, which equals the entire space: $\mathbb{R}^n = S_1 \oplus S_2$. (This follows since vectors $\mathbf{v}$ for which $\mathbf{\Phi'}$ is not positive definite, if they exist, span a linear subspace, but its intersection to S_1, S_2, and

therefore $S_1 \oplus S_2$, is trivial (Hefferon, 2020).) In both subspaces we will prove that $\mathbf{v} \in S_i$ and $\mathbf{v} \neq \mathbf{0}$ implies $\mathbf{v}^\intercal \boldsymbol{\Phi}' \mathbf{v} > 0$ which proves that $\boldsymbol{\Phi}'$ is positive definite in that subspace.

We begin by considering $\mathbf{v} \in S_2$, which satisfy $\mathbf{1}^\intercal \mathbf{v} = \mathbf{0}$, and by hypothesis $\mathbf{v} \neq \mathbf{0}$. Therefore $\mathbf{v}^\intercal \mathbf{J} \mathbf{v} = 0$ in this subspace, which results in

$$\mathbf{v}^\intercal \boldsymbol{\Phi}' \mathbf{v} = \frac{1}{1 - \tilde{\varphi}} \mathbf{v}^\intercal \boldsymbol{\Phi} \mathbf{v} > 0,$$

where the final inequality follows since the original kinship matrix is positive definite and $1 - \tilde{\varphi} > 0$.

Lastly, we consider $\mathbf{v} \in S_1$, which are necessarily of the form $\mathbf{v} = v\mathbf{1}$, and by hypothesis $v \neq 0$. Therefore

$$\mathbf{v}^\intercal \boldsymbol{\Phi}' \mathbf{v} = \frac{v^2}{1 - \tilde{\varphi}} (\mathbf{1}^\intercal \boldsymbol{\Phi} \mathbf{1} - \tilde{\varphi} n^2) = \frac{v^2 n^2}{1 - \tilde{\varphi}} (\bar{\varphi} - \tilde{\varphi}),$$

where $\bar{\varphi}$ is the overall mean kinship value, while $\tilde{\varphi}$ is the mean of the off-diagonal kinship values only (Eq. (7)). Note that $v^2, n^2, 1 - \tilde{\varphi} > 0$, so the desired result follows if $\tilde{\varphi} < \bar{\varphi}$, which is proven in Appendix D. In general it is true that $\tilde{\varphi} \leq \bar{\varphi}$, and $\tilde{\varphi} = \bar{\varphi}$ occurs if and only if the kinship matrix has the degenerate form $\boldsymbol{\Phi} = \bar{\varphi} \mathbf{J}$, which is a singular matrix not expected in practice (in this case $\boldsymbol{\Phi}'$ is a matrix full of zeroes).

D Mean kinship inequalities

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

$$0 \leq \tilde{\varphi} \leq \bar{\varphi} \leq \bar{d} \leq 1,$$

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

The inequalities $0 \leq \bar{\varphi} \leq \bar{d} \leq 1$ 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). Recall that $\tilde{\varphi}$ is defined in Eq. (7). The lower bound $0 \leq \tilde{\varphi}$ follows since every kinship value is non-negative. Note that $\bar{\varphi}$ and $\tilde{\varphi}$ are related

by

$$\bar{\varphi} = \frac{\tilde{\varphi}(n-1) + \bar{d}}{n}. \tag{13}$$

Applying $\bar{\varphi} \leq \bar{d}$ to Eq. (13) and simplifying yields $\tilde{\varphi} \leq \bar{d}$. Lastly, since $\bar{\varphi} - \tilde{\varphi} = (\bar{d} - \tilde{\varphi})/n$ (from rearranging Eq. (13)), it also follows that $\tilde{\varphi} \leq \bar{\varphi}$, as desired. Furthermore, $\tilde{\varphi} = \bar{\varphi}$ holds if and only if all $\varphi_{jk} = \bar{d}$, since that is necessary and sufficient for $\bar{\varphi} = \bar{d}$.

Supplemental figures

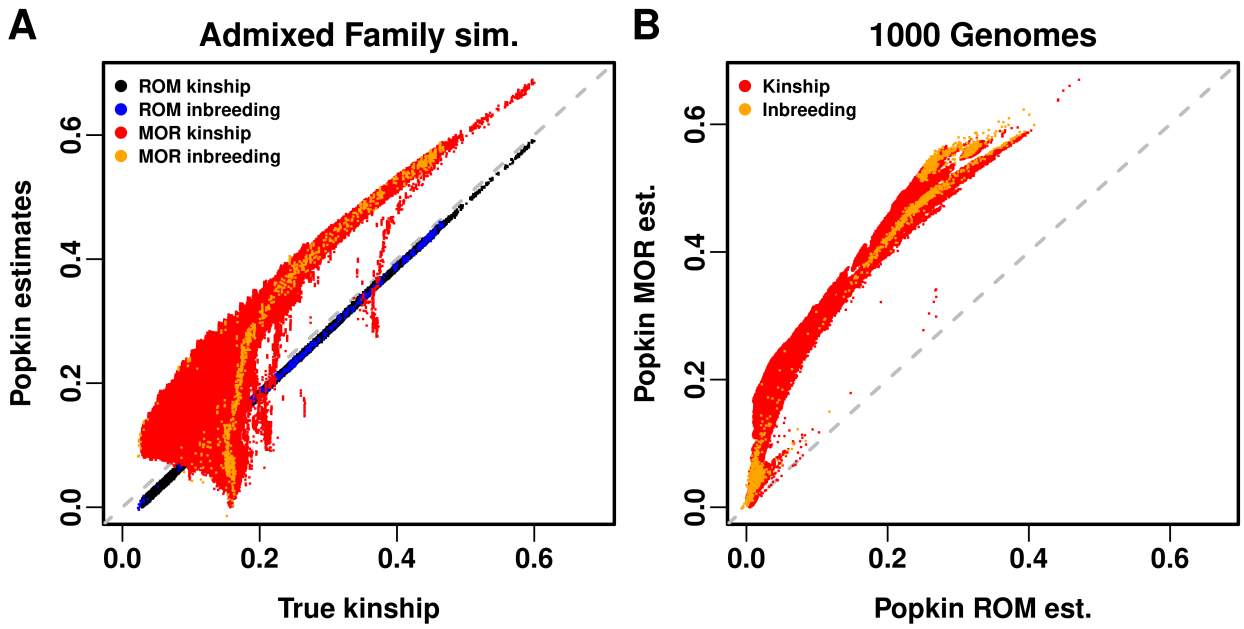

Figure S1: **Comparison of popkin ROM and MOR estimates.** Kinship (off-diagonal of matrix) and inbreeding (transformed diagonal) are plotted in different colors, which shows that their biases (if any) overlap. **A.** In admixed family simulation, both estimates are compared against true kinship. Popkin 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.

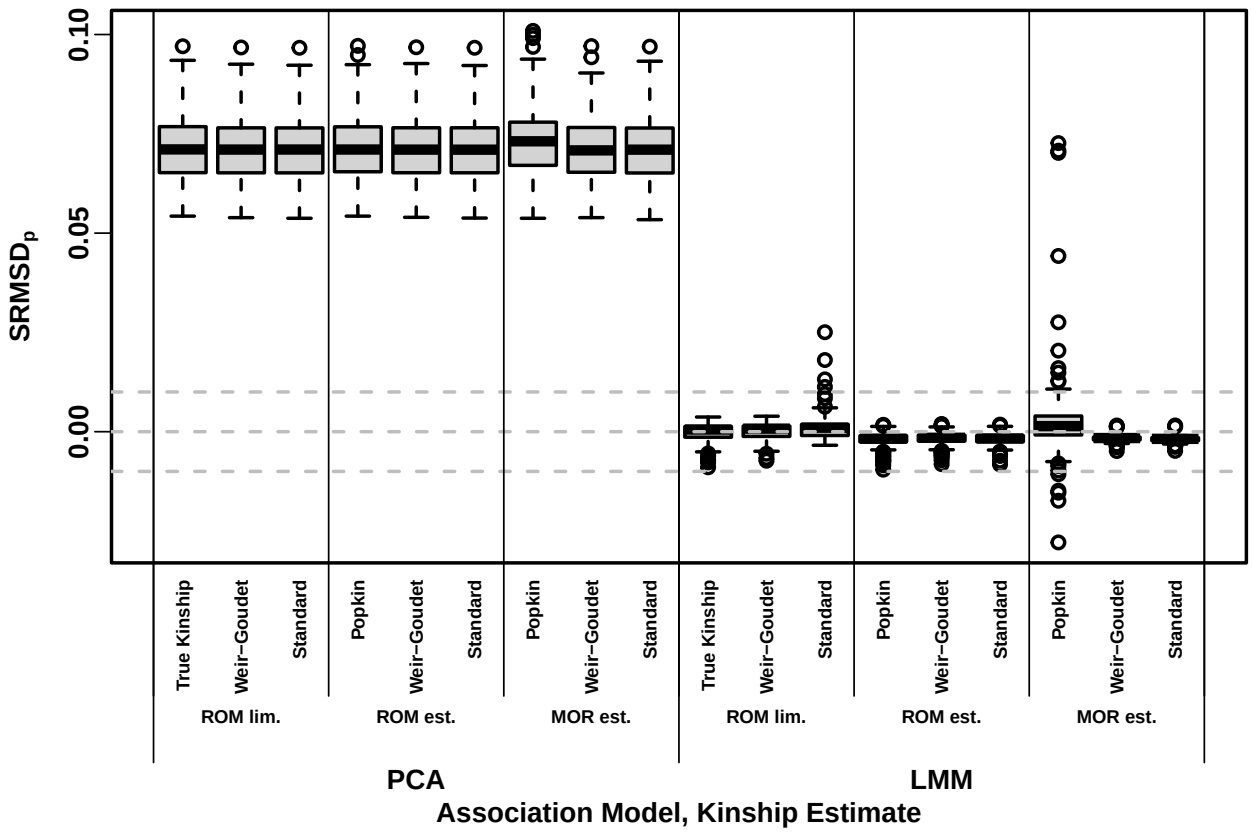

Figure S2: **Signed Root Mean Square Deviation of null p-values (SRMSD_p) on the admixed family simulation.** Same methods and simulation as Fig. 2, see that for more information. $|\text{SRMSD}_p| < 0.01$ (area between gray dashed lines) is considered calibrated. All PCA runs are miscalibrated by similar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

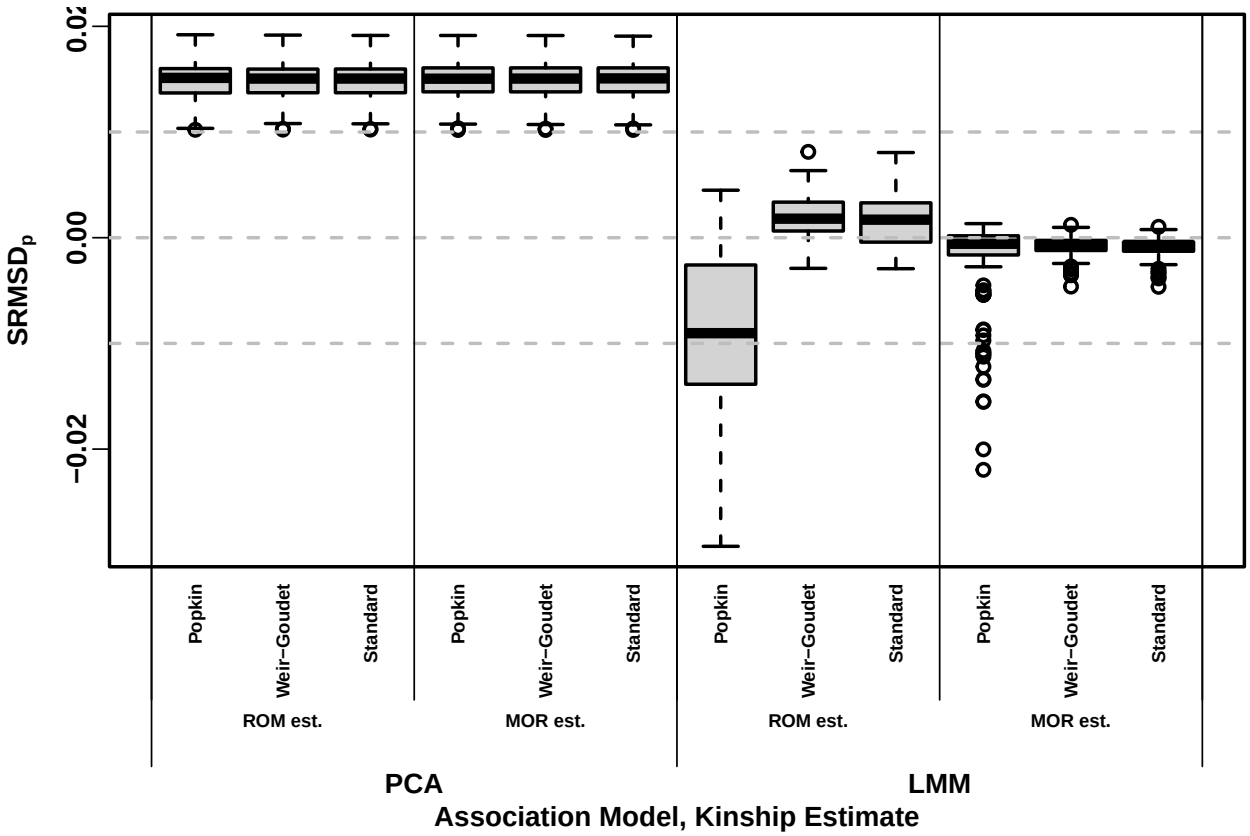

Figure S3: **Signed Root Mean Square Deviation of null p-values (SRMSD$_p$) on 1000 Genomes.** Same methods and simulation as Fig. 5, and y-axis statistic and conclusions of Fig. S2, see those for more information.

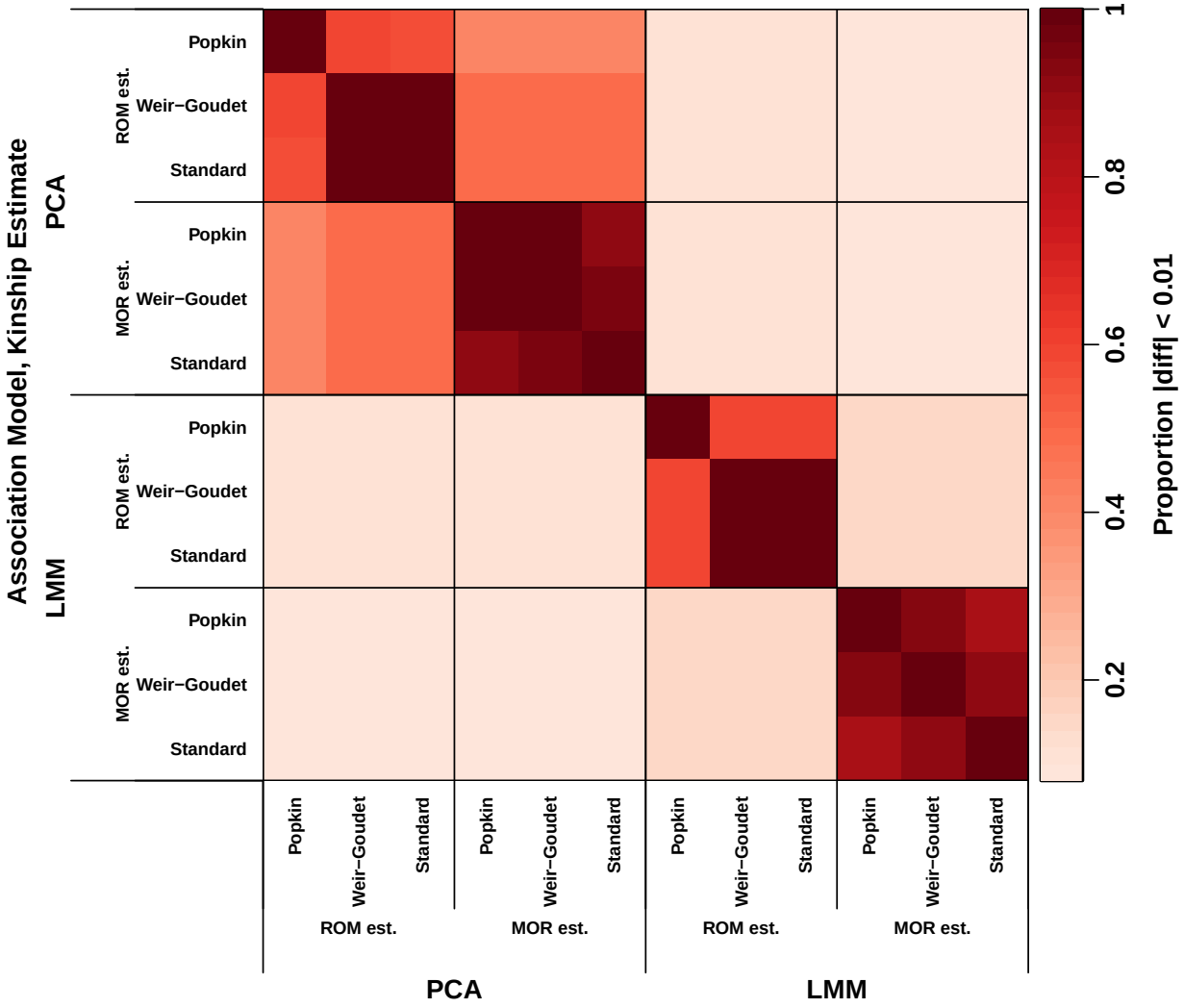

Figure S4: **Approximate agreement between p-values on 1000 Genomes.** See Fig. 3 for more details.