Fixed and mixed-effect 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

Genetic association studies for structured populations model the correlation structure between individuals, most often using population kinship matrices. The two most common association models, Principal Components Analysis (PCA) association and the Linear Mixed-effects Model (LMM), are expressed in terms of kinship matrices. However, the most common kinship estimators can have severe biases that were only recently characterized. The goal of this work is to characterize the effect of these kinship biases on genetic association. We employed a large simulated admixed family and data from the 1000 Genomes Project, both with simulated traits, to evaluate the performance of each of PCA and LMM when a variety of kinship matrices is used. We tested several common kinship estimators, and in the simulated admixed family we also included the true kinship matrix and the the theoretical limits of the biased estimators (as the number of loci approaches infinity). Remarkably, we find negligible differences in association statistics and overall performance between all kinship matrices tested. Our empirical observations led us to the hypothesis that the association tests are invariant to whether the true kinship matrix or two of the biased limits are employed. We prove, using linear algebra, that our hypothesis holds exactly for LMM and approximately for PCA. Our constructive proof shows that both biased kinship matrices result in regression covariates (PCs in PCA, random effects in LMM) whose biases are compensated for by fitting the intercept, suggesting that only models with this precise arrangement will be robust to these kinship biases.

1 Introduction

Genome-Wide Association Studies (GWAS) are commonly used to detect loci that are related specific traits. For GWAS based on a structured population, which includes admixed individuals and multiethnic cohorts, controlling for population structure is crucial. If the population structure is not properly taken into account, the analysis will lead to spurious associations and lack of power (Devlin and Roeder, 1999; Voight and Pritchard, 2005; Astle and Balding, 2009). Linear mixed models (LMM) and Principal component analysis (PCA) are two popular approaches for GWAS in a structured population. These two models are closely related, and the main difference is that LMM use random effects whereas PCA use fixed effects (Astle and Balding, 2009; Hoffman, 2013).

Kinship is utilized in PCA and LMM to correct for structure in GWAS (Xie et al., 1998; Yu et al., 2006; Aulchenko et al., 2007; Price et al., 2006; Astle and Balding, 2009; Kang et al., 2008; Kang et al., 2010; Yang et al., 2011; Zhou and Stephens, 2012; Yang et al., 2014; Loh et al., 2015; Sul et al., 2018). LMM incorporates kinship matrix in its random effect term, while for PCA, the PCs are the eigenvectors of the kinship matrix.

The most commonly-used kinship estimator (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) (standard kinship estimator) was recently determined to have a complex bias (Weir and Goudet, 2017; Ochoa and Storey, 2021). The bias could vary for every pairs of individual and may cause issues for the downstream analysis that is based on kinship estimation. Weir-Goudet kinship estimator was proposed recently and also has a uniformly downward bias (Weir and Goudet, 2017; Ochoa and Storey, 2021). We previously proposed an unbiased estimator called the popkin estimator (Ochoa and Storey, 2021). The popkin estimator could be generalized by assigning different weights for loci, and connected to the standard kinship estimator.

In this paper, we originally hypothesized that the bias from kinship estimation affect association testing. We explore our findings 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), evaluate the performance of different kinship estimator using AUC_{PR} and SRMSD_p, and plot correlation between p-values of different approaches. Furthermore, we provide theoretical justification of the simulation results for two association models (LMM and PCA) and two biased kinship estimators (the standard estimator and Weir-Goudet estimator). Overall, we found that the intercept and population structure coefficients compensate for the kinship bias, leading to a unbiased association testing statistics 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

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

where $\mathbf{x}_i$ is the length-n column vector of genotypes at locus i, $\mathbf{X} = (\mathbf{x}_i^\intercal)$ is the complete $m \times n$ genotype matrix, $\mathbf{\Phi} = (\varphi_{jk})$ is the $n \times n$ kinship matrix, $\mathbf{p} = (p_i)$ is a length-m column vector of ancestral allele frequencies, $\mathbf{1} = (1)$ is a length-n column vector where every element is 1, and the $\intercal$ superscript denotes matrix transposition. Both kinship ($\mathbf{\Phi}$) and ancestral allele frequencies ($\mathbf{p}$) 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 estimation

2.2.1 Standard kinship estimator

The "standard" kinship estimator is the most common estimator employed across various applications for population structure (Astle and Balding, 2009; Yang et al., 2014; Speed and Balding, 2015; Wang et al., 2017), including heritability estimation (Speed et al., 2012; Yang et al., 2014; Speed and Balding, 2015; Speed et al., 2017) and genetic association tests based on PCA (Price et al., 2006), LMMs (Astle and Balding, 2009; Zhou and Stephens, 2012; Yang et al., 2014; Loh et al., 2015; Sul et al., 2018) and other models (Rakovski and Stram, 2009; Thornton and McPeek, 2010).

There are two versions of this standard kinship estimator, namely the mean-of-ratios (MOR) and ratio-of-means (ROM) version (Ochoa and Storey, 2021). Most approaches implement the MOR version. However, the ROM version has more favorable convergence properties relevant to our overall theoretical argument.

The ROM version of the standard kinship estimator, and its almost sure limit as the number of

loci m go to infinity (Ochoa and Storey, 2021), are given by

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

$$\xrightarrow[m\to\infty]{\text{a.s.}} \frac{\varphi_{jk} - \bar{\varphi}_j - \bar{\varphi}_k + \bar{\varphi}}{1 - \bar{\varphi}}, \tag{2}$$

where $\hat{\varphi}_{jk}^{\text{std-rom}}$ is the estimated kinship of individuals j and k, $\hat{p}_i = \frac{1}{2n}\sum_{j=1}^{n} x_{ij}$ is the standard ancestral allele frequency estimator, $\bar{\varphi}_j = \frac{1}{n}\sum_{k=1}^{n} \varphi_{jk}$ is the mean kinship of individual j with all others, and $\bar{\varphi} = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n} \varphi_{jk}$ is the overall mean kinship. This is a complex bias that varies for every pair of individuals, and which is on average a downward bias. (Note that the mean estimate per row and column, or $\frac{1}{n}\sum_{k=1}^{n} \hat{\varphi}_{jk}^{\text{std-rom}}$ for every j, is algebraically zero, regardless of the true value of the row mean kinship; the same is true for the MOR version below.)

The MOR version of the standard estimator, which again is the most common form of the estimator, is given by

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

This estimator does not have closed-form limit, but it is well approximated by Eq. (2) in practice, especially when loci with small minor allele frequencies are excluded prior to calculating this estimate.

Variants of this approach that weigh loci according to linkage disequilibrium (Speed et al., 2017; Wang et al., 2017) do not alter the bias calculated in Eq. (2), since the same bias is present in each individual locus (Ochoa and Storey, 2021). Our previous work also considered a more general form where the ancestral allele frequency estimator $\hat{p}_i = \frac{1}{2}\sum_{j=1}^{n} w_j x_{ij}$ is calculated with weights w_j per individual j (such that $\sum_{j=1}^{n} w_j = 1$), and found that these weights alter the values of the bias terms $\bar{\varphi}_j$ and $\bar{\varphi}$ to be weighted averages, but no choice of weights eliminates these biases (Ochoa and Storey, 2021). Such weighted $\hat{p}_i$ estimates encompass the best unbiased linear estimator (Astle and Balding, 2009; Thornton and McPeek, 2010), with weights corresponding to $\mathbf{w} = (\mathbf{1}^\intercal \mathbf{\Phi}^{-1}\mathbf{1})^{-1}\mathbf{1}^\intercal\mathbf{\Phi}^{-1}$.

2.2.2 Popkin kinship estimator

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

$$\begin{aligned} A_{jk} &= \frac{1}{m}\sum_{i=1}^{m} w_i((x_{ij}-1)(x_{ik}-1)-1), \\ \hat{\varphi}_{jk}^{\text{popkin}} &= 1 - \frac{A_{jk}}{\hat{A}_{\min}}, \end{aligned} \tag{4}$$

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 *ratio-of-means* (ROM) estimator, has $w_i = 1$ and satisfies

$$\hat{\varphi}_{jk}^{\text{popkin-ROM}} \xrightarrow[m\to\infty]{\text{a.s.}} \varphi_{jk},$$

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

2.2.3 Weir-Goudet kinship estimator

The Weir-Goudet (WG) kinship estimator and its limit are given by (Weir and Goudet, 2017; Ochoa and Storey, 2021)

$$\hat{\varphi}_{jk}^{\text{WG}} = 1 - \frac{A_{jk}}{\hat{A}_{\text{avg}}} \tag{5}$$

$$\xrightarrow[m\to\infty]{\text{a.s.}} \frac{\varphi_{jk} - \tilde{\varphi}}{1 - \tilde{\varphi}}, \tag{6}$$

where A_{jk} is as in Eq. (4), and

$$\begin{aligned} \hat{A}_{\text{avg}} &= \frac{2}{n(n-1)} \sum_{j=2}^{n} \sum_{k=1}^{j-1} A_{jk}, \\ \tilde{\varphi} &= \frac{2}{n(n-1)} \sum_{j=2}^{n} \sum_{k=1}^{j-1} \varphi_{jk}. \end{aligned} \tag{7}$$

Thus, the WG estimator resembles the popkin estimator in Eq. (4), except it replaces $\hat{A}_{\min}$ with $\hat{A}_{\text{avg}}$ and that results in a uniform downward bias given by $\tilde{\varphi}$, which is the mean kinship between all different individual pairs (it excludes the diagonal, or self-kinship values, compared to the $\bar{\varphi}$ that appears in the standard estimator). 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.

2.3 Simulations

2.3.1 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 $p_i(1-p_i)f_{S_u}$. 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`.

2.3.2 Trait simulation algorithm

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. 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 estimator $\hat{v}_i^T = \hat{p}_i(1-\hat{p}_i)/(1-\bar{\varphi}^T)$, 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 $\mathbf{p}$ 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.

2.4 Real genotype data processing

To compare the results from 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.

2.5 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.

2.6 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 demonstration of robustness to kinship bias in PCA and LMM genetic association studies

To quantify the effect of the various kinship matrix estimators, and their limiting biases, we simulated genotypes and a trait, and calculated association p-values for all kinship variants of the PCA and LMM methods. 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 results in a large difference in performance between PCA and LMM approaches (Yao and Ochoa, 2022).

The kinship estimates on this simulation, and the theoretical limits of these estimators, are shown in Fig. 1. The true kinship matrix (Fig. 1A) shows the family relatedness as high values concentrated near the diagonal and the ancestry-driven population structure as the broad patterns off-diagonal. This simulation illustrates the severity of the biases of these estimators, which were previously characterized theoretically (Ochoa and Storey, 2021). Only the Popkin estimator is unbiased (Fig. 1B), the rest presenting large negative biases which in turn result in abundant negative estimates. The Standard kinship estimator (the mean-of-ratios, or MOR, version; Fig. 1E), which is the most common estimator in these applications, and the closely related ratio-of-means (ROM) estimator (Fig. 1D), both have severe biases that are not only overall downwardly biased, but these biases also vary for every pair of individuals. The limit of the Standard ROM estimator (Fig. 1C), calculated from the true kinship matrix and its known functional form (see Methods), closely matches the previously mentioned estimates obtained from genotypes. Lastly, the Weir-Goudet estimator has a uniform downward bias (Fig. 1H-I).

We then performed both LMM and PCA association tests in order to determine if the kinship biases carry over to association biases. Surprisingly, we found that none of these kinship biases have discernible effects on association performance, as summarized by the Area Under the Precision-Recall Curve (AUC_{PR}; Fig. 2) and SRMSD_p (Fig. S1). The largest difference in performance is explained by the association model used (LMM vs PCA), as expected due to our use of a family simulation, where PCA is expected to perform less well than LMM. For PCA only, there are no clear differences between the performance of any of the kinship matrices. In contrast, the LMM results are relatively more sensitive to the use of noisy kinship estimates versus the (noiseless) limits of these estimates, a difference certainly increased by our deliberate use of a small number

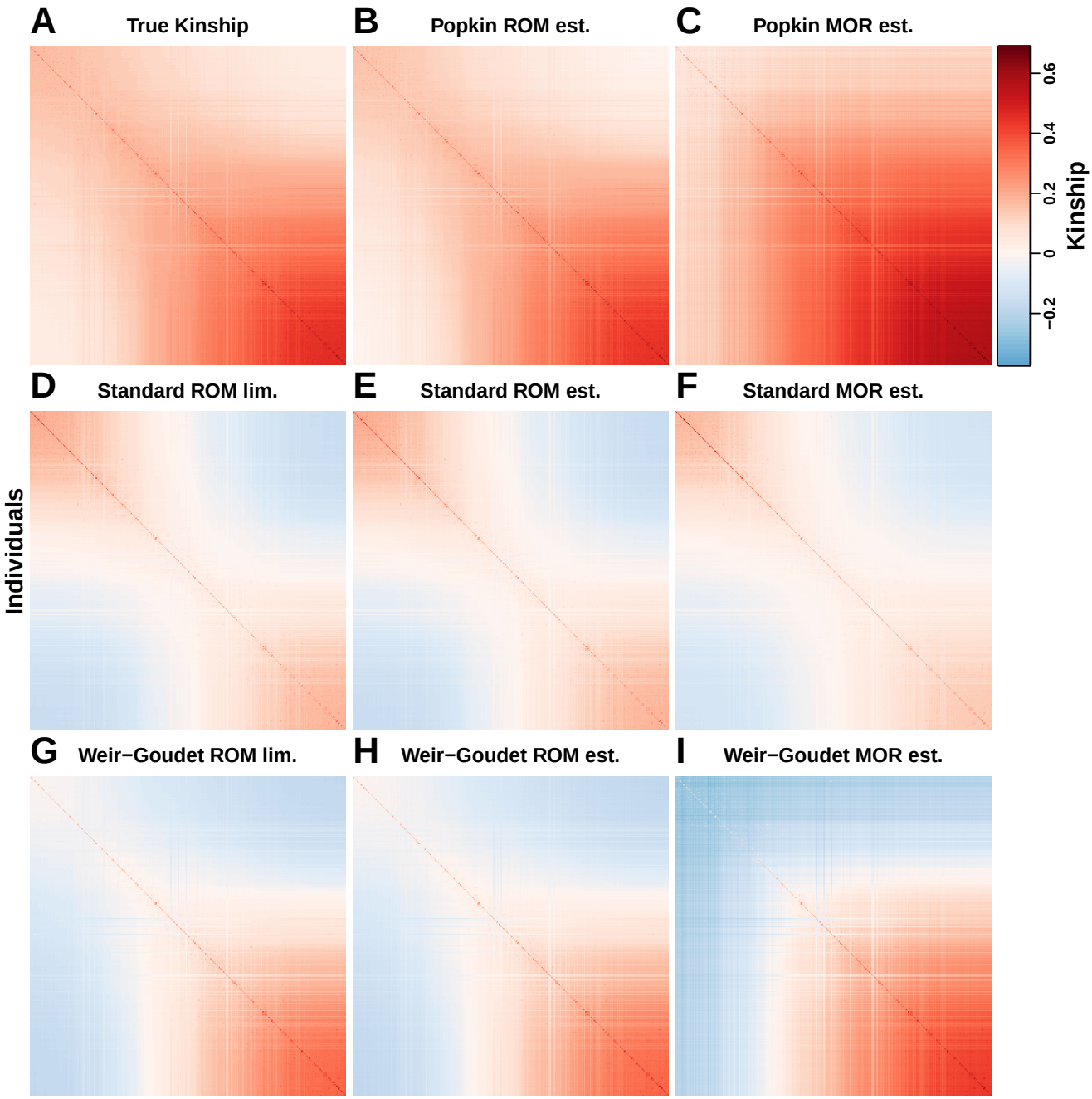

Figure 1: **Kinship estimates and their limits on the admixed family simulation.** Each panel represents a kinship matrix as a heatmap, with each of the $n = 1000$ individuals along both x and y axes, and the kinship value presented as color: positive estimates are in red, negative estimates in blue. The estimators considered are Popkin, Standard, and Weir-Goudet. Each estimator has its limit in the first column (the limit of Popkin is Truth). Standard has two variants: ROM (ratio of means) and MOR (mean of ratios). Standard MOR does not have a closed-form limit, but in practice matches that of Standard ROM.

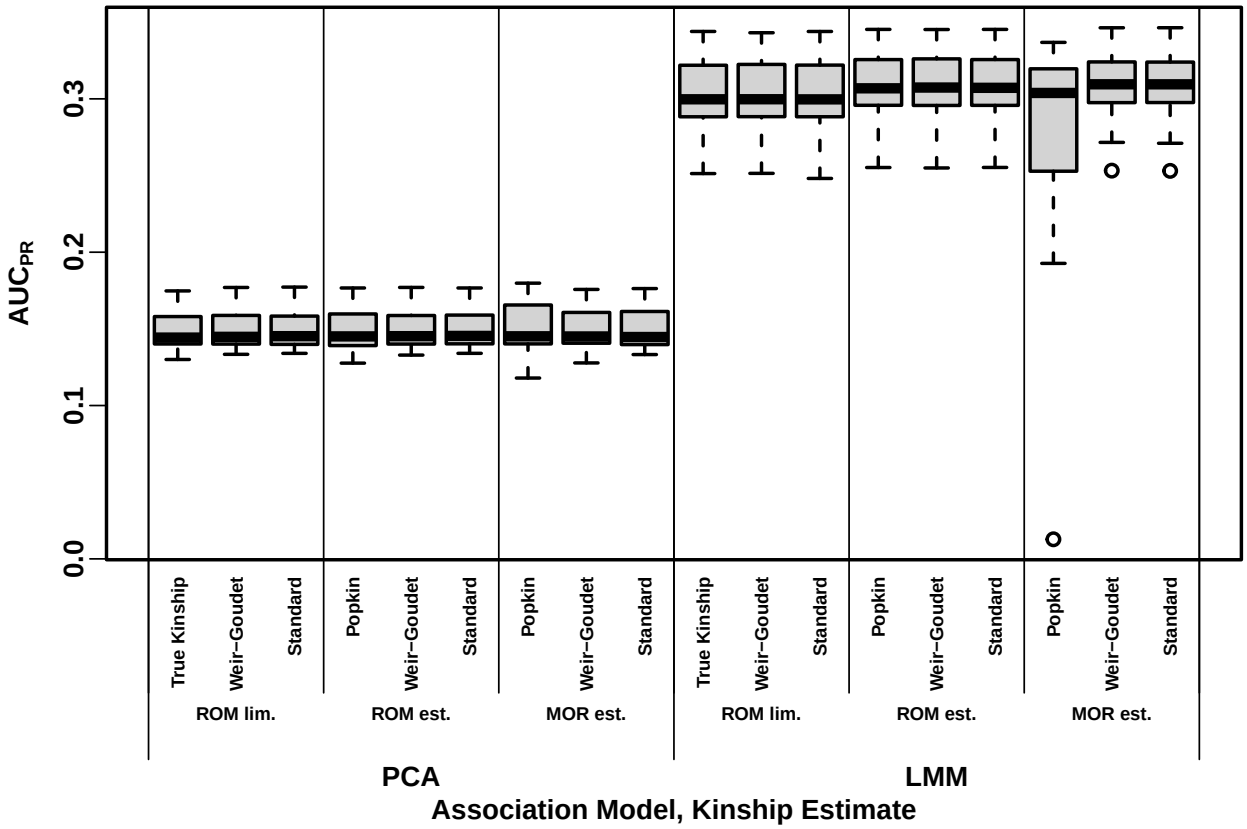

Figure 2: **Area Under the Precision-Recall Curves ($\mathbf{AUC_{PR}}$) for every combination of association model and kinship estimate on the admixed family simulation.** Results based on a single replicate of the random genotype matrix and trait vector. Approaches appear to cluster primarily by association model (LMM vs PCA) and whether a kinship estimate was used or not, and do not depend much at all on the form of the bias.

of simulated loci. For AUC_{PR}, among the kinship estimator limits, use of the true kinship matrix, the limit of the Weir-Goudet estimator, or the limit of the Standard ROM estimator all result in practically the same performance. Similarly, among the estimates, the popkin, Weir-Goudet, and Standard ROM estimates result in the same performance, whereas the Standard MOR estimates have a slightly different performance. For SRMSD_p, all estimates perform similarly except the popkin ROM estimates perform slightly different with a larger variance.

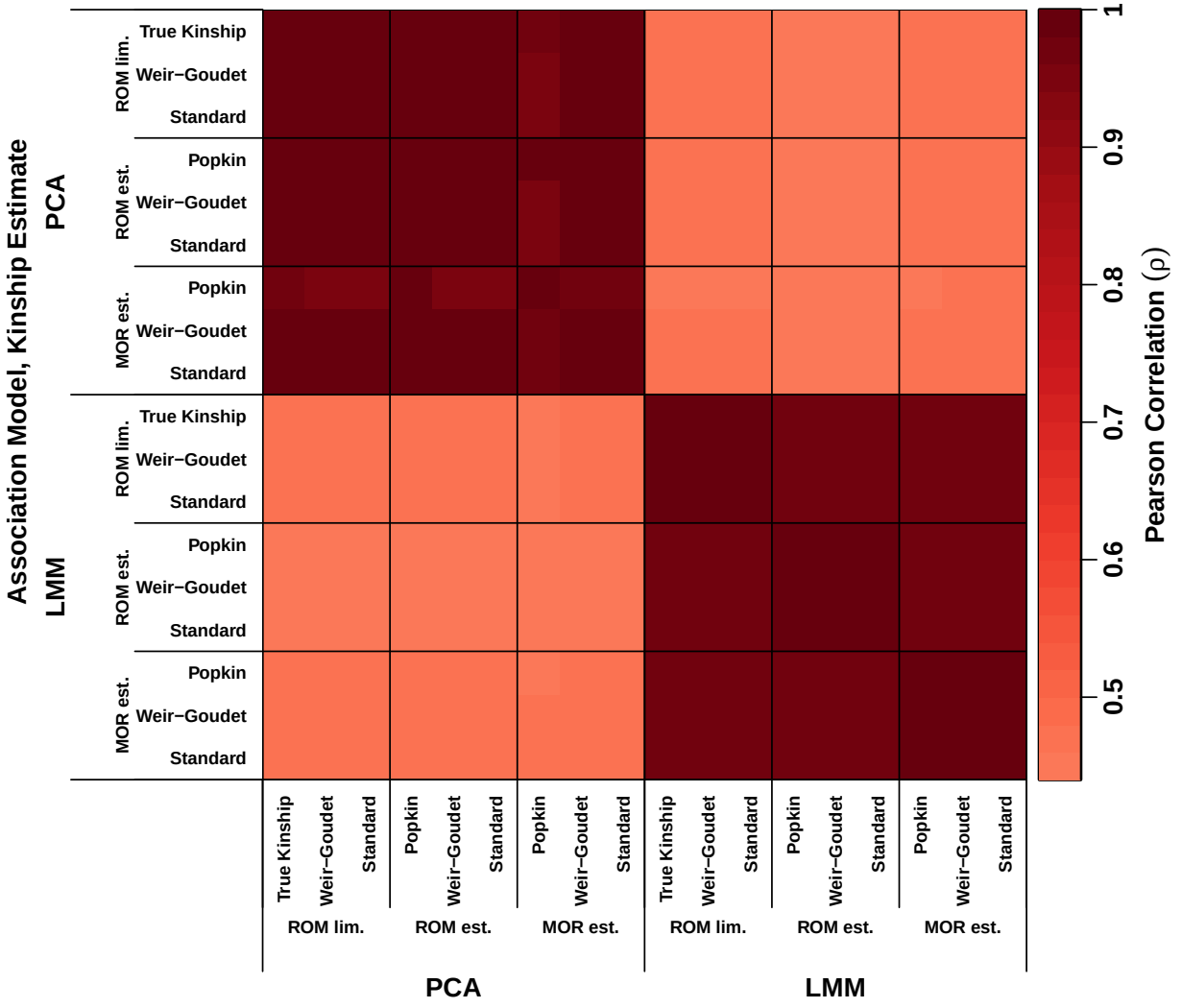

Figure 3: **Correlation between p-values of different association models and kinship estimates on the admixed family simulation.** The association p-value vector (one value per tested locus) produced by each combination of association model (LMM vs PCA) and kinship matrix (x and y axes) was used to compute Pearson correlations (color). Within same type of association model (LMM or PCA), methods are strongly correlated with correlation close to 1 regardless of kinship matrix used.

To better understand the degree of agreement between the association results of the various kinship matrices, we next measured the agreement between methods at the level of the individual association p-values, summarized using pearson correlation coefficients (ρ; Fig. 3). We found that all the kinship matrices yield highly correlated p-values when applied to the same association model (LMM vs PCA), and otherwise mirroring our previous observations based on AUC_{PR}. The minimum correlation among PCA methods was 0.86, and among LMMs it was 0.84. Within PCA, the cluster that excludes the true kinship matrix and the popkin estimate has nearly identical p-values ($\rho > 0.99$), and inclusion of popkin lowers the minimum correlation to $\rho > 0.93$. Among LMMs, the cluster that includes the true kinship and the Weir-Goudet and Standard ROM limits also results in practically identical p-values ($\rho \approx 1$). The LMM cluster that includes the Popkin, Weir-Goudet, and Standard ROM estimators also has $\rho \approx 1$, and separately the cluster with the Standard MOR estimators has $\rho \approx 1$, while the LMM cluster that includes all estimates has $\rho > 0.96$. Thus, several sets of kinship matrices with different biases result in identical association statistics, within both LMMs and PCA models, while differences are largely driven by the association model used and whether the kinship matrices were estimates or not.

3.2 Results using 1000 Genomes Project real genotypes

Now we replicate our previous findings using the real genotypes of 1000 Genomes. Kinship estimates are shown in Fig. 4. 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 attain higher values and 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, Weir-Goudet estimates are uniformly smaller than popkin's and attain large negative values (Fig. 4E-F).

When we perform PCA and LMM association tests using simulated traits and these kinship estimates, the results are similar to our simulation study: there are no differences between kinship estimators with different biases but of the same type (MOR or ROM) and association model (Fig. 5, Fig. S2), with the sole exception of popkin ROM. However, unlike the admixed family simulation results, here the MOR estimates greatly outperform ROM estimates (both using LMM) for both AUC_{PR} and SRMSD_p. We also measured correlations of one between p-values from the same association model and estimator type (MOR or ROM), regardless of bias type (Fig. S3).

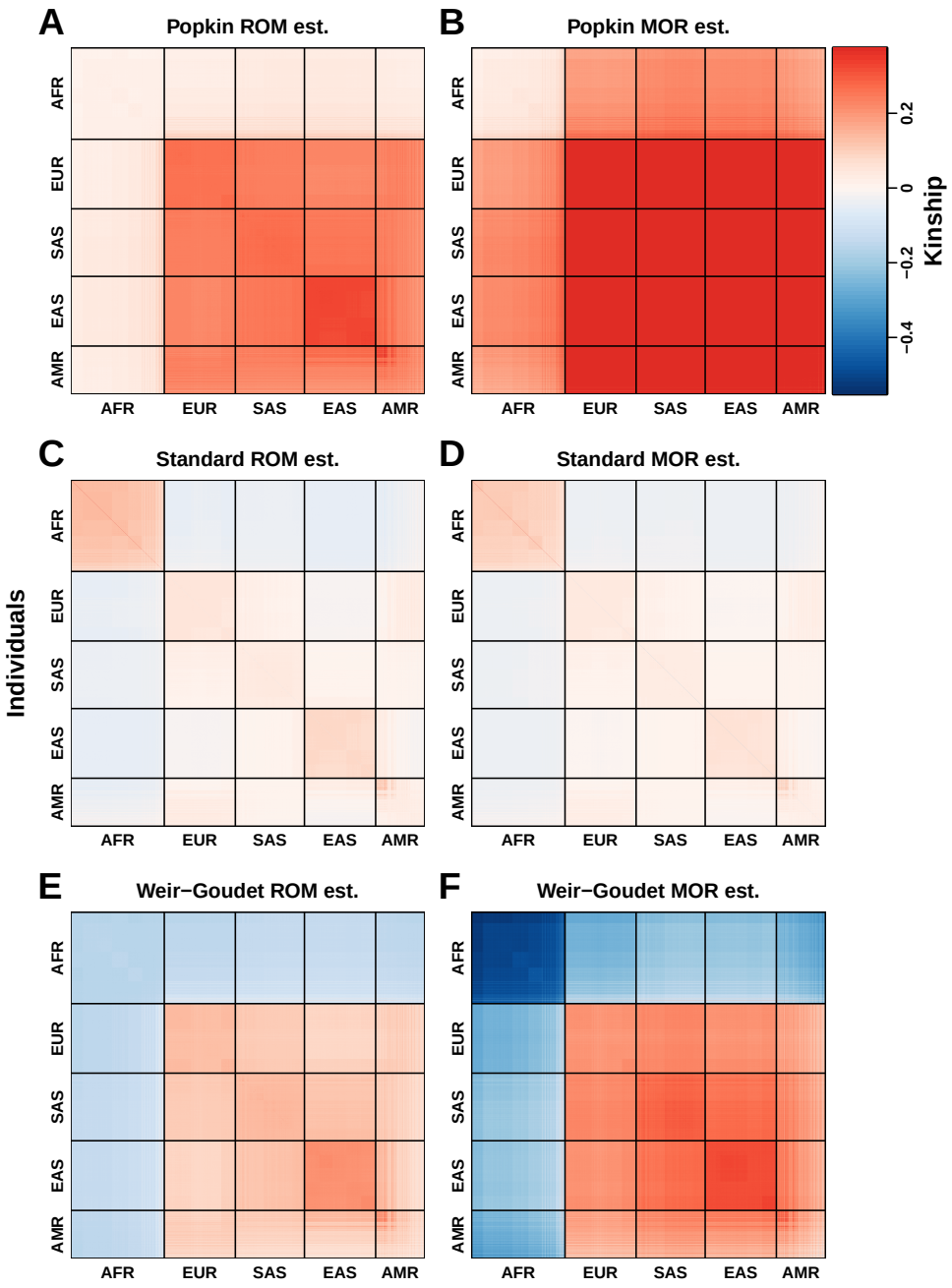

Figure 4: **Kinship estimates and their limits on 1000 Genomes.** Each panel represents a kinship matrix as a heatmap, with each individual along both x and y axes, and the kinship value presented as color: positive estimates are in red, negative estimates in blue. Superpopulation codes: AFR = African, EUR = European, SAS = South Asian, EAS = East Asian, AMR = Admixed Americans (Hispanics). Each estimator (Popkin, Standard, and Weir-Goudet) has two variants: 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.

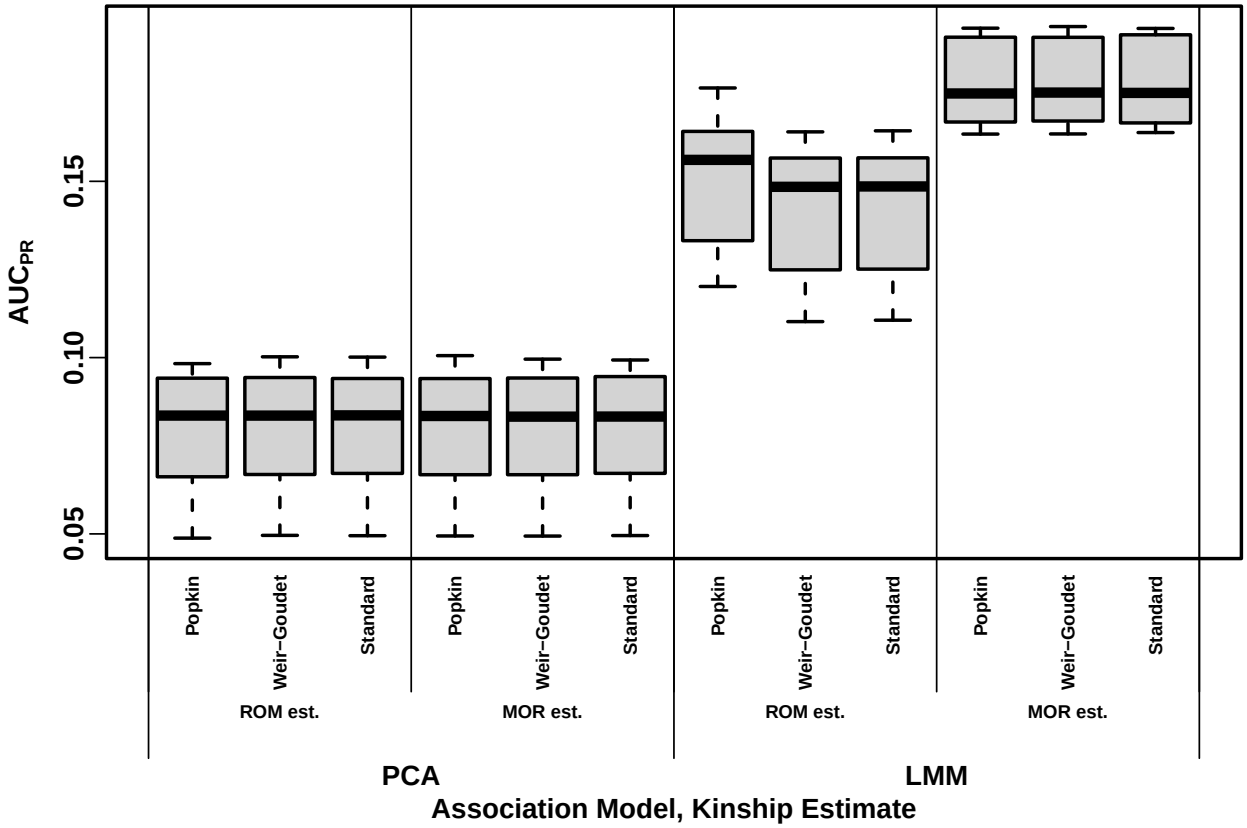

Figure 5: **Area Under the Precision-Recall Curves ($\mathbf{AUC_{PR}}$) for every combination of association model and kinship estimate on 1000 Genomes.** Results based on 10 simulated trait replicates (real genotype matrix is fixed). Approaches cluster primarily by association model (LMM vs PCA) and whether a kinship estimate was used or not, and do not depend much at all on the form of the bias.

3.3 Theoretical justification of empirical observations under standard kinship bias

Our empirical observations suggested that the biases of the Standard kinship estimator do not alter association statistics whatsoever; here we provide proof that this is indeed the case. To eliminate random estimation noise from the analysis (which our empirical evaluations suggest play a minor role), we shall focus exclusively on the limiting bias of the Standard (ROM) kinship estimator. Our constructive proof identifies the exact conditions that allows the kinship bias to be compensated for, namely that population structure is modeled as a (fixed or random effect) covariate and an intercept term is fit jointly. More specifically, we find that the coefficients of the intercept and the random effect or PCs (for LMM and PCA, respectively) adapt to the bias, and no other coefficients change. This is fortunate, as the intercept and random effects or PCs coefficients are nuisance parameters that are most often unreported, while the focal genetic association coefficient and its p-value are completely unchanged by this precise bias.

Our theoretical results only consider the true kinship matrix $\mathbf{\Phi}$ and the limit of the standard kinship estimator (see **Methods**, Eq. (2)), which can be stated in matrix notation as

$$\hat{\mathbf{\Phi}}^{\text{std-lim}} = \frac{1}{1-\bar{\varphi}} \left(\mathbf{\Phi} + \bar{\varphi}\mathbf{J} - \boldsymbol{\varphi}\mathbf{1}^\intercal - \mathbf{1}\boldsymbol{\varphi}^\intercal \right),$$

where $\mathbf{1}$ is a length-n column vector of ones, $\mathbf{J} = \mathbf{1}\mathbf{1}^\intercal$ is the $n \times n$ matrix full 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 overall mean kinship (scalar). The two kinship matrices are related more succinctly using the $n \times n$ centering matrix,

$$\mathbf{C} = \mathbf{I} - \frac{1}{n}\mathbf{J},$$

where $\mathbf{I}$ is the $n \times n$ identity matrix. The limit of the standard kinship estimator is given in terms of a transformation of the true kinship matrix by

$$\hat{\mathbf{\Phi}}^{\text{std-lim}} = \frac{1}{1-\bar{\varphi}}\mathbf{C}\mathbf{\Phi}\mathbf{C}. \tag{8}$$

The centering matrix has been well studied, and we review its properties here. For any length-n vector $\mathbf{v}$ we have

$$\mathbf{C}\mathbf{v} = \mathbf{v} - \mathbf{1}\bar{v},$$

where $\bar{v} = \frac{1}{n}\mathbf{1}^\intercal\mathbf{v}$ is the mean value of the elements of $\mathbf{v}$. Therefore, $\mathbf{v} = \mathbf{1}$ gets transformed to the zero vector, so it is an eigenvector with an eigenvalue of zero:

$$\mathbf{C}\mathbf{1} = \mathbf{0}.$$

Moreover, any vector $\mathbf{v}$ orthogonal to $\mathbf{1}$ has a zero mean element ($\bar{v} = 0$) by hypothesis and it is not altered by $\mathbf{C}$ ($\mathbf{C}\mathbf{v} = \mathbf{v}$). Therefore, the nullspace of $\mathbf{C}$ is spanned by $\mathbf{1}$.

This centering matrix provides the key insight as to why LMM and PCA approaches are robust to this specific kinship bias, namely that the bias in the random effects (for LMM) or eigenvectors (for PCA) of $\hat{\mathbf{\Phi}}^{\text{std-lim}}$ results in removing the mean values of these covariates only, so fitting the intercept term $\mathbf{1}\alpha$ compensates exactly for this bias. In the remaining sections we detail this argument, where we construct equivalent solutions under the true and biased kinship matrices.

3.3.1 Kinship matrix square root

Here we shall consider decompositions of positive semidefinite matrices of the form $\mathbf{\Sigma} = \mathbf{B}\mathbf{B}^\intercal$, which are guaranteed to exist. We denote such a $\mathbf{B}$ as a square root of $\mathbf{\Sigma}$, or in short $\mathbf{B} = \mathbf{\Sigma}^{\frac{1}{2}}$. These square roots of $\mathbf{\Sigma}$ are not unique, which is not a problem for our following argument; any such square root will work. (Note that there are alternate definitions of matrix square roots, such as $\mathbf{\Sigma} = \mathbf{B}\mathbf{B}$, but due to its connection to covariance, the definition $\mathbf{\Sigma} = \mathbf{B}\mathbf{B}^\intercal$ we adopted is most natural for this work and the notation $\mathbf{B} = \mathbf{\Sigma}^{\frac{1}{2}}$ simplifies our arguments.)

Given a square root of the true kinship matrix, $\mathbf{\Phi}^{\frac{1}{2}}$, we can construct a square root of the limit of the standard kinship estimator as

$$\left(\hat{\mathbf{\Phi}}^{\text{std-lim}}\right)^{\frac{1}{2}} = \frac{1}{\sqrt{1-\bar{\varphi}}}\mathbf{C}\mathbf{\Phi}^{\frac{1}{2}}. \tag{9}$$

This matrix square root can be verified to satisfy $\left(\hat{\mathbf{\Phi}}^{\text{std-lim}}\right)^{\frac{1}{2}}\left(\left(\hat{\mathbf{\Phi}}^{\text{std-lim}}\right)^{\frac{1}{2}}\right)^\intercal = \hat{\mathbf{\Phi}}^{\text{std-lim}}$ as given in Eq. (8).

3.3.2 The LMM genetic association model

In genetic association we are given data for n individuals, namely a length-n vector of trait values $\mathbf{y}$, which correspond to a quantitative trait, and a length-n vector $\mathbf{x}_i$ of genotypes at each locus i. These genotypes are encoded as dosages for a given risk allele that varies for each locus i, so it takes on the values of 0, 1, or 2 for diploid individuals. The goal is to determine if there is a significant association between the trait and the genotype vectors. Most genetic association models are formulated as, or are equivalent to, regression models.

The LMM regression model is given by

$$\mathbf{y} = \alpha\mathbf{1} + \beta\mathbf{x}_i + \mathbf{s} + \boldsymbol{\epsilon}, \tag{10}$$

where α is the intercept coefficient, β is the genetic effect coefficient, $\boldsymbol{\epsilon}$ are random independent residuals ($\boldsymbol{\epsilon} \sim \text{Normal}(\mathbf{0}, \sigma_\epsilon^2\mathbf{I})$ for some σ_ϵ^2 fit to the data), and the random effect satisfies (Sul et al., 2018)

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

where σ^2 is also fit to the data. The dependence of the model on $\mathbf{\Phi}$ is clearer by writing it equivalently as

$$\mathbf{y} = \alpha\mathbf{1} + \beta\mathbf{x}_i + \sigma\mathbf{\Phi}^{\frac{1}{2}}\mathbf{r} + \boldsymbol{\epsilon}, \tag{11}$$

where $\mathbf{r} \sim \text{Normal}\,(\mathbf{0}, \mathbf{I})$. The equivalence of Eq. (10) and Eq. (11) follows since $\mathbf{r}$ being Multivariate Normal implies that the affine transformation $\mathbf{s} = \sigma\mathbf{\Phi}^{\frac{1}{2}}\mathbf{r}$ is also Multivariate Normal, with matching mean and covariance of the desired $\mathbf{s}$, namely

$$\begin{aligned}\text{E}[\mathbf{s}] &= \sigma\mathbf{\Phi}^{\frac{1}{2}}\,\text{E}[\mathbf{r}] = \mathbf{0},\\ \text{Cov}(\mathbf{s}) &= \left(\sigma\mathbf{\Phi}^{\frac{1}{2}}\right)\text{Cov}(\mathbf{r})\left(\sigma\mathbf{\Phi}^{\frac{1}{2}}\right)^{\mathsf{T}} = \sigma^2\mathbf{\Phi}.\end{aligned}$$

Note that the equivalence holds for every square root of $\mathbf{\Phi}$ (all $\mathbf{s} = \sigma\mathbf{\Phi}^{\frac{1}{2}}\mathbf{r}$ are equal in distribution), since the only requirement for equivalence, $\left(\mathbf{\Phi}^{\frac{1}{2}}\right)\left(\mathbf{\Phi}^{\frac{1}{2}}\right)^{\mathsf{T}} = \mathbf{\Phi}$, is satisfied by hypothesis.

3.3.3 Equivalent LMM fit under standard kinship bias

In the LMM of Eq. (11), $\mathbf{y}$, $\mathbf{x}_i$, and $\mathbf{\Phi}$ are given, while the coefficients α, β, σ, and the random effects $\mathbf{r}$ and $\boldsymbol{\epsilon}$ are fit to this data. This data is typically fit using maximum likelihood (ML) or restricted maximum likelihood (REML) (Kang et al., 2008); our argument covers any likelihood-based approach, since we will establish a parameter map between both models that results in equal likelihoods for every parameter set in the map.

We shall consider two model fits, one where $\mathbf{\Phi}$ is given, while in the other we provide $\mathbf{\Phi}' = \hat{\mathbf{\Phi}}^{\text{std-lim}}$ instead. The first fit will result in the unprimed variables below, while we distinguish the second fit using primed variables, namely:

$$\begin{aligned}\mathbf{y} &= \alpha\mathbf{1} + \beta\mathbf{x}_i + \sigma\mathbf{\Phi}^{\frac{1}{2}}\mathbf{r} + \boldsymbol{\epsilon}\\ &= \alpha'\mathbf{1} + \beta'\mathbf{x}_i + \sigma'\left(\mathbf{\Phi}'\right)^{\frac{1}{2}}\mathbf{r}' + \boldsymbol{\epsilon}'.\end{aligned}$$

Now we shall construct coefficients for the second model that result in the same fit to the data, including the same likelihood, as the first model. We achieve this by first setting $\beta' = \beta$, $\boldsymbol{\epsilon}' = \boldsymbol{\epsilon}$, and $\mathbf{r}' = \mathbf{r}$. Note that the previous equal random effects (including residuals) immediately results in the same likelihood for both models. The only parameters left to construct are α' and σ', which must satisfy

$$\alpha\mathbf{1} + \sigma\mathbf{\Phi}^{\frac{1}{2}}\mathbf{r} = \alpha'\mathbf{1} + \sigma'\left(\mathbf{\Phi}'\right)^{\frac{1}{2}}\mathbf{r}.$$

Next we substitute $\mathbf{\Phi}' = \hat{\mathbf{\Phi}}^{\text{std-lim}}$ using the matrix square root determined in terms of $\mathbf{\Phi}$ in Eq. (9), which results in

$$\alpha\mathbf{1} + \sigma\mathbf{\Phi}^{\frac{1}{2}}\mathbf{r} = \alpha'\mathbf{1} + \sigma'\frac{1}{\sqrt{1-\bar{\varphi}}}\mathbf{C}\mathbf{\Phi}^{\frac{1}{2}}\mathbf{r}.$$

The unknowns are solved for by left-multiplying, in turns, by $\mathbf{C}$ (which makes terms with $\mathbf{1}$ vanish) and by $\mathbf{1}^{\mathsf{T}}$ (which makes the term with $\mathbf{C}$ vanish). In the first case, left-multiplying by $\mathbf{C}$ results in

$$\sigma\mathbf{C}\mathbf{\Phi}^{\frac{1}{2}}\mathbf{r} = \sigma'\frac{1}{\sqrt{1-\bar{\varphi}}}\mathbf{C}\mathbf{\Phi}^{\frac{1}{2}}\mathbf{r},$$

so the only value of the scalar σ' that satisfies this equation is

$$\sigma' = \sigma\sqrt{1-\bar{\varphi}}.$$

In the second case, left-multiplying by $\mathbf{1}^\intercal$, while noting that $\mathbf{1}^\intercal\mathbf{1} = n$, and solving for α' results in

$$\alpha' = \alpha + \sigma\frac{1}{n}\mathbf{1}^\intercal\boldsymbol{\Phi}^{\frac{1}{2}}\mathbf{r}.$$

Note that the last equation can also be written as

$$\alpha' = \alpha + \eta, \qquad \eta \sim \text{Normal}(0, \sigma^2\bar{\varphi}),$$

since $\eta = \sigma\frac{1}{n}\mathbf{1}^\intercal\boldsymbol{\Phi}^{\frac{1}{2}}\mathbf{r}$ is a scalar random Normal variable (because it is a sum of marginals of a Multivariate Normal, each of which are Normal) with mean zero (because $\mathbf{r}$ had mean zero) and the stated variance (since $\text{Var}(\sigma\frac{1}{n}\mathbf{1}^\intercal\boldsymbol{\Phi}^{\frac{1}{2}}\mathbf{r}) = (\sigma\frac{1}{n}\mathbf{1}^\intercal\boldsymbol{\Phi}^{\frac{1}{2}})\,\text{Cov}(\mathbf{r})(\sigma\frac{1}{n}\mathbf{1}^\intercal\boldsymbol{\Phi}^{\frac{1}{2}})^\intercal$, $\text{Cov}(\mathbf{r}) = \mathbf{I}$, and $\frac{1}{n^2}\mathbf{1}^\intercal\boldsymbol{\Phi}\mathbf{1} = \bar{\varphi}$).

We just determined that every solution in terms of the true kinship matrix (including every combination of fixed coefficients and random effects) has a corresponding solution in terms of the limit of the standard kinship estimator, which has equal likelihood and equal fit to the data. This includes the optimal solution, whether determined by the ML or REML criteria. In both cases, the coefficient for the genetic effect is identical ($\beta' = \beta$ above), and because the fit to the data is also equal (in terms of likelihood and/or residuals), the association p-value is also equal (whether determined from the likelihood or from residuals). Thus, while two coefficients (the intercept α and the random effect variance scale σ^2) vary depending on whether the true or the limit of the biased standard kinship estimator are used, these are both nuisance parameters as far as the association test is concerned. The focal genetic effect coefficient and significance statistic are both invariant under this particular kinship bias compared to using the true kinship matrix.

3.3.4 The PCA genetic association model

The argument we just presented for LMM equivalence can be made with small changes for the PCA regression model, since these two models are so similar. Here we shall state the PCA model and elaborate on its strong connection to the LMM model, which has been presented before in similar forms (Astle and Balding, 2009; Hoffman, 2013).

The PCA regression model is

$$\mathbf{y} = \alpha\mathbf{1} + \beta\mathbf{x}_i + \mathbf{U}_d\boldsymbol{\gamma}_d + \boldsymbol{\epsilon}, \tag{12}$$

where d is the number of principal components, $\mathbf{U}_d$ is the $n \times d$ matrix of top eigenvectors (often refered to as "principal components" in genetics), and $\boldsymbol{\gamma}_d$ is a length-d vector of coefficients for each eigenvector. Note that the only difference from the LMM model (Eq. (11)) is the replacement of $\sigma\boldsymbol{\Phi}^{\frac{1}{2}}\mathbf{r}$ with $\mathbf{U}_d\boldsymbol{\gamma}_d$ here.

Before proceeding with our proof for invariance under the PCA model, to enhance our intuition of these two models, we present the relationship between eigendecomposition and matrix square roots, which helps us connect the PCA model firmly to the LMM. Denote the eigendecomposition of the true kinship matrix as

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

where $\mathbf{U}$ is the complete $n \times n$ matrix of eigenvectors, and $\mathbf{\Lambda}$ is the $n \times n$ diagonal matrix of eigenvalues. Therefore, one square root of $\mathbf{\Phi}$ is given by

$$\mathbf{\Phi}^{\frac{1}{2}} = \mathbf{U}\mathbf{\Lambda}^{\frac{1}{2}},$$

where $\mathbf{\Lambda}^{\frac{1}{2}}$ contains the square roots of each eigenvalue along the diagonal. This equation reveals that the LMM model in Eq. (11) can be written in terms of the eigendecomposition, and thus resemble the PCA model even more closely, since

$$\sigma\mathbf{\Phi}^{\frac{1}{2}}\mathbf{r} = \sigma\mathbf{U}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{r} = \mathbf{U}\boldsymbol{\gamma},$$

so that the length-n vector $\boldsymbol{\gamma}$ of coefficients for all the n eigenvectors is given by $\boldsymbol{\gamma} = \sigma\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{r}$. Thus, the PCA model attempts to fit coefficients only for the top d eigenvectors, whereas the LMM model effectively fits all of these coefficients by constraining them to a distribution.

3.3.5 Approximately equivalent PCA fit under standard kinship bias

We shall again consider two alternate model fits, here based on the PCA model of Eq. (12), one where the eigenvector matrix $\mathbf{U}_d$ corresponds to the true kinship matrix, and in the other $\mathbf{U}'_d$ corresponds to the biased limit of the standard kinship estimator. They key approximation is that

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

which is not strictly equal (since $\mathbf{CU}$ is not orthogonal, as eigenvectors must be), but we have found it to be a good approximation in practice.

The two model fits we are considering are

$$\begin{aligned}\mathbf{y} &= \alpha\mathbf{1} + \beta\mathbf{x}_i + \mathbf{U}_d\boldsymbol{\gamma}_d + \boldsymbol{\epsilon} \\ &= \alpha'\mathbf{1} + \beta'\mathbf{x}_i + \mathbf{U}'_d\boldsymbol{\gamma}'_d + \boldsymbol{\epsilon}',\end{aligned}$$

and we again assume that the focal parameter $\beta' = \beta$ and the residuals $\boldsymbol{\epsilon}' = \boldsymbol{\epsilon}$ are equal. Eliminating the resulting shared terms, and replacing $\mathbf{U}'_d = \mathbf{C}\mathbf{U}_d$ (assuming that our approximation holds exactly) results in the remaining coefficients having to satisfy

$$\alpha\mathbf{1} + \mathbf{U}_d\boldsymbol{\gamma}_d = \alpha'\mathbf{1} + \mathbf{C}\mathbf{U}_d\boldsymbol{\gamma}'_d.$$

We solve for the missing coefficients by left-multiplying by $\mathbf{C}$ and $\mathbf{1}^\intercal$ as before, which here ultimately results in

$$\boldsymbol{\gamma}'_d = \boldsymbol{\gamma}_d, \qquad \alpha' = \alpha + \frac{1}{n}\mathbf{1}^\intercal \mathbf{U}_d \boldsymbol{\gamma}_d.$$

Thus, as for LMM, here the nuisance intercept coefficient compensates for the bias in the eigenvectors.

The PCA regression is an ordinary multiple linear regression, which is fit by minimizing the sum of square residuals. Since the residuals were equal in both models, then the optimal solution in one model maps to the optimal solution in the other model as well. The p-value of the genetic effect is usually calculated using a chi-squared test or an F-test, both of which depend only on the residuals and the degrees of freedom, all of which are invariant under the solution parameter map we constructed. Therefore, both the focal genetic effect coefficient β and its p-value are invariant under this particular kinship bias compared to using the true kinship matrix. However, the result for PCA relies on an approximation, whereas for LMM it was exact.

3.4 Proof that association based on Weir-Goudet and true kinship are equivalent

The limiting bias of the Weir-Goudet (WG) estimator (see Methods, Eq. (6)) is given in terms of the true kinship matrix by

$$\hat{\boldsymbol{\Phi}}^{\text{WG-lim}} = \frac{1}{1-\tilde{\varphi}}(\boldsymbol{\Phi} - \tilde{\varphi}\mathbf{J}). \tag{13}$$

Our strategy for proving the association equivalence under both kinship matrices is to construct a random effect structured as the true kinship matrix from a random effect structured as WG. Although we were unable to calculate a matrix square root for the WG limit in terms of the true kinship matrix, we found that this alternate derivation for LMMs proves the desired result and leads to strikingly analogous equations for the variance scale σ^2 and intercept α between both models, where the only difference ultimately is the replacement of the mean kinship $\bar{\varphi}$ in the standard estimator results with $\tilde{\varphi}$ for WG. We were unable to prove the PCA case with this strategy.

3.4.1 Constructing random effects with true kinship structure from Weir-Goudet structure

Solving for the true kinship matrix in Eq. (13), we obtain

$$\boldsymbol{\Phi} = (1-\tilde{\varphi})\hat{\boldsymbol{\Phi}}^{\text{WG-lim}} + \tilde{\varphi}\mathbf{J}. \tag{14}$$

Let the random effect from the LMM using the WG limit be

$$\mathbf{s}^{\text{WG}} \sim \text{Normal}\left(\mathbf{0}, \sigma^2_{\text{WG}}\hat{\boldsymbol{\Phi}}^{\text{WG-lim}}\right),$$

where σ^2_{WG} is a parameter fit to the data. The Normal distribution above requires that $\hat{\mathbf{\Phi}}^{\text{WG-lim}}$ be positive definite, a fact that is proven for arbitrary true kinship matrices (except for one degenerate case) in Appendix C. We desire to construct a random effect such that $\mathbf{s} \sim \text{Normal}\left(\mathbf{0}, \sigma^2\mathbf{\Phi}\right)$, where again σ^2 has been fit to the data. Eq. (14) suggest a relationship of the form

$$\mathbf{s} = \mathbf{s}^{\text{WG}} + \eta\mathbf{1},$$

where η is a random scalar drawn independently from a Normal distribution,

$$\eta \sim \text{Normal}(0, \sigma^2_\eta),$$

where σ^2_η is yet to be determined. This constructed $\mathbf{s}$ is indeed Multivariate Normal since $\eta\mathbf{1}$ is a (degenerate) Multivariate Normal and the sum of two Multivariate Normal variables is itself a Multivariate Normal variable. All that is left is to match the mean of covariance of the desired $\mathbf{s}$, namely

$$\begin{aligned}
\text{E}[\mathbf{s}] &= \text{E}\left[\mathbf{s}^{\text{WG}}\right] + \text{E}[\eta]\mathbf{1} = \mathbf{0}, \\
\text{Cov}(\mathbf{s}) &= \text{Cov}\left(\mathbf{s}^{\text{WG}}\right) + \text{Var}(\eta)\mathbf{J} \\
&= \sigma^2_{\text{WG}}\hat{\mathbf{\Phi}}^{\text{WG-lim}} + \sigma^2_\eta\mathbf{J} = \sigma^2\mathbf{\Phi},
\end{aligned}$$

which is achieved with $\sigma^2 = \sigma^2_{\text{WG}}/(1-\tilde{\varphi})$ and $\sigma^2_\eta = \tilde{\varphi}\sigma^2$.

3.4.2 Equivalent LMM fit under Weir-Goudet kinship bias

Similarly to the case for the standard kinship estimator, we consider two equivalent LMM fits, one with $\mathbf{s}$ from the true kinship matrix, the other with $\mathbf{s}' = \mathbf{s}^{\text{WG}}$ from the WG limit:

$$\mathbf{y} = \alpha\mathbf{1} + \beta\mathbf{x}_i + \mathbf{s} + \boldsymbol{\epsilon} = \alpha'\mathbf{1} + \beta'\mathbf{x}_i + \mathbf{s}' + \boldsymbol{\epsilon}'.$$

As before, our empirical results motivate that $\beta = \beta', \boldsymbol{\epsilon} = \boldsymbol{\epsilon}'$, and only α' is left, which must satisfy

$$\alpha\mathbf{1} + \mathbf{s} = \alpha'\mathbf{1} + \mathbf{s}'.$$

We now substitute our constructed $\mathbf{s} = \mathbf{s}' + \eta\mathbf{1}$, where η is as determined in the previous subsection, which after simplifying yields

$$\alpha\mathbf{1} + \eta\mathbf{1} = \alpha'\mathbf{1}.$$

Note $\mathbf{s}'$ drops out of the equation, which follows from having previously selected its variance appropriately to match that of the model for the true kinship matrix, namely

$$\sigma' = \sigma\sqrt{1-\tilde{\varphi}},$$

where $\sigma' = \sigma_{\text{WG}}$, which agrees with the solution we found earlier for the standard kinship estimator, expect replacing $\bar{\varphi}$ with $\tilde{\varphi}$. Lastly, the relation between intercepts is

$$\alpha' = \alpha + \eta,$$

which again resembles the solution for the standard model, where both draw a random variable with mean zero, except the earlier variance of $\sigma^2\bar{\varphi}$ is replaced with $\sigma^2\tilde{\varphi}$ here. As before, these results imply that the association statistics are invariant to the choice between the true and WG limit kinship matrices.

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. S1) and popkin ROM (1000 Genomes, Fig. S2). 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), 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.

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. 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.

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.

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.

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.

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.

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.

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. (4) 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

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

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:

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

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

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

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 always positive definite.

Starting from the fact that a true kinship matrix is positive definite by definition (it is a covariance matrix), we shall prove that the limit of the Weir-Goudet estimator is also positive definite, with the exception of one degenerate case. Recall from Eq. (13) that these two matrices are related by $\hat{\mathbf{\Phi}}^{\text{WG-lim}} = \frac{1}{1-\tilde{\varphi}}(\mathbf{\Phi} - \tilde{\varphi}\boldsymbol{J})$. 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 $\varphi_{jk} < 1$. Below we will 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 $\hat{\mathbf{\Phi}}^{\text{WG-lim}}$ 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 $\hat{\mathbf{\Phi}}^{\text{WG-lim}}$ 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 $\mathbf{v}^\intercal\hat{\mathbf{\Phi}}^{\text{WG-lim}}\mathbf{v} > 0$ which proves that $\hat{\mathbf{\Phi}}^{\text{WG-lim}}$ 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

$$\mathbf{v}^\intercal\hat{\mathbf{\Phi}}^{\text{WG-lim}}\mathbf{v} = \frac{1}{1-\tilde{\varphi}}\mathbf{v}^\intercal\mathbf{\Phi}\mathbf{v} > 0,$$

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

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

$$\mathbf{v}^\intercal \hat{\mathbf{\Phi}}^{\text{WG-lim}} \mathbf{v} = \frac{\beta^2}{1-\tilde{\varphi}}(\mathbf{1}^\intercal \mathbf{\Phi} \mathbf{1} - \tilde{\varphi} n^2) = \frac{\beta^2 n^2}{1-\tilde{\varphi}}(\bar{\varphi} - \tilde{\varphi}),$$

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 $\beta^2, n^2, 1 - \tilde{\varphi} > 0$, 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 true kinship matrix has the degenerate form $\mathbf{\Phi} = \bar{\varphi}\mathbf{J}$, which is a singular matrix not expected to be observed in any real scenarios (in this case $\hat{\mathbf{\Phi}}^{\text{WG-lim}}$ 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

$$\bar{\varphi} = \frac{\tilde{\varphi}(n-1) + \bar{d}}{n}. \tag{15}$$

Applying $\bar{\varphi} \leq \bar{d}$ to Eq. (15) and simplifying yields $\tilde{\varphi} \leq \bar{d}$. Lastly, since $\bar{\varphi} - \tilde{\varphi} = (\bar{d} - \tilde{\varphi})/n$ (from rearranging Eq. (15)), 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

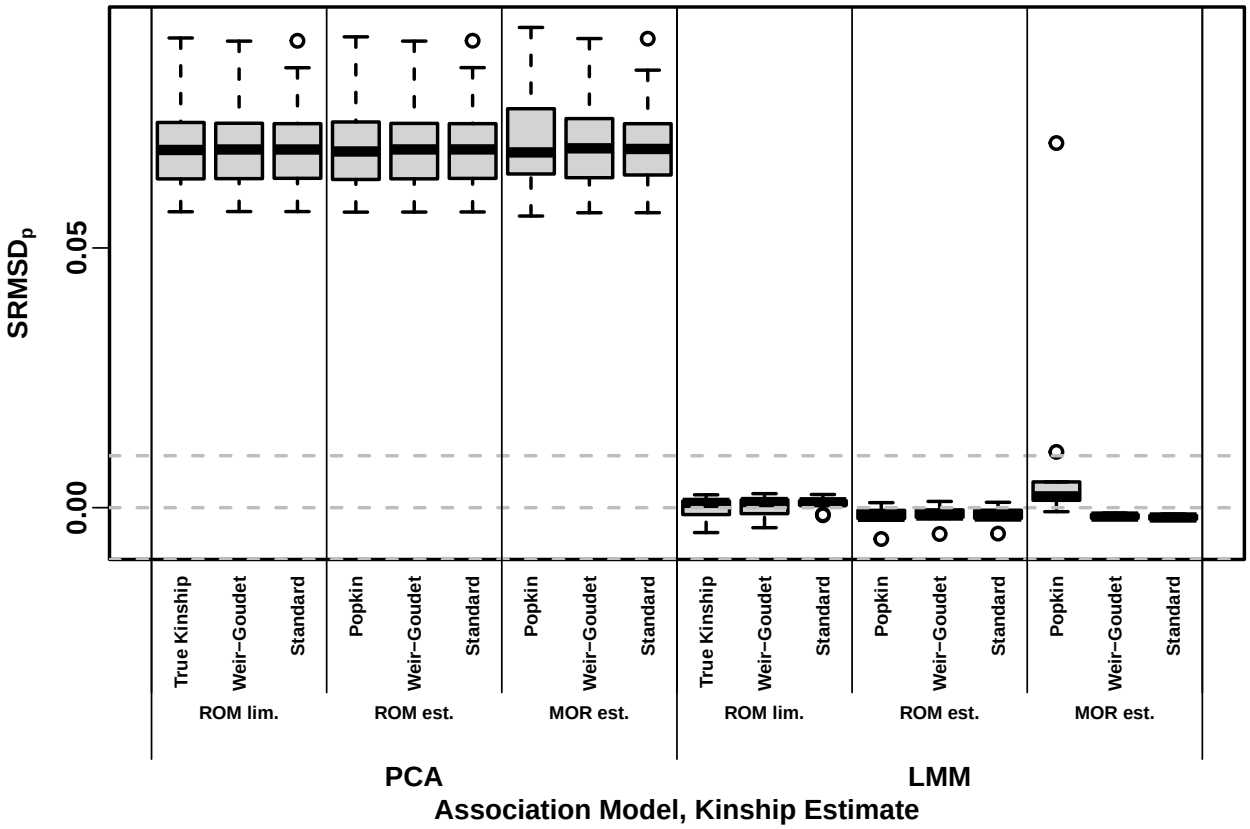

Figure S1: **Signed Root Mean Square Deviation of null p-values (SRMSD$_p$) for every combination of association model and kinship estimate on the admixed family simulation.**

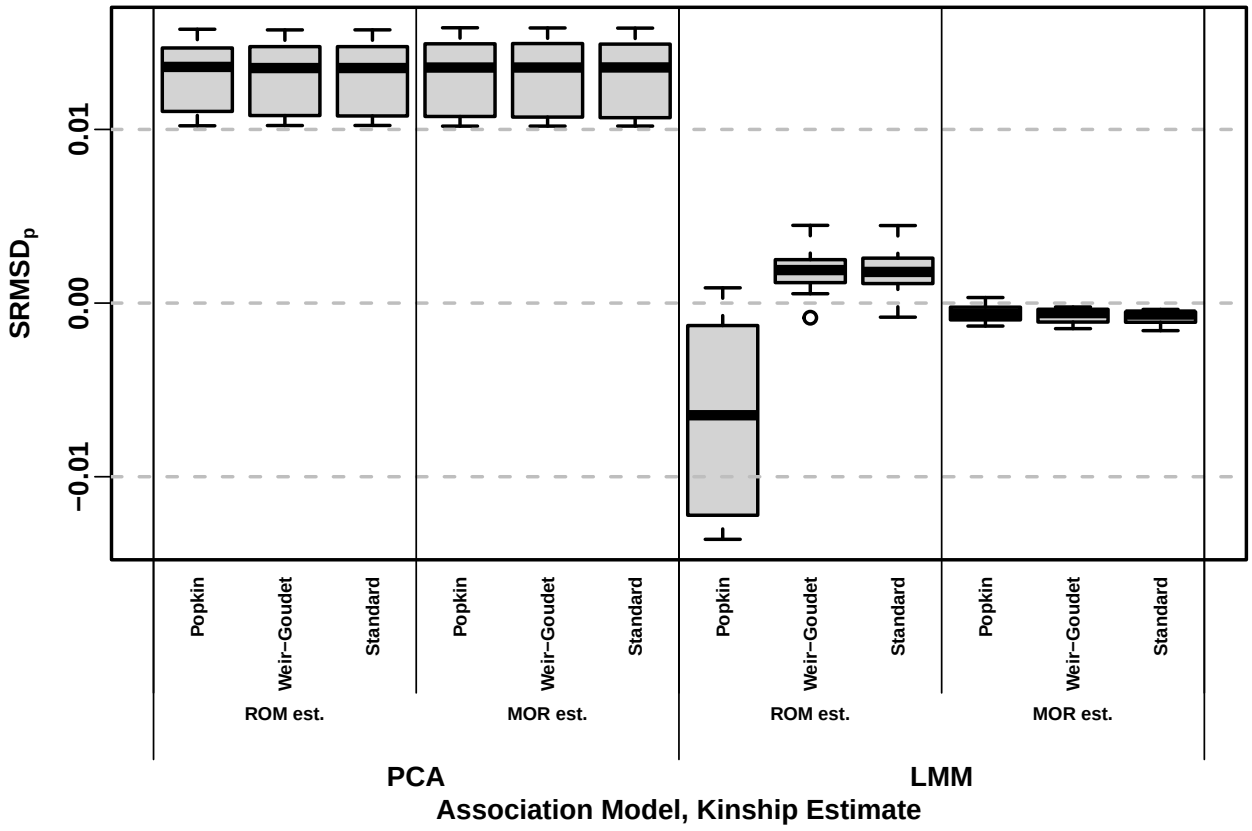

Figure S2: **Signed Root Mean Square Deviation of null p-values ($SRMSD_p$) for every combination of association model and kinship estimate on 1000 Genomes.**

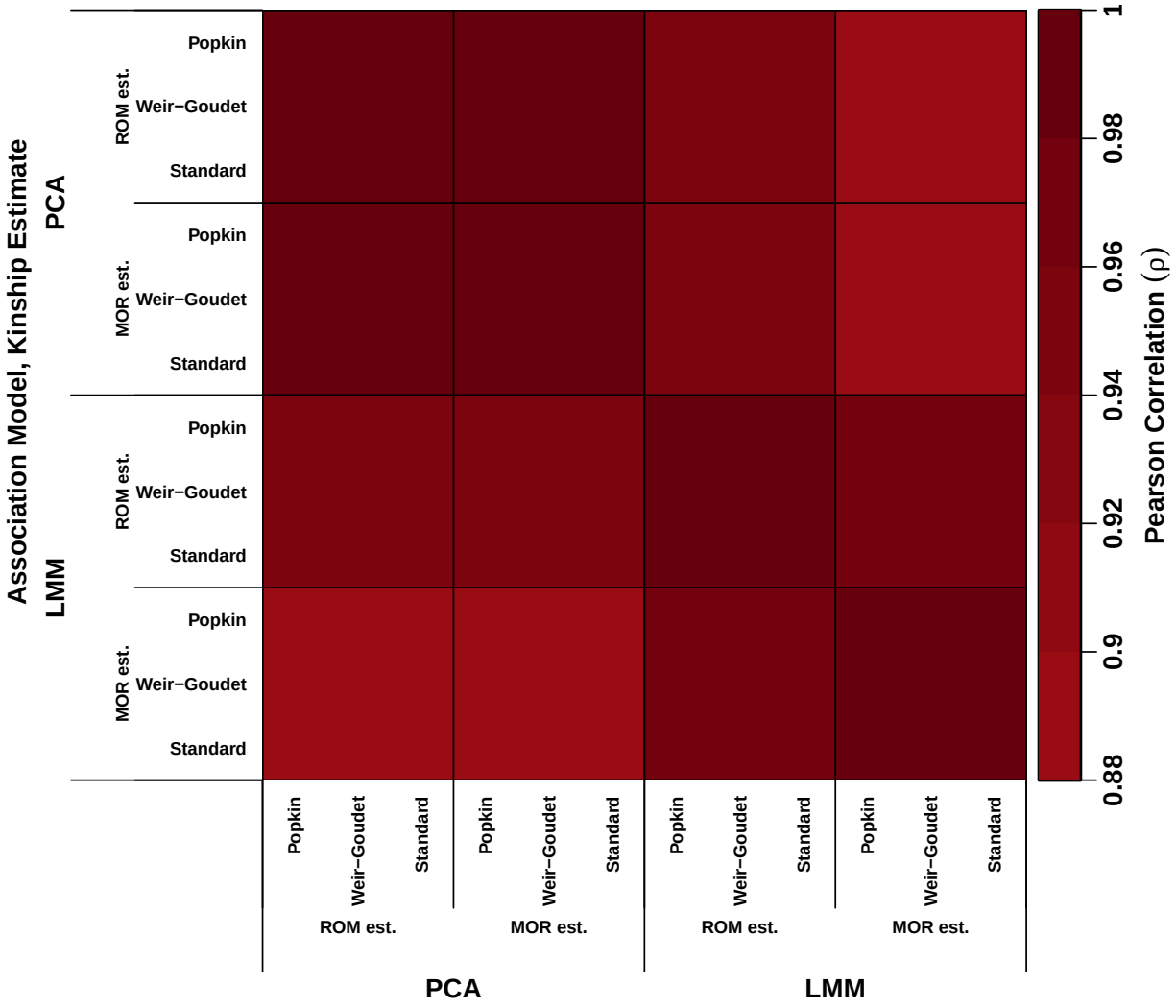

Figure S3: **Correlation between p-values of different association models and kinship estimates on 1000 Genomes.** The association p-value vector (one value per tested locus) produced by each combination of association model (LMM vs PCA) and kinship matrix (x and y axes) was used to compute Pearson correlations (color). P-values between all kinship matrices or association model (PCA or LMM) are highly correlated (including biased and unbiased estimates), and often the maximum correlation of 1 is achieved.