OXFORD GENETICS

Genetic association models are robust to common population kinship estimation biases

Zhuoran Hou[1] and 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: Duke Center for Statistical Genetics and Genomics, Duke University, Durham, NC 27705, USA. Email: alejandro.ochoa@duke.edu

Abstract

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

Keywords: kinship matrices; estimation bias; genetic association studies; linear mixed effects models; principal components analysis

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, parameterize 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, 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, 2011; Zhou and Stephens 2012; Speed *et al.* 2012; Yang *et al.* 2014; Speed and Balding 2015; Loh *et al.* 2015; Wang *et al.* 2017; Sul *et al.* 2018). The impetus of our work is the recent characterization of a complex bias for this standard estimator, which varies for every pair of individuals (Weir and Goudet 2017; Ochoa and Storey 2021). These recent works also produced two new kinship estimators, which we are interested in characterizing in the context of association. The Weir-Goudet (WG) estimator constitutes a key improvement

in that it has a uniformly downward bias (Weir and Goudet 2017; Ochoa and Storey 2021). Lastly, the popkin estimator is the only unbiased estimator under arbitrary relatedness (Ochoa and Storey 2021). To the best of our knowledge, the new WG and popkin estimators have not been used in association studies before, but represent potential improvements over the use of the standard estimator for association.

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

In this work, we originally hypothesized that kinship estimation bias would affect association testing. We perform evaluations using an admixed family simulation (Yao and Ochoa 2022) as well as real genotypes from the 1000 Genomes project (1000 Genomes Project Consortium 2010; 1000 Genomes Project Consortium *et al.* 2012; Fairley *et al.* 2020), in both cases with simulated traits, to characterize type I error control and power using robust statistics. Surprisingly, we find that both LMM and PCA association statistics are largely invariant to kinship estimation bias. We theoretically characterize the conditions under which these kinship biases result in invariant association statistics, which encompass changing ancestral population in the kinship matrix too. As we discover that most kinship estimates are non-positive semidefinite (non-PSD), breaking a key modeling assumption, we perform additional empirical validations and discover that some LMMs can handle these improper covariance matrices surprisingly well. Overall, we find that long-used association approaches are unaffected by the most common kinship estimation biases, and develop theory that may help improve association and related approaches such as heritability estimation.

Materials and methods

Genetic model

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

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

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

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

Kinship estimators

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

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

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

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

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

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

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

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

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

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

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

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

$$\begin{aligned}\hat{\mathbf{\Phi}}^{T,\text{std-ROM}} &= F^{\text{std}}\left(\hat{\mathbf{\Phi}}^{T,\text{popkin-ROM}}\right),\\ \hat{\mathbf{\Phi}}^{T,\text{std-MOR}} &= F^{\text{std}}\left(\hat{\mathbf{\Phi}}^{T,\text{popkin-MOR}}\right).\end{aligned}$$

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

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

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

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

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

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

In Appendix C we prove that

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

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

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

Association models

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

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

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

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

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

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

Simulations

Admixed family genotype simulation: An admixed family is simulated following previous work (Yao and Ochoa 2022), except here only $K = 3$ ancestries are simulated and $F_{ST} = 0.3$ for the admixed individuals, which more closely resembles Hispanics and African Americans. Briefly, our admixture model first simulates $n = 1000$ founder individuals with $m = 100,000$ loci, which was purposefully reduced compared to previous work to increase the difference between estimated kinship matrices (which will be noisier) and their limits. Random ancestral allele frequencies p_i^T, subpopulation allele frequencies $p_i^{S_u}$, individual-specific allele frequencies π_{ij}, and genotypes x_{ij} are drawn from this hierarchical model:

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

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

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

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 heritability h^2 (0.8 or 0.3 in this work) and the number of causal loci m_1 (here chosen using $m_1 = \text{round}(nh^2/8)$, which empirically balances power as sample size and heritability are varied), the goal is to choose causal coefficients $\boldsymbol{\beta}$ and the intercept α that result in zero mean and the

desired trait heritability. Here, we use the "fixed effect sizes" trait simulation model described in (Yao and Ochoa 2022). Briefly, first m_1 causal loci are randomly selected, and for these steps only $\mathbf{X}$ is subset to these loci and reindexed. For known p_i^T, causal coefficients are constructed as:

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

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

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

where $\mathbf{1}_{m_1}$ is a length-m_1 column vector of ones. Genotypes were simulated from the admixed family model separately per heritability value and every replicate.

Real genotype data processing

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

Evaluation of performance

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

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

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

AUC_{PR} (Area Under the Precision and Recall Curve) is a binary classification measure that reflects calibrated power (Yao and Ochoa 2022), which is calculated from the total numbers of true positives (TP), false positives (FP) and false negatives (FN) at some threshold or parameter t:

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

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

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

Software

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

GCTA, which implements the model of Equations (8) and (9), is used to run all LMM associations (Yang *et al.* 2011, 2014). We pass $2\boldsymbol{\Phi}^T$ for all kinship matrices tested (the same scale as its own kinship estimate). `plink2`, which implements the model of Equations (10) and (11), performs the PCA associations (Chang *et al.* 2015). We use $r = K - 1 = 2$ PCs for the admixed family simulations, and $r = 10$ PCs for 1000 Genomes.

Results

Empirical analysis using admixed family simulation

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

Kinship estimates and limits for this simulation are shown in Figure 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 (Figure S1A in File S1). In contrast, the Standard and Weir-Goudet (WG) estimates have large downward biases overall, resulting in abundant negative values; Standard biases vary for every pair of individuals (as described in Equation (4); Figure S2A in File S1), while WG has a uniform bias (following Equation (6)). The difference is most noticeable near the diagonal: the true kinship matrix has monotonically increasing values, WG has smaller values but which are still monotonically increasing, and Standard estimates follow a U-shaped pattern (decreasing at first, then increasing again in Figure 1D-F).

We perform LMM and PCA association tests to determine how kinship biases affect association performance. Surprisingly, we find that kinship bias type does not have a discernible effect on association performance, as summarized by AUC_{PR} (a robust proxy for power; high and low heritability in Figure 2 and Figure S3 in File S1, respectively) and SRMSD_p (measures null statistic calibration; Figures S4, S5 in File S1). The largest differences in performance are explained by the association model (LMM vs PCA), as expected due to our use of a family simulation where PCA performs poorly. Within association models, there are no clear differences between the performance of any of the kinship matrices, in fact many appear to have identical distributions (both statistics), the only clear exception being LMM popkin MOR with $h^2 = 0.8$, which has a few outlier replicates where

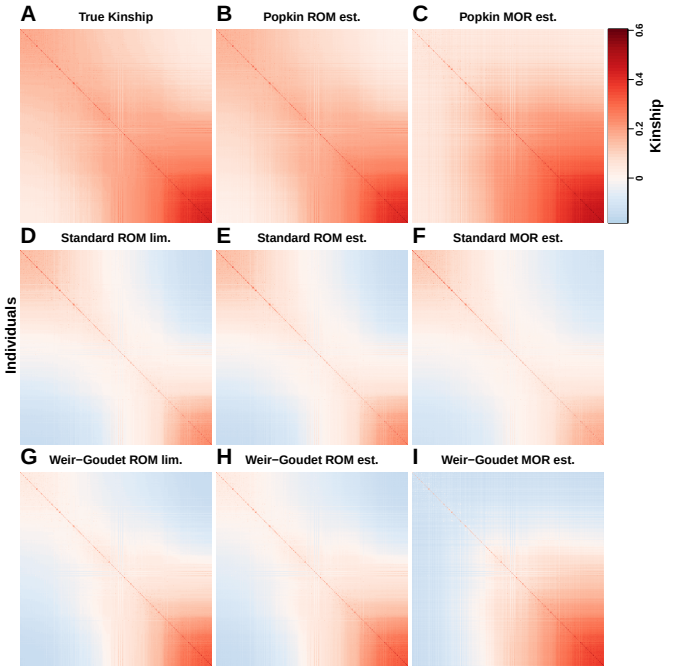

Figure 1 Kinship estimates and limits on the admixed family simulation. Each panel shows a kinship matrix as a heatmap, with each of the $n = 1000$ individuals along both x and y axes, color represents kinship: positive values in red, negative in blue. Diagonal contains inbreeding values. Each estimator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

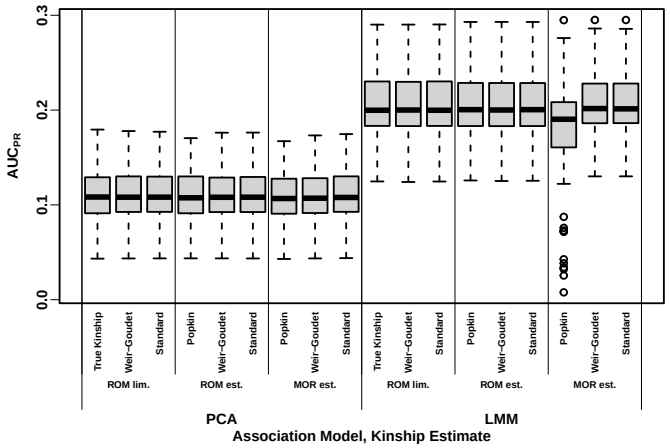

Figure 2 Distributions of Area Under the Precision-Recall Curve (AUC_{PR}) on the admixed family simulation with $h^2 = 0.8$. Higher AUC_{PR} is better performance. Results for 100 replicates (each a random genotype matrix and trait vector). Approaches cluster primarily by association model (LMM or PCA), and vary little across bias types.

performance is exceedingly poor (at the end of the results we show these are due to limited numerical precision exacerbated by high condition numbers of trait covariance matrices).

To better characterize the nearly identical performance distribution just observed, we next measure the agreement between individual association p-values. We measure high correlations ρ between p-values, near $\rho = 1$ for comparisons involving the same model (between LMM methods or between PCA methods, both heritabilities), and across models around $\rho = 0.6$ for $h^2 = 0.8$, which increases to $\rho = 0.88$ for $h^2 = 0.3$ (Figures S6, S7 in File S1). To measure numerical agreement more stringently, we calculate the proportion of loci between two methods with p-values within 0.01 of each other, and find a remarkably high agreement between estimators of different bias types after matching association model and locus weight type or limit (Figure 3, and Figure S8 in File S1). This is in contrast to low agreement between PCA and LMM statistics, and between LMM statistics with different locus weight types or limits. Minimum agreements are higher across PCA methods, though here the true kinship or popkin estimates disagree more from Standard and WG matrices. Overall, kinship matrices with different bias types (otherwise matched) result in nearly identical association statistics.

Empirical analysis using 1000 Genomes

Now we repeat our analysis using the real genotypes of 1000 Genomes. Kinship estimates are shown in Figure 4 (note real data have no true kinship or estimator limits). Popkin ROM estimates display an approximate nested block structure that arises from the tree relationships between subpopulations (Figure 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 (Figure 4B, and Figure S1B in File S1). Standard estimates have values closer to zero, and a different bias for each pair of individuals (Figure S2B in File S1), resulting in higher relative kinship for African compared to non-African populations (Figure 4C-D), whereas kinship in African populations is the lowest in the unbiased estimate (Figure 4A). Lastly, WG estimates are uniformly smaller than popkin's and attain large negative values (Figure 4E-F).

Our association test conclusion are similar to our simulation study: AUC_{PR} and $SRMSD_p$ distributions are nearly identical for estimators of different bias types but same locus weight type (ROM or MOR) and association model. However, unlike the simulation, here for $h^2 = 0.8$ (but not 0.3) the MOR estimates noticeably outperform ROM estimates (LMM only), in terms of both AUC_{PR} (Figure 5, and Figure S9 in File S1) and $SRMSD_p$ (Figures S10, S11 in File S1). P-values are even more highly correlated in this case (Figures S12, S13 in File S1), and again nearly identical at a large proportion of loci between approaches with matched association model and locus weight type (MOR or ROM), regardless of bias type (Figures S14, S15 in File S1).

Proof of association invariability to common kinship biases

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

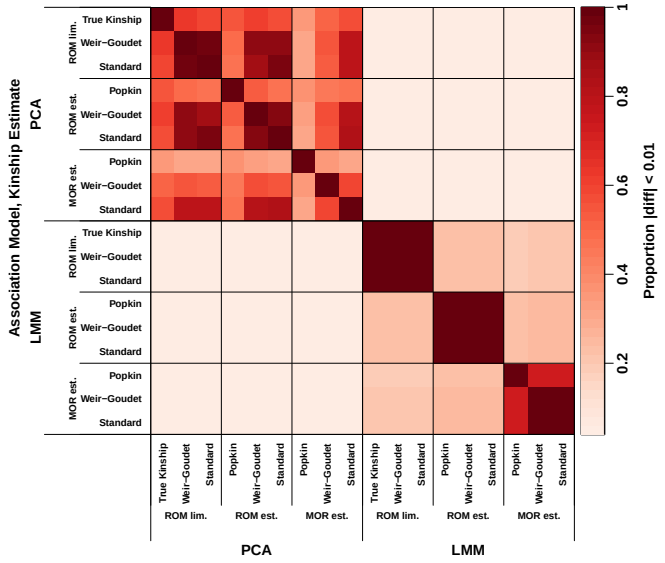

Figure 3 Agreement between p-values on the admixed family simulation with $h^2 = 0.8$. Calculated agreement (absolute difference under 0.01) averaged over loci (color) of association p-values between association models (LMM vs PCA) and kinship matrices (x and y axes). All 100 replicates are used. Different bias types (matched for association model and locus weight type) have large proportions of nearly identical p-values.

association coefficient and its p-value are unchanged by these biases.

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

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

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

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

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

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

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

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

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

The derivation of this factorization is given in Appendix D. The determinant of the quadratic solution q would be non-negative

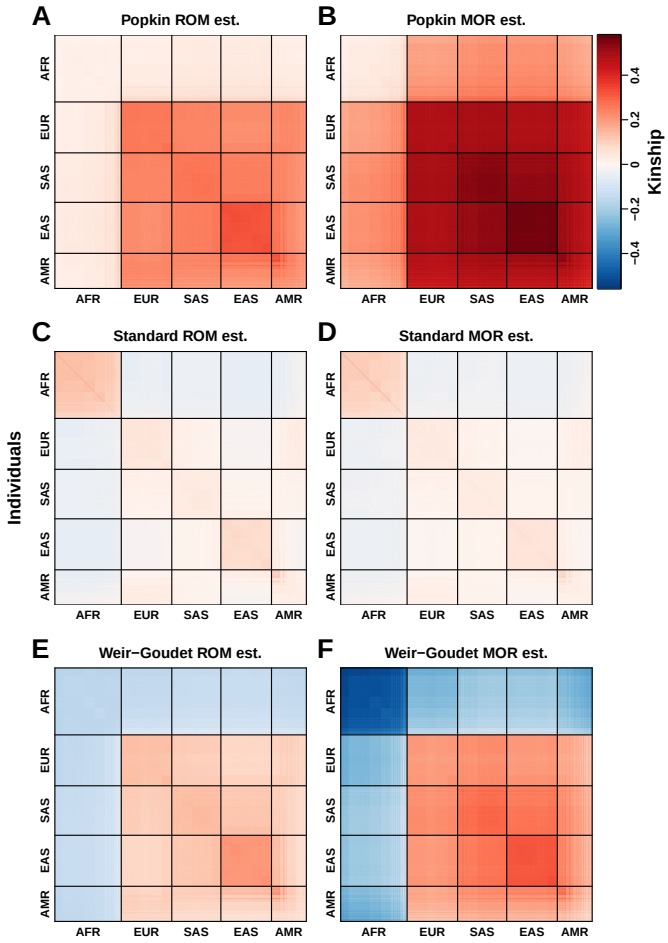

Figure 4 Kinship estimates on 1000 Genomes. Each panel represents a kinship matrix as a heatmap, as in Figure 1. Superpopulation codes: AFR = African, EUR = European, SAS = South Asian, EAS = East Asian, AMR = Admixed Americans (Hispanics). Each estimator bias type (Popkin, Standard, and Weir-Goudet;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

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

Proof for LMM case: Consider a random effect $\mathbf{s}$ drawn using $\mathbf{\Phi}^T$, as given in Equation (9). Using the affine transformation property of Multivariate Normal distributions (which holds even

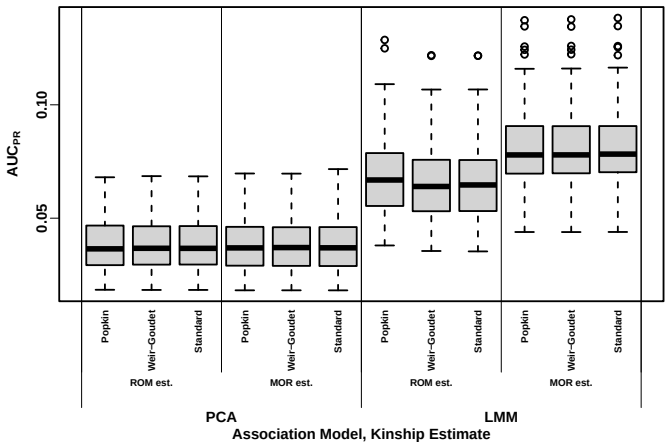

Figure 5 Distributions of Area Under the Precision-Recall Curve (AUC_{PR}) on 1000 Genomes with $h^2 = 0.8$. Higher AUC_{PR} is better performance. Results based on 100 simulated trait replicates (real genotype matrix is fixed). Approaches cluster primarily by association model (LMM or PCA) and locus weight type (ROM or MOR), and do not depend much at all on the bias type.

if $\mathbf{B}$ below is singular) and Equation (12), then

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

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

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

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

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

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

The LMM association p-value does not change in several common tests, including the F-test, since it only depends on the residuals and these do not change, as well as the likelihood ratio test, because although covariance determinants change, they cancel out in the ratio. The Wald test used by GCTA (Yang *et al.* 2014) is also invariant to these kinship biases given our empirical results in Figure 3, and Figure S14 in File S1, and proven explicitly for WG bias in Appendix F. Lastly, using an implementation in R and simulated data, we confirmed that the LMM Score test is also invariant to these kinship biases. These arguments hold whether variance components are fit with maximum likelihood or restricted maximum likelihood (Kang *et al.* 2008, 2010; Yang *et al.* 2014), since multiplying the estimated genetic variance component σ^2 by c and adjusting the intercept compensates for the bias regardless of how $\sigma^2, \sigma_\epsilon^2$ are estimated.

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 Equations (10) and (11), let $\mathbf{U}_r$ be the top eigenvectors of $\mathbf{\Phi}^T$, and $\mathbf{U}'_r$ those of $\mathbf{\Phi}^{T\prime}$. They key approximation is that

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

which is not strictly equal (since $\mathbf{B}\mathbf{U}_r$ is not generally orthogonal, as eigenvectors must be), but we have found it to be a good approximation in practice. In this case the eigenvector coefficients need not change, $\gamma'_r = \gamma_r$, since the difference in scale of the kinship matrices (c in Equation (12)) is absorbed by the eigenvalues not present in this model. Applying the shift of Equation (13) shows that

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

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

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

Proof of association invariability to change in ancestral population

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

Start from a kinship matrix $\mathbf{\Phi}^S$ in terms of ancestral population S, and let T be a population ancestral to S. If the inbreeding coefficient of S when T is the reference ancestral population is f_S^T, then the kinship matrix $\mathbf{\Phi}^T$ in terms of T is given by (Ochoa and Storey 2021)

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

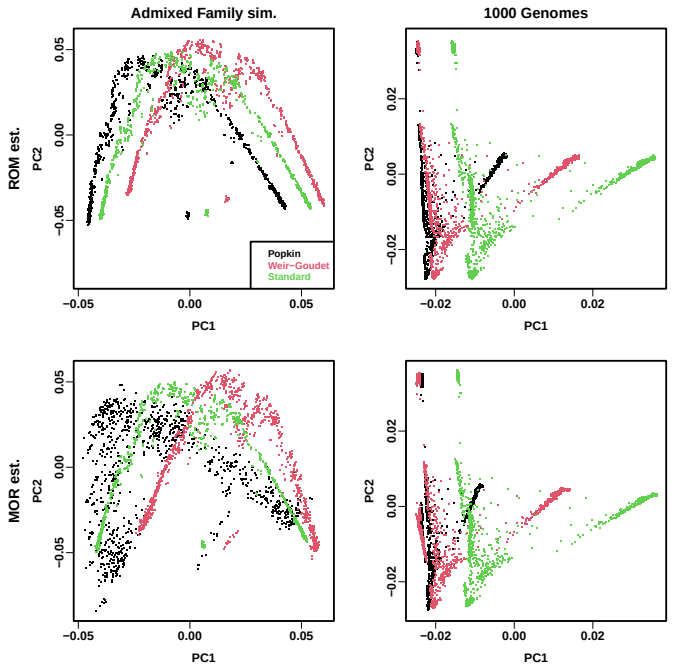

Figure 6 Visualization of PC shift due to kinship biases. Each panel shows three estimates (bias types): Popkin, Standard, and Weir-Goudet. ROM estimates are in first row, MOR in second row. (In admixed family, ROM limits are very similar to ROM estimates (not shown).) Columns show estimates from each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

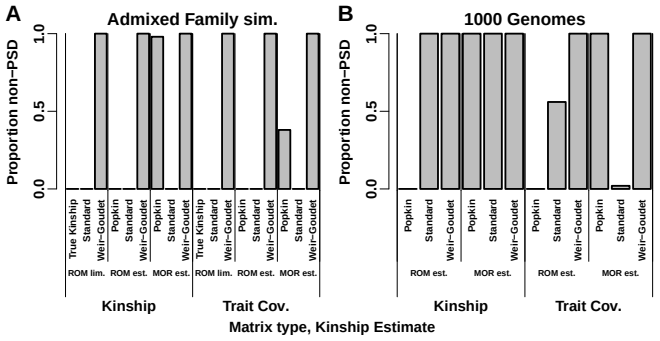

Figure 7 Proportion of kinship and trait covariance (V) matrices with $h^2 = 0.8$ that are not positive semidefinite (PSD). A matrix is non-PSD if it has negative eigenvalues (below -10^{-7} to allow for limited machine precision). Proportion is calculated over 100 replicates (1000 Genomes kinship has one value since genotypes are fixed, but $\mathbf{V}$ varies per replicate). **A.** In admixed family simulation, which does not have missing genotypes, all WG matrices and most popkin MOR estimates are non-PSD. All non-PSD kinship matrices results in non-PSD $\mathbf{V}$ except some popkin ROM estimates yield PSD $\mathbf{V}$. **B.** In 1000 Genomes, which has missingness, all kinship estimates are non-PSD except popkin ROM. Of the non-PSD kinship matrices, only some Standard estimates yield PSD $\mathbf{V}$.

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

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

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

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

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

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

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

We find that all WG matrices have very large negative minimum eigenvalues, and popkin MOR estimates also have smaller negative minimum eigenvalues (Figures S16, S17 in File S1). Moreover, besides all WG matrices and most popkin MOR estimates, Standard matrices are also often non-PSD but only in 1000 Genomes (Figure 7, and Figure S18 in File S1), which has missing genotypes (the admixed family simulation does not have missing genotypes). Each of these non-PSD matrices only has one negative eigenvalue. Notably, all popkin ROM estimates are PSD in every evaluation, including under missingness in 1000 Genomes.

In order to quantify matrix singularity, as well as numerical accuracy problems caused by multiplying by inverses of nearly-singular matrices, we calculate condition numbers, which equal the maximum absolute eigenvalue divided by the minimum absolute eigenvalue of our covariance matrices. As expected, we see that Standard kinship matrices are singular on our admixed family simulation (which lacks missingness), as reflected by extremely high condition numbers, but their trait covariances have small condition numbers (Figures S19, S20 in File S1). No other matrices are singular, but popkin MOR estimates in the admixed family simulation have relatively high condition numbers for kinship, as well as trait covariance for $h^2 = 0.8$ but not 0.3.

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

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

where $\mathbf{U}$ and $\mathbf{\Lambda}$ are the eigenvectors and eigenvalues of $\mathbf{\Phi}^T$, re-

spectively (Equation (11)), so the eigenvectors of $\mathbf{V}$ are also $\mathbf{U}$ and its eigenvalues are $2\sigma^2\mathbf{\Lambda} + \sigma_\epsilon^2\mathbf{I}$. Therefore, since $\sigma^2, \sigma_\epsilon^2 \geq 0$, then if $\mathbf{\Phi}^T$ is positive definite (all of its eigenvalues are positive) then so is $\mathbf{V}$, and the condition number of $\mathbf{V}$ is always smaller (better) or equal than that of $\mathbf{\Phi}^T$. A negative kinship eigenvalue λ_k may become positive for $\mathbf{V}$ only if $\lambda_k > -\sigma_\epsilon^2/(2\sigma^2) = -(1-h^2)/(2h^2)$, so very large negative λ_k values as observed for WG do not become positive in $\mathbf{V}$, in fact they can become more negative for high heritability (Figure S16 in File S1), though they are less negative at lower heritability (Figure S17 in File S1). $\mathbf{V}$ is always invertible and well-conditioned even when $\mathbf{\Phi}^T$ is singular PSD (has zero eigenvalues), as the Standard estimator is under no missingness, since a kinship zero eigenvalue becomes σ_ϵ^2 for $\mathbf{V}$. Conversely, the above equation explains why some non-PSD kinship matrices are particularly problematic: negative eigenvectors near the heritability-dependent value $-\sigma_\epsilon^2/(2\sigma^2)$ can result in ill-conditioned $\mathbf{V}$. We see that popkin MOR estimates are non-PSD (Figure S16 in File S1) in such a way that some of their $\mathbf{V}$ are ill-conditioned under high heritability (Figure S19 in File S1) but not the lower heritability (Figure S20 in File S1), and this explains its poorer performance in the admixed family evaluations with high heritability (Figure 2, and Figure S4 in File S1), as shown in the next subsection.

Further empirical validation of theoretical predictions

Seeing that WG is always non-PSD, and to query other instances where predictions are not fully met, here we analyze estimation accuracy of various parameters to better understand theoretically and empirically how broken assumptions affect them. With PCA, no deviations from expectation of AUC_{PR} and SRMSD_p are observed for WG (Figures 2, 5, and Figures S4, S10 in File S1), which makes sense since PCA simply ignores eigenvectors with negative or zero eigenvalues. Therefore, our analysis focuses on LMM and $h^2 = 0.8$ only, where large deviations are observed, and clarification regarding WG is needed.

LMMs such as GCTA perform association testing in two steps. First is the restricted maximum likelihood step used to fit variance components. Although the eigendecomposition approaches (Kang *et al.* 2008; Lippert *et al.* 2011; Svishcheva *et al.* 2012; Zhou and Stephens 2012; Sul *et al.* 2018) require positive definite $\mathbf{V}$ (lest the determinant of $\mathbf{V}$ be negative), surprisingly the GCTA average information algorithm only requires in practice that $\mathbf{V}$ be invertible (Yang *et al.* 2011). Thus, the relationship between WG, Standard, and True or Popkin variance components are largely as expected from our theoretical prediction $\sigma^{2\prime} = c\sigma^2$ in Equation (16), with the exception of popkin ROM on 1000 Genomes $h^2 = 0.8$ only, whose genetic variance estimates are slightly smaller than expected (Figures S21, S22 in File S1).

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

Lastly, we explain the largest deviations from our predictions of the performance metrics AUC_{PR} and SRMSD_p for $h^2 = 0.8$ (for $h^2 = 0.3$ there are no large prediction deviations). We find that the small performance errors of popkin ROM in 1000 Genomes (Figure 5, and Figure S10 in File S1) are driven by errors in genetic variance component estimation σ^2 (Figure S23 in File S1). However, the larger performance errors of popkin MOR in the admixed family simulation (Figure 2, and Figure S4 in File S1) are instead explained by the condition number of $\mathbf{V}$ (Figure S24 in File S1). This result makes sense since the condition number by definition quantifies regression coefficient estimation accuracy.

Discussion

Previous research showed that commonly used kinship estimators are biased, and that these biases can be large (Ochoa and Storey (2021); Figure 1, and Figure S2 in File S1). Our initial hypothesis was that these kinship biases would affect association testing, but surprisingly found that association is unaffected. We then proved theoretically that it is the intercept and relatedness effect (random effect or PCs) coefficients that compensate for the bias, and result in identical association coefficients and significance statistics.

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

Given that kinship bias type is not important for association studies, we are free to choose a kinship estimator based on other properties. Ideally, kinship matrices result in well conditioned trait covariance matrices, since that has the largest effect in numerical accuracy and power in LMMs. Well-conditioned association is guaranteed for PSD kinship matrices, and popkin ROM is the only estimator that produces PSD matrices consistently across our evaluations (Figure 7). Popkin ROM is also the only unbiased kinship estimator (Ochoa and Storey 2021). We observed that Standard kinship estimates are also not PSD when genotypes are missing, a well understood phenomenon for related sample covariance estimators outside genetics (Jurczak and Rohde 2017). Fortunately, non-PSD kinship estimators often perform well for association. Nevertheless, in our admixed family simulation we did see the other popkin estimator (the MOR version) perform particularly poorly due to being non-PSD, which in a heritability-dependent manner results in ill-conditioned association tests and substantial loss of accuracy and power (Figure 2, and Figures S4, S24 in File S1). Theory predicts that the same can happen with any non-PSD estimator, depending on unknowns such as the heritability and the value of the negative eigenvalues of the kinship estimator, so

it is risky to use MOR estimators (all of which are non-PSD in 1000 Genomes), as well as the WG estimator generally (which is non-PSD in all replicates of all of our evaluations). We also observe smaller numerical inaccuracies for popkin ROM, the estimator we recommend, in 1000 Genomes with $h^2 = 0.8$ only, although the result is mixed: performance is slightly better (Figure 5) although null p-value calibration is slightly worse (Figure S10 in File S1). The cause is variance components are poorly estimated (Figure S21 in File S1), but we did not find a more fundamental explanation. Overall, our assessment suggests that the popkin ROM estimator is the safest choice due to its guarantee of well-conditioned associations that other estimators cannot make.

Despite being non-PSD, we observe better performance for MOR versus ROM estimators in LMM association of 1000 Genomes with $h^2 = 0.8$ (Figure 5; the approaches were tied under lower heritability in Figure S9). Perhaps this is expected because we simulated larger coefficients for rare variants, while MOR estimators upweigh rare variants. This effect is not observed in the admixed family simulation, where MOR and ROM versions give similar kinship estimates (Figure 1) and performed similarly (Figure 2), compared to 1000 Genomes where kinship estimates are also strikingly different (Figure 4). However, only popkin ROM is unbiased (Figure 1B, and Figure S1 in File S1). One potential explanation is that our kinship model assumes that all variants existed in the MRCA population, whereas rare variants in human data are known to be more recent mutations, and thus their effective kinship matrix is different than that of ancestral variants. Therefore, despite its biases, the popkin MOR estimator may better capture the covariance of rare variants and thus model them better in association tests, particularly in LMMs where the effect is most pronounced. Future work should focus on better approaches for upweighing rare variants or otherwise estimating their covariance structure while resulting in positive definite kinship estimates.

Our conclusions that common kinship biases do not affect association studies extend to variations of the Standard kinship estimator that weigh loci according to linkage disequilibrium (Speed *et al.* 2017; Wang *et al.* 2017), which also have the Standard bias type since this bias is present in each locus (Ochoa and Storey 2021). As shown in our theoretical results, another form of the Standard kinship estimator that weighs individuals to estimate ancestral allele frequencies $\hat{p}_i^T$, including the best unbiased linear estimator in Appendix E (Astle and Balding 2009; Thornton and McPeek 2010), is also subject to the same conclusions. Our proof holds for arbitrary kinship matrices and their biased counterparts, making no assumptions about how they are related to the loci being tested for association, so they hold whether the test locus was included in the kinship estimate (the scenario tested empirically) or not, such as in the leave-one-chromosome-out variant of association testing with LMMs (Lippert *et al.* 2011; Yang *et al.* 2014). In addition, the proof does not make any assumptions about the tested loci, so it does not depend on allele frequency, and thus holds for rare and common variants.

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

Data 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/ and detailed processing instructions are found on GitHub at https://github.com/OchoaLab/data.

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

Funding

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.

Conflicts of interest

The authors declare no conflicts of interest.

Literature cited

1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. Nature. 467:1061–1073.

1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature. 491:56–65.

Altschul SF, Carroll RJ, Lipman DJ. 1989. Weights for data related by a tree. Journal of Molecular Biology. 207:647–653.

Astle W, Balding DJ. 2009. Population Structure and Cryptic Relatedness in Genetic Association Studies. Statist. Sci. 24:451–471.

Aulchenko YS, de Koning DJ, Haley C. 2007. Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. Genetics. 177:577–585.

Balding DJ, Nichols RA. 1995. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica. 96:3–12.
Bhatia G, Patterson N, Sankararaman S, Price AL. 2013. Estimating and interpreting FST: the impact of rare variants. Genome Res. 23:1514–1521.
Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 4:7.
Devlin B, Roeder K. 1999. Genomic Control for Association Studies. Biometrics. 55:997–1004.
Emik LO, Terrill CE. 1949. Systematic procedures for calculating inbreeding coefficients. J Hered. 40:51–55.
Fairley S, Lowy-Gallego E, Perry E, Flicek P. 2020. The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. Nucleic Acids Res. 48:D941–D947.
García-Cortés LA. 2015. A novel recursive algorithm for the calculation of the detailed identity coefficients. Genetics Selection Evolution. 47:33.
Hoffman GE. 2013. Correcting for population structure and kinship using the linear mixed model: theory and extensions. PLoS ONE. 8:e75707.
Jacquard A. 1970. *Structures génétiques des populations*. Masson et Cie. Paris.
Jurczak K, Rohde A. 2017. Spectral analysis of high-dimensional sample covariance matrices with missing observations. Bernoulli. 23:2466–2532.
Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E. 2010. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 42:348–354.
Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E. 2008. Efficient control of population structure in model organism association mapping. Genetics. 178:1709–1723.
Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. 2011. FaST linear mixed models for genome-wide association studies. Nat Methods. 8:833–835.
Loh PR, Tucker G, Bulik-Sullivan BK, Vilhjálmsson BJ, Finucane HK, Salem RM, Chasman DI, Ridker PM, Neale BM, Berger B *et al*. 2015. Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat Genet. 47:284–290.
Malécot G. 1948. *Mathématiques de l'hérédité*. Masson et Cie.
Ochoa A, Storey JD. 2021. Estimating FST and kinship for arbitrary population structures. PLoS Genet. 17:e1009241.
Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 38:904–909.
Rakovski CS, Stram DO. 2009. A kinship-based modification of the armitage trend test to address hidden population structure and small differential genotyping errors. PLoS ONE. 4:e5825.
Sherman J, Morrison WJ. 1950. Adjustment of an Inverse Matrix Corresponding to a Change in One Element of a Given Matrix. Ann Math Stat. 21:124–127.
Speed D, Balding DJ. 2015. Relatedness in the post-genomic era: is it still useful? Nat Rev Genet. 16:33–44.
Speed D, Cai N, the UCLEB Consortium, Johnson MR, Nejentsev S, Balding DJ. 2017. Reevaluation of SNP heritability in complex human traits. Nat Genet. 49:986–992.
Speed D, Hemani G, Johnson MR, Balding DJ. 2012. Improved heritability estimation from genome-wide SNPs. Am J Hum Genet. 91:1011–1021.
Sul JH, Martin LS, Eskin E. 2018. Population structure in genetic studies: Confounding factors and mixed models. PLoS Genet. 14:e1007309.
Svishcheva GR, Axenovich TI, Belonogova NM, van Duijn CM, Aulchenko YS. 2012. Rapid variance components–based method for whole-genome association analysis. Nat Genet. 44:1166–1170.
Thornton T, McPeek MS. 2010. ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure. Am J Hum Genet. 86:172–184.
Voight BF, Pritchard JK. 2005. Confounding from Cryptic Relatedness in Case-Control Association Studies. PLOS Genetics. 1:e32.
Wang B, Sverdlov S, Thompson E. 2017. Efficient estimation of realized kinship from single nucleotide polymorphism genotypes. Genetics. 205:1063–1078.
Weir BS, Goudet J. 2017. A Unified Characterization of Population Structure and Relatedness. Genetics. 206:2085–2103.
Wright S. 1922. Coefficients of Inbreeding and Relationship. The American Naturalist. 56:330–338.
Wright S. 1949. The Genetical Structure of Populations. Annals of Eugenics. 15:323–354.
Xie C, Gessler DD, Xu S. 1998. Combining different line crosses for mapping quantitative trait loci using the identical by descent-based variance component method. Genetics. 149:1139–1146.
Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW *et al*. 2010. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 42:565–569.
Yang J, Lee SH, Goddard ME, Visscher PM. 2011. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 88:76–82.
Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. 2014. Advantages and pitfalls in the application of mixed-model association methods. Nat Genet. 46:100–106.
Yao Y, Ochoa A. 2022. Limitations of principal components in quantitative genetic association models for human studies. bioRxiv. https://www.biorxiv.org/content/10.1101/2022.03.25.485885v1.
Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB *et al*. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet. 38:203–208.
Zhou X, Stephens M. 2012. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 44:821–824.

Appendices

A. Justification for popkin generalizations

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

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

then for fixed w_i we get

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

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

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

as desired.

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

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

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

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

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

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

B. Connection between popkin and standard kinship estimator

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

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

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

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

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

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

Therefore, the combination of interest is:

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

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

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

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

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

also requires $n_i = n$.

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

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

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

C. Mean kinship inequalities

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

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

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

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

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

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

D. Derivation of WG bias factorization

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

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

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

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

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

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

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

E. Minimum weighted mean kinship

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

The last factor above vanishes:

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

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

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