Limitations of principal components in quantitative genetic association models for human studies

Yiqi Yao[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

Principal Component Analysis (PCA) and the Linear Mixed-Effects Model (LMM), sometimes in combination, are the most common modern models for genetic association. Previous PCA-LMM comparisons give mixed results and unclear guidance, and have several limitations, including not varying the number of principal components (PCs), simulating overly simple population structures, and inconsistent use of real data and power evaluations. In this work, we thoroughly evaluate PCA and LMM both with varying number of PCs in new realistic genotype and complex trait simulations including admixed families, trees, and large real multiethnic human genotype datasets (1000 Genomes Project, the Human Genome Diversity Panel, and Human Origins) with simulated traits. We find that LMM without PCs performs best in all cases, with the largest effects in the family simulation and all real human datasets. We determined that the large gaps in PCA to LMM performance on the real human datasets is due to the high-dimensional family structure stemming from large numbers of distant relatives, and not from the smaller number of highly related pairs. While it was known that PCA fails on family data, here we report a strong effect on association of cryptic family relatedness in several genetically diverse human datasets, a problem that is not avoided with the common practice of pruning high-relatedness individual pairs. Overall, this work better characterizes the severe limitations of PCA compared to LMM in modeling the complex relatedness structures present in real multiethnic human data and its impact in association studies.

Abbreviations: PCA: principal component analysis; PCs: principal components; LMM: linear mixed-effects model; FES: fixed effect sizes; RC: random coefficients; WGS: whole genome sequencing.

1 Introduction

The goal of a genetic association study is to identify loci whose genotype variation is significantly correlated to given trait. An important, implicit assumption by naive association tests is that, under the null hypothesis, genotypes are unstructured: drawn independently from a common allele frequency. This assumption does not hold for structured populations, which includes multiethnic cohorts and admixed individuals, and for family data. When naive or insufficient approaches are applied to structured populations or family data, association statistics become miscalibrated relative to the null expectation, resulting in greater numbers of false positives than expected and loss of power (Devlin and Roeder, 1999; Voight and Pritchard, 2005; Astle and Balding, 2009). Therefore, many specialized approaches have been developed for genetic association in structured data. Here we focus on extensively evaluating the two most popular association models: PCA and LMM.

Genetic association with PCA consists of including the top eigenvectors of the population kinship matrix as covariates in a generalized linear model (Zhang et al., 2003; Price et al., 2006; Bouaziz et al., 2011). These top eigenvectors are commonly referred to as PCs in genetics (Patterson et al., 2006), the convention adopted here, but it is worth noting that in other fields PCs instead denote the projections of loci onto eigenvectors (Jolliffe, 2002). The direct ancestor of PCA association is structured association, in which inferred ancestry or admixture proportions are used as regression covariates (Pritchard et al., 2000). These models are deeply connected because PCs map to ancestry empirically (*e.g.*, Alexander et al., 2009; Zhou et al., 2016) and theoretically (McVean, 2009; Zheng and Weir, 2016; Cabreros and Storey, 2019; Chiu et al., 2022), and they work as well as global ancestry in association studies but are estimated more easily (Patterson et al., 2006; Zhao et al., 2007; Alexander et al., 2009; Bouaziz et al., 2011). The strength of PCA is its simplicity, which as covariates can be readily integrated into more complex models, such as haplotype association (Xu and Guan, 2014) and polygenic models (Qian et al., 2020). However, PCA fundamentally

assumes that relatedness is low-dimensional, which may limit its applicability. PCA is known to be inadequate for data containing family structure (Patterson et al., 2006; Thornton and McPeek, 2010; Price et al., 2010), which is called "cryptic relatedness" when it is unknown to the researchers, but no other specific troublesome scenarios have been confidently identified. Recent work has focused on developing more scalable versions of the PCA algorithm (Lee et al., 2012; Abraham and Inouye, 2014; Galinsky et al., 2016; Abraham et al., 2017; Agrawal et al., 2020). PCA remains a popular and powerful approach for association studies.

The other dominant association model for structured populations is the LMM, in which this structure is a random effect drawn from a multivariate Normal model parametrized by the kinship matrix. Unlike PCA, LMM does not assume that relatedness is low-dimensional, and explicitly models family structure via the kinship matrix. Early LMMs required kinship matrices estimated from known pedigrees or which otherwise captured family-level relatedness only (Yu et al., 2006; Zhao et al., 2007). Modern LMMs estimate kinship from genotypes using a non-parametric estimator, often referred to as a genetic relationship matrix, that captures the combined covariance due to recent family relatedness and ancestral population structure (Kang et al., 2008; Astle and Balding, 2009; Ochoa and Storey, 2021). The classic LMM assumes a quantitative (continuous) complex trait, the focus of our work. Although case-control (binary) traits and their underlying ascertainment are theoretically a challenge (Yang et al., 2014), LMMs have been applied successfully to balanced case-control studies (Astle and Balding, 2009; Kang et al., 2010) and simulations (Price et al., 2010; Wu et al., 2011; Sul and Eskin, 2013), and have been adapted for unbalanced case-control studies (Zhou et al., 2018). However, LMMs tend to be considerably slower than PCA and other models, so much effort has been devoted to improving their runtime and scalability (Aulchenko et al., 2007; Kang et al., 2008; Kang et al., 2010; Zhang et al., 2010; Lippert et al., 2011; Yang et al., 2011; Listgarten et al., 2012; Zhou and Stephens, 2012; Svishcheva et al., 2012; Loh et al., 2015; Zhou et al., 2018).

An LMM variant that incorporates PCs as fixed covariates is tested thoroughly in our work. Since PCs are the top eigenvectors of the same kinship matrix estimate used in modern LMMs (Astle and Balding, 2009; Hoffman, 2013), then population structure is modeled twice in an LMM

with PCs. However, some previous work has found the apparent redundancy of an LMM with PCs beneficial (Price et al., 2010; Tucker et al., 2014), while others did not (Liu et al., 2011), and the approach continues to be used (Zeng et al., 2018). It is worth noting that early LMMs had a different arrangement, as their kinship matrices captured family relatedness only, so population structure had to be modeled separately, in practice as admixture fractions instead of PCs (Yu et al., 2006; Zhao et al., 2007).

LMM and PCA are closely related models (Astle and Balding, 2009; Hoffman, 2013), which suggests similar performance in some cases, particularly low-dimensional relatedness. Direct comparisons have yielded mixed results, with several studies finding superior performance for LMM (notably from papers promoting advances in LMMs) while many others report comparable performance (Table 1). None of these papers find that PCA outperforms LMM decisively, although PCA occasionally performs better in isolated and artificial cases or individual measures (often with unknown significance). Several patterns emerged in our literature search, which may explain these

Table 1: **Previous PCA-LMM evaluations in the literature.**

	Sim. Genotypes							
Publication	Type[a]	K[b]	F_{ST} [c]	Real[d]	Trait[e]	Power	PCs (r)	Best
Zhao et al., 2007				✓	Q	✓	8	LMM
Astle and Balding, 2009	I	3	0.10		CC	✓	10	Tie
Kang et al., 2010				✓	Both		2-100	LMM
Price et al., 2010	I, F	2	0.01		CC		1	Mixed
Wu et al., 2011	I, A	2-4	0.01		CC	✓	10	Mixed
Liu et al., 2011	S, A	2-3	R		Q	✓	10	Tie
Sul and Eskin, 2013	I	2	0.01		CC		1	Tie
Tucker et al., 2014	I	2	0.05	✓	Both	✓	5	Tie
Yang et al., 2014				✓	CC	✓	5	Tie
Song et al., 2015	S, A	2-3	R		Q		3	LMM
Loh et al., 2015				✓	Q	✓	10	LMM
Liu et al., 2016				✓	Q	✓	3-6	LMM
Sul et al., 2018				✓	Q		100	LMM
This work	A, T, F	10-243	$\leq$0.25	✓	Q	✓	0-90	LMM

[a]Genotype simulation types. I: Independent subpopulations; S: subpopulations (with parameters drawn from real data); A: Admixture; T: Tree; F: Family.
[b]Model dimensionality (number of subpopulations or ancestries)
[c]R: simulated parameters based on real data, F_{ST} not reported.
[d]Evaluations using unmodified real genotypes.
[e]Q: quantitative; CC: case-control.

discrepancies. Previous studies were generally divided into two types: those that employed exclusively simulated genotypes, versus exclusively real genotypes (only one study used both). We find that the simulated genotype studies, which tended to have low dimensionalities and differentiation (F_{ST}), were more likely to report ties or mixed results (6/7), whereas real genotypes tended to clearly favor LMMs (5/7). Similarly, 6/8 papers that use quantitative traits favor LMMs, whereas 5/7 papers that used case-control traits gave ties or mixed results (the only factor we do not explore). Other limitations of previous evaluations are that, although all measured type I error (or proxies such as inflation factors or QQ plots), a large fraction (5/13) did not measure power (including proxies such as ROC curves), and only two reported trying more than one number of PCs when evaluating PCA. Lastly, no consensus has emerged as to why LMM might outperform PCA or vice versa (Price et al., 2010; Sul and Eskin, 2013; Price et al., 2013; Hoffman, 2013), or which features of the real datasets are critical for the LMM advantage, with the exception that LMM handles family relatedness whereas PCA does not, leaving users confused as to when it is appropriate to use PCA. To better evaluate PCA and LMM, our work includes more complex, high-dimensional genotype simulations with differentiation matching that of multiethnic human cohorts, as well as real genotypes, we vary the number of PCs, and consistently measure robust proxies for type I error and power.

In this work, we study the performance of the PCA and LMM association models, characterizing their behavior under various numbers of PCs (included in LMM too). We use genotype simulations (admixture, family, and tree models) and three real datasets: the 1000 Genomes Project (Consortium, 2010; 1000 Genomes Project Consortium et al., 2012), the Human Genome Diversity Panel (HGDP) (Cann et al., 2002; Rosenberg et al., 2002; Bergström et al., 2020), and Human Origins (Patterson et al., 2012; Lazaridis et al., 2014; Lazaridis et al., 2016; Skoglund et al., 2016). We simulate quantitative traits from two models: fixed effect sizes (FES; coefficients inverse to allele frequency) that matches real data (Park et al., 2011; Zeng et al., 2018; O'Connor et al., 2019) and corresponds to high pleiotropy and strong balancing selection (Simons et al., 2018) and strong negative selection (Zeng et al., 2018; O'Connor et al., 2019), which are appropriate assumptions for diseases; and random coefficients (RC; independent of allele frequency) that corresponds to neutral

traits (Zeng et al., 2018; Simons et al., 2018). Across all tests, LMM without PCs consistently performs best, and greatly outperforms PCA in the family simulation and in all real datasets. The tree simulations do not recapitulate the real data results, suggesting that family-like structure in real data is the reason for poor PCA performance. Lastly, removing up to 4th degree relatives in the real datasets recapitulates poor PCA performance, showing that the more numerous distant relatives explain the result. All together, we find that LMMs without PCs are generally a preferable association model, and present novel simulation and evaluation approaches to measure the performance of these and other genetic association approaches.

2 Materials and Methods

2.1 The complex trait model and PCA and LMM approximations

Let $x_{ij} \in \{0, 1, 2\}$ be the genotype at the biallelic locus i for individual j, which counts the number of reference alleles. Suppose there are n individuals and m loci, $\mathbf{X} = (x_{ij})$ is their $m \times n$ genotype matrix, and $\mathbf{y}$ is the length-n (column) vector of individual trait values. The additive linear model for a quantitative (continuous) trait is:

$$\mathbf{y} = \mathbf{1}\alpha + \mathbf{X}^\intercal\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{1}$ is a length-n vector of ones, α is the scalar intercept coefficient, $\boldsymbol{\beta}$ is the length-m vector of locus coefficients, $\boldsymbol{\epsilon}$ is a length-n vector of residuals, and $\intercal$ denotes matrix transposition. The residuals follow $\epsilon_j \sim \text{Normal}(0, \sigma^2)$ independently per individual j, for some σ^2. For simplicity, non-genetic covariates are omitted from this model (and the PCA and LMM counterparts) but are trivial to include without changing any of our theoretical results.

The full model of Eq. (1), which has a coefficient for each of the m loci, is overdetermined in current datasets where $m \gg n$. The PCA and LMM models, respectively, approximate the full

model fit at a single locus i:

$$\text{PCA:} \quad \mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta_i + \mathbf{U}_r\boldsymbol{\gamma}_r + \boldsymbol{\epsilon}, \tag{2}$$

$$\text{LMM:} \quad \mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta_i + \mathbf{s} + \boldsymbol{\epsilon}, \qquad \mathbf{s} \sim \text{Normal}\left(\mathbf{0}, 2\sigma_s^2\boldsymbol{\Phi}^T\right), \tag{3}$$

where $\mathbf{x}_i$ is the length-n vector of genotypes at locus i only, β_i is the locus coefficient, $\mathbf{U}_r$ is an $n \times r$ matrix of PCs, $\boldsymbol{\gamma}_r$ is the length-r vector of PC coefficients, $\mathbf{s}$ is a length-n vector of random effects, $\boldsymbol{\Phi}^T = (\varphi_{jk}^T)$ is the $n \times n$ kinship matrix conditioned on the ancestral population T, and σ_s^2 is a variance factor (do not confuse the ancestral population superscript T with the matrix transposition symbol $\intercal$). Both models condition the regression of the focal locus i on an approximation of the total polygenic effect $\mathbf{X}^\intercal\boldsymbol{\beta}$ (assumes the contribution of the test i is infinitesimal) with the same covariance structure, which is parametrized by the kinship matrix. Under the kinship model, genotypes are random variables obeying

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

where p_i^T is the ancestral allele frequency of locus i (Malécot, 1948; Wright, 1951; Jacquard, 1970; Astle and Balding, 2009). Assuming independent loci, the covariance of the polygenic effect is

$$\text{Cov}(\mathbf{X}^\intercal\boldsymbol{\beta}) = 2\sigma_s^2\boldsymbol{\Phi}^T, \qquad \sigma_s^2 = \sum_{i=1}^{m} 2p_i^T(1-p_i^T)\beta_i^2,$$

which is readily modeled by the LMM random effect $\mathbf{s}$. (The difference in mean is absorbed by the intercept.) Alternatively, considering the eigendecomposition of the kinship matrix $\boldsymbol{\Phi}^T = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\intercal$ where $\mathbf{U}$ is the $n \times n$ eigenvector matrix and $\boldsymbol{\Lambda}$ is the $n \times n$ diagonal matrix of eigenvalues. The random effect can be written as

$$\mathbf{s} = \mathbf{U}\boldsymbol{\gamma}_{\text{LMM}}, \qquad \boldsymbol{\gamma}_{\text{LMM}} \sim \text{Normal}(\mathbf{0}, 2\sigma_s^2\boldsymbol{\Lambda}),$$

which follows from the affine transformation property of multivariate normal distributions. Therefore, the PCA term $\mathbf{U}_r\boldsymbol{\gamma}_r$ can be derived from the above equation under the additional assumption

that the kinship matrix has rank r (is low dimensional) and the coefficients $\boldsymbol{\gamma}_r$ are fit without constraints. In contrast, the LMM uses all eigenvectors, while effectively shrinking their coefficients $\boldsymbol{\gamma}_{\text{LMM}}$ as all random effects models do, although these parameters are marginalized (Astle and Balding, 2009; Hoffman, 2013). Thus, LMM has fewer parameters to fit compared to PCA: ignoring the shared terms in Eq. (2) and Eq. (3), PCA has r parameters to fit (length of $\boldsymbol{\gamma}$), whereas LMMs fit only one (σ_s^2). Therefore, PCA may overfit more than LMM—and thus lose power—when r is very large and the sample size small.

The null hypothesis of $\beta_j = 0$ (no association) is evaluated in PCA with a standard regression t-test, yielding a two-sided p-value. Some PCA implementations trade this t-test for an asymptotic χ^2 test, which requires the number of parameters to be much smaller than the number of individuals. Statistical significance in LMMs has been performed with various tests, including likelihood ratio tests, F tests, score tests, and Wald tests.

2.1.1 Kinship estimates from genotype data

In practice, given the kinship model of Eq. (4), the kinship matrix used for PCA and LMM is estimated with the method-of-moments formula applied to standardized genotypes $\mathbf{X}_S$:

$$\mathbf{X}_S = \left(\frac{x_{ij} - 2\hat{p}_i^T}{\sqrt{4\hat{p}_i^T \left(1 - \hat{p}_i^T\right)}} \right), \qquad \hat{\boldsymbol{\Phi}}^T = \frac{1}{m} \mathbf{X}_S^\intercal \mathbf{X}_S, \tag{5}$$

where the unknown p_i^T is replaced by the estimate $\hat{p}_i^T = \frac{1}{2n} \sum_{j=1}^{n} x_{ij}$, as well as minor variations (Price et al., 2006; Kang et al., 2008; Kang et al., 2010; Yang et al., 2011; Zhou and Stephens, 2012; Loh et al., 2015; Sul et al., 2018; Zhou et al., 2018). However, this kinship estimator has a complex bias that differs for every individual pair, which arises due to the use of the estimated $\hat{p}_i^T$ (Ochoa and Storey, 2021; Ochoa and Storey, 2019). Nevertheless, in PCA and LMM these biased estimates perform as well as using unbiased estimates, an observation that will be explored in future work (data not shown).

2.1.2 PCA and LMM implementations

We selected fast and robust software implementing the basic PCA and LMM models based on internal benchmarks.

PCA association was performed with `plink2` (Chang et al., 2015). The quantitative trait association model is a linear regression with covariates, evaluated using the t-test. PCs were calculated with `plink2`, which equal the top eigenvectors of Eq. (5) after removing loci with MAF < 0.1.

LMM association was performed using GCTA (Yang et al., 2011; Yang et al., 2014). Kinship equals Eq. (5) except self-kinship uses a different formula. PCs were calculated using GCTA from its kinship estimate. Association significance is evaluated with a score test. GCTA with large numbers of PCs in the small simulation only had convergence and singularity errors in some replicates, where SRMSD_p and AUC_{PR} were treated as missing.

2.2 Simulations

Below we use the notation f_A^B for the inbreeding coefficient of a subpopulation A from another subpopulation B ancestral to A. In the special case of the *total* inbreeding of A, f_A^T, T is an overall ancestral population (ancestral to every individual under consideration, such as the most recent common ancestor (MRCA) population).

2.2.1 Genotype simulation from the admixture model

We consider three admixture simulation scenarios: Large, Small, and Family. The basic admixture model is as described previously (Ochoa and Storey, 2021) and is implemented in the R package `bnpsd`. Large and Family have $n = 1,000$ individuals, while Small has $n = 100$. The number of loci is $m = 100,000$. Individuals are admixed from $K = 10$ intermediate subpopulations, or ancestries. Each subpopulation S_u ($u \in \{1, ..., K\}$) lies at coordinate u and has an inbreeding coefficient $f_{S_u}^T = u\tau$ for some τ. Ancestry proportions q_{ju} for individual j and subpopulation S_u arise from a random walk model on the given 1-dimensional geography with spread σ, and the free parameters τ and σ are fit to result in $F_{\text{ST}} = 0.1$ and mean kinship $\bar{\theta}^T = 0.5F_{\text{ST}}$ for the admixed individuals, as before (Ochoa and Storey, 2021). 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 Beta is the Balding-Nichols distribution with mean p_i^T and variance $p_i^T\left(1 - p_i^T\right) f_{S_u}^T$ (Balding and Nichols, 1995). Fixed loci (i where $x_{ij} = 0$ for all j, or $x_{ij} = 2$ for all j) are drawn again from the model, starting from p_i^T, iterating until no loci are fixed. Each replicate draws a new genotype matrix starting from p_i^T.

As a brief aside, we prove that global ancestry proportions as covariates is equivalent in expectation to using PCs under the admixture model. Note that the latent space of $\mathbf{X}$, given by (π_{ij}), has K dimensions (number of columns of $\mathbf{Q} = (q_{ju})$), so the top K PCs span this space. Since associations include an intercept term ($\mathbf{1}\alpha$ in Eq. (2)), estimated PCs are orthogonal to $\mathbf{1}$ (note $\hat{\mathbf{\Phi}}^T\mathbf{1} = \mathbf{0}$ because $\mathbf{X}_S\mathbf{1} = \mathbf{0}$), and the sum of rows of $\mathbf{Q}$ sums to one, then only $K - 1$ PCs (plus intercept) are needed to span the latent space of this admixture model.

2.2.2 Genotype simulation from random admixed families

We simulated a pedigree with admixed founders, no close relative pairings, assortative mating based on a 1D geography (to preserve admixture structure), and arbitrary numbers of generations (20 in this work). This simulation is implemented in the R package `simfam`. Generations are drawn iteratively. Generation 1 has $n = 1000$ individuals from the above admixture simulation ordered by their 1D geography. The local kinship matrix measures the pedigree relatedness; in the first generation, everybody is locally unrelated and outbred. Individuals are randomly assigned to male or female. In the next generation, individuals are paired iteratively, removing random males from

the pool of available males and pairing them with the nearest available female with local kinship $< 1/4^3$ (stay unpaired if there are no matches), until there are no more available males or females. Let $n = 1000$ be the desired population size, $n_m = 1$ the minimum number of children and n_f the number of families (paired parents) in the current generation, then the number of additional children (beyond the minimum) is drawn from Poisson$(n/n_f - n_m)$. Let δ be the difference between desired and current population sizes. If $\delta > 0$, then δ random families are incremented by 1. If $\delta < 0$, then $|\delta|$ random families with at least $n_m + 1$ children are decremented by 1. If $|\delta|$ exceeds the number of families, all families are incremented or decremented as needed and the process is iterated. Children are assigned sex randomly, and are reordered by the average coordinate of their parents. Children draw alleles from their parents independently per locus. A new random pedigree is drawn for each replicate, as well as new founder genotypes from the admixture model.

2.2.3 Genotype simulation from a tree model

This model draws subpopulations allele frequencies from a hierarchical model parametrized by a tree, which is also implemented in `bnpsd` and relies on `ape` for general tree data structures and methods (Paradis and Schliep, 2019). The ancestral population T is the root, and each node is a subpopulation S_w indexed arbitrarily. Each edge between S_w and its parent population P_w has an inbreeding coefficient $f_{S_w}^{P_w}$. p_i^T are drawn from a given distribution (constructed to mimic each real dataset in Appendix A). Given the allele frequencies $p_i^{P_w}$ of the parent population, S_w's allele frequencies are drawn from:

$$p_i^{S_w}|p_i^{P_w} \sim \text{Beta}\left(p_i^{P_w}\left(\frac{1}{f_{S_w}^{P_w}} - 1\right), \left(1 - p_i^{P_w}\right)\left(\frac{1}{f_{S_w}^{P_w}} - 1\right)\right).$$

Individuals j in S_w draw genotypes from its allele frequency: $x_{ij}|p_i^{S_w} \sim \text{Binomial}\left(2, p_i^{S_w}\right)$. Loci with MAF $= \min\left\{\hat{p}_i^T, 1 - \hat{p}_i^T\right\} < 0.01$ are drawn again starting from the p_i^T distribution, iterating until no such loci remain.

2.2.4 Fitting tree to real data

We developed new methods to fit trees to real data based on unbiased kinship estimates from `popkin`, implemented in `bnpsd`. A tree with given inbreeding edges $f_{S_w}^{P_w}$ gives rise to a coancestry matrix ϑ_{uv}^T for a subpopulation pair (S_u, S_v), and the goal is to recover the inbreeding edges from coancestry estimates. Coancestry values are total inbreeding coefficients of the MRCA population of each subpopulation pair. Therefore, we calculate $f_{S_w}^T$ for every S_w recursively from the root as follows. Nodes with parent $P_w = T$ are already as desired. Given $f_{P_w}^T$, the desired $f_{S_w}^T$ is calculated via the additive edge δ_w (Ochoa and Storey, 2021):

$$f_{S_w}^T = f_{P_w}^T + \delta_w, \qquad \delta_w = f_{S_w}^{P_w}\left(1 - f_{P_w}^T\right). \tag{6}$$

These $\delta_w \geq 0$ because $0 \leq f_{S_w}^{P_w}, f_{P_w}^T \leq 1$ for every w. Inbreeding edges can be recovered from additive edges: $f_{S_w}^{P_w} = \delta_w/(1 - f_{P_w}^T)$. Overall, coancestry values are sums of δ_w over common ancestor nodes,

$$\vartheta_{uv}^T = \sum_w \delta_w I_w(u, v), \tag{7}$$

where the sum includes all w, and $I_w(u, v)$ equals 1 if S_w is a common ancestor of S_u, S_v, 0 otherwise. Note that $I_w(u, v)$ reflects tree topology and δ_w edge values.

To estimate population-level coancestry, first kinship ($\hat{\varphi}_{jk}^T$) is estimated using `popkin` (Ochoa and Storey, 2021). Individual coancestry ($\hat{\theta}_{jk}^T$) is estimated from kinship using

$$\hat{\theta}_{jk}^T = \begin{cases} \hat{\varphi}_{jk}^T & \text{if} \quad k \neq j, \\ \hat{f}_j^T = 2\hat{\varphi}_{jj}^T - 1 & \text{if} \quad k = j. \end{cases} \tag{8}$$

Lastly, coancestry $\hat{\vartheta}_{uv}^T$ between subpopulations are averages of individual coancestry values:

$$\hat{\vartheta}_{uv}^T = \frac{1}{|S_u||S_v|} \sum_{j \in S_u} \sum_{k \in S_v} \hat{\theta}_{jk}^T.$$

Topology is estimated with hierarchical clustering using the weighted pair group method with

arithmetic mean (Sokal and Michener, 1958), with distance function $d(S_u, S_v) = \max\left\{\hat{\vartheta}^T_{uv}\right\} - \hat{\vartheta}^T_{uv}$, which succeeds due to the monotonic relationship between node depth and coancestry (Eq. (7)). This algorithm recovers the true topology from the true coancestry values, and performs well for estimates from genotypes.

To estimate tree edge lengths, first δ_w are estimated from $\hat{\vartheta}^T_{uv}$ and the topology using Eq. (7) and non-negative least squares linear regression (Lawson and Hanson, 1974) (implemented in `nnls`; Mullen and Stokkum, 2012) to yield non-negative δ_w, and $f^{P_w}_{S_w}$ are calculated from δ_w by reversing Eq. (6). To account for small biases in coancestry estimation, an intercept term δ_0 is included ($I_0(u, v) = 1$ for all u, v), and when converting δ_w to $f^{P_w}_{S_w}$, δ_0 is treated as an additional edge to the root, but is ignored when drawing allele frequencies from the tree.

2.2.5 Trait Simulation

Traits are simulated from the quantitative trait model of Eq. (1), with novel bias corrections for simulating the desired heritability from real data relying on the unbiased kinship estimator `popkin` (Ochoa and Storey, 2021). This simulation is implemented in the R package `simtrait`. All simulations have a narrow-sense heritability of $h^2 = 0.8$ and $\epsilon_j \sim \text{Normal}(0, 1 - h^2)$. To balance power while varying n, the number of causal loci is $m_1 = n/10$. The set of causal loci C is drawn anew for each replicate, from loci with MAF ≥ 0.01 to avoid rare causal variants (inappropriate for PCA and LMM). Letting $v^T_i = p^T_i\left(1 - p^T_i\right)$, the effect size of locus i equals $2v^T_i\beta^2_i$, its contribution of the trait variance (Park et al., 2010). Under the *fixed effect sizes* (FES) model, initial causal coefficients are

$$\beta_i = \frac{1}{\sqrt{2v^T_i}}$$

for known p^T_i; otherwise v^T_i is replaced by the unbiased estimator (Ochoa and Storey, 2021) $\hat{v}^T_i = \hat{p}^T_i\left(1 - \hat{p}^T_i\right)/(1 - \bar{\varphi}^T)$, where $\bar{\varphi}^T$ is the mean kinship estimated with `popkin`. Each causal locus is multiplied by -1 with probability 0.5. Alternatively, under the *random coefficients* (RC) model, initial causal coefficients are drawn independently from $\beta_i \sim \text{Normal}(0, 1)$. For both models, the initial genetic variance is $\sigma^2_0 = \sum_{i \in C} 2v^T_i\beta^2_i$, replacing v^T_i with $\hat{v}^T_i$ for unknown p^T_i (so σ^2_0 is an

unbiased estimate), so we multiply every initial β_i by $\frac{h}{\sigma_0}$ to have the desired heritability. Lastly, for known p_i^T, the intercept coefficient is $\alpha = -\sum_{i \in C} 2p_i^T \beta_i$. When p_i^T are unknown, $\hat{p}_i^T$ should not replace p_i^T since that distorts the trait covariance (for the same reason the standard kinship estimator in Eq. (5) is biased), which is avoided with

$$\alpha = -\frac{2}{m_1} \left(\sum_{i \in C} \hat{p}_i^T \right) \left(\sum_{i \in C} \beta_i \right).$$

2.3 Real human genotype datasets

The three datasets were processed as before (Ochoa and Storey, 2019) (summarized below), except with an additional filter so loci are in approximate linkage equilibrium and rare variants are removed. All processing was performed with `plink2` (Chang et al., 2015), and analysis was uniquely enabled by the R packages `BEDMatrix` (Grueneberg and Campos, 2019) and `genio`. Each dataset groups individuals in a two-level hierarchy, which we call continental and fine-grained subpopulations, respectively. Final dataset sizes are in Table 2.

We obtained the full (including non-public) Human Origins by contacting the authors and agreeing to their usage restrictions. The Pacific data (Skoglund et al., 2016) was obtained separately from the rest (Lazaridis et al., 2014; Lazaridis et al., 2016), and datasets were merged using the intersection of loci. We removed ancient individuals, and individuals from singleton and non-native subpopulations. Non-autosomal loci were removed. Our analysis of the WGS version of HGDP (Bergström et al., 2020) was restricted to autosomal biallelic SNP loci with filter "PASS". Our analysis of the high-coverage NYGC version of 1000 Genomes (Fairley et al., 2020) was restricted to autosomal biallelic SNP loci with filter "PASS".

Since our evaluations assume uncorrelated loci, we filtered each dataset with `plink2` using parameters "`--indep-pairwise 1000kb 0.3`", which iteratively removes loci that have a greater than 0.3 correlation coefficient with another locus that is within 1000kb, stopping until no such loci remain. Since all real datasets have numerous rare variants, while PCA and LMM are not able to detect associations involving rare variants, we removed all loci with MAF < 0.01. Kinship rank and eigenvalues were calculated from `popkin` kinship estimates. Eigenvalues were assigned p-values with

`twstats` of the Eigensoft package (Patterson et al., 2006), and kinship rank was estimated as the largest number of consecutive eigenvalue from the start that all satisfy $p < 0.01$ (p-values did not increase monotonically). For the evaluation with close relatives removed, each dataset was filtered with `plink2` with option "`--king-cutoff`" with cutoff 0.02209709 ($= 2^{-11/2}$) for removing up to 4th degree relatives using KING-robust (Manichaikul et al., 2010), and MAF < 0.01 is reapplied (Table S1).

2.4 Evaluation of performance

All approaches are evaluated in two orthogonal dimensions: SRMSD_p quantifies p-value uniformity, and AUC_{PR} measures causal locus classification performance and reflects power while ranking miscalibrated models fairly. These measures are more robust alternatives to previous measures from the literature (see Appendix B), and are implemented in `simtrait`.

P-values for continuous test statistics have a uniform distribution when the null hypothesis holds, a crucial assumption for type I error and FDR control (Storey, 2003; Storey and Tibshirani, 2003). We use the Signed Root Mean Square Deviation (SRMSD_p) to measure 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, here 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). Thus, $\text{SRMSD}_p = 0$ corresponds to calibrated p-values, $\text{SRMSD}_p > 0$ indicate anti-conservative p-values, and $\text{SRMSD}_p < 0$ are conservative p-values. The maximum SRMSD_p is achieved when all p-values are zero (the limit of anti-conservative p-values), which for infinite loci approaches

$$\text{SRMSD}_p \to \sqrt{\int_0^1 u^2 du} = \frac{1}{\sqrt{3}} \approx 0.577.$$

The same worst-case value (with negative sign) occurs for all p-values of 1.

Precision and recall are standard performance measures for binary classifiers that do not require calibrated p-values (Grau et al., 2015). Given the total numbers of true positives (TP), false positives (FP) and false negatives (FN) at some threshold or parameter t, precision and recall are

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

Precision and Recall trace a curve as t is varied, and the area under this curve is AUC_{PR}. We use the R package `PRROC` to integrate the correct non-linear piecewise function when interpolating between points. A model obtains the maximum $\text{AUC}_{\text{PR}} = 1$ if there is a t that classifies all loci perfectly. In contrast, the worst models, which classify at random, have an expected precision ($= \text{AUC}_{\text{PR}}$) equal to the overall proportion of causal loci: $\frac{m_1}{m}$.

3 Results

The success of our investigation hinges on simulating a variety of population structures and quantitative trait models, introduced first, which have the goal of capturing all the essential features present in genetically diverse human studies, and compared directly to evaluations based on real genotypes. Then we summarize the evaluation methods and present the results.

3.1 Overview of genotype simulations and real datasets

We utilized three real genotype datasets and simulated genotypes from six population structure scenarios to cover various features of interest (Table 2). We will introduce them here in sets of three, as they appear in the rest of our results. The population structures are also conveniently visualized in Fig. 1 using `popkin` to estimate population kinship matrices (which combine family and population relatedness) without bias (Ochoa and Storey, 2021).

The first set of three simulated genotypes are based on an admixture model from 10 subpopulations (Fig. 1A) (Ochoa and Storey, 2021; Gopalan et al., 2016; Cabreros and Storey, 2019). The "large" version has 1000 individuals and illustrates asymptotic performance, while the "small"

simulation has 100 individuals to illustrate model overfitting. The "family" simulation has admixed founders and draws a 20-generation random pedigree with assortative mating, resulting in a complex joint family and ancestry structure in the last generation (Fig. 1B).

The second set of three are the real human datasets: Human Origins (Fig. 1D), HGDP (Fig. 1G), and 1000 Genomes (Fig. 1J). All of these represent global human diversity with varying resolutions, making them of great interest as representatives of proposed multiethnic studies. These datasets had loci filtered to avoid linkage disequilibrium, to simplify our evaluation, and are enriched for small minor allele frequencies, even after excluding rare variants (MAF $< 1\%$; Fig. 1C).

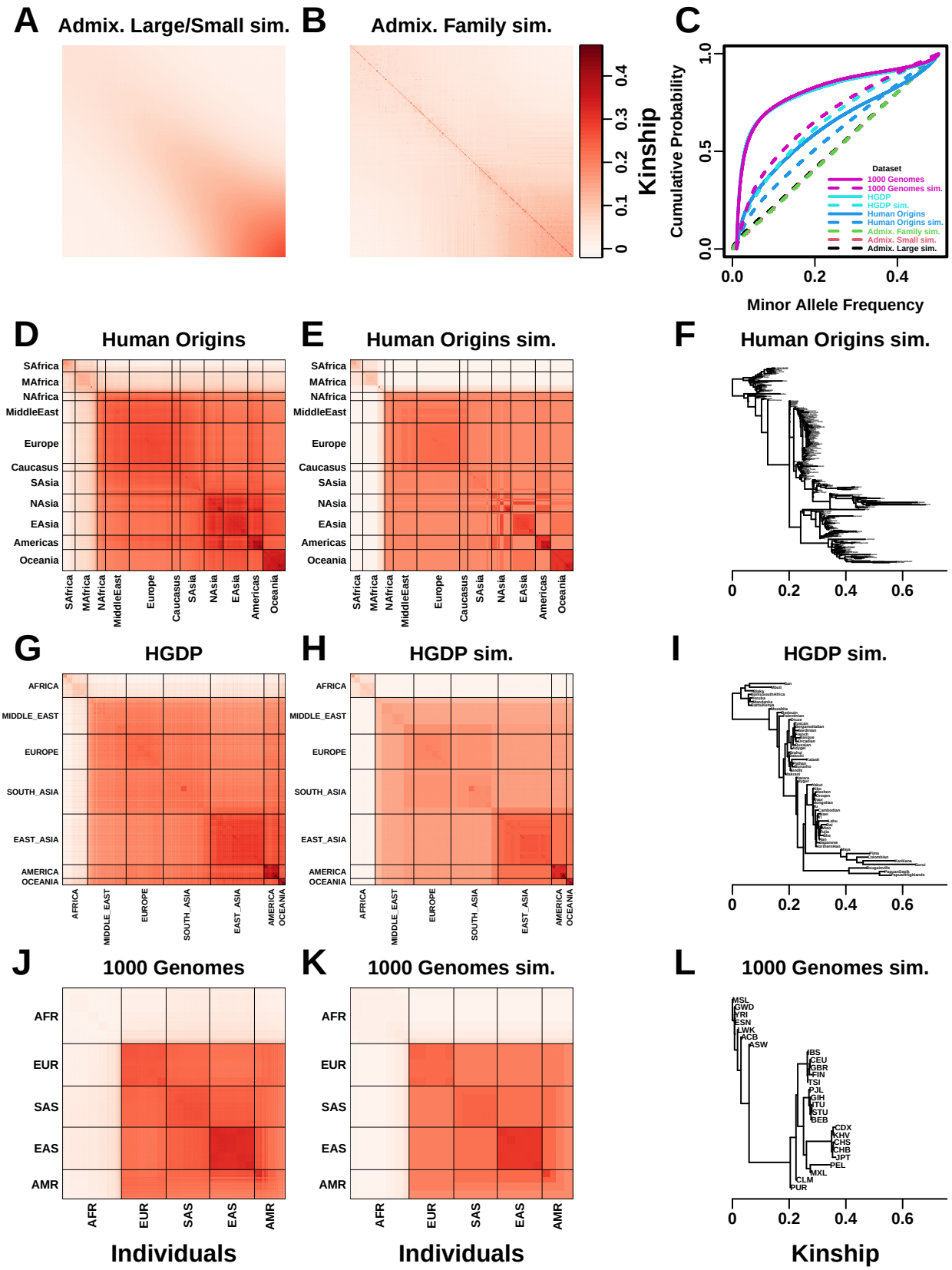
Last are tree simulations (Fig. 1F,I,L), fit to the kinship of each real human dataset (Fig. 1E,H,K), and used to draw genotypes. Ancestral allele frequencies were constructed to mimic the real allele

Table 2: **Features of simulated and real human genotype datasets.**

Dataset	Type	Loci (m)	Ind. (n)	Subpops.[a] (K)	Causal loci[b] (m_1)	F_{ST} [c]
Admix. Large sim.	Admix.	100,000	1000	10	100	0.1
Admix. Small sim.	Admix.	100,000	100	10	10	0.1
Admix. Family sim.	Admix.+Pedig.	100,000	1000	10	100	0.1
Human Origins	Real	190,394	2922	11-243	292	0.28
HGDP	Real	924,892	929	7-54	93	0.28
1000 Genomes	Real	1,111,266	2504	5-26	250	0.22
Human Origins sim.	Tree	190,394	2922	243	292	0.23
HGDP sim.	Tree	924,892	929	54	93	0.25
1000 Genomes sim.	Tree	1,111,266	2504	26	250	0.21

[a]For admixed family, ignores dimensionality of 20 generation pedigree structure. For real datasets, lower range is continental subpopulations, upper range is number of fine-grained subpopulations.
[b]$m_1 = n/10$ in all cases to balance power across dataset.
[c]Model parameter for simulations, estimated value on real datasets.

Figure 1 *(following page)*: **Population structures of simulated and real human genotype datasets.** First two columns are population kinship matrices as heatmaps: Individuals are placed along both x- and y-axes, kinship represented with color (lighter is closer to zero, darker red are higher values). Diagonal shows inbreeding values. **A.** Admixture scenario for both Large and Small simulations. **B.** Last generation of 20-generation admixed family, shows larger kinship values near diagonal corresponding to siblings, first cousins, etc. **C.** Minor allele frequency (MAF) distributions. Real datasets and tree simulations had MAF ≥ 0.01 filter. **D.** Human Origins is an array dataset from a large diversity of humans from around the world. **G.** Human Genome Diversity Panel (HGDP) is a WGS dataset from native populations around the world. **J.** 1000 Genomes Project is a WGS dataset sampling cosmopolitan populations around the world. **F,I,L.** Trees between subpopulations fit to real data, used to draw genotypes in simulations. **E,H,K.** Simulations from trees fit to the real data recapitulate structure at the subpopulation level.


A
Admix. Large/Small sim.
B
Admix. Family sim.
Kinship
0
0.1
0.2
0.3
0.4
C
Cumulative Probability
0.0
0.5
1.0
0.0
0.2
0.4
Minor Allele Frequency
Dataset
1000 Genomes
1000 Genomes sim.
HGDP
HGDP sim.
Human Origins
Human Origins sim.
Admix. Family sim.
Admix. Small sim.
Admix. Large sim.
D
Human Origins
E
Human Origins sim.
SAfrica
MAfrica
NAfrica
MiddleEast
Europe
Caucasus
SAsia
NAsia
EAsia
Americas
Oceania
F
Human Origins sim.
0
0.2
0.4
0.6
G
HGDP
H
HGDP sim.
AFRICA
MIDDLE_EAST
EUROPE
SOUTH_ASIA
EAST_ASIA
AMERICA
OCEANIA
I
HGDP sim.
0
0.2
0.4
0.6
J
1000 Genomes
K
1000 Genomes sim.
AFR
EUR
SAS
EAS
AMR
Individuals
Individuals
L
1000 Genomes sim.
MSL
GWD
YRI
ESN
LWK
ACB
ASW
IBS
CEU
GBR
FIN
TSI
PJL
GIH
ITU
STU
BEB
CDX
KHV
CHS
CHB
JPT
PEL
MXL
CLM
PUR
0
0.2
0.4
0.6
Kinship

frequency skews (Fig. 1C). By design, these tree simulations do not contain any family structure.

3.2 Overview of trait simulation models

All traits in this work are simulated, starting from (real or simulated) genotypes and picking causal loci randomly. We repeated all of our tests for two additive quantitative trait models, *fixed effect sizes* (FES) and *random coefficients* (RC), which differ in how causal coefficients are constructed. In both cases coefficients are scaled to yield a heritability of 0.8.

The FES simulation selects coefficients β_i such that the effect size $2\beta_i^2 p_i^T(1-p_i^T)$ have the same value at every locus i, where p_i^T is the ancestral allele frequency of the simulation (T is the ancestral population). Our simple model captures the rough inverse relationship between coefficient and minor allele frequency that arises under strong negative and balancing selection and has been observed in numerous diseases and other traits (Park et al., 2011; Zeng et al., 2018; Simons et al., 2018; O'Connor et al., 2019), and thus the focus of our results.

The RC simulation selects coefficients randomly, independent of allele frequency, which corresponds to neutral traits (Zeng et al., 2018; Simons et al., 2018). The resulting effect size distributions are wider, which reduces association power and effective polygenicity compared to FES.

3.3 Overview of evaluations

Since our quantitative traits are simulated, true causal loci are known, resulting in known true positives, false positives, and false negatives. We employ two complementary measures: (1) SRMSD_p (p-value signed root mean square deviation) measures null p-value uniformity (closer to zero is better), and (2) AUC_{PR} (precision-recall area under the curve) measures causal locus classification performance (higher is better; Fig. 2). SRMSD_p is a more robust alternative to the common inflation factor λ and type I error measures; we found a good correspondence between λ and SRMSD_p, and determined that the threshold $\text{SRMSD}_p > 0.01$ corresponds to $\lambda > 1.06$ (Fig. S1) and thus evidence of miscalibration close to the rule of thumb of $\lambda > 1.05$ (Price et al., 2010). AUC_{PR} has been used to evaluate association models (Rakitsch et al., 2013), reflects statistical power for calibrated models (see Models and Methods), and is a more robust alternative to statistical power for miscalibrated

models. Reducing the complexity of null p-value distributions and precision-recall curves to two scalars is crucial for our extensive evaluations.

Both PCA and LMM were evaluated in each dataset including a number r of PCs as fixed covariates, varying r between 0 and 90, and every test was repeated 50 times. Our overall statistical evaluation will be summarized first, followed by detailed evaluations in each datasets in the rest of the results.

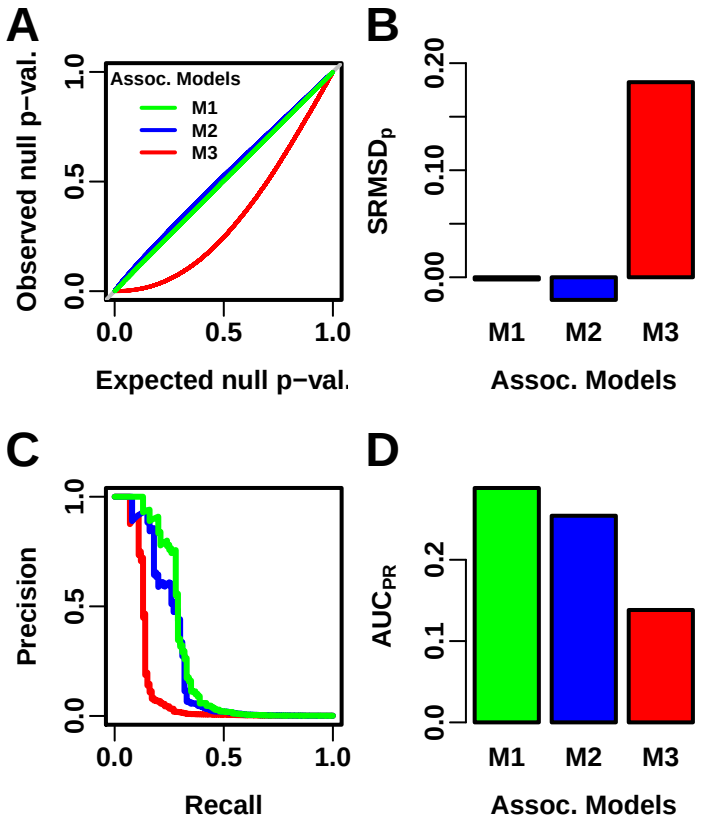


Figure 2: **Illustration of evaluation measures.** Three archetypal models illustrate our two complementary measures: M1 is well calibrated, M2 overfits slightly, and M3 does not model relatedness. **A.** QQ plot of p-values of "null" (non-causal) loci. M1 has uniform null p-values as desired (overlaps $y = x$). M2/M3 have null p-values larger/smaller than expected. **B.** SRMSD$_p$ (p-value Signed Root Mean Square Deviation) is a distance between the observed null p-values and their uniform expectations, negative if their median is larger than expected (closer to zero is better). **C.** Precision-Recall (PR) measure causal locus classification performance across thresholds without assuming calibrated p-values (higher is better). **D.** AUC$_{\text{PR}}$ (Area Under the PR Curve) reflects power (higher is better).

First we describe the null p-value uniformity $|\text{SRMSD}_p|$ results (Table 3). Here the sign of SRMSD_p was ignored, so smaller is better and Wilcoxon paired 1-tailed tests were used to determine if a suboptimal distribution was significantly different. For PCA, the optimal number of PCs r is typically large across all datasets (up to $r = 90$, the largest tested), but we found that much smaller "min" r values often performed as well (numbers in parentheses in Table 3). However, even the min r values for PCA tended to be large on the family simulation and real datasets. Most cases had a mean $|\text{SRMSD}_p| < 0.01$ (marked with asterisks), whose p-values are effectively calibrated. Mean $|\text{SRMSD}_p| > 0.01$ (miscalibrated) PCA cases were observed on the family simulation and real datasets. In contrast, for LMM, $r = 0$ (no PCs) was always the optimal choice, and was always calibrated. Lastly, comparing LMM with $r = 0$ to PCA with its best r, LMM was either always significantly better or statistically tied to PCA.

Next we turn to classification performance (AUC_{PR}; Table 3). For PCA, the best r for AUC_{PR} was always smaller than the best r for $|\text{SRMSD}_p|$, and also for the respective "min" r comparisons. Thus, for PCA there is often a tradeoff between calibrated p-values versus classification performance. For LMM there is no such tradeoff, as $r = 0$ (no PCs) resulted in AUC_{PR} distributions not significantly different from the best r in all tests except two (the min r was 2 for both 1000 Genomes simulation with FES trait and 1000 Genomes real dataset with RC trait). Lastly, LMM with its best r always had significantly greater AUC_{PR} distributions than PCA with its best r except for one statistical tie.

3.4 Evaluations in admixture simulations

Now we look more closely at the results of every individual evaluation. The SRMSD_p and AUC_{PR} distributions for the first three admixture simulations and FES traits are in Fig. 3. We repeated the evaluation with RC traits, which gave qualitatively similar results (Fig. S2).

In the large admixture simulation, the SRMSD_p of PCA is largest when $r = 0$ (no PCs) and decreases rapidly to near zero at $r = 3$, where it stays for up to $r = 90$ (Fig. 3A). Thus, PCA has calibrated p-values for $r \geq 3$, which is smaller than the theoretical optimum for this simulation of $r = K - 1 = 9$. In contrast, the SRMSD_p for LMM starts near zero for $r = 0$, but becomes negative

as r increases (p-values are conservative). The AUC_{PR} distribution of PCA is similarly worst at $r = 0$, increases rapidly and peaks at $r = 3$, then decreases slowly for $r > 3$, while the AUC_{PR} distribution for LMM starts near its maximum at $r = 0$, and decreases for larger r. Although the AUC_{PR} distributions for LMM and PCA overlap considerably at each r, LMM with $r = 0$ has significantly greater AUC_{PR} values than PCA with $r = 3$ (Table 3). However, qualitatively PCA closely matches LMM's performance in this simulation. Both LMM and PCA are robust to extreme values of r.

The observed robustness to large r led us to consider smaller sample sizes. Our expectation is

Table 3: **Overview of PCA and LMM evaluation results**

	Metric:	$\lvert\text{SRMSD}_p\rvert$			AUC_{PR}		
		Best (min[b]) PCs			Best (min[b]) PCs		
Dataset	Trait model[a]	PCA	LMM	Best[c]	PCA	LMM	Best[c]
Admix. Large sim.	FES	84* (3*)	0*	tie	3	3 (0)	LMM
Admix. Small sim.	FES	4* (2*)	0*	LMM	4 (1)	0	LMM
Admix. Family sim.	FES	90 (87)	0*	LMM	83 (34)	0	LMM
Human Origins	FES	90 (87)	0*	LMM	34 (9)	1 (0)	LMM
HGDP	FES	87* (34*)	0*	LMM	19 (16)	1 (0)	LMM
1000 Genomes	FES	39 (32)	0*	LMM	8	1 (0)	LMM
Human Origins sim.	FES	90* (80*)	0*	tie	47 (36)	0	LMM
HGDP sim.	FES	43* (20*)	0*	LMM	17 (15)	0	LMM
1000 Genomes sim.	FES	77* (15*)	0*	LMM	16 (6)	2	LMM
Admix. Large sim.	RC	89* (3*)	0*	tie	3	2 (0)	LMM
Admix. Small sim.	RC	8* (2*)	0*	tie (LMM)	1 (0)	0	LMM
Admix. Family sim.	RC	90 (88)	0*	LMM	74 (28)	0	LMM
Human Origins	RC	89* (79*)	0*	LMM	34 (18)	5 (0)	LMM
HGDP	RC	77* (30*)	0*	LMM	19 (13)	3 (0)	tie (LMM)
1000 Genomes	RC	37* (27*)	0*	LMM	19 (4)	9 (2)	LMM
Human Origins sim.	RC	89* (85*)	0*	tie	45 (25)	0	LMM
HGDP sim.	RC	30* (23*)	0*	LMM	18 (15)	5 (0)	LMM
1000 Genomes sim.	RC	90* (16)	0*	LMM	10 (6)	2 (0)	LMM

[a]FES: Fixed Effect Sizes, RC: Random Coefficients.
[b]Parentheses: smallest r (number of PCs) whose distribution ($\lvert\text{SRMSD}_p\rvert$ or AUC_{PR}) was not significantly different (Wilcoxon paired 1-tailed $p > 0.01$) from the r with best mean value (if any).
[c]Tie if distributions of best PCA and LMM version (previous two columns) did not differ significantly (Wilcoxon paired 1-tailed $p > 0.01$). Result was always the same whether "best" or "min" (in parenthesis) cases were compared, except in two cases in parentheses.
*r for which mean $\lvert\text{SRMSD}_p\rvert < 0.01$ ($\lvert\text{SRMSD}_p\rvert$ columns only).

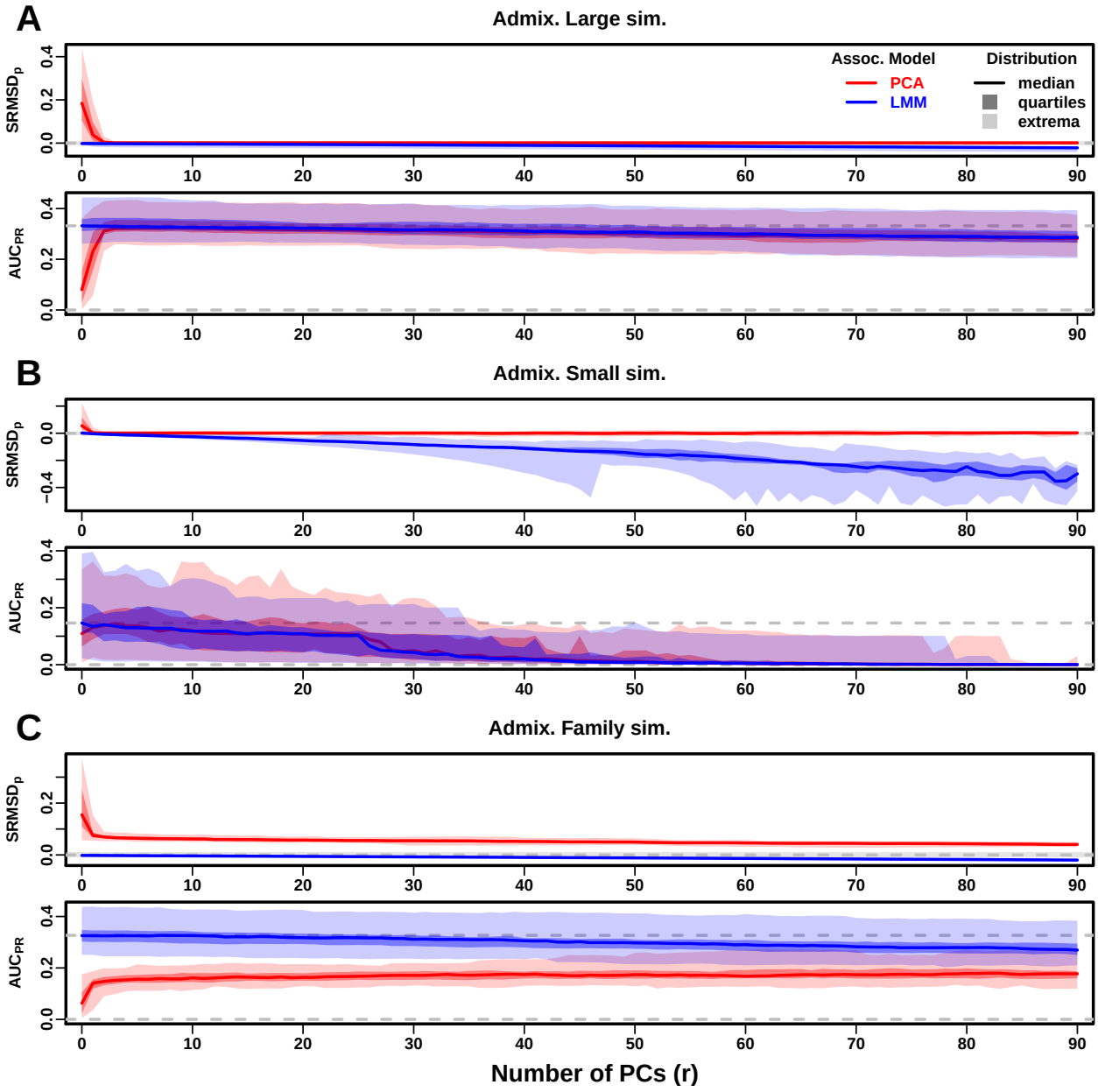


Figure 3: **Evaluations in admixture simulations.** Traits simulated from FES model. PCA and LMM models have varying number of PCs ($r \in \{0, ..., 90\}$ on x-axis), with the distributions (y-axis) of SRMSD_p (top subpanel) and AUC_{PR} (bottom subpanel) for 50 replicates. Best performance is zero SRMSD_p and large AUC_{PR}. Zero values and maximum median AUC_{PR} values marked with horizontal gray dashed lines, and $|\text{SRMSD}_p| < 0.01$ band is marked with a light gray area. LMM always performs best with $r = 0$, PCA with r between 1-4. **A.** The large simulation has $n = 1,000$ individuals. **B.** The small simulation has $n = 100$ individuals, shows overfitting for large r. **C.** The family simulation has $n = 1,000$ individuals from a family with admixed founders and large numbers of close relatives from a realistic random 20-generation pedigree. PCA performs poorly compared to LMM: $\text{SRMSD}_p > 0$ for all r and large AUC_{PR} gap.

that a model with large numbers of parameters r should overfit more as r approaches the sample size n. Rather than increase r beyond 90, which is not done in practice, we reduce individuals to $n = 100$, which is small for typical association studies but may occur in studies of rare diseases, pilot studies, or other constraints. To compensate for the loss of power due to reducing n, we also reduce the number of causal loci from 100 to $m_1 = 10$, (fixed ratio $n/m_1 = 10$) to increase per-locus effect sizes. As expected, we found a large decrease in performance for both PCA and LMM as r increases, with optimal performance attained near $r = 1$ for PCA and $r = 0$ for LMM (Fig. 3B). Remarkably, LMM attains much larger negative SRMSD_p values than in our other evaluations. While LMM with $r = 0$ is significantly better than PCA ($r = 1$ to 4) in both metrics (Table 3), qualitatively the difference is negligible.

The family simulation adds a 20-generation random family to our admixture simulation. Only the last generation is studied for association, which contains numerous siblings, first cousins, etc., with the initial admixture structure preserved by geographically-biased mating. Previous work has reported, in limited settings, that PCA performs poorly with family structure, whereas LMM is formulated in terms of kinship so it is expected to perform better. Our evaluation reveals a sizable gap in both metrics between LMM and PCA across all r (Fig. 3C). LMM again performs best with $r = 0$ and achieves mean $|\text{SRMSD}_p| < 0.01$. However, PCA does not achieve mean $|\text{SRMSD}_p| < 0.01$ at any r, and its best mean AUC_{PR} across r is considerably worse than that of LMM. Thus, LMM is conclusively superior to PCA, and the only calibrated model, when there is family structure.

3.5 Evaluations in real human genotype datasets

Next we repeat our evaluations with real human genotype data, which differs from our simulations in allele frequency distributions and more complex population structures with greater differentiation, numerous correlated subpopulations, and potential cryptic family relatedness. We chose three datasets that span global human diversity and include both array and whole genome sequencing (WGS) genotyping platforms. Loci in high linkage disequilibrium were removed to simplify our evaluation, and traits were simulated from these genotypes and FES (Fig. 4) and RC (Fig. S3) models.

Human Origins has the greatest number and diversity of subpopulations. The SRMSD_p and AUC_{PR} distributions in this dataset and FES traits (Fig. 4A) most resemble those from the family simulation (Fig. 3C). In particular, while LMM with $r = 0$ performed optimally (both metrics) and satisfies mean $|\text{SRMSD}_p| < 0.01$, PCA maintained $\text{SRMSD}_p > 0.01$ for all r and its AUC_{PR} were all considerably smaller than the best AUC_{PR} of LMM.

HGDP has the fewest individuals among real datasets, but compared to Human Origins it contains many more loci and low-frequency variants. The SRMSD_p and AUC_{PR} distributions (Fig. 4B) are intermediate between the admixture and family simulations. In particular, here both LMM ($r = 0$) and PCA ($r \geq 31$) achieve mean $|\text{SRMSD}_p| < 0.01$ (p-values are calibrated). However, there is a sizable AUC_{PR} gap between LMM and PCA. Maximum AUC_{PR} values were lowest in HGDP compared to the two other real datasets.

1000 Genomes has the fewest subpopulations but largest number of individuals per subpopulation, and is WGS. Thus, although this dataset is expected to have the simplest population structure among the real datasets, we find SRMSD_p and AUC_{PR} distributions (Fig. 4C) that again most resemble our earlier family simulation, with mean $|\text{SRMSD}_p| < 0.01$ for LMM only and large AUC_{PR} gaps between LMM and PCA.

Our results are qualitatively different for RC traits, which had smaller AUC_{PR} gaps between LMM and PCA (Fig. S3). Maximum AUC_{PR} were smaller in RC compared to FES in Human Origins and 1000 Genomes, suggesting lower power for RC traits across association models. Nevertheless, for RC traits we found LMM with $r = 0$ significantly better than PCA for all metrics in the real datasets (Table 3).

3.6 Evaluations in tree simulations fit to human data

To better understand which features of the real datasets lead to the large differences in performance between LMM and PCA, we carried out additional simulations. Human subpopulations are related roughly by trees, which induce the strongest correlations, so we fit trees to each real dataset and tested if data simulated from these trees could recapitulate our previous results (Fig. 1). These tree simulations also feature non-uniform ancestral allele frequency distributions, which recapitulated

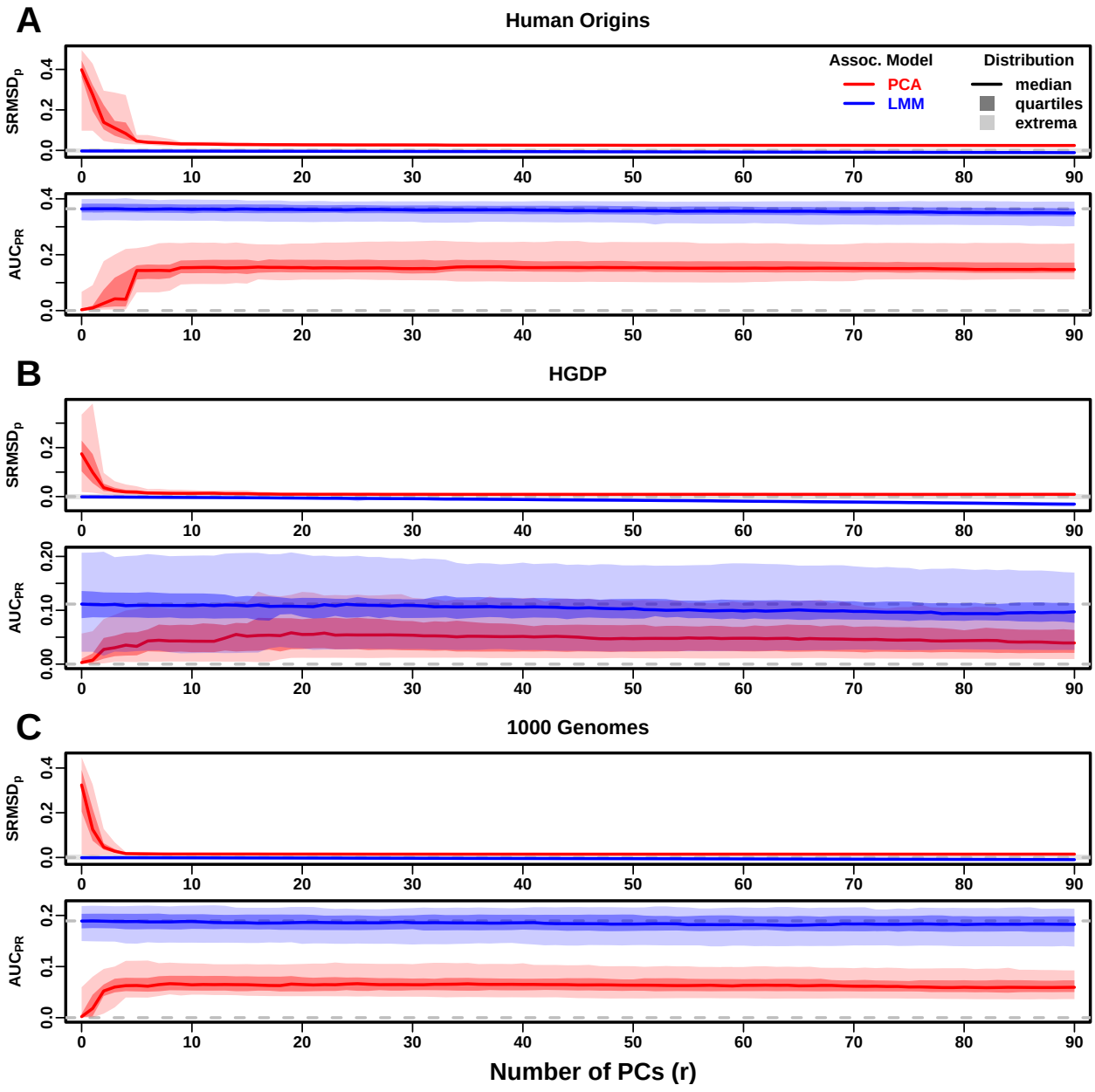


Figure 4: **Evaluations in real human genotype datasets.** Traits simulated from FES model. Same setup as Fig. 3, see that for details. These datasets strongly favor LMM with $r = 0$ PCs over PCA, with distributions that most resemble the previous admixed family simulation. **A.** The Human Origins dataset. **B.** The Human Genome Diversity Panel (HGDP) dataset. **C.** The 1000 Genomes Project dataset.

some of the skew for smaller minor allele frequencies of the real datasets (Fig. 1C).

The SRMSD_p and AUC_{PR} distributions for these tree simulations (Fig. 5) resembled our admixture simulation more than either the family simulation (Fig. 3) or real data results (Fig. 4). In all these simulations, both LMM with $r = 0$ and PCA (various r) achieve mean $|\text{SRMSD}_p| < 0.01$ (Table 3). The AUC_{PR} distributions of both LMM and PCA track closely as r is varied, although there is a small gap resulting in LMM ($r = 0$) besting PCA in all three simulations. The results are qualitatively similar for the random coefficients trait model (Fig. S4 and Table 3). Overall, these tree simulations do not recapitulate the large LMM advantage over PCA observed in the real data results.

3.7 Estimated eigenvalues do not explain PCA performance

A first-principles hypothesis for explaining PCA performance is the dimensionality of the population structure, since PCA assumes a low-dimensional genetic structure whereas LMM can model high-dimensional structures without overfitting. We applied the Tracy-Widom test (Patterson et al., 2006) with $p < 0.01$ to estimate the number of significant PCs in each dataset, which corresponds to the kinship matrix rank (Fig. S5A). These estimates slightly underestimated the true dimensionality of our simulations (Table 2), but agree that the family simulation has the greatest rank by far, and estimates greater ranks for the real datasets than their respective tree simulations. However, these estimated ranks do not differentiate datasets with good PCA performance from those with poor performance, particularly between the real and tree simulations, which span a similar range of ranks. Moreover, the 1000 Genomes rank estimate is lower than 90, yet PCA performed poorly for all $r \geq 90$ numbers of PCs tested (Fig. 4). Performance might also depend on the sample size, but these datasets do not differ greatly in size (at most 3×, excluding the small simulation), while our analysis does not reveal a line that separates datasets by PCA performance.

We also compared eigenvalues across datasets expressed as variance explained to facilitate comparisons across datasets (Fig. S5C). The top eigenvalue explained a proportion of variance proportional to F_{ST} (Table 2), but the rest of the top 10 eigenvalues show no clear differences between datasets, except the small simulation had larger variances explained per eigenvalue (as expected

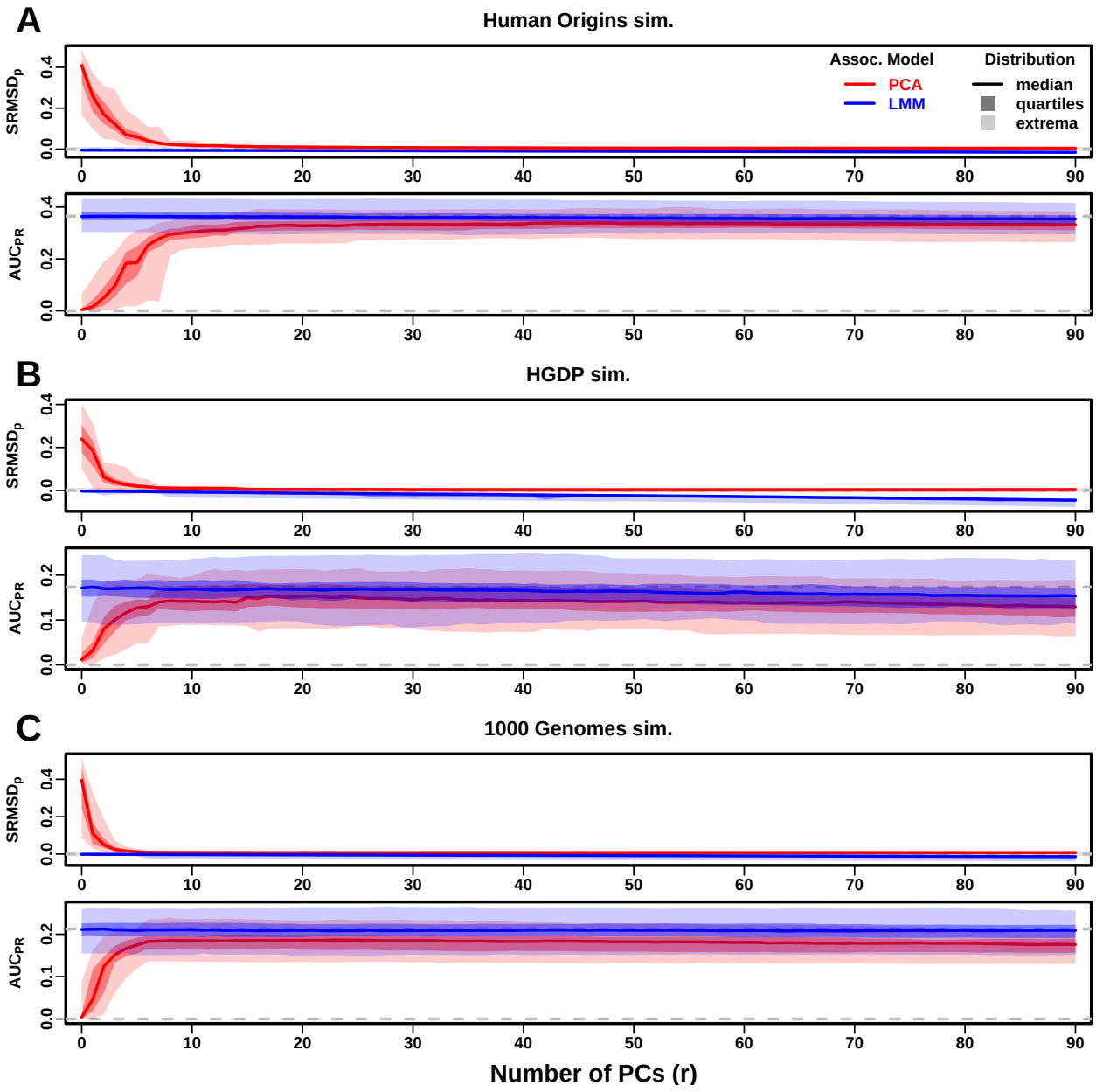


Figure 5: **Evaluations in tree simulations fit to human data.** Traits simulated from FES model. Same setup as Fig. 3, see that for details. These tree simulations, which exclude family structure by design, do not explain the large gaps in LMM-PCA performance observed in the real datasets. **A.** The Human Origins simulation. **B.** The Human Genome Diversity Panel (HGDP) simulation. **C.** The 1000 Genomes Project simulation.

since it has fewer eigenvalues). Lastly, we compared cumulative variance explained versus eigenvalue rank fraction (Fig. S5B). Each dataset has a different starting point, but all increase almost linearly from there until they reach 1, except the family simulation has much greater variance explained by mid-rank eigenvalues. There is again no clear separation between real datasets (where PCA performed poorly) from the corresponding tree simulations (where PCA performed relatively well).

3.8 Local kinship explains PCA performance

Local kinship, which is relatedness due to family structure excluding population structure, is the presumed cause of the LMM to PCA performance gap observed in real datasets but not their tree simulation counterparts. Instead of inferring local kinship through increased dimensionality, as attempted in the last section, here we measure it directly using the KING-robust estimator (Manichaikul et al., 2010). As expected, we observe more large local kinship values in the real datasets and the family simulation compared to the admixture and tree simulations (Fig. 6). However, for real datasets this distribution depends on the subpopulation structure, since locally related pairs are most likely in the same subpopulation while pairs between subpopulations tend to be negative for this estimator. Therefore, the only comparable curve to each real dataset is their corresponding tree simulation, which matches subpopulation structure.

In all real datasets we identified highly related individual pairs with kinship above the 4th degree relative threshold of 0.022 (Manichaikul et al., 2010; Conomos et al., 2016). However, these highly related pairs are vastly outnumbered by pairs who are less related but have greater than zero kinship (Fig. 6). Non-zero kinship can be inferred using the tree simulations' distribution as a null, which allowing for a small fraction of false positives suggests kinship above 0.001 or 0.01 for these datasets.

To try to improve PCA performance, we removed 4th degree relatives, which reduced sample sizes between 5% and 10% (Table S1). Only $r = 0$ for LMM and $r = 20$ for PCA were tested, as these performed well in our earlier evaluation. Only FES traits were tested, which earlier showed a large performance gap between association models. LMM significantly outperforms PCA in all these cases (Wilcoxon paired 1-tailed $p < 0.01$; Fig. 7). Notably, PCA still had miscalibrated p-values in

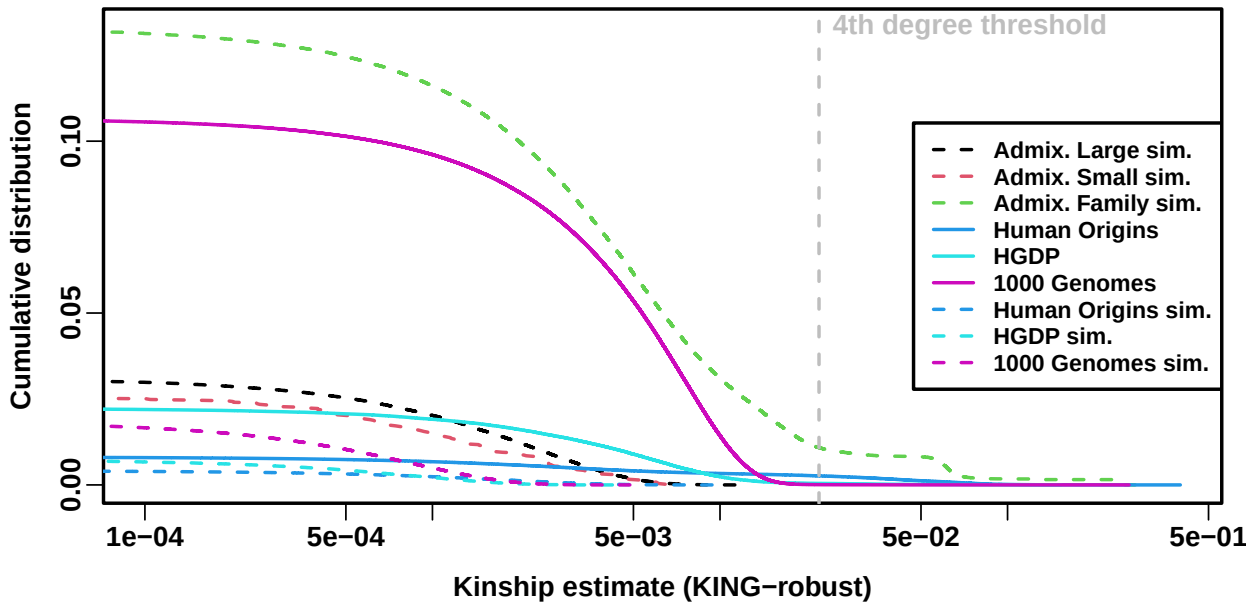


Figure 6: **Local kinship estimate distribution.** Curves are the complementary cumulative distribution of all lower triangular kinship matrix values from the KING-robust estimator. Self kinship is excluded. Note log scale of x-axis; negative estimates are counted but not shown. Most values in all datasets are below the 4th degree relative threshold value. Each real dataset has a greater cumulative than its tree simulations.

Human Origins and 1000 Genomes ($|\text{SRMSD}_p| > 0.01$). Otherwise, AUC_{PR} and SRMSD_p ranges were similar here as in the test with all individuals. Therefore, the small number of highly related individual pairs had a negligible effect in PCA performance, so the larger number of more distantly related pairs explain the poor PCA performance compared to LMM in the real datasets.

4 Discussion

Our evaluations conclusively determined that LMM without PCs performs better than PCA (for any number of PCs) across all scenarios, including all real and simulated genotypes and two trait simulation models. Although the addition of a few PCs does not greatly hurt the performance of LMM (except for small sample sizes), such additions rarely improved performance (Table 3), which

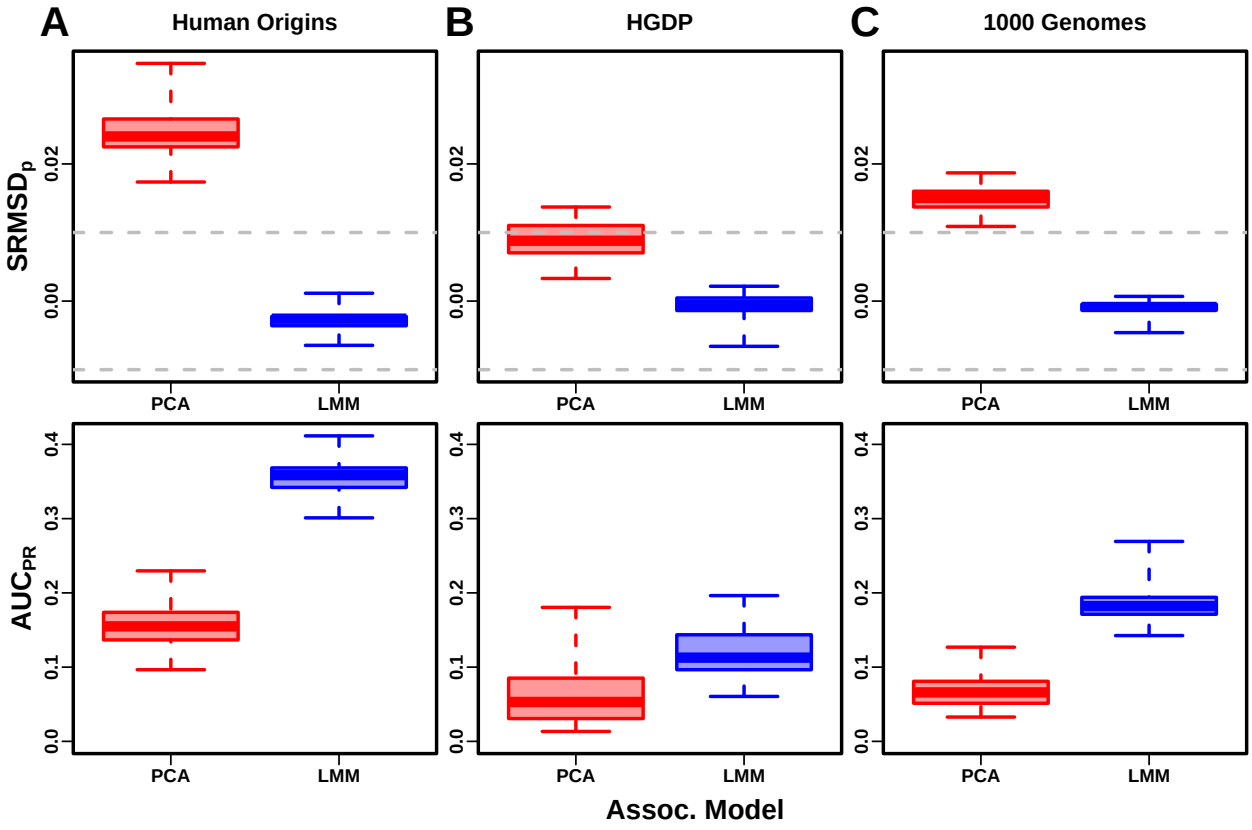


Figure 7: **LMM and PCA performance in real datasets excluding 4th degree relatives.** LMM had $r = 0$ PCs and PCA had $r = 20$. FES traits only. Each dataset is a column, rows are metrics. First row has $|\text{SRMSD}_p| < 0.01$ band marked with dashed gray lines.

agrees with previous observations (Liu et al., 2011) but contradicts others (Zhao et al., 2007; Price et al., 2010). Our findings make sense since PCs are the eigenvectors of the kinship matrix used to model the random effects, so including both is redundant.

Previous studies had found that PCA was better calibrated than LMM in a hypothetical setting, namely unusually differentiated markers (Price et al., 2010; Wu et al., 2011; Yang et al., 2014), which as simulated are an artificial scenario not based on a population genetics model, and are otherwise believed to be unusual (Sul and Eskin, 2013; Price et al., 2013). Our evaluations on real human data, which contain such loci in relevant proportions if they exist, do not replicate that result. The presence of cryptic relatedness strongly favors LMM, an advantage that probably always outweighs this potential PCA benefit in real datasets.

Relative to LMM, the behavior of PCA fell between two extremes. When PCA performed well, there was a small number of PCs with both SRMSD_p near zero and AUC_{PR} near that of LMM without PCs. Conversely, when PCA performed poorly, no number of PCs had either acceptably low SRMSD_p or acceptably large AUC_{PR}. PCA performed well in the admixture simulations (without families, both trait models), real human genotypes with RC traits, and, to a lesser extent, the tree simulations (both trait models). Conversely, PCA performed poorly in the admixed family simulation (both trait models) and the real human genotypes with FES traits.

PCA assumes that genetic structure is low-dimensional, whereas LMM can handle high-dimensional structures. Thus, PCA performs well in the admixture simulation, which is explicitly low-dimensional (see Models and Methods), and our tree simulations, which had few nodes with long branches so a low-dimensional approximation suffices. Conversely, PCA performs poorly under family structure because its kinship matrix is high-dimensional (Fig. S5). One theoretical inconvenience is that true kinship matrices are always full rank: for example, an unstructured population has a kinship matrix of $\mathbf{I}/2$, whose eigenvalues are all equal to $1/2$ (none are zero). Population structure induces an unbalanced eigenvalue distribution with few large eigenvalues (Fig. S5), so in practice dimensionality is given by the number of eigenvalues above some small threshold. Estimating the dimensionality of real datasets is challenging because estimated kinship matrices have noisy eigenvalues with biased distributions. We used the Tracy-Widom test to estimate dimensionality (Patterson et al., 2006),

which slightly underestimates the dimensionality of our simulations (an expected loss of power to detect eigenvalues which are also less important in our association tests). Estimated eigenvalues alone do not predict when PCA will perform poorly. Estimated local kinship finds considerable cryptic relatedness in all real human datasets, better explaining why LMM outperforms PCA there. The trait model also influences the relative performance of PCA, so genotype-only eigenvalues or local kinship alone do not tell the full story.

The real human genotype results, which are the most relevant in practice, suggests that PCA is at best underpowered relative to LMMs, and at worst miscalibrated regardless of the numbers of PCs included. Among our simulations, such poor performance occurred only in the admixed family simulation. Local kinship estimates reveal considerable family relatedness in the real datasets absent in the corresponding tree simulations. Admixture is absent in our tree simulations, but our simulations and theory show that admixture is handled well by PCA. Hundreds of close relative pairs have been identified in 1000 Genomes (Gazal et al., 2015; Al-Khudhair et al., 2015; Fedorova et al., 2016; Schlauch et al., 2017), but their removal does not improve PCA performance sufficiently in our tests, so the larger number of more distantly related pairs are PCA's most serious obstacle in practice. Distant relatives are expected to be numerous in any large human dataset (Henn et al., 2012; Shchur and Nielsen, 2018). Furthermore, our FES trait tests show that the challenges of cryptic relatedness are exacerbated when rarer variants have larger coefficients. Overall, the high dimensionality induced by cryptic relatedness is the key challenge for PCA association in modern datasets that is readily overcome by LMM.

Our tests also found PCA robust to large numbers of PCs, far beyond the optimal choice, agreeing with previous anecdotal observations (Price et al., 2006; Kang et al., 2010), in contrast to using too few PCs for which there is a large performance penalty. The exception was the small sample size simulation, where only small numbers of PCs performed well. In contrast, LMM is simpler since there is no need to choose the number of PCs. However, an LMM with a large number of covariates may have conservative p-values (as we observed for LMM with large numbers of PCs), a weakness of the asymptotic test used by this LMM that can be overcome with a more accurate test such as the t-test used by PCA. Simulations or post hoc evaluations remain crucial (for any

association model) to ensure that statistics are calibrated.

The largest limitation of our work is that we only considered quantitative traits. We noted that previous evaluations involving case-control traits tended to report PCA-LMM ties or mixed results, but also tended to employ low-dimensional simulations which do not feature cryptic relatedness (Table 1). An additional concern in these studies is case-control ascertainment bias, which appears to affect LMMs more severely although recent work appears to solve this problem (Yang et al., 2014; Zhou et al., 2018). Future work should aim to ask these questions in the context of our new genotype simulations and real datasets, to ensure that previous results were not biased in favor of PCA by employing unrealistic low-dimensional genotype simulations, or by not simulating large coefficients for rare variants that are predicted for diseases by various selection models.

Overall, our results lead us to always recommend LMM over PCA for association studies. Although PCA offer flexibility and speed compared to LMM, in practice much additional work is required to ensure that PCA is adequate, including removal of close relatives (lowering sample size and wasting resources) followed by simulations or other evaluations of statistics, and even then there is no guarantee that PCA will perform nearly as well as LMM, in terms of both type I error control and power. The large numbers of distant relatives expected of any real dataset all but ensures that PCA will perform poorly in practice compared to LMM for association studies. Our findings also suggest that related applications such as polygenic models may enjoy gains in power and accuracy by employing an LMM instead of PCA to model relatedness (Rakitsch et al., 2013; Qian et al., 2020). PCA remains indispensable across population genetics, from visualizing population structure and performing quality control to its deep connection to admixture models, but the time has come to limit its use in association testing in favor of LMM or other, richer models capable of modeling all forms of relatedness.

5 Appendices

5.1 Appendix A: Fitting ancestral allele frequency distribution to real data

We calculated $\hat{p}_i^T$ distributions of each real dataset. However, differentiation increases the variance of $\hat{p}_i^T$ relative to the true p_i^T (Ochoa and Storey, 2021). We present a new algorithm for constructing an "undifferentiated" distribution based on the input data but with the lower variance of the true ancestral distribution. Suppose the p_i^T distribution over loci i satisfies $\mathrm{E}\left[p_i^T\right] = \frac{1}{2}$ and $\mathrm{Var}\left(p_i^T\right) = V^T$. The sample allele frequency $\hat{p}_i^T$, conditioned on p_i^T, satisfies

$$\mathrm{E}\left[\hat{p}_i^T \middle| p_i^T\right] = p_i^T, \qquad \mathrm{Var}\left(\hat{p}_i^T \middle| p_i^T\right) = p_i^T\left(1 - p_i^T\right)\bar{\varphi}^T,$$

where $\bar{\varphi}^T = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}\varphi_{jk}^T$ is the mean kinship over all individual (Ochoa and Storey, 2021). The unconditional moments of $\hat{p}_i^T$ follow from the laws of total expectation and variance: $\mathrm{E}\left[\hat{p}_i^T\right] = \frac{1}{2}$ and

$$W^T = \mathrm{Var}\left(\hat{p}_i^T\right) = \bar{\varphi}^T\frac{1}{4} + \left(1 - \bar{\varphi}^T\right)V^T.$$

Since $V^T \leq \frac{1}{4}$ and $\bar{\varphi}^T \geq 0$, then $W^T \geq V^T$. Thus, the goal is to construct a new distribution with the original, lower variance of

$$V^T = \frac{W^T - \frac{1}{4}\bar{\varphi}^T}{1 - \bar{\varphi}^T}. \tag{9}$$

We use the unbiased estimator $\hat{W}^T = \frac{1}{m}\sum_{i=1}^{m}(\hat{p}_i^T - \frac{1}{2})^2$, while $\bar{\varphi}^T$ is calculated from the tree parameters: the subpopulation coancestry matrix (Eq. (7)), expanded from subpopulations to individuals, the diagonal converted to kinship (reversing Eq. (8)), and the matrix averaged. However, since our model ignores the MAF filters imposed in our simulations, $\bar{\varphi}^T$ was adjusted. For Human Origins the true model $\bar{\varphi}^T$ of 0.143 was used. For 1000 Genomes and HGDP the true $\bar{\varphi}^T$ are 0.126 and 0.124, respectively, but 0.4 for both produced a better fit.

Lastly, we construct new allele frequencies,

$$p' = w\hat{p}_i^T + (1 - w)q,$$

by a weighted average of $\hat{p}_i^T$ and $q \in (0,1)$ drawn independently from a different distribution. $\mathrm{E}[q] = \frac{1}{2}$ is required to have $\mathrm{E}[p'] = \frac{1}{2}$. The resulting variance is

$$\mathrm{Var}(p') = w^2 W^T + (1-w)^2 \mathrm{Var}(q),$$

which we equate to the desired V^T (Eq. (9)) and solve for w. For simplicity, we set $\mathrm{Var}(q) = V^T$, which is achieved with:

$$q \sim \mathrm{Beta}\left(\frac{1}{2}\left(\frac{1}{4V^T} - 1\right), \frac{1}{2}\left(\frac{1}{4V^T} - 1\right)\right).$$

Although $w = 0$ yields $\mathrm{Var}(p') = V^T$, we use the second root of the quadratic equation to use $\hat{p}_i^T$:

$$w = \frac{2V^T}{W^T + V^T}.$$

5.2 Appendix B: comparisons between SRMSD_p, $\mathrm{AUC}_{\mathrm{PR}}$, and evaluation measures from the literature

5.2.1 The inflation factor λ

Test statistic inflation has been used to measure model calibration (Astle and Balding, 2009; Price et al., 2010). The inflation factor λ is defined as the median χ^2 association statistic divided by theoretical median under the null hypothesis (Devlin and Roeder, 1999). λ can be calculated from p-values using

$$\lambda = \frac{F^{-1}(1 - p_{\mathrm{median}})}{F^{-1}(1 - u_{\mathrm{median}})},$$

where p_{median} is the median observed p-value (includes causal and non-causal loci), $u_{\mathrm{median}} = \frac{1}{2}$ is its null expectation, and F is the χ^2 cumulative density function (F^{-1} is the quantile function). We use this equation to compare p-values from non-χ^2 tests (such as t statistics).

To compare λ and SRMSD_p directly, for simplicity assume that all p-values are null. In this case, calibrated p-values give $\lambda = 1$ and $\mathrm{SRMSD}_p = 0$. However, non-uniform p-values with the expected median, such as genomic control (Devlin and Roeder, 1999), result in $\lambda = 1$, but $\mathrm{SRMSD}_p \neq 0$

except for uniform p-values, a key flaw of λ that SRMSD_p overcomes. Inflated statistics (anti-conservative p-values) give $\lambda > 1$ and $\text{SRMSD}_p > 0$. Deflated statistics (conservative p-values) give $\lambda < 1$ and $\text{SRMSD}_p < 0$. Thus, $\lambda \neq 1$ always implies $\text{SRMSD}_p \neq 0$ (where $\lambda - 1$ and SRMSD_p have the same sign), but not the other way around. Overall, λ depends only on the median p-value, while SRMSD_p uses the complete distribution. However, SRMSD_p requires knowing which loci are null, so unlike λ it is only applicable to simulated traits.

5.2.2 Empirical comparison of SRMSD_p and λ

One advantage of SRMSD_p is that its range is bounded, while λ is unbounded. There is a near one-to-one correspondence between λ and SRMSD_p (Fig. S1). PCA tended to be inflated ($\lambda > 1$ and $\text{SRMSD}_p > 0$) whereas LMM tended to be deflated ($\lambda < 1$ and $\text{SRMSD}_p < 0$), otherwise the data for both models fall on the same contiguous curve. We fit the following sigmoidal function to this data,

$$\text{SRMSD}_p(\lambda) = a\frac{\lambda^b - 1}{\lambda^b + 1}, \tag{10}$$

which for $a, b > 0$ satisfies $\text{SRMSD}_p(\lambda = 1) = 0$ and reflects $\log(\lambda)$ about zero ($\lambda = 1$):

$$\text{SRMSD}_p(\log(\lambda) = -x) = -\text{SRMSD}_p(\log(\lambda) = x).$$

We fit this model to $\lambda > 1$ only since it was less noisy and of greater interest, and obtained the curve shown in Fig. S1 with $a = 0.566$ and $b = 0.616$. The value $\lambda = 1.05$, a common threshold for benign inflation (Price et al., 2010), corresponds to $\text{SRMSD}_p = 0.0085$ according to Eq. (10). Conversely, $\text{SRMSD}_p = 0.01$, serving as a simpler rule of thumb, corresponds to $\lambda = 1.06$.

5.2.3 Type I error rate

The type I error rate is the proportion of null p-values with $p \leq t$. Calibrated p-values have type I error rate near t, which may be evaluated with a binomial test. This measure may give different results for different t, for example be significantly miscalibrated only for large t (due to lack of power for smaller t). In contrast, $\text{SRMSD}_p = 0$ guarantees calibrated type I error rates at all t, while large

$|\text{SRMSD}_p|$ indicates incorrect type I errors for a range of t.

5.2.4 Statistical power and comparison to AUC_{PR}

Power is the probability that a test is declared significant when the alternative hypothesis H_1 holds. At a p-value threshold t, power equals

$$F(t) = \Pr(p < t|H_1).$$

$F(t)$ is a cumulative function, so it is monotonically increasing and has an inverse. Like type I error, power may rank models differently depending on t.

Power is hard to interpret when p-values are not calibrated. To establish a clear connection to AUC_{PR}, assume calibrated (uniform) null p-values: $\Pr(p < t|H_0) = t$. TPs, FPs, and FNs at t are

$$\begin{aligned}
\text{TP}(t) &= m\pi_1 F(t), \\
\text{FP}(t) &= m\pi_0 t, \\
\text{FN}(t) &= m\pi_1(1 - F(t)),
\end{aligned}$$

where $\pi_0 = \Pr(H_0)$ is the proportion of null cases and $\pi_1 = 1 - \pi_0$ of alternative cases. Therefore,

$$\begin{aligned}
\text{Precision}(t) &= \frac{\pi_1 F(t)}{\pi_1 F(t) + \pi_0 t}, \\
\text{Recall}(t) &= F(t).
\end{aligned}$$

Noting that $t = F^{-1}(\text{Recall})$, precision can be written as a function of recall, the power function, and constants:

$$\text{Precision}(\text{Recall}) = \frac{\pi_1 \text{Recall}}{\pi_1 \text{Recall} + \pi_0 F^{-1}(\text{Recall})}.$$

This last form leads most clearly to $\text{AUC}_{\text{PR}} = \int_0^1 \text{Precision}(\text{Recall}) d\text{Recall}$.

Now lets consider a simple yet common case in which model A is uniformly more powerful than

model B, so $F_A(t) > F_B(t)$ for every t. Therefore $F_A^{-1}(\text{Recall}) < F_B^{-1}(\text{Recall})$ for every recall value. This ensures that the precision of A is greater than that of B at every recall value, so AUC_{PR} is greater for A than B. Thus, in this case AUC_{PR} ranks models according to their power.

Declaration of interests

The authors declare no competing interests.

Acknowledgments

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/`

Eigensoft, `https://github.com/DReichLab/EIG`

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

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

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

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

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

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

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

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

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

Data and code availability

The data and code generated during this study are available on GitHub at `https://github.com/OchoaLab/pca-assoc-paper`. The public subset of Human Origins is available on the Reich Lab website at `https://reich.hms.harvard.edu/datasets`; non-public samples have to be requested from David Reich. The WGS version of HGDP was downloaded from the Wellcome Sanger Institute FTP site at `ftp://ngs.sanger.ac.uk/production/hgdp/hgdp_wgs.20190516/`. 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.

Abraham, Gad and Michael Inouye (2014). "Fast Principal Component Analysis of Large-Scale Genome-Wide Data". *PLOS ONE* 9(4), e93766.

Abraham, Gad, Yixuan Qiu, and Michael Inouye (2017). "FlashPCA2: principal component analysis of Biobank-scale genotype datasets". *Bioinformatics* 33(17), pp. 2776–2778.

Agrawal, Aman et al. (2020). "Scalable probabilistic PCA for large-scale genetic variation data". *PLOS Genetics* 16(5), e1008773.

Al-Khudhair, Ahmed et al. (2015). "Inference of Distant Genetic Relations in Humans Using "1000 Genomes"". *Genome Biology and Evolution* 7(2), pp. 481–492.

Alexander, David H., John Novembre, and Kenneth Lange (2009). "Fast model-based estimation of ancestry in unrelated individuals". *Genome Res.* 19(9), pp. 1655–1664.

Astle, William and David J. Balding (2009). "Population Structure and Cryptic Relatedness in Genetic Association Studies". *Statist. Sci.* 24(4), 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.

Bergström, Anders et al. (2020). "Insights into human genetic variation and population history from 929 diverse genomes". *Science* 367(6484).

Bouaziz, Matthieu, Christophe Ambroise, and Mickael Guedj (2011). "Accounting for Population Stratification in Practice: A Comparison of the Main Strategies Dedicated to Genome-Wide Association Studies". *PLOS ONE* 6(12), e28845.

Cabreros, Irineo and John D. Storey (2019). "A Likelihood-Free Estimator of Population Structure Bridging Admixture Models and Principal Components Analysis". *Genetics* 212(4), pp. 1009–1029.

Cann, Howard M. et al. (2002). "A human genome diversity cell line panel". *Science* 296(5566), pp. 261–262.

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

Chiu, Alec M. et al. (2022). "Inferring population structure in biobank-scale genomic data". *The American Journal of Human Genetics* 0(0).

Conomos, Matthew P. et al. (2016). "Model-free Estimation of Recent Genetic Relatedness". *The American Journal of Human Genetics* 98(1), pp. 127–148.

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.

Fedorova, Larisa et al. (2016). "Atlas of Cryptic Genetic Relatedness Among 1000 Human Genomes". *Genome Biology and Evolution* 8(3), pp. 777–790.

Galinsky, Kevin J. et al. (2016). "Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia". *The American Journal of Human Genetics* 98(3), pp. 456–472.

Gazal, Steven et al. (2015). "High level of inbreeding in final phase of 1000 Genomes Project". *Sci Rep* 5(1), p. 17453.

Gopalan, Prem et al. (2016). "Scaling probabilistic models of genetic variation to millions of humans". *Nat. Genet.* 48(12), pp. 1587–1590.

Grau, Jan, Ivo Grosse, and Jens Keilwagen (2015). "PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R". *Bioinformatics* 31(15), pp. 2595–2597.

Grueneberg, Alexander and Gustavo de los Campos (2019). "BGData - A Suite of R Packages for Genomic Analysis with Big Data". *G3: Genes, Genomes, Genetics* 9(5), pp. 1377–1383.

Henn, Brenna M. et al. (2012). "Cryptic Distant Relatives Are Common in Both Isolated and Cosmopolitan Genetic Samples". *PLOS ONE* 7(4), e34267.

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

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

Jolliffe, Ian T. (2002). *Principal Component Analysis*. 2nd ed. New York: Springer-Verlag.

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.

Lawson, Charles L. and R. J. Hanson (1974). "Solving least squares problems prentice-hall". *Englewood Cliffs*.

Lazaridis, Iosif et al. (2014). "Ancient human genomes suggest three ancestral populations for present-day Europeans". *Nature* 513(7518), pp. 409–413.

Lazaridis, Iosif et al. (2016). "Genomic insights into the origin of farming in the ancient Near East". *Nature* 536(7617), pp. 419–424.

Lee, Seokho et al. (2012). "Sparse Principal Component Analysis for Identifying Ancestry-Informative Markers in Genome-Wide Association Studies". *Genetic Epidemiology* 36(4), pp. 293–302.

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

Listgarten, Jennifer et al. (2012). "Improved linear mixed models for genome-wide association studies". *Nat Methods* 9(6), pp. 525–526.

Liu, Nianjun et al. (2011). "Controlling Population Structure in Human Genetic Association Studies with Samples of Unrelated Individuals". *Stat Interface* 4(3), pp. 317–326.

Liu, Xiaolei et al. (2016). "Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies". *PLOS Genet* 12(2), e1005767.

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

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

Manichaikul, Ani et al. (2010). "Robust relationship inference in genome-wide association studies". *Bioinformatics* 26(22), pp. 2867–2873.

McVean, Gil (2009). "A genealogical interpretation of principal components analysis". *PLoS Genet* 5(10), e1000686.

Mullen, Katharine M. and Ivo H. M. van Stokkum (2012). *nnls: The Lawson-Hanson algorithm for non-negative least squares (NNLS)*.

Ochoa, Alejandro and John D. Storey (2019). "New kinship and F_{ST} estimates reveal higher levels of differentiation in the global human population". *bioRxiv* (10.1101/653279).

— (2021). "Estimating FST and kinship for arbitrary population structures". *PLoS Genet* 17(1), e1009241.

O'Connor, Luke J. et al. (2019). "Extreme Polygenicity of Complex Traits Is Explained by Negative Selection". *The American Journal of Human Genetics* 0(0).

Paradis, E. and K. Schliep (2019). "ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R". *Bioinformatics* 35, pp. 526–528.

Park, Ju-Hyun et al. (2010). "Estimation of effect size distribution from genome-wide association studies and implications for future discoveries". *Nature Genetics* 42(7), pp. 570–575.

Park, Ju-Hyun et al. (2011). "Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants". *PNAS* 108(44), pp. 18026–18031.

Patterson, Nick, Alkes L Price, and David Reich (2006). "Population Structure and Eigenanalysis". *PLoS Genet* 2(12), e190.

Patterson, Nick et al. (2012). "Ancient admixture in human history". *Genetics* 192(3), pp. 1065–1093.

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

Price, Alkes L. et al. (2010). "New approaches to population stratification in genome-wide association studies". *Nature Reviews Genetics* 11(7), pp. 459–463.

— (2013). "Response to Sul and Eskin". *Nature Reviews Genetics* 14(4), p. 300.

Pritchard, Jonathan K. et al. (2000). "Association Mapping in Structured Populations". *The American Journal of Human Genetics* 67(1), pp. 170–181.

Qian, Junyang et al. (2020). "A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK Biobank". *PLOS Genetics* 16(10), e1009141.

Rakitsch, Barbara et al. (2013). "A Lasso multi-marker mixed model for association mapping with population structure correction". *Bioinformatics* 29(2), pp. 206–214.

Rosenberg, Noah A. et al. (2002). "Genetic Structure of Human Populations". *Science* 298(5602), pp. 2381–2385.

Schlauch, Daniel, Heide Fier, and Christoph Lange (2017). "Identification of genetic outliers due to sub-structure and cryptic relationships". *Bioinformatics* 33(13), pp. 1972–1979.

Shchur, Vladimir and Rasmus Nielsen (2018). "On the number of siblings and p-th cousins in a large population sample". *J Math Biol* 77(5), pp. 1279–1298.

Simons, Yuval B. et al. (2018). "A population genetic interpretation of GWAS findings for human quantitative traits". *PLOS Biology* 16(3), e2002985.

Skoglund, Pontus et al. (2016). "Genomic insights into the peopling of the Southwest Pacific". *Nature* 538(7626), pp. 510–513.

Sokal, Robert R. and Charles D. Michener (1958). "A statistical method for evaluating systematic relationships." *Univ. Kansas, Sci. Bull.* 38, pp. 1409–1438.

Song, Minsun, Wei Hao, and John D. Storey (2015). "Testing for genetic associations in arbitrarily structured populations". *Nat. Genet.* 47(5), pp. 550–554.

Storey, John D. (2003). "The positive false discovery rate: a Bayesian interpretation and the q-value". *Ann. Statist.* 31(6), pp. 2013–2035.

Storey, John D. and Robert Tibshirani (2003). "Statistical significance for genomewide studies". *Proceedings of the National Academy of Sciences of the United States of America* 100(16), pp. 9440–9445.

Sul, Jae Hoon and Eleazar Eskin (2013). "Mixed models can correct for population structure for genomic regions under selection". *Nature Reviews Genetics* 14(4), p. 300.

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.

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

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.

Tucker, George, Alkes L. Price, and Bonnie Berger (2014). "Improving the Power of GWAS and Avoiding Confounding from Population Stratification with PC-Select". *Genetics* 197(3), pp. 1045–1049.

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

Wright, S. (1951). "The genetical structure of populations". *Ann Eugen* 15(4), pp. 323–354.

Wu, Chengqing et al. (2011). "A Comparison of Association Methods Correcting for Population Stratification in Case–Control Studies". *Annals of Human Genetics* 75(3), pp. 418–427.

Xu, Hanli and Yongtao Guan (2014). "Detecting Local Haplotype Sharing and Haplotype Association". *Genetics* 197(3), pp. 823–838.

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.

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.

Zeng, Jian et al. (2018). "Signatures of negative selection in the genetic architecture of human complex traits". *Nature Genetics* 50(5), pp. 746–753.

Zhang, Shuanglin, Xiaofeng Zhu, and Hongyu Zhao (2003). "On a semiparametric test to detect associations between quantitative traits and candidate genes using unrelated individuals". *Genetic Epidemiology* 24(1), pp. 44–56.

Zhang, Zhiwu et al. (2010). "Mixed linear model approach adapted for genome-wide association studies". *Nat Genet* 42(4), pp. 355–360.

Zhao, Keyan et al. (2007). "An Arabidopsis Example of Association Mapping in Structured Samples". *PLOS Genetics* 3(1), e4.

Zheng, Xiuwen and Bruce S. Weir (2016). "Eigenanalysis of SNP data with an identity by descent interpretation". *Theor Popul Biol* 107, pp. 65–76.

Zhou, Quan, Liang Zhao, and Yongtao Guan (2016). "Strong Selection at MHC in Mexicans since Admixture". *PLoS Genet.* 12(2), e1005847.

Zhou, Wei et al. (2018). "Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies". *Nat Genet* 50(9), pp. 1335–1341.

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

S1 Supplementary figures

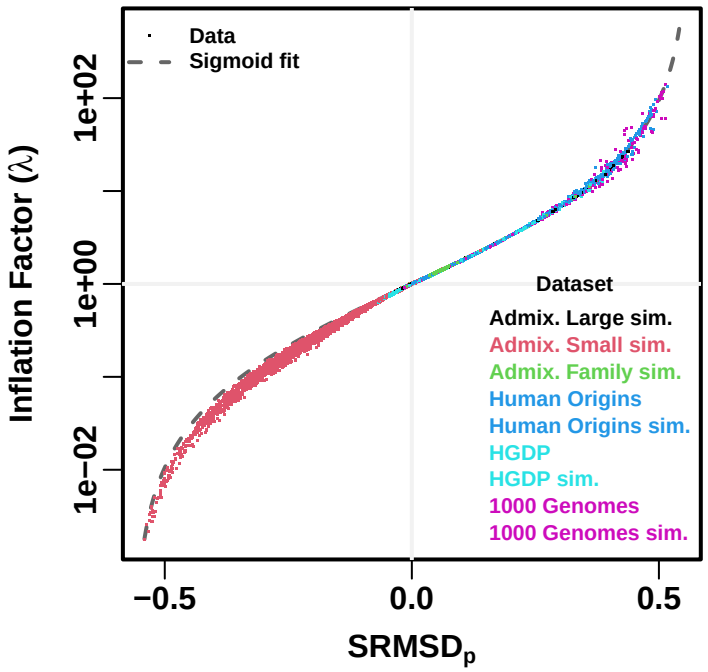


Figure S1: **Comparison between SRMSD$_p$ and inflation factor.** Each point is a pair of statistics for one replicate, one association model (PCA or LMM with some number of PCs r), one trait model (FES vs RC), and one dataset (color coded by dataset). Note y-axis (λ) is on a log scale. The sigmoidal curve in Eq. (10) is fit to the data.

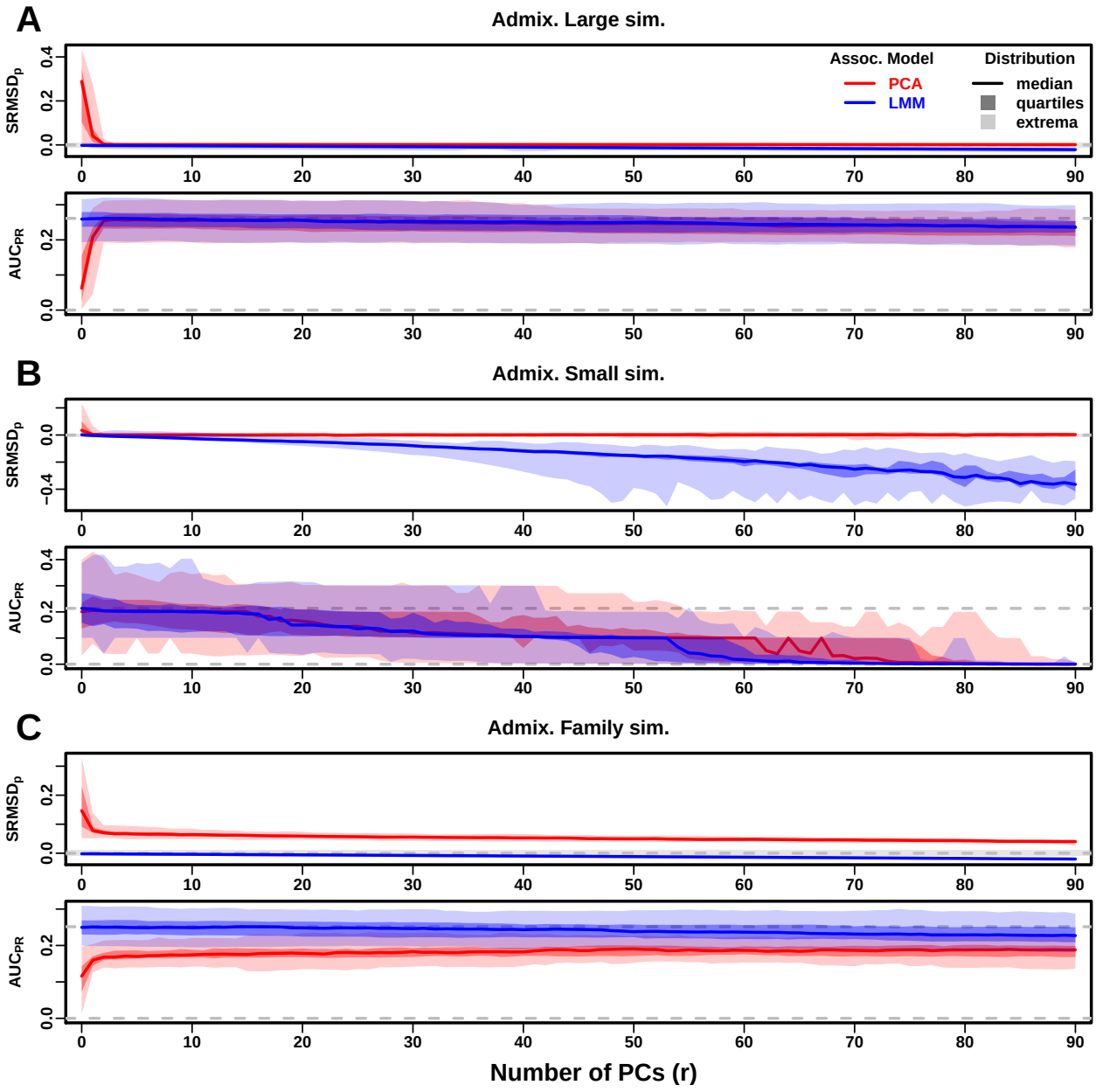


Figure S2: **Evaluations in admixture simulations.** Traits simulated from RC model, otherwise the same as Fig. 3.

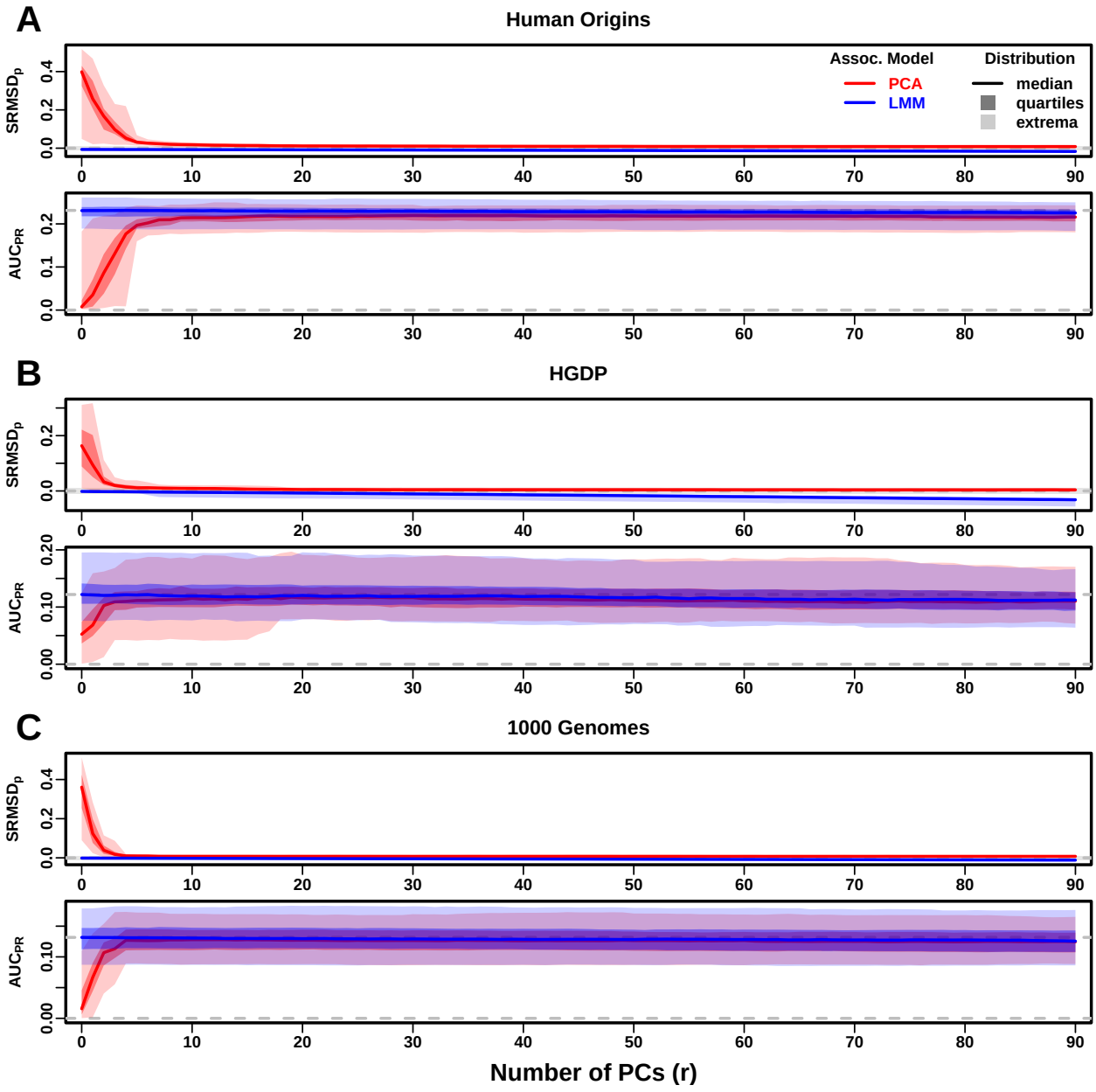


Figure S3: **Evaluations in real human genotype datasets.** Traits simulated from RC model, otherwise the same as Fig. 4.

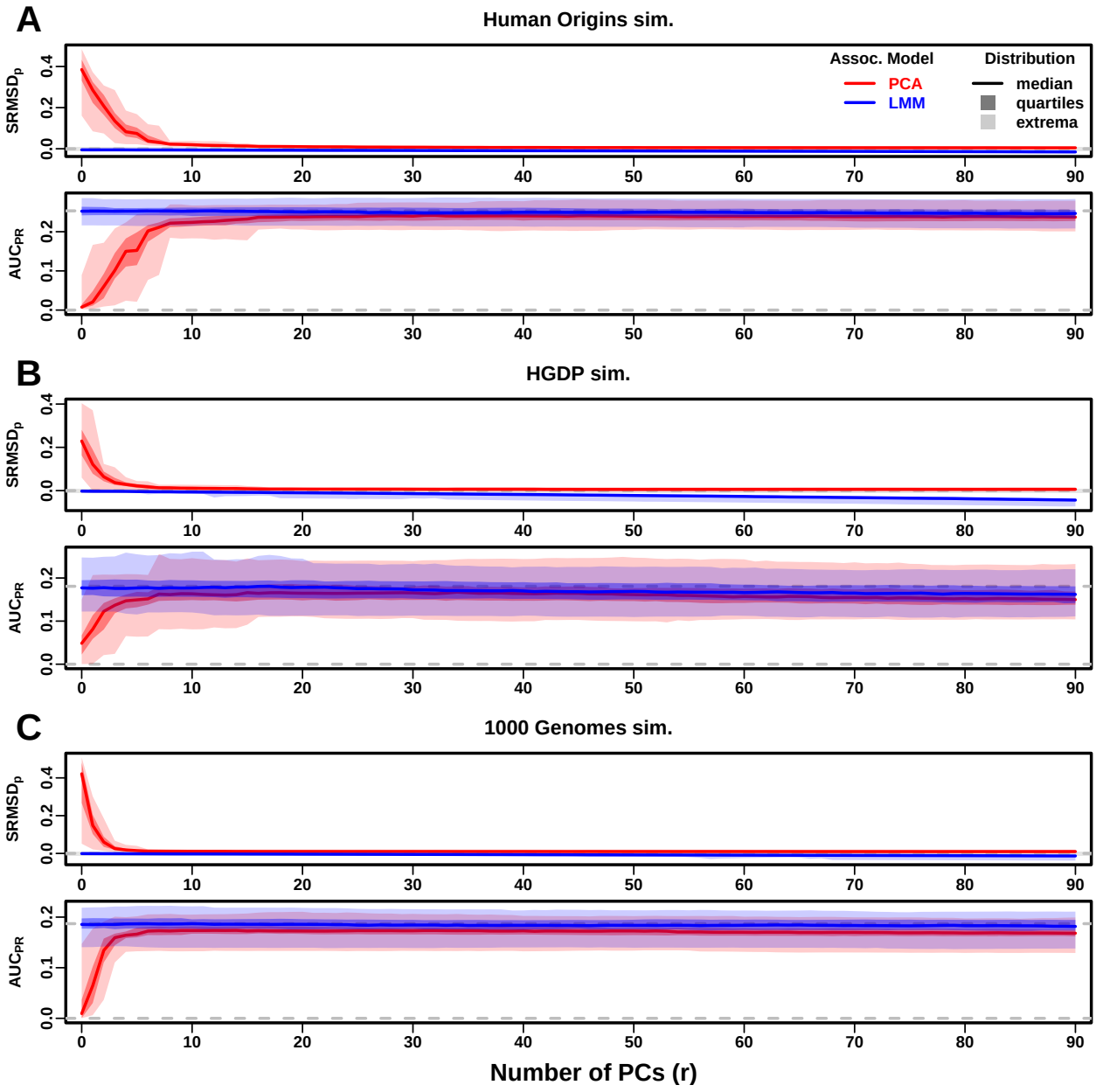


Figure S4: **Evaluations in tree simulations fit to human data.** Traits simulated from RC model, otherwise the same as Fig. 5.

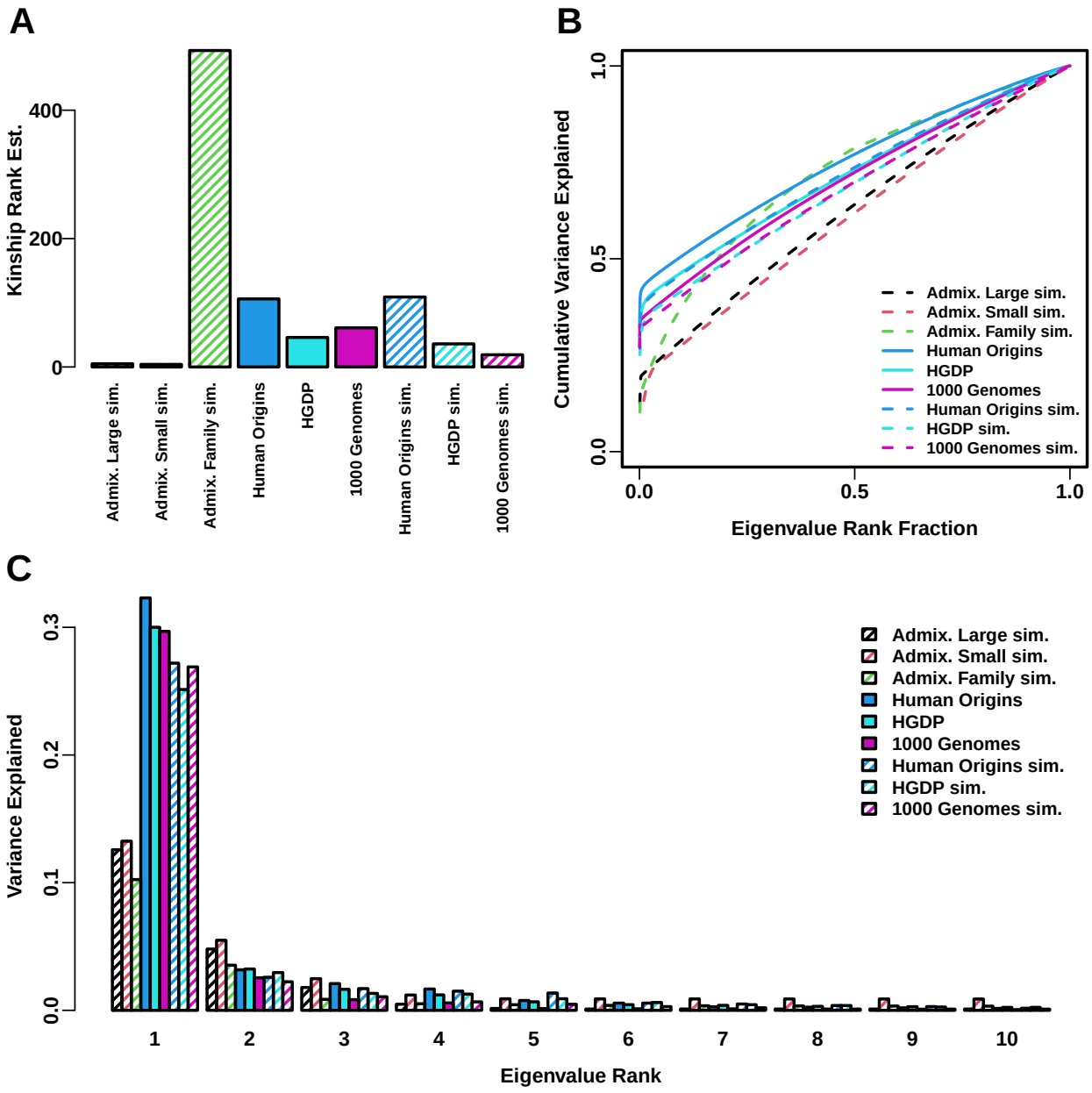


Figure S5: **Estimated dimensionality of datasets. A.** Kinship matrix ranks estimated with the Tracy-Widom test with $p < 0.01$. **B.** Cumulative variance explained versus eigenvalue rank fraction. **C.** Variance explained by first 10 eigenvalues.

S2 Supplementary tables

Table S1: **Dataset sizes after 4th degree relative filter.**

Dataset	Loci (m)	Ind. (n)	Ind. removed (%)
Human Origins	189,722	2636	9.8
HGDP	905,838	842	9.4
1000 Genomes	1,097,415	2390	4.6