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

Yiqi Yao,[1,3] 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

[3] Present address: BenHealth Consulting, Shanghai, Shanghai, 200023, China

[*] Correspondence: `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 (trait model); RC: random coefficients (trait model); MAF: minor allele frequency; 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. Naive association tests assume that genotypes are drawn independently from a common allele frequency. This assumption does not hold for structured populations, which includes multiethnic cohorts and admixed individuals (ancient relatedness), and for family data (recent relatedness).[1] When insufficient approaches are applied to data with relatedness, their association statistics are miscalibrated, resulting in excess false positives and loss of power.[1–3] Therefore, many specialized approaches have been developed for genetic association under relatedness, of which PCA and LMM are the most popular.

Genetic association with PCA consists of including the top eigenvectors of the population kinship matrix as covariates in a generalized linear model.[4–6] These top eigenvectors are commonly referred to as PCs in genetics,[7] the convention adopted here, but in other fields PCs denote the projections of loci onto eigenvectors.[8] The direct ancestor of PCA association is structured association, in which inferred ancestry or admixture proportions are used as regression covariates.[9] These models are deeply connected because PCs map to ancestry empirically[10,11] and theoretically,[12–15] and they work as well as global ancestry in association studies but are estimated more easily.[6,7,10,16] The strength of PCA is its simplicity, which as covariates can be readily included in more complex models, such as haplotype association[17] and polygenic models.[18] However, PCA assumes that relatedness is low-dimensional, which may limit its applicability. PCA is known to be inadequate for family data,[7,19,20] which is called "cryptic relatedness" when it is unknown to the researchers, but no other troublesome cases have been confidently identified. Recent work has focused on developing more scalable versions of the PCA algorithm.[21–25] PCA remains a popular and powerful approach for association studies.

The other dominant association model under relatedness is the LMM, which includes a random

effect parametrized by the kinship matrix. Unlike PCA, LMM does not assume that relatedness is low-dimensional, and explicitly models families via the kinship matrix. Early LMMs required kinship matrices estimated from known pedigrees or which otherwise captured recent relatedness only.[16,26] 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.[1,27,28] 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,[29] LMMs have been applied successfully to balanced case-control studies[1,30] and simulations,[20,31,32] and have been adapted for unbalanced case-control studies.[33] However, LMMs tend to be considerably slower than PCA and other models, so much effort has focused on improving their runtime and scalability.[27,30,33–41]

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,[1,42] 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,[20,43] while others did not,[44] and the approach continues to be used.[45] Recall that early LMMs used kinship to model family relatedness only, so population structure had to be modeled separately, in practice as admixture fractions instead of PCs.[16,26]

LMM and PCA are closely related models,[1,42] so similar performance is expected particularly under 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). No papers find that PCA outperforms LMM decisively, although PCA occasionally performs better in isolated and artificial cases or individual measures (often with unknown significance). Previous studies were generally divided those that employed simulated versus real genotypes (only one study used both). 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 with quantitative traits favor LMMs, whereas 5/7 papers with case-

control traits gave ties or mixed results (the only factor we do not explore). Additionally, although all previous evaluations 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 used more than one number of PCs for PCA. Lastly, no consensus has emerged as to why LMM might outperform PCA or vice versa,[20,32,42,49] or which features of the real datasets are critical for the LMM advantage other than cryptic relatedness, resulting in unclear guidance for using PCA. Hence, our work includes real and simulated genotypes with higher dimensionalities and differentiation matching that of multiethnic human cohorts, we vary the number of PCs, and measure robust proxies for type I error control and power.

In this work, we evaluate the PCA and LMM association models 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,[50,51] the Human Genome Diversity Panel (HGDP),[52–54] and Human Origins.[55–58] We simulate quantitative traits from two models: fixed

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

Publication	Sim. Genotypes Type[a]	K[b]	F_{ST} [c]	Real[d]	Trait[e]	Power	PCs (r)	Best
Zhao et al.[16]				✓	Q	✓	8	LMM
Astle and Balding[1]	I	3	0.10		CC	✓	10	Tie
Kang et al.[30]				✓	Both		2-100	LMM
Price et al.[20]	I, F	2	0.01		CC		1	Mixed
Wu et al.[31]	I, A	2-4	0.01		CC	✓	10	Mixed
Liu et al.[44]	S, A	2-3	R		Q	✓	10	Tie
Sul and Eskin[32]	I	2	0.01		CC		1	Tie
Tucker, Price, and Berger[43]	I	2	0.05	✓	Both	✓	5	Tie
Yang et al.[29]				✓	CC	✓	5	Tie
Song, Hao, and Storey[46]	S, A	2-3	R		Q		3	LMM
Loh et al.[41]				✓	Q	✓	10	LMM
Liu et al.[47]				✓	Q	✓	3-6	LMM
Sul, Martin, and Eskin[48]				✓	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.

effect sizes (FES; coefficients inverse to allele frequency) that matches real data[45,59,60] and corresponds to high pleiotropy and strong balancing selection[61] and strong negative selection,[45,60] which are appropriate assumptions for diseases; and random coefficients (RC; independent of allele frequency) that corresponds to neutral traits.[45,61] 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 relatedness 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, and suggesting that PCA is generally not an appropriate model for real data. 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 $\mathbf{X}^\intercal\boldsymbol{\beta}$ 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.[1,62–64] 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, consider 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 dimensionality r 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.[1,42] PCA has more parameters than LMM, so it may overfit more: ignoring the shared terms in Eqs. (2) and (3), PCA fits r parameters (length of $\boldsymbol{\gamma}$), whereas LMMs fit only one (σ_s^2).

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

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

where the unknown p_i^T is estimated by $\hat{p}_i^T = \frac{1}{2n} \sum_{j=1}^n x_{ij}$.[5,27,30,33,37,39,41,48] However, this kinship estimator has a complex bias that differs for every individual pair, which arises due to the use of this estimated $\hat{p}_i^T$.[28,65] Nevertheless, in PCA and LMM these biased estimates perform as well as unbiased ones, an observation that will be explored in future work (data not shown).

We selected fast and robust software implementing the basic PCA and LMM models. PCA association was performed with `plink2`.[66] 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 minor allele frequency MAF < 0.1.

LMM association was performed using GCTA.[29,37] 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 (small simulation only) had convergence and singularity errors in some replicates, which were treated as missing data.

2.2 Simulations

Every simulation was replicated 50 times, drawing anew all genotypes (except for real datasets) and traits. 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

The basic admixture model is as described previously[28] 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 ($u \in \{1, ..., K\}$) is at coordinate u and has an inbreeding coefficient $f^T_{S_u} = u\tau$ for some τ. Ancestry proportions q_{ju} for individual j and S_u arise from a random walk with spread σ on the 1D geography, and τ and σ are fit to give $F_{\mathrm{ST}} = 0.1$ and mean kinship $\bar{\theta}^T = 0.5F_{\mathrm{ST}}$ for the admixed individuals.[28] Random ancestral allele frequencies p^T_i, subpopulation allele frequencies $p^{S_u}_i$, individual-specific allele frequencies π_{ij}, and genotypes x_{ij} are drawn from this hierarchical model:

$$
\begin{aligned}
p^T_i &\sim \mathrm{Uniform}(0.01, 0.5), \\
p^{S_u}_i | p^T_i &\sim \mathrm{Beta}\left(p^T_i \left(\frac{1}{f^T_{S_u}} - 1\right), \left(1 - p^T_i\right)\left(\frac{1}{f^T_{S_u}} - 1\right)\right), \\
\pi_{ij} &= \sum_{u=1}^{K} q_{ju} p^{S_u}_i, \\
x_{ij} | \pi_{ij} &\sim \mathrm{Binomial}(2, \pi_{ij}),
\end{aligned}
$$

where this Beta is the Balding-Nichols distribution[67] with mean p^T_i and variance $p^T_i\left(1 - p^T_i\right) f^T_{S_u}$. 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^T_i, iterating until no loci are fixed. Each replicate draws a genotypes starting from p^T_i.

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), random family sizes, and arbitrary numbers of generations (20 here). 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. Local kinship measures pedigree relatedness; in the first generation, everybody is locally unrelated and outbred. Individuals are randomly assigned sex. 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.[68] 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 < 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:[28]

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

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`.[28] Indi-

vidual coancestry ($\hat{\theta}^T_{jk}$) is estimated from kinship using

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

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

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

Topology is estimated with hierarchical clustering using the weighted pair group method with arithmetic mean,[69] 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[70] (implemented in `nnls`[71]) 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`.[28] 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.[72]

Under the *fixed effect sizes* (FES) model, initial causal coefficients are

$$\beta_i = \frac{1}{\sqrt{2v_i^T}}$$

for known p_i^T; otherwise v_i^T is replaced by the unbiased estimator[28] $\hat{v}_i^T = \hat{p}_i^T \left(1 - \hat{p}_i^T\right) / (1 - \bar{\varphi}^T)$, 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 $\sigma_0^2 = \sum_{i \in C} 2v_i^T \beta_i^2$, replacing v_i^T with $\hat{v}_i^T$ for unknown p_i^T (so σ_0^2 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[65] (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`,[66] and analysis was uniquely enabled by the R packages `BEDMatrix`[73] 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[58] was obtained separately from the rest,[56,57] 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 whole-genome sequencing (WGS) version of HGDP[54] was restricted to autosomal biallelic SNP loci with filter "PASS". Our analysis of the high-coverage NYGC version of 1000 Genomes[74] 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 dimensionality and eigenvalues were calculated from `popkin` kinship estimates. Eigenvalues were assigned p-values with `twstats` of the Eigensoft package,[7] and dimensionality 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,[75] 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 mis-calibrated 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.[76,77] 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.[78] 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

3.1 Overview of evaluations

We use three real genotype datasets and simulated genotypes from six population structure scenarios to cover various features of interest (Table 2). We introduce them in sets of three, as they appear in the rest of our results. Population kinship matrices, which combine population and family relatedness, are estimated without bias using `popkin`[28] (Fig. 1). The first set of three simulated genotypes are based on an admixture model with 10 ancestries (Fig. 1A).[14,28,79] The "large" version (1000 individuals) illustrates asymptotic performance, while the "small" simulation (100 individuals) illustrates 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 representing global human diversity: Human Origins (Fig. 1D), HGDP (Fig. 1G), and 1000 Genomes (Fig. 1J), which are enriched for small minor allele frequencies even after MAF $< 1\%$ filter (Fig. 1C). Last are tree simulations (Fig. 1F,I,L) fit to the kinship (Fig. 1E,H,K) and MAF (Fig. 1C) of each real human dataset, which by design do not have family structure.

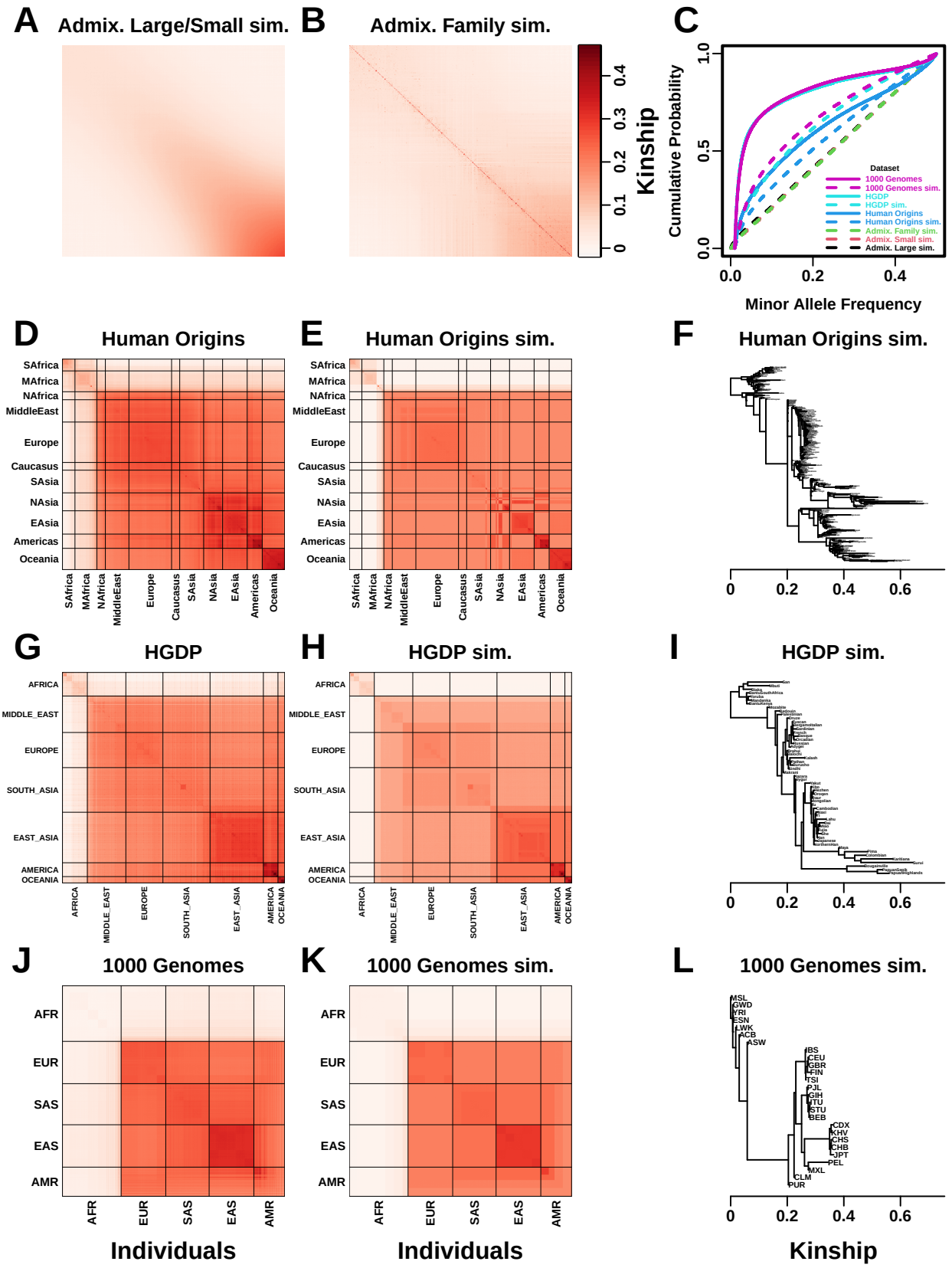
All traits in this work are simulated. We repeated all evaluations on two additive quantitative trait models, *fixed effect sizes* (FES) and *random coefficients* (RC), which differ in how causal coefficients are constructed. The FES model captures the rough inverse relationship between coefficient and minor allele frequency that arises under strong negative and balancing selection and has been observed in numerous diseases and other traits,[45,59–61] so it is the focus of our results. The RC

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]$m_1 = n/10$ to balance power across datasets.
[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 along x- and y-axes, kinship as color. 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 of a large diversity of global populations. **G.** Human Genome Diversity Panel (HGDP) is a WGS dataset from global native populations. **J.** 1000 Genomes Project is a WGS dataset of global cosmopolitan populations. **F,I,L.** Trees between subpopulations fit to real data. **E,H,K.** Simulations from trees fit to the real data recapitulate subpopulation structure.


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
Minor Allele Frequency
0.0
0.2
0.4
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

model draws coefficients independent of allele frequency, corresponding to neutral traits,[45,61] which results in a wider effect size distribution that reduces association power and effective polygenicity compared to FES.

We evaluate using two complementary measures: (1) SRMSD_p (p-value signed root mean square deviation) measures p-value calibration (closer to zero is better), and (2) $\mathrm{AUC}_{\mathrm{PR}}$ (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 control measures; there is a correspondence between λ and SRMSD_p, with $\mathrm{SRMSD}_p > 0.01$ giving $\lambda > 1.06$ (Fig. S1) and thus evidence of miscalibration close to the rule of thumb of $\lambda > 1.05$.[20] $\mathrm{AUC}_{\mathrm{PR}}$ has been used to evaluate association models,[80] and reflects statistical power while being robust to miscalibrated models (Appendix B).

Both PCA and LMM were evaluated in each replicate dataset including a number of PCs r between 0 and 90 as fixed covariates. In terms of p-value calibration, for PCA the best number of PCs r (minimizing mean $|\mathrm{SRMSD}_p|$ over replicates) is typically large across all datasets, but much smaller "min" r values often performed as well (numbers in parentheses in Table 3). Most cases had a mean $|\mathrm{SRMSD}_p| < 0.01$ (marked with asterisks in Table 3), whose p-values are effectively calibrated. However, PCA best and min r values tended to be large on the family simulation and real datasets, and those cases were often miscalibrated. In contrast, for LMM, $r = 0$ (no PCs) was always best, and was always calibrated. Comparing LMM with $r = 0$ to PCA with its best r, LMM always had significantly smaller $|\mathrm{SRMSD}_p|$ than PCA or was statistically tied. For $\mathrm{AUC}_{\mathrm{PR}}$ and PCA, the best r was always smaller than the best r for $|\mathrm{SRMSD}_p|$, so there is often a tradeoff between calibrated p-values versus classification performance. For LMM there is no tradeoff, as $r = 0$ had $\mathrm{AUC}_{\mathrm{PR}}$ not significantly different from the best r in all cases 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 $\mathrm{AUC}_{\mathrm{PR}}$ than PCA with its best r except for one tie (HGDP with RC trait).

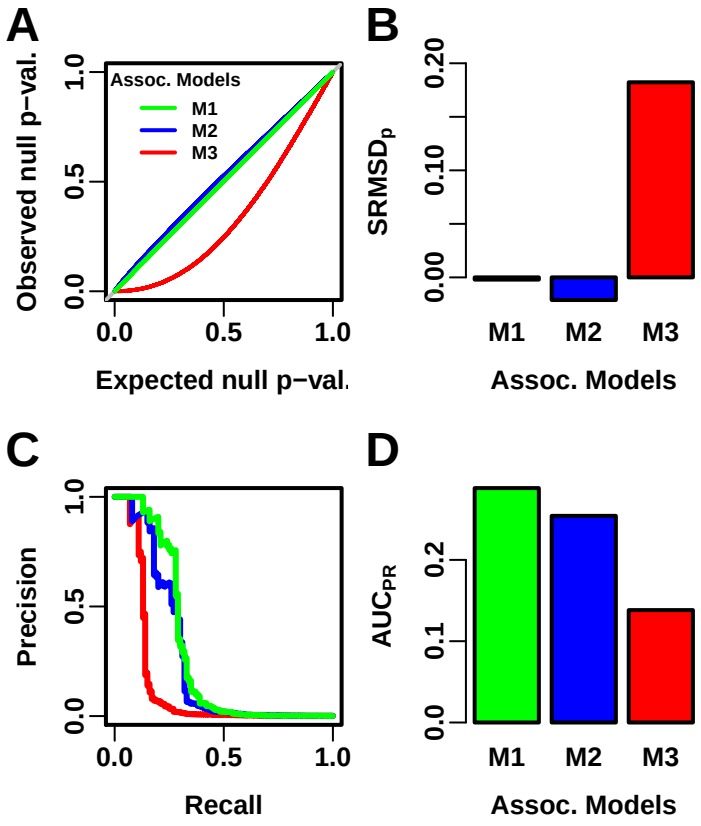


Figure 2: **Illustration of evaluation measures.** Three archetypal models illustrate our complementary measures: M1 is ideal, M2 overfits slightly, M3 is naive. **A.** QQ plot of p-values of "null" (non-causal) loci. M1 has desired uniform p-values, M2/M3 are miscalibrated. **B.** SRMSD_p (p-value Signed Root Mean Square Deviation) measures signed distance between observed and expected null p-values (closer to zero is better). **C.** Precision and Recall (PR) measure causal locus classification performance (higher is better). **D.** $\mathrm{AUC}_{\mathrm{PR}}$ (Area Under the PR Curve) reflects power (higher is better).

3.2 Evaluations in admixture simulations

Now we look more closely at results per dataset. The complete SRMSD_p and AUC_{PR} distributions for the admixture simulations and FES traits are in Fig. 3. RC traits 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 $r \geq 3$, 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

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

		$\lvert\text{SRMSD}_p\rvert$[a]			AUC_{PR}		
		Best (min[c]) PCs			Best (min[c]) PCs		
Dataset	Trait model[b]	PCA	LMM	Best[d]	PCA	LMM	Best[d]
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] r for which mean $\lvert\text{SRMSD}_p\rvert < 0.01$ marked with asterisk.
[b] FES: Fixed Effect Sizes, RC: Random Coefficients.
[c] 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 best r (if any).
[d] Tie if distributions of best PCA and LMM version did not differ significantly (Wilcoxon paired 1-tailed $p > 0.01$). Same result for "min" except cases in parentheses.

as r increases (p-values are conservative). The $\mathrm{AUC_{PR}}$ distribution of PCA is similarly worst at $r = 0$, increases rapidly and peaks at $r = 3$, then decreases slowly for $r > 3$, while the $\mathrm{AUC_{PR}}$ distribution for LMM starts near its maximum at $r = 0$ and decreases with r. Although the $\mathrm{AUC_{PR}}$ distributions for LMM and PCA overlap considerably at each r, LMM with $r = 0$ has significantly greater $\mathrm{AUC_{PR}}$ values than PCA with $r = 3$ (Table 3). However, qualitatively PCA performs nearly as well as LMM in this simulation.

The observed robustness to large r led us to consider smaller sample sizes. A model with large numbers of parameters r should overfit more as r approaches the sample size n. Rather than increase r beyond 90, 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 (fixed ratio $n/m_1 = 10$), which increases per-locus effect sizes. We found a large decrease in performance for both models as r increases, and best performance for $r = 1$ for PCA and $r = 0$ for LMM (Fig. 3B). Remarkably, LMM attains much larger negative SRMSD_p values than in our other evaluations. LMM with $r = 0$ is significantly better than PCA ($r = 1$ to 4) in both measures (Table 3), but qualitatively the difference is negligible.

The family simulation adds a 20-generation random family to our large 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. Our evaluation reveals a sizable gap in both measures between LMM and PCA across all r (Fig. 3C). LMM again performs best with $r = 0$ and achieves mean $|\mathrm{SRMSD}_p| < 0.01$. However, PCA does not achieve mean $|\mathrm{SRMSD}_p| < 0.01$ at any r, and its best mean $\mathrm{AUC_{PR}}$ is considerably worse than that of LMM. Thus, LMM is conclusively superior to PCA, and the only calibrated model, when there is family structure.

3.3 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,

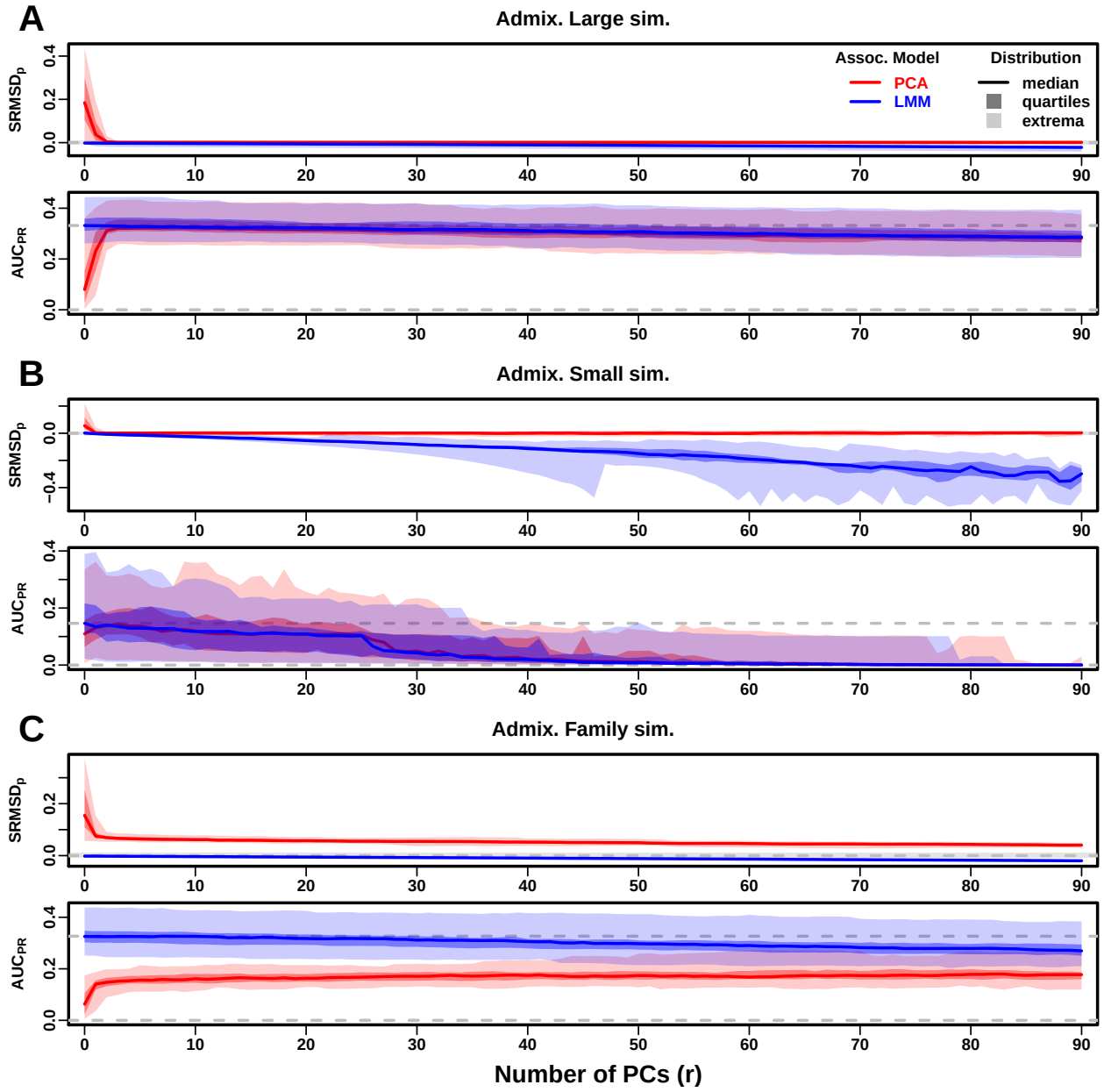


Figure 3: **Evaluations in admixture simulations.** Traits simulated from FES model. PCA and LMM models have varying number of PCs ($r \in \{0, ..., 90\}$ on x-axis), with the distributions (y-axis) of SRMSD_p (top subpanel) and AUC_{PR} (bottom subpanel) for 50 replicates. Best performance is zero SRMSD_p and large AUC_{PR}. Zero and maximum median AUC_{PR} values are marked with horizontal gray dashed lines, and $|\text{SRMSD}_p| < 0.01$ is marked with a light gray area. LMM performs best with $r = 0$, PCA with various r. **A.** Large simulation ($n = 1,000$ individuals). **B.** Small simulation ($n = 100$) shows overfitting for large r. **C.** Family simulation ($n = 1,000$) has 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.

numerous correlated subpopulations, and potential cryptic family relatedness.

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 measures) 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 contains more loci and low-frequency variants. Performance (Fig. 4B) was 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. Thus, although this dataset has the simplest subpopulation 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, LMM with $r = 0$ was significantly better than PCA for all measures in the real datasets and RC traits (Table 3).

3.4 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 tree simulations. Human subpopulations are related roughly by trees, which induce the strongest correlations and have numerous tips, so we fit trees to each real dataset and tested if data simulated from these complex tree structures could recapitulate our

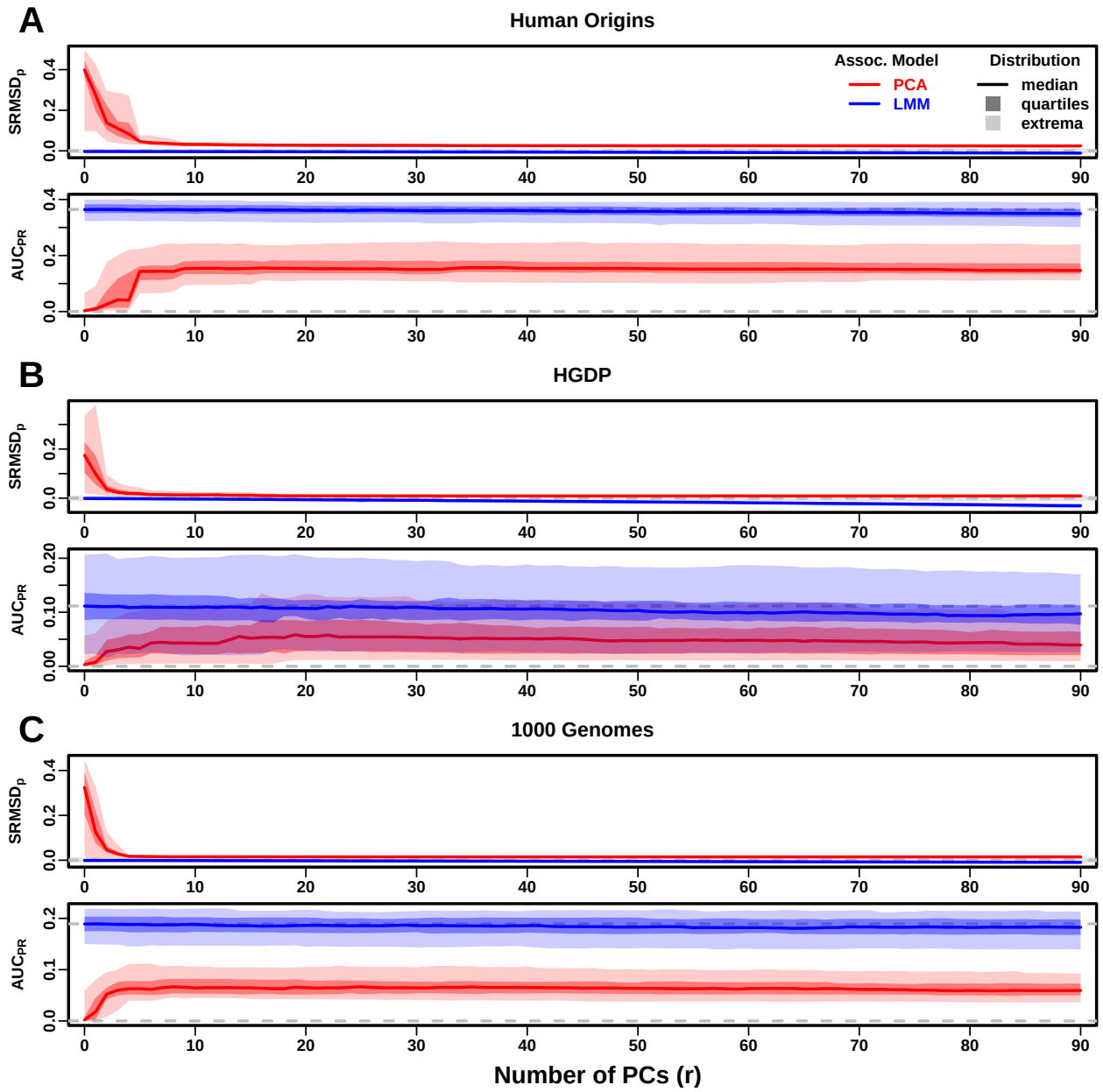


Figure 4: **Evaluations in real human genotype datasets.** Traits simulated from FES model. Same setup as Fig. 3, see that for details. These datasets strongly favor LMM with no PCs over PCA, with distributions that most resemble the family simulation. **A.** Human Origins. **B.** Human Genome Diversity Panel (HGDP). **C.** 1000 Genomes Project.

previous results (Fig. 1). These tree simulations also feature non-uniform ancestral allele frequency distributions, which recapitulated some of the skew for smaller minor allele frequencies of the real datasets (Fig. 1C). The SRMSD_p and $\mathrm{AUC}_{\mathrm{PR}}$ 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). Both LMM with $r = 0$ and PCA (various r) achieve mean $|\mathrm{SRMSD}_p| < 0.01$ (Table 3). The $\mathrm{AUC}_{\mathrm{PR}}$ 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 RC traits (Fig. S4 and Table 3). Overall, these tree simulations do not recapitulate the large LMM advantage over PCA observed on the real data.

3.5 Numerous distant relatives explain poor PCA performance in real data

In principle, PCA performance should be determined by the dimensionality of relatedness, since PCA is a low-dimensional model whereas LMM can model high-dimensional relatedness without overfitting. We used the Tracy-Widom test[7] with $p < 0.01$ to estimate dimensionality as the number of significant PCs (Fig. S5A). The true dimensionality of our simulations is slightly underestimated (Table 2), but we confirm that the family simulation has the greatest dimensionality, and real datasets have greater estimates than their respective tree simulations, which confirms our hypothesis to some extent. However, estimated dimensionalities do not separate real datasets from tree simulations, as required to predict the observed PCA performance. Moreover, the HGDP and 1000 Genomes dimensionality estimates are 46 and 61, respectively, yet PCA performed poorly for all $r \leq 90$ numbers of PCs (Fig. 4). 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 (expected since it has fewer eigenvalues; Fig. S5C). Comparing cumulative variance explained versus rank fraction across all eigenvalues, all datasets increase from their starting point almost linearly until they reach 1, except the family simulation has much greater variance explained by mid-rank eigenvalues (Fig. S5B). Overall, there is no separation between real datasets (where PCA performed poorly) and tree simulations (where PCA performed relatively well) in terms of their eigenvalues or

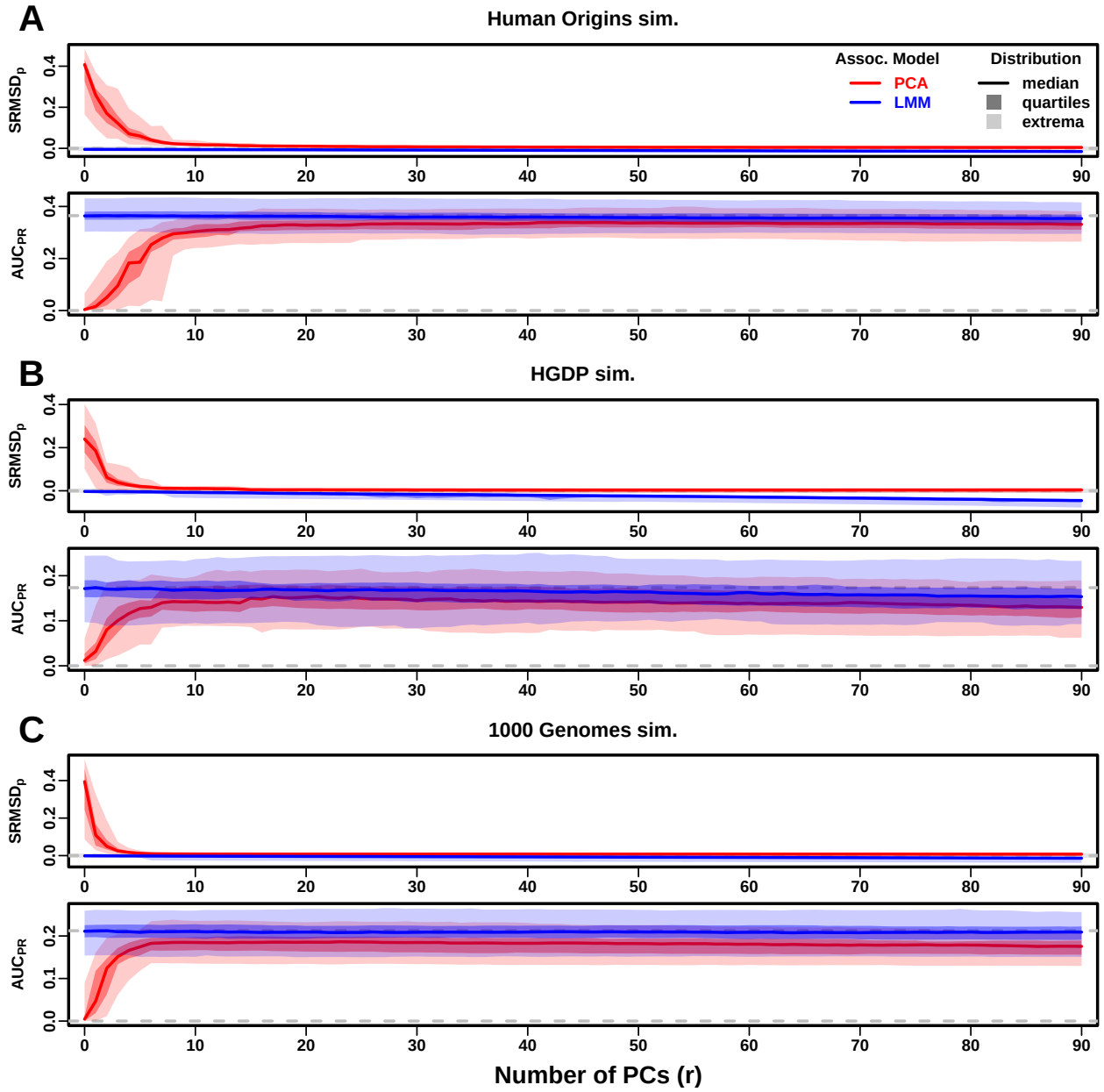


Figure 5: **Evaluations in tree simulations fit to human data.** Traits simulated from FES model. Same setup as Fig. 3, see that for details. These tree simulations, which exclude family structure by design, do not explain the large gaps in LMM-PCA performance observed in the real data. **A.** Human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

dimensionality estimates.

Local kinship, which is recent 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.[75] We observe more large local kinship in the real datasets and the family simulation compared to the other simulations (Fig. 6). However, for real data this distribution depends on the subpopulation structure, since locally related pairs are most likely in the same subpopulation. 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.[75,81] However, these highly related pairs are vastly outnumbered by more distant pairs with evident non-zero local kinship as compared to the extreme tree simulation values.

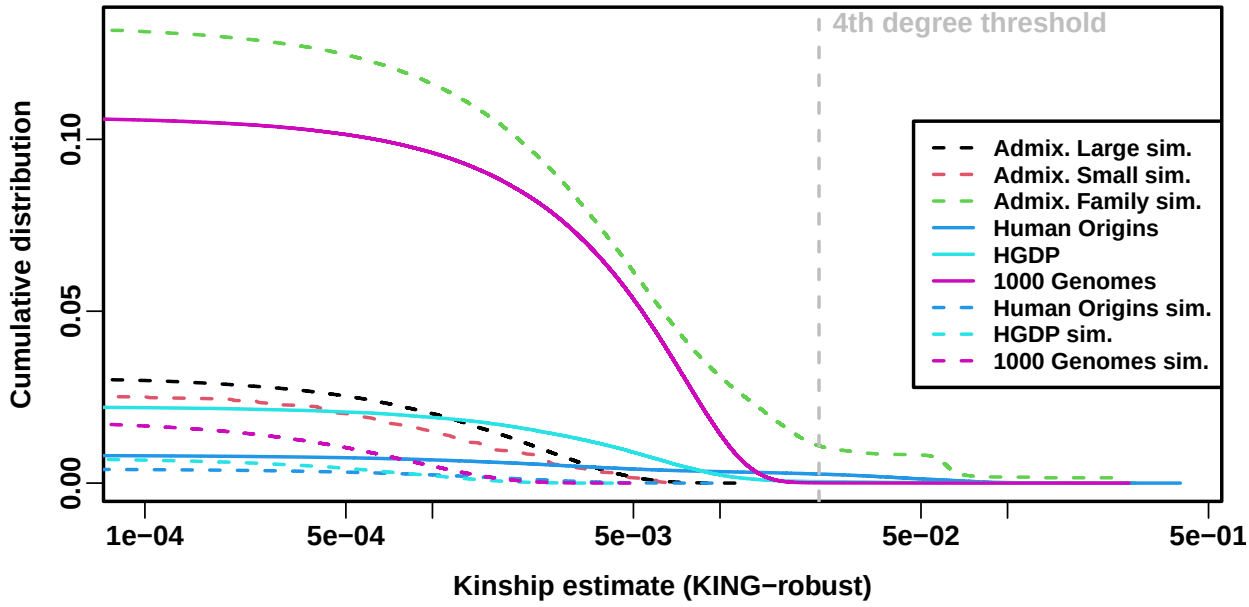


Figure 6: **Local kinship distributions.** Curves are complementary cumulative distribution of lower triangular kinship matrix (self kinship excluded) from KING-robust estimator. Note log x-axis; negative estimates are counted but not shown. Most values are below 4th degree relative threshold. Each real dataset has a greater cumulative than its tree simulations.

To try to improve PCA performance, we followed the standard practice of removing 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, and only FES traits were tested because they previously displayed the large PCA-LMM performance gap. 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 Human Origins and 1000 Genomes ($|\text{SRMSD}_p| > 0.01$). Otherwise, AUC_{PR} and SRMSD_p ranges were similar here as in our earlier evaluation. Therefore, the removal of 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 in the real datasets.

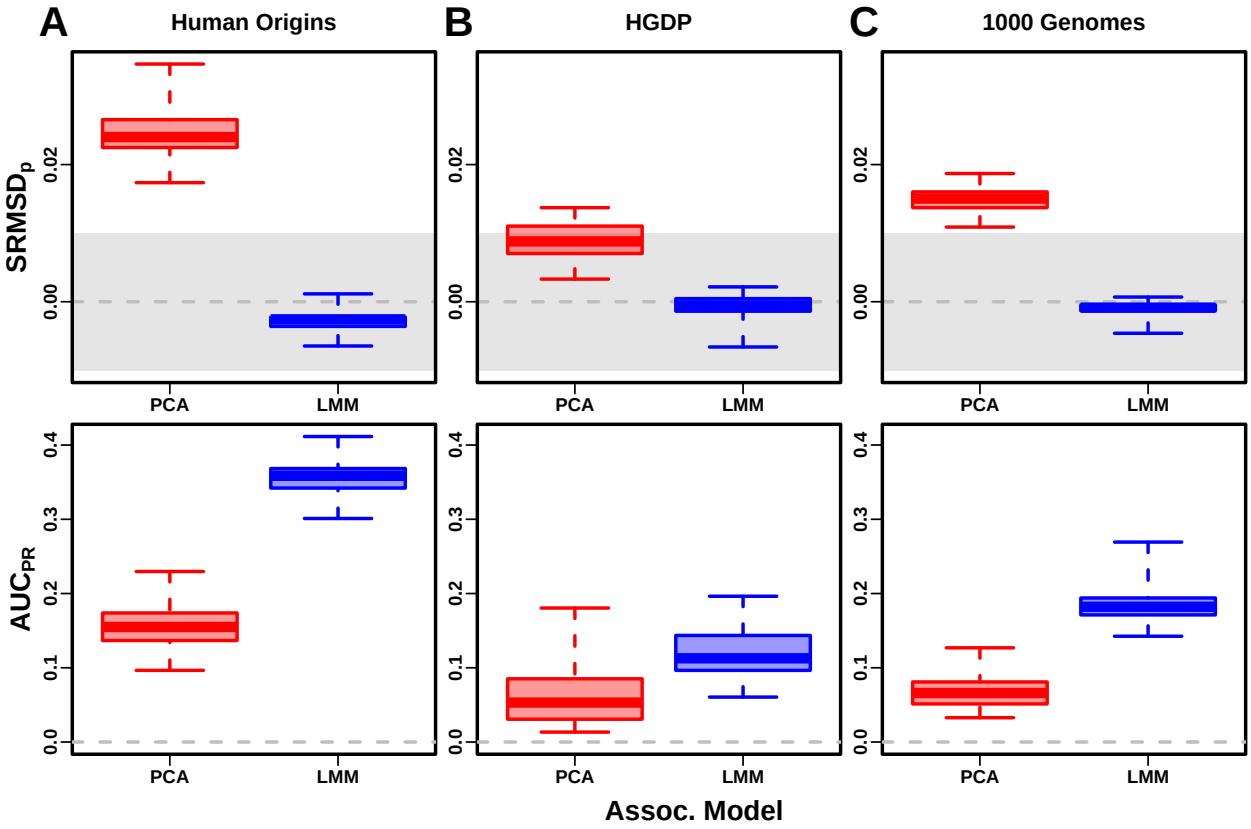


Figure 7: **Evaluation in real datasets excluding 4th degree relatives.** LMM had $r = 0$ PCs and PCA had $r = 20$. FES traits only. Each dataset is a column, rows are measures. First row has $|\text{SRMSD}_p| < 0.01$ band marked as gray area.

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 to LMM does not greatly hurt its performance (except for small sample sizes), they generally did not improve it either (Table 3), which agrees with previous observations[44] but contradicts others.[16,20] Our findings make sense since PCs are the eigenvectors of the same kinship matrix that parametrizes random effects, so including both is redundant.

Previous studies found that PCA was better calibrated than LMM for unusually differentiated markers,[20,29,31] which as simulated were an artificial scenario not based on a population genetics model, and are otherwise believed to be unusual.[32,49] Our evaluations on real human data, which contain such loci in relevant proportions if they exist, do not replicate that result. Cryptic relatedness strongly favors LMM, an advantage that probably outweighs this potential PCA benefit in real data.

Relative to LMM, the behavior of PCA fell between two extremes. When PCA performed well, there was a small number of PCs with both calibrated p-values and $\mathrm{AUC_{PR}}$ near that of LMM without PCs. Conversely, PCA performed poorly when no number of PCs had either calibrated p-values or acceptably large $\mathrm{AUC_{PR}}$. There were no cases where high numbers of PCs optimized an acceptable $\mathrm{AUC_{PR}}$, or cases with miscalibrated p-values but high $\mathrm{AUC_{PR}}$. 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 relatedness is low-dimensional, whereas LMM can handle high-dimensional relatedness. Thus, PCA performs well in the admixture simulation, which is explicitly low-dimensional (see Models and Methods), and our tree simulations, which, although complex in principle due to the large number of nodes, had few long branches so a low-dimensional approximation suffices. Conversely, PCA performs poorly under family structure because its kinship matrix is

high-dimensional (Fig. S5). However, estimating the dimensionality of real datasets is challenging because estimated eigenvalues have biased distributions. Dimensionality estimated using the Tracy-Widom test[7] did not fully predict the datasets that PCA performs well on. In contrast, estimated local kinship finds considerable cryptic relatedness in all real human datasets and better explains why PCA performs poorly there. The trait model also influences the relative performance of PCA, so genotype-only parameters (eigenvalues or local kinship) alone do not tell the full story.

PCA is at best underpowered relative to LMMs, and at worst miscalibrated regardless of the numbers of PCs included, in real human genotype tests. Among our simulations, such poor performance occurred only in the admixed family. Local kinship estimates reveal considerable family relatedness in the real datasets absent in the corresponding tree simulations. Admixture is also 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,[82–85] 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.[86,87] Our FES trait tests show that cryptic relatedness is more challenging 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,[5,30] 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 observed for LMM with large numbers of PCs), a weakness of the score test used by the LMM we evaluated that may be overcome with other statistical tests. Simulations or post hoc evaluations remain crucial for ensuring 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,

an observation potentially confounded by the use of low-dimensional simulations without family relatedness (Table 1). An additional concern is case-control ascertainment bias, which appears to affect LMMs more severely, although recent work appears to solve this problem.[29,33] Future evaluations should aim to include our 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 expected for diseases by various selection models.

Overall, our results lead us to recommend LMM over PCA for association studies in general. Although PCA offer flexibility and speed compared to LMM, 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 PCA may perform poorly 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 compared to LMM. 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.[18,80] 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 these sample $\hat{p}_i^T$ relative to the true p_i^T.[28] 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.[28] 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 $\hat{W}^T = \frac{1}{m}\sum_{i=1}^{m}\left(\hat{p}_i^T - \frac{1}{2}\right)^2$, 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 $\mathrm{E}\left[p'\right] = \frac{1}{2}$. The resulting variance is

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

which we equate to the desired V^T (Eq. (9)) and solve for w. For simplicity, we also 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.[1,20] The inflation factor λ is defined as the median χ^2 association statistic divided by theoretical median under the null hypothesis.[2] To compare p-values from non-χ^2 tests (such as t-statistics), λ can be calculated from p-values using

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

where p_{median} is the median observed p-value (including 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).

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 from genomic control,[2] 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 $\mathrm{SRMSD}_p > 0$. Deflated statistics (conservative p-values) give $\lambda < 1$ and $\mathrm{SRMSD}_p < 0$. Thus, $\lambda \neq 1$ always implies $\mathrm{SRMSD}_p \neq 0$ (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

There is a near one-to-one correspondence between λ and SRMSD_p in our data (Fig. S1). PCA tended to be inflated ($\lambda > 1$ and $\mathrm{SRMSD}_p > 0$) 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 a 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,[20] 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 control, power may rank models differently depending on t.

Power is not meaningful when p-values are not calibrated. To establish a clear connection to

$\mathrm{AUC_{PR}}$, assume calibrated (uniform) null p-values: $\Pr(p < t|H_0) = t$. TPs, FPs, and FNs at t are

$$\mathrm{TP}(t) = m\pi_1 F(t),$$
$$\mathrm{FP}(t) = m\pi_0 t,$$
$$\mathrm{FN}(t) = m\pi_1(1 - F(t)),$$

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

$$\mathrm{Precision}(t) = \frac{\pi_1 F(t)}{\pi_1 F(t) + \pi_0 t},$$
$$\mathrm{Recall}(t) = F(t).$$

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

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

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

Lastly, consider a simple yet common case in which model A is uniformly more powerful than model B: $F_A(t) > F_B(t)$ for every t. Therefore $F_A^{-1}(\mathrm{Recall}) < F_B^{-1}(\mathrm{Recall})$ for every recall value. This ensures that the precision of A is greater than that of B at every recall value, so $\mathrm{AUC_{PR}}$ is greater for A than B. Thus, $\mathrm{AUC_{PR}}$ ranks calibrated models according to 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

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

[1] Astle, W. and Balding, D. J. (2009). Population Structure and Cryptic Relatedness in Genetic Association Studies. Statist. Sci. *24*, 451–471.

[2] Devlin, B. and Roeder, K. (1999). Genomic Control for Association Studies. Biometrics. *55*, 997–1004.

[3] Voight, B. F. and Pritchard, J. K. (2005). Confounding from Cryptic Relatedness in Case-Control Association Studies. PLOS Genetics. *1*, e32.

[4] Zhang, S., Zhu, X., and Zhao, H. (2003). On a semiparametric test to detect associations between quantitative traits and candidate genes using unrelated individuals. Genetic Epidemiology. *24*, 44–56.

[5] Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909.

[6] Bouaziz, M., Ambroise, C., and Guedj, M. (2011). Accounting for Population Stratification in Practice: A Comparison of the Main Strategies Dedicated to Genome-Wide Association Studies. PLOS ONE. *6*, e28845.

[7] Patterson, N., Price, A. L., and Reich, D. (2006). Population Structure and Eigenanalysis. PLoS Genet. *2*, e190.

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

[9] Pritchard, J. K., Stephens, M., Rosenberg, N. A., and Donnelly, P. (2000). Association Mapping in Structured Populations. The American Journal of Human Genetics. *67*, 170–181.

[10] Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664.

[11] Zhou, Q., Zhao, L., and Guan, Y. (2016). Strong Selection at MHC in Mexicans since Admixture. PLoS Genet. *12*, e1005847.

[12] McVean, G. (2009). A genealogical interpretation of principal components analysis. PLoS Genet. *5*, e1000686.

[13] Zheng, X. and Weir, B. S. (2016). Eigenanalysis of SNP data with an identity by descent interpretation. Theor Popul Biol. *107*, 65–76.

[14] Cabreros, I. and Storey, J. D. (2019). A Likelihood-Free Estimator of Population Structure Bridging Admixture Models and Principal Components Analysis. Genetics. *212*, 1009–1029.

[15] Chiu, A. M., Molloy, E. K., Tan, Z., Talwalkar, A., and Sankararaman, S. (2022). Inferring population structure in biobank-scale genomic data. The American Journal of Human Genetics. *0*.

[16] Zhao, K., Aranzana, M. J., Kim, S., Lister, C., Shindo, C., Tang, C., Toomajian, C., Zheng, H., Dean, C., Marjoram, P., et al. (2007). An Arabidopsis Example of Association Mapping in Structured Samples. PLOS Genetics. *3*, e4.

[17] Xu, H. and Guan, Y. (2014). Detecting Local Haplotype Sharing and Haplotype Association. Genetics. *197*, 823–838.

[18] Qian, J., Tanigawa, Y., Du, W., Aguirre, M., Chang, C., Tibshirani, R., Rivas, M. A., and Hastie, T. (2020). A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK Biobank. PLOS Genetics. *16*, e1009141.

[19] Thornton, T. and McPeek, M. S. (2010). ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure. Am. J. Hum. Genet. *86*, 172–184.

[20] Price, A. L., Zaitlen, N. A., Reich, D., and Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. Nature Reviews Genetics. *11*, 459–463.

[21] Lee, S., Epstein, M. P., Duncan, R., and Lin, X. (2012). Sparse Principal Component Analysis for Identifying Ancestry-Informative Markers in Genome-Wide Association Studies. Genetic Epidemiology. *36*, 293–302.

[22] Abraham, G. and Inouye, M. (2014). Fast Principal Component Analysis of Large-Scale Genome-Wide Data. PLOS ONE. *9*, e93766.

[23] Galinsky, K., Bhatia, G., Loh, P.-R., Georgiev, S., Mukherjee, S., Patterson, N., and Price, A. (2016). Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia. The American Journal of Human Genetics. *98*, 456–472.

[24] Abraham, G., Qiu, Y., and Inouye, M. (2017). FlashPCA2: principal component analysis of Biobank-scale genotype datasets. Bioinformatics. *33*, 2776–2778.

[25] Agrawal, A., Chiu, A. M., Le, M., Halperin, E., and Sankararaman, S. (2020). Scalable probabilistic PCA for large-scale genetic variation data. PLOS Genetics. *16*, e1008773.

[26] Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. *38*, 203–208.

[27] Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., and Eskin, E. (2008). Efficient control of population structure in model organism association mapping. Genetics. *178*, 1709–1723.

[28] Ochoa, A. and Storey, J. D. (2021). Estimating FST and kinship for arbitrary population structures. PLoS Genet. *17*, e1009241.

[29] Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M., and Price, A. L. (2014). Advantages and pitfalls in the application of mixed-model association methods. Nat Genet. *46*, 100–106.

[30] Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-Y., Freimer, N. B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. *42*, 348–354.

[31] Wu, C., DeWan, A., Hoh, J., and Wang, Z. (2011). A Comparison of Association Methods Correcting for Population Stratification in Case–Control Studies. Annals of Human Genetics. *75*, 418–427.

[32] Sul, J. H. and Eskin, E. (2013). Mixed models can correct for population structure for genomic regions under selection. Nature Reviews Genetics. *14*, 300.

[33] Zhou, W., Nielsen, J. B., Fritsche, L. G., Dey, R., Gabrielsen, M. E., Wolford, B. N., LeFaive, J., VandeHaar, P., Gagliano, S. A., Gifford, A., et al. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat Genet. *50*, 1335–1341.

[34] Aulchenko, Y. S., Koning, D.-J. de, and Haley, C. (2007). Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. Genetics. *177*, 577–585.

[35] Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., Bradbury, P. J., Yu, J., Arnett, D. K., Ordovas, J. M., et al. (2010). Mixed linear model approach adapted for genome-wide association studies. Nat Genet. *42*, 355–360.

[36] Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., and Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. Nat. Methods. *8*, 833–835.

[37] Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. *88*, 76–82.

[38] Listgarten, J., Lippert, C., Kadie, C. M., Davidson, R. I., Eskin, E., and Heckerman, D. (2012). Improved linear mixed models for genome-wide association studies. Nat Methods. *9*, 525–526.

[39] Zhou, X. and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. *44*, 821–824.

[40] Svishcheva, G. R., Axenovich, T. I., Belonogova, N. M., Duijn, C. M. van, and Aulchenko, Y. S. (2012). Rapid variance components–based method for whole-genome association analysis. Nat Genet. *44*, 1166–1170.

[41] Loh, P.-R., Tucker, G., Bulik-Sullivan, B. K., Vilhjálmsson, B. J., Finucane, H. K., Salem, R. M., Chasman, D. I., Ridker, P. M., Neale, B. M., Berger, B., et al. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat. Genet. *47*, 284–290.

[42] Hoffman, G. E. (2013). Correcting for population structure and kinship using the linear mixed model: theory and extensions. PLoS ONE. *8*, e75707.

[43] Tucker, G., Price, A. L., and Berger, B. (2014). Improving the Power of GWAS and Avoiding Confounding from Population Stratification with PC-Select. Genetics. *197*, 1045–1049.

[44] Liu, N., Zhao, H., Patki, A., Limdi, N. A., and Allison, D. B. (2011). Controlling Population Structure in Human Genetic Association Studies with Samples of Unrelated Individuals. Stat Interface. *4*, 317–326.

[45] Zeng, J., Vlaming, R., Wu, Y., Robinson, M. R., Lloyd-Jones, L. R., Yengo, L., Yap, C. X., Xue, A., Sidorenko, J., McRae, A. F., et al. (2018). Signatures of negative selection in the genetic architecture of human complex traits. Nature Genetics. *50*, 746–753.

[46] Song, M., Hao, W., and Storey, J. D. (2015). Testing for genetic associations in arbitrarily structured populations. Nat. Genet. *47*, 550–554.

[47] Liu, X., Huang, M., Fan, B., Buckler, E. S., and Zhang, Z. (2016). Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies. PLOS Genet. *12*, e1005767.

[48] Sul, J. H., Martin, L. S., and Eskin, E. (2018). Population structure in genetic studies: Confounding factors and mixed models. PLoS Genet. *14*, e1007309.

[49] Price, A. L., Zaitlen, N. A., Reich, D., and Patterson, N. (2013). Response to Sul and Eskin. Nature Reviews Genetics. *14*, 300.

[50] Consortium, T. G. P. (2010). A map of human genome variation from population-scale sequencing. Nature. *467*, 1061–1073.

[51] 1000 Genomes Project Consortium, Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T., and McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature. *491*, 56–65.

[52] Cann, H. M., Toma, C. de, Cazes, L., Legrand, M.-F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W. F., Bonne-Tamir, B., Cambon-Thomsen, A., et al. (2002). A human genome diversity cell line panel. Science. *296*, 261–262.

[53] Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., and Feldman, M. W. (2002). Genetic Structure of Human Populations. Science. *298*, 2381–2385.

[54] Bergström, A., McCarthy, S. A., Hui, R., Almarri, M. A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., et al. (2020). Insights into human genetic variation and population history from 929 diverse genomes. Science. *367*.

[55] Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. (2012). Ancient admixture in human history. Genetics. *192*, 1065–1093.

[56] Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., Sudmant, P. H., Schraiber, J. G., Castellano, S., Lipson, M., et al. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature. *513*, 409–413.

[57] Lazaridis, I., Nadel, D., Rollefson, G., Merrett, D. C., Rohland, N., Mallick, S., Fernandes, D., Novak, M., Gamarra, B., Sirak, K., et al. (2016). Genomic insights into the origin of farming in the ancient Near East. Nature. *536*, 419–424.

[58] Skoglund, P., Posth, C., Sirak, K., Spriggs, M., Valentin, F., Bedford, S., Clark, G. R., Reepmeyer, C., Petchey, F., Fernandes, D., et al. (2016). Genomic insights into the peopling of the Southwest Pacific. Nature. *538*, 510–513.

[59] Park, J.-H., Gail, M. H., Weinberg, C. R., Carroll, R. J., Chung, C. C., Wang, Z., Chanock, S. J., Fraumeni, J. F., and Chatterjee, N. (2011). Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. PNAS. *108*, 18026–18031.

[60] O'Connor, L. J., Schoech, A. P., Hormozdiari, F., Gazal, S., Patterson, N., and Price, A. L. (2019). Extreme Polygenicity of Complex Traits Is Explained by Negative Selection. The American Journal of Human Genetics. *0*.

[61] Simons, Y. B., Bullaughey, K., Hudson, R. R., and Sella, G. (2018). A population genetic interpretation of GWAS findings for human quantitative traits. PLOS Biology. *16*, e2002985.

[62] Malécot, G. (1948). *Mathématiques de l'hérédité.* (Masson et Cie).

[63] Wright, S. (1951). The genetical structure of populations. Ann Eugen. *15*, 323–354.

[64] Jacquard, A. (1970). *Structures génétiques des populations.* (Paris: Masson et Cie).

[65] Ochoa, A. and Storey, J. D. (2019). New kinship and F_{ST} estimates reveal higher levels of differentiation in the global human population. bioRxiv.

[66] Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. *4*, 7.

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

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

[69] Sokal, R. R. and Michener, C. D. (1958). A statistical method for evaluating systematic relationships. Univ. Kansas, Sci. Bull. *38*, 1409–1438.

[70] Lawson, C. L. and Hanson, R. J. (1974). Solving least squares problems prentice-hall. Englewood Cliffs.

[71] Mullen, K. M. and Stokkum, I. H. M. v. (2012). *nnls: The Lawson-Hanson algorithm for non-negative least squares (NNLS)*. 2012.

[72] Park, J.-H., Wacholder, S., Gail, M. H., Peters, U., Jacobs, K. B., Chanock, S. J., and Chatterjee, N. (2010). Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nature Genetics. *42*, 570–575.

[73] Grueneberg, A. and Campos, G. d. l. (2019). BGData - A Suite of R Packages for Genomic Analysis with Big Data. G3: Genes, Genomes, Genetics. *9*, 1377–1383.

[74] Fairley, S., Lowy-Gallego, E., Perry, E., and Flicek, P. (2020). The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. Nucleic Acids Research. *48*, D941–D947.

[75] Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., and Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics. *26*, 2867–2873.

[76] Storey, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. Ann. Statist. *31*, 2013–2035.

[77] Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences of the United States of America. *100*, 9440–9445.

[78] Grau, J., Grosse, I., and Keilwagen, J. (2015). PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. Bioinformatics. *31*, 2595–2597.

[79] Gopalan, P., Hao, W., Blei, D. M., and Storey, J. D. (2016). Scaling probabilistic models of genetic variation to millions of humans. Nat. Genet. *48*, 1587–1590.

[80] Rakitsch, B., Lippert, C., Stegle, O., and Borgwardt, K. (2013). A Lasso multi-marker mixed model for association mapping with population structure correction. Bioinformatics. *29*, 206–214.

[81] Conomos, M., Reiner, A., Weir, B., and Thornton, T. (2016). Model-free Estimation of Recent Genetic Relatedness. The American Journal of Human Genetics. *98*, 127–148.

[82] Gazal, S., Sahbatou, M., Babron, M.-C., Génin, E., and Leutenegger, A.-L. (2015). High level of inbreeding in final phase of 1000 Genomes Project. Sci Rep. *5*, 17453.

[83] Al-Khudhair, A., Qiu, S., Wyse, M., Chowdhury, S., Cheng, X., Bekbolsynov, D., Saha-Mandal, A., Dutta, R., Fedorova, L., and Fedorov, A. (2015). Inference of Distant Genetic Relations in Humans Using "1000 Genomes". Genome Biology and Evolution. *7*, 481–492.

[84] Fedorova, L., Qiu, S., Dutta, R., and Fedorov, A. (2016). Atlas of Cryptic Genetic Relatedness Among 1000 Human Genomes. Genome Biology and Evolution. *8*, 777–790.

[85] Schlauch, D., Fier, H., and Lange, C. (2017). Identification of genetic outliers due to substructure and cryptic relationships. Bioinformatics. *33*, 1972–1979.

[86] Henn, B. M., Hon, L., Macpherson, J. M., Eriksson, N., Saxonov, S., Pe'er, I., and Mountain, J. L. (2012). Cryptic Distant Relatives Are Common in Both Isolated and Cosmopolitan Genetic Samples. PLOS ONE. *7*, e34267.

[87] Shchur, V. and Nielsen, R. (2018). On the number of siblings and p-th cousins in a large population sample. J Math Biol. *77*, 1279–1298.

S1 Supplemental figures

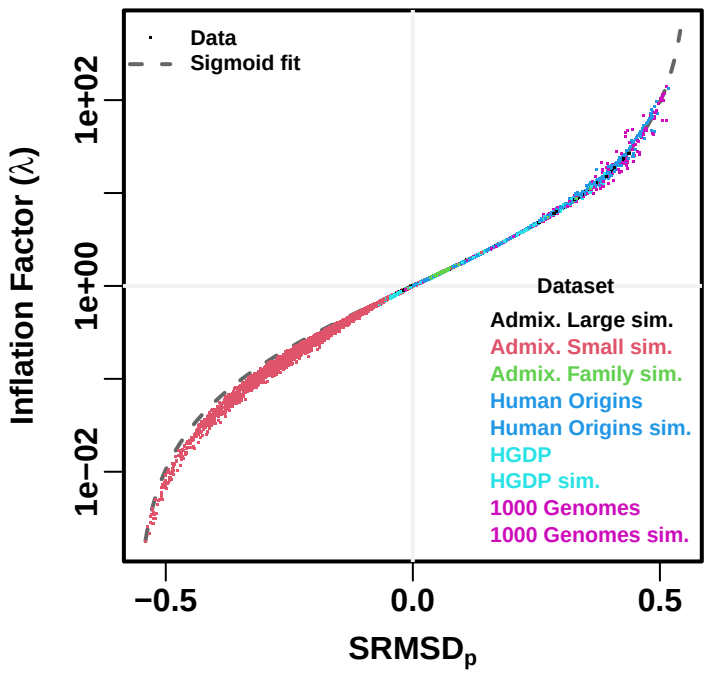


Figure S1: **Comparison between SRMSD$_p$ and inflation factor.** Each point is a pair of statistics for one replicate, one association model (PCA or LMM with some number of PCs r), one trait model (FES vs RC), and one dataset (color coded by dataset). Note log y-axis (λ). The sigmoidal curve in Eq.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

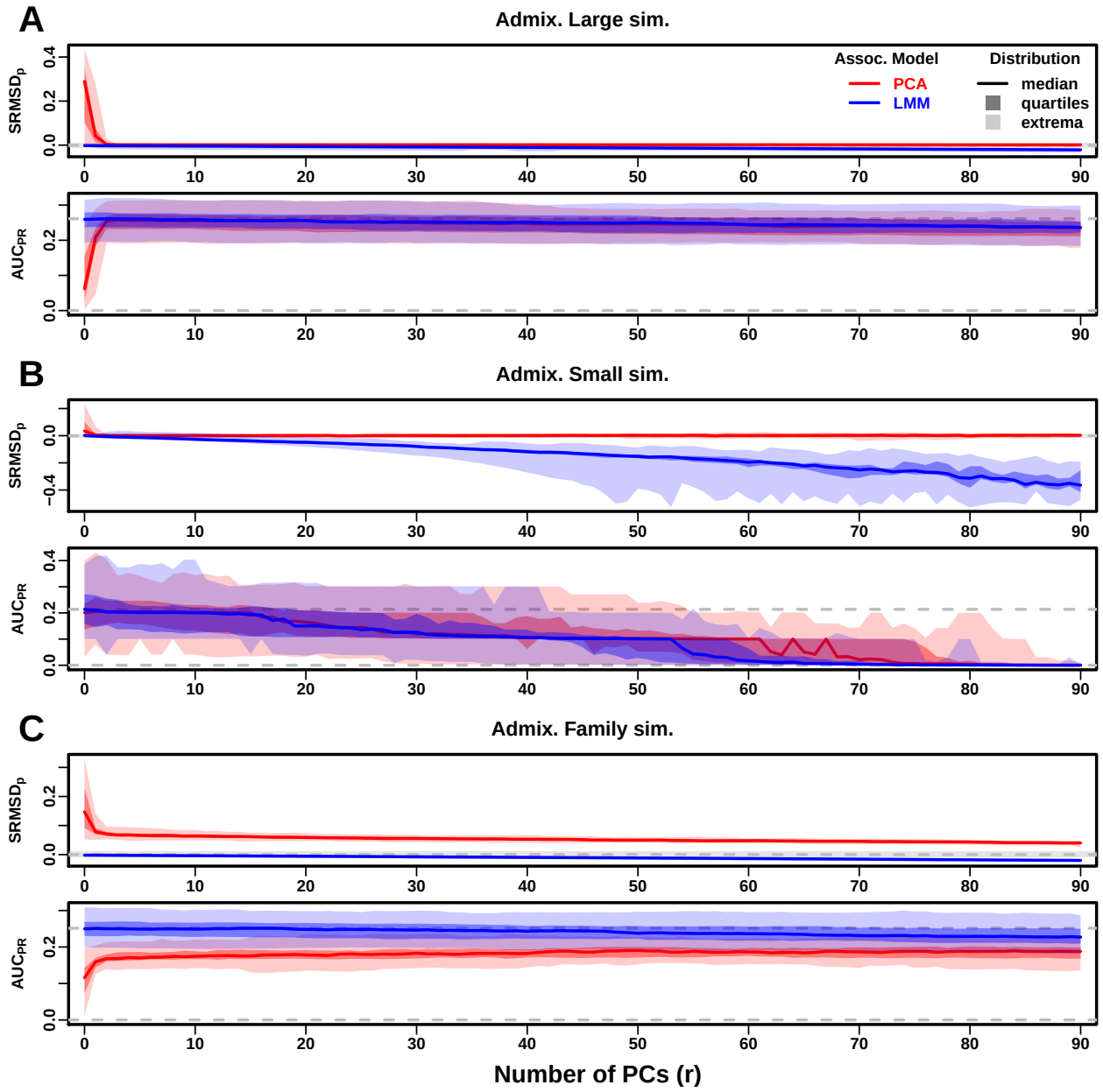


Figure S2: **Evaluations in admixture simulations with RC traits.** Traits simulated from RC model, otherwise the same as Fig. 3.

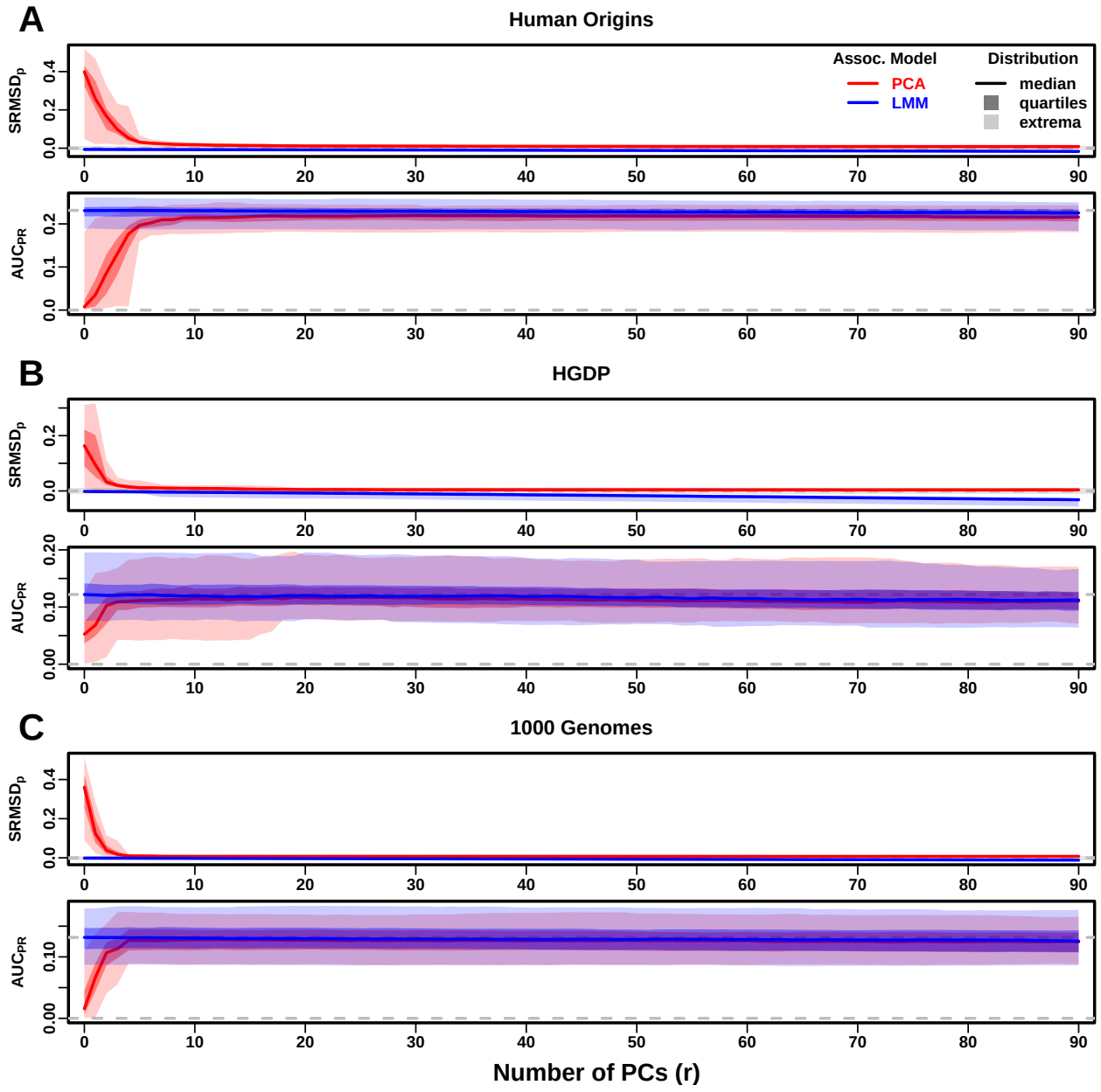


Figure S3: **Evaluations in real human genotype datasets with RC traits.** Traits simulated from RC model, otherwise the same as Fig. 4.

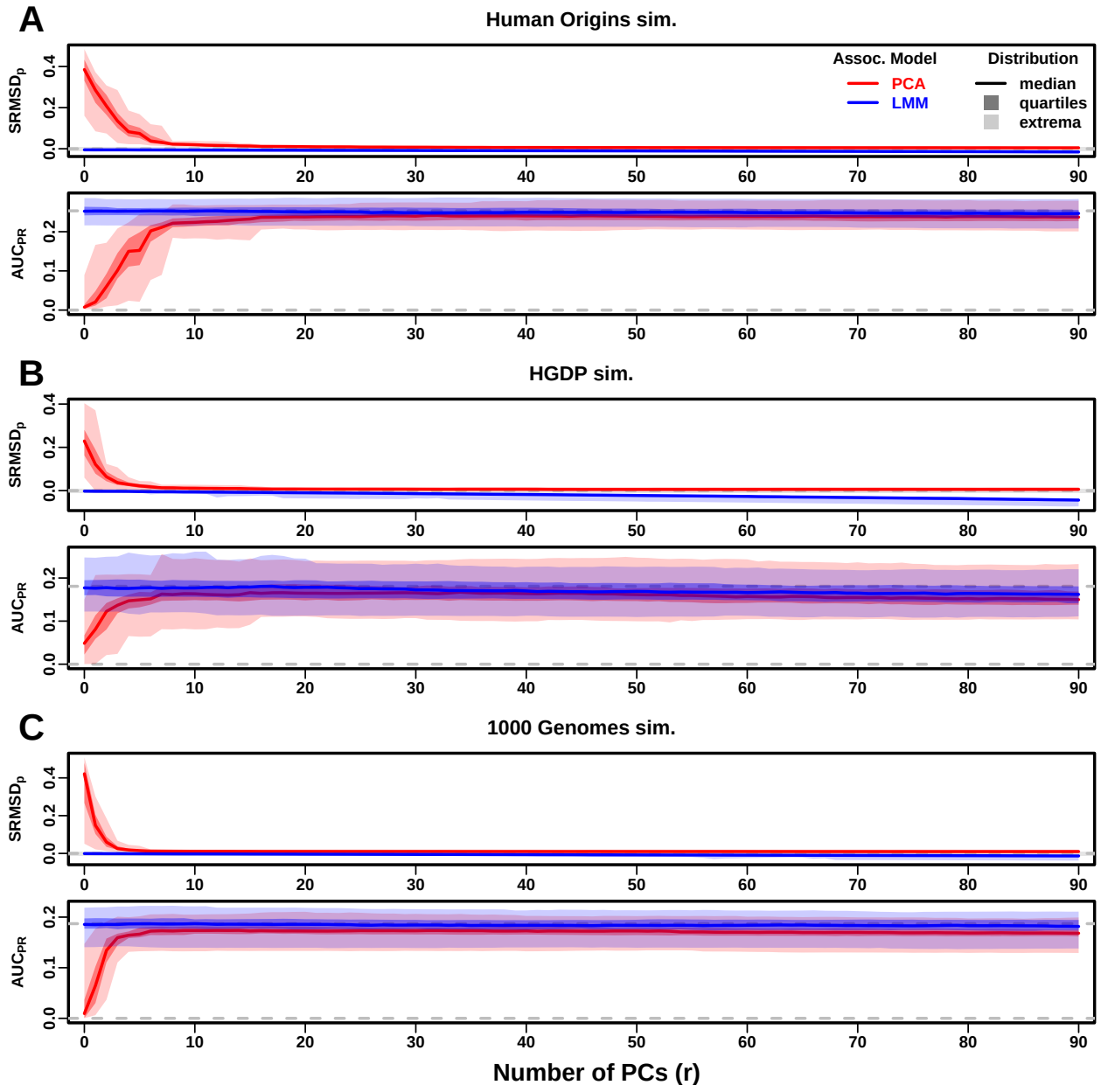


Figure S4: **Evaluations in tree simulations fit to human data with RC traits.** Traits simulated from RC model, otherwise the same as Fig. 5.

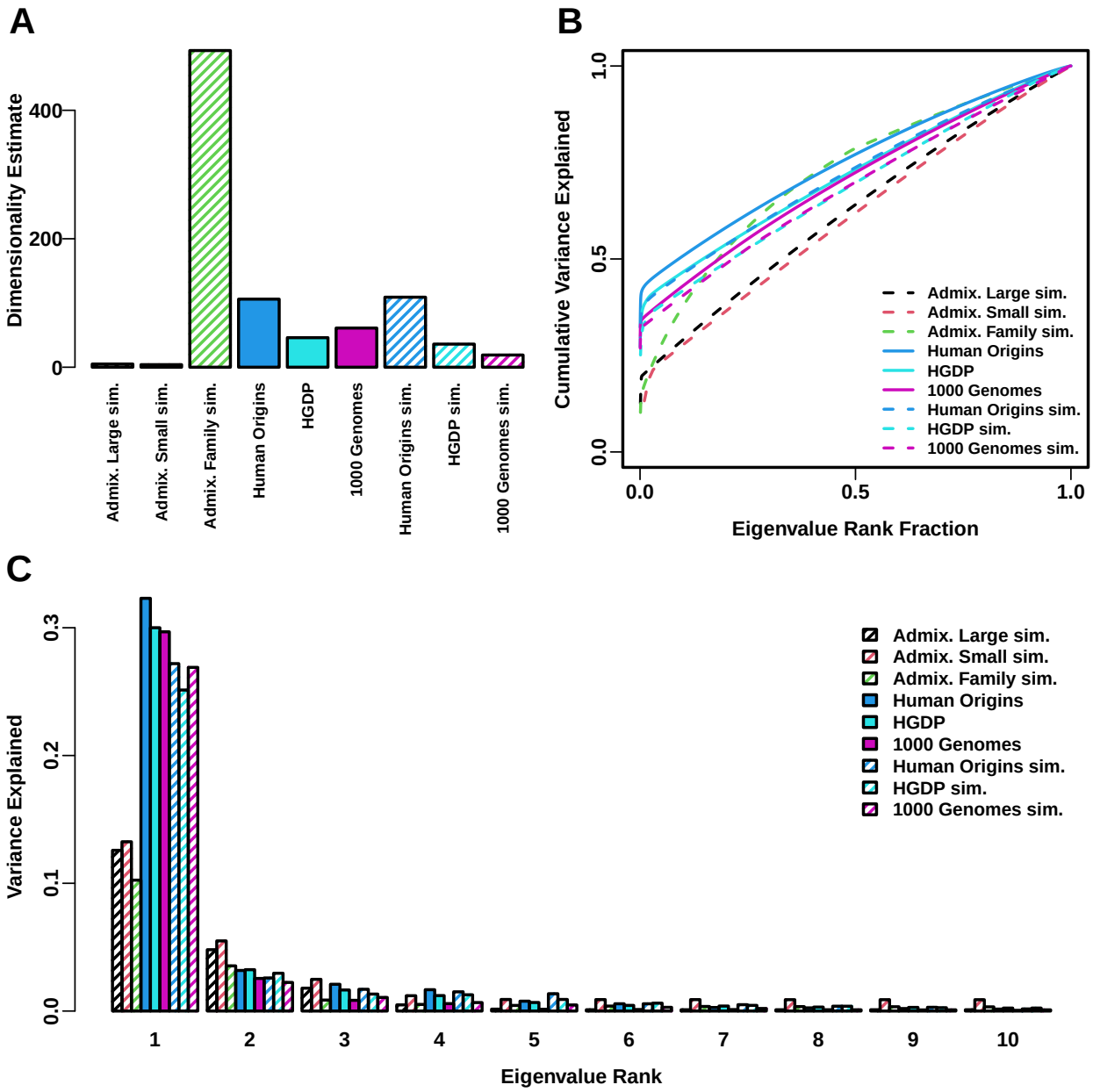


Figure S5: **Estimated dimensionality of datasets. A.** Kinship dimensionalities estimated with the Tracy-Widom test with $p < 0.01$. **B.** Cumulative variance explained versus eigenvalue rank fraction. **C.** Variance explained by first 10 eigenvalues.

S2 Supplemental 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