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 genetic association models. Previous PCA-LMM comparisons give mixed results, unclear guidance, and have several limitations, including not varying the number of principal components (PCs), simulating simple population structures, and inconsistent use of real data and power evaluations. We evaluate PCA and LMM both varying number of PCs in realistic genotype and complex trait simulations including admixed families, subpopulation trees, and real multiethnic human datasets with simulated traits. We find that LMM without PCs usually performs best, with the largest effects in family simulations and real human datasets and traits without environment effects. Poor PCA performance on human datasets is driven by large numbers of distant relatives more than the smaller number of closer relatives. While PCA was known to fail on family data, we report strong effects of family relatedness in genetically diverse human datasets, not avoided by pruning close relatives. Environment effects driven by geography and ethnicity are better modeled with LMM including those labels instead of PCs. This work better characterizes the severe limitations of PCA compared to LMM in modeling the complex relatedness structures of multiethnic human data for 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; LD: linkage disequilibrium.

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]. Association studies of admixed and multiethnic cohorts, the focus of this work, are becoming more common, are believed to be more powerful, and are necessary to bring more equity to genetic medicine [2–22]. When insufficient approaches are applied to data with relatedness, their association statistics are miscalibrated, resulting in excess false positives and loss of power [1, 23, 24]. 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 [25–27]. These top eigenvectors are a new set of coordinates for individuals that are commonly referred to as PCs in genetics [28], the convention adopted here, but in other fields PCs instead denote what in genetics would be the projections of loci onto eigenvectors, which are new independent coordinates for loci [29]. The direct ancestor of PCA association is structured association, in which inferred ancestry (genetic cluster membership, often corresponding with labels such as “European”, “African”, “Asian”, etc.) or admixture proportions of these ancestries are used as regression covariates [30]. These models are deeply connected because PCs map to ancestry empirically [31, 32] and theoretically [33–36], and they work as well as global ancestry in association studies but are estimated more easily [27, 28, 31, 37]. Another approach closely related to PCA is nonmetric multidimensional scaling [38]. PCs are also proposed for modeling environment effects that are correlated to ancestry, for example, through geography [20, 39, 40]. The strength of PCA is its simplicity, which as covariates can be readily included in more complex models, such as haplotype association [41] and polygenic models [42]. However, PCA assumes that the underlying relatedness space is low dimensional (or low rank), so it can be well modeled with a small number of PCs, which may limit its applicability. PCA is known to be inadequate for family data [28, 38, 43, 44], 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 [45–49]. 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 parameterized 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 used kinship matrices estimated from known pedigrees or using methods that captured recent relatedness only, and modeled population structure (ancestry) as fixed effects [37, 38, 50]. 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 family relatedness and ancestry [1, 51, 52]. Like PCA, LMM has also been proposed for modeling environment correlated to genetics [53, 54]. 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 [55], LMMs have been applied successfully to balanced case-control studies [1, 56] and simulations [44, 57, 58], and have been adapted for unbalanced case-control studies [59]. However, LMMs tend to be considerably slower than PCA and other models, so much effort has focused on improving their runtime and scalability [51, 56, 59–67].

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, 3, 40, 68], 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 [40, 44, 69], while others did not [68, 70], and the approach continues to be used [71, 72] though not always [73]. (Recall that early LMMs used kinship to model family relatedness only, so population structure had to be modeled separately in those models, in practice as admixture fractions instead of PCs [37, 38, 50].) The LMM with PCs (vs no PCs) is also believed to help better model loci that have experienced selection [44, 53] and environment effects correlated with genetics [40].

LMM and PCA are closely related models [1, 3, 40, 68], 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 generally used either only simulated or only real genotypes, with only two studies using both. The simulated genotype studies, which tended to have low model dimensions and F_{ST}, were more likely to report ties or mixed results (6/8), whereas real genotypes tended to clearly favor LMMs (9/11). Similarly, 10/12 papers with quantitative traits favor LMMs, whereas 6/9 papers with case-control traits gave

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

Publication	Sim. Genotypes Type[a]	Sim. Genotypes K[b]	Sim. Genotypes F_{ST} [c]	Real[d]	Trait[e]	Power	PCs (r)	Best
Zhao et al. [37]				✓	Q	✓	8	LMM
Zhu and Yu [38]	I, A, F	3, 8	≤0.15	✓	Q	✓	1-22	LMM
Astle and Balding [1]	I	3	0.10		CC	✓	10	Tie
Kang et al. [56]				✓	Both		2-100	LMM
Price et al. [44]	I, F	2	0.01		CC		1	Mixed
Wu et al. [57]	I, A	2-4	0.01		CC	✓	10	Mixed
Liu et al. [70]	S, A	2-3	R		Q	✓	10	Tie
Sul and Eskin [58]	I	2	0.01		CC		1	Tie
Tucker, Price, and Berger [69]	I	2	0.05	✓	Both	✓	5	Tie
Yang et al. [55]				✓	CC	✓	5	Tie
Song, Hao, and Storey [74]	S, A	2-3	R		Q		3	LMM
Loh et al. [67]				✓	Q	✓	10	LMM
Zhang and Pan [40]				✓	Q	✓	20-100	LMM
Liu et al. [75]				✓	Q	✓	3-6	LMM
Sul, Martin, and Eskin [76]				✓	Q		100	LMM
Loh et al. [77]				✓	Both	✓	20	LMM
Mbatchou et al. [72]				✓	Both		1	LMM
This work	A, T, F	10-243	≤0.25	✓	Q	✓	0-90	LMM

[a]Genotype simulation types. I: Independent subpopulations; S: subpopulations (with parameters drawn from real data); A: Admixture; T: Subpopulation Tree; F: Family.
[b]Model dimension (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.

ties or mixed results—the only factor we do not explore in this work. Additionally, although all previous evaluations measured type I error (or proxies such as genomic inflation factors [23] or QQ plots), a large fraction (6/17) did not measure power (or proxies such as ROC curves), and only four used more than one number of PCs for PCA. Lastly, no consensus has emerged as to why LMM might outperform PCA or vice versa [3, 44, 58, 78], or which features of the real datasets are critical for the LMM advantage other than family relatedness, resulting in unclear guidance for using PCA. Hence, our work includes real and simulated genotypes with higher model dimensions and F_{ST} matching that of multiethnic human cohorts [52, 79], we vary the number of PCs, and measure robust proxies for type I error control and calibrated power.

In this work, we evaluate the PCA and LMM association models under various numbers of PCs, which are included in LMMs too. We use genotype simulations (admixture, family, and subpopulation tree models) and three real datasets: the 1000 Genomes Project [80, 81], the Human Genome Diversity Panel (HGDP) [82–84], and Human Origins [85–88]. We simulate quantitative traits from two models: fixed effect sizes (FES) construct coefficients inverse to allele frequency, which matches real data [71, 89, 90] and corresponds to high pleiotropy and strong balancing selection [91] and strong negative selection [71, 90], which are appropriate assumptions for diseases; and random coefficients (RC), which are drawn independent of allele frequency, and corresponds to neutral traits [71, 91]. LMM without PCs consistently performs best in simulations without environment, and greatly outperforms PCA in the family simulation and in all real datasets. The tree simulations, which model subpopulations with the tree but exclude family structure, 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. We find that both LMM and PCA are able to model environment effects correlated with genetics, and LMM with PCs gains a small advantage in this setting only, but direct modeling of environment performs much better. 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:

$$\mathbf{y} = \mathbf{1}\alpha + \mathbf{X}'\boldsymbol{\beta} + \mathbf{Z}'\boldsymbol{\eta} + \boldsymbol{\epsilon}, \tag{1}$$

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, $\mathbf{Z}$ is a design matrix of environment effects and other covariates, $\boldsymbol{\eta}$ is the vector of environment coefficients, $\boldsymbol{\epsilon}$ is a length-n vector of residuals, and the prime symbol ($'$) denotes matrix transposition. The residuals follow $\epsilon_j \sim \text{Normal}(0, \sigma_\epsilon^2)$ independently per individual j, for some σ_ϵ^2.

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

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

$$\text{LMM:} \quad \mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta_i + \mathbf{s} + \mathbf{Z}'\boldsymbol{\eta} + \boldsymbol{\epsilon}, \qquad \mathbf{s} \sim \text{Normal}\left(\mathbf{0}, 2\sigma_s^2\boldsymbol{\Phi}^T\right), \tag{3}$$

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. Both models condition the regression of the focal locus i on an approximation of the total polygenic effect $\mathbf{X}'\boldsymbol{\beta}$ with the same covariance structure, which is parameterized by the kinship matrix. Under the kinship model, genotypes are random variables obeying

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

where p_i^T is the ancestral allele frequency of locus i [1, 92–94]. Assuming independent loci, the covariance of the polygenic effect is

$$\mathrm{Cov}(\mathbf{X}'\boldsymbol{\beta}) = 2\sigma_s^2 \mathbf{\Phi}^T, \qquad \sigma_s^2 = \sum_{i=1}^{m} 2p_i^T(1-p_i^T)\beta_i^2,$$

which is readily modeled by the LMM random effect $\mathbf{s}$, where the difference in mean is absorbed by the intercept. Alternatively, consider the eigendecomposition of the kinship matrix $\mathbf{\Phi}^T = \mathbf{U}\mathbf{\Lambda}\mathbf{U}'$ where $\mathbf{U}$ is the $n \times n$ eigenvector matrix and $\mathbf{\Lambda}$ is the $n \times n$ diagonal matrix of eigenvalues. The random effect can be written as

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

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 approximate rank r and the coefficients $\boldsymbol{\gamma}_r$ are fit without constraints. In contrast, the LMM uses all eigenvectors, while effectively shrinking their coefficients $\boldsymbol{\gamma}_{\mathrm{LMM}}$ as all random effects models do, although these parameters are marginalized [1, 3, 40, 68]. 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):

$$\mathbf{X}_S = \left(\frac{x_{ij} - 2\hat{p}_i^T}{\sqrt{4\hat{p}_i^T\left(1-\hat{p}_i^T\right)}} \right), \qquad \hat{\mathbf{\Phi}}^T = \frac{1}{m}\mathbf{X}_S'\mathbf{X}_S, \tag{5}$$

where the unknown p_i^T is estimated by $\hat{p}_i^T = \frac{1}{2n}\sum_{j=1}^n x_{ij}$ [26, 51, 55, 56, 59, 63, 65, 67, 76]. However, this kinship estimator has a complex bias that differs for every individual pair, which arises due to the use of this estimated $\hat{p}_i^T$ [52, 79]. Nevertheless, in PCA and LMM these biased estimates perform as well as unbiased ones [95].

We selected fast and robust software implementing the basic PCA and LMM models. PCA association was performed with `plink2` [96]. 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 [55, 63]. Its kinship estimator equals Eq. (5). PCs were calculated using GCTA from its kinship estimate. Association significance is evaluated with a score test. In the small simulation only, GCTA with large numbers of PCs 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, which is 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 [52] and is implemented in the R package `bnpsd`. Both Large and Family simulations 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 $f_{S_u}^T = u\tau$ 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 [52]. Random ancestral allele frequencies p_i^T, subpopulation allele frequencies $p_i^{S_u}$, individual-specific allele frequencies π_{ij}, and genotypes x_{ij}

are drawn from this hierarchical model:

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

where this Beta is the Balding-Nichols distribution [97] with mean p_i^T and variance $p_i^T \left(1 - p_i^T\right) f_{S_u}^T$. 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 genotypes 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 $\mathbf{X}$, which is the subspace to which the data is constrained by the admixture model, is given by (π_{ij}), which 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 the 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 per family 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 subpopulation tree model

This model draws subpopulations allele frequencies from a hierarchical model parameterized by a tree, which is also implemented in `bnpsd` and relies on the R package `ape` for general tree data structures and methods [98]. 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 $f^{P_w}_{S_w}$. p^T_i are drawn from a given distribution, which is constructed to mimic each real dataset in Appendix A. Given the allele frequencies $p^{P_w}_i$ of the parent population, S_w's allele frequencies are drawn from:

$$p^{S_w}_i | p^{P_w}_i \sim \text{Beta}\left(p^{P_w}_i \left(\frac{1}{f^{P_w}_{S_w}} - 1 \right), \left(1 - p^{P_w}_i\right) \left(\frac{1}{f^{P_w}_{S_w}} - 1 \right) \right).$$

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

2.2.4 Fitting subpopulation 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 coefficients $f_{S_w}^{P_w}$ for its edges (between subpopulation S_w and its parent 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 these edge inbreeding coefficients 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 [52]:

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

These $\delta_w \geq 0$ because $0 \leq f_{S_w}^{P_w}, f_{P_w}^T \leq 1$ for every w. Edge inbreeding coefficients 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` [52]. 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 [99], 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 [100] (implemented in `nnls` [101]) 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` [52]. This simulation is implemented in the R package `simtrait`. All simulations have a fixed narrow-sense heritability of h^2, a variance proportion due to environment effects σ^2_η, and residuals are drawn from $\epsilon_j \sim \text{Normal}(0, \sigma^2_\epsilon)$ with $\sigma^2_\epsilon = 1 - h^2 - \sigma^2_\eta$. The number of causal loci m_1, which determines the average coefficient size, is chosen with the heuristic formula $m_1 = \text{round}(nh^2/8)$, which empirically balances power well with varying n and h^2. The set of causal loci C is drawn anew for each replicate, from loci with MAF ≥ 0.01 to avoid rare causal variants, which are not discoverable by PCA or LMM at the sample sizes we considered. 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 [102]. 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 [52] $\hat{v}^T_i = \hat{p}^T_i \left(1 - \hat{p}^T_i\right) / (1 - \bar{\varphi}^T)$, 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 $\sigma_0^2 = \sum_{i \in C} 2 v_i^T \beta_i^2$, 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 $\alpha = -\sum_{i \in C} 2 p_i^T \beta_i$. 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

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

Simulations optionally included multiple environment group effects, similarly to previous models [40, 54], as follows. Each independent environment i has predefined groups, and each group g has random coefficients drawn independent from $\eta_{gi} \sim \text{Normal}(0, \sigma_{\eta i}^2)$ where $\sigma_{\eta i}^2$ is a specified variance proportion for environment i. $\mathbf{Z}$ has individuals along columns and environment-groups along rows, and it contains indicator variables: 1 if the individual belongs to the environment-group, 0 otherwise.

We performed trait simulations with the following variance parameters (Table 2): *high heritability* used $h^2 = 0.8$ and no environment effects; *low heritability* used $h^2 = 0.3$ and no environment effects; lastly, *environment* used $h^2 = 0.3, \sigma_{\eta 1}^2 = 0.3, \sigma_{\eta 2}^2 = 0.2$ (total $\sigma_\eta^2 = \sigma_{\eta 1}^2 + \sigma_{\eta 2}^2 = 0.5$). For real genotype datasets, the groups are the continental (environment 1) and fine-grained (environment 2) subpopulation labels given (see next subsection). For simulated genotypes, we created these labels by grouping by the index j (geographical coordinate) of each simulated individual, assigning group $g = \text{ceiling}(j k_i / n)$ where k_i is the number of groups in environment i, and we selected $k_1 = 5$ and $k_2 = 25$ to mimic the number of groups in each level of 1000 Genomes (Table 3).

Table 2: **Variance parameters of trait simulations.**

Trait variance type	h^2	σ_η^2	σ_ϵ^2
High heritability	0.8	0.0	0.2
Low heritability	0.3	0.0	0.7
Environment	0.3	0.5	0.2

2.3 Real human genotype datasets

The three datasets were processed as before [79] (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` [96], and analysis was uniquely enabled by the R packages `BEDMatrix` [103] and `genio`. Each dataset groups individuals in a two-level hierarchy: continental and fine-grained subpopulations. Final dataset sizes are in Table 3.

We obtained the full (including non-public) Human Origins by contacting the authors and agreeing to their usage restrictions. The Pacific data [88] was obtained separately from the rest [86, 87], 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 both the whole-genome sequencing (WGS) version of HGDP [84] and the high-coverage NYGC version of 1000 Genomes [104] was restricted to autosomal biallelic SNP loci with filter "PASS".

Since our evaluations assume uncorrelated loci, we filtered each real dataset with `plink2` using parameters "`--indep-pairwise 1000kb 0.3`", which iteratively removes loci that have a greater than 0.3 squared 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. Lastly,

Table 3: **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	771,322	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	771,322	929	54	93	0.25
1000 Genomes sim.	Tree	1,111,266	2504	26	250	0.21

[a]For admixed family, ignores additional model dimension of 20 generation pedigree structure. For real datasets, lower range is continental subpopulations, upper range is number of fine-grained subpopulations.
[b]$m_1 = \text{round}(nh^2/8)$ to balance power across datasets, shown for $h^2 = 0.8$ only.
[c]Model parameter for simulations, estimated value on real datasets.

only HGDP had loci with over 10% missingness removed, as they were otherwise 17% of remaining loci (for Human Origins and 1000 Genomes they were under 1% of loci so they were not removed). Kinship matrix rank and eigenvalues were calculated from `popkin` kinship estimates. Eigenvalues were assigned p-values with `twstats` of the Eigensoft package [28], and kinship matrix 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 [105], and MAF < 0.01 filter is reapplied (Table S1).

2.4 Evaluation of performance

All approaches are evaluated using two complementary metrics: SRMSD_p quantifies p-value uniformity, and $\mathrm{AUC}_{\mathrm{PR}}$ 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 [106, 107]. 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:

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

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, $u_{\mathrm{median}} = \frac{1}{2}$ is its expectation, and sgn is the sign function (1 if $u_{\mathrm{median}} \geq p_{\mathrm{median}}$, -1 otherwise). Thus, $\mathrm{SRMSD}_p = 0$ corresponds to calibrated p-values, $\mathrm{SRMSD}_p > 0$ indicate anti-conservative p-values, and $\mathrm{SRMSD}_p < 0$ 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

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

The same value with a 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 [108]. 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

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

Precision and Recall trace a curve as t is varied, and the area under this curve is $\mathrm{AUC}_{\mathrm{PR}}$. We use the R package `PRROC` to integrate the correct non-linear piecewise function when interpolating between points. A model obtains the maximum $\mathrm{AUC}_{\mathrm{PR}} = 1$ if there is a t that classifies all loci perfectly. In contrast, the worst models, which classify at random, have an expected precision ($= \mathrm{AUC}_{\mathrm{PR}}$) 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 3). 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` [52] (Fig. 1). The first set of three simulated genotypes are based on an admixture model with 10 ancestries (Fig. 1A) [35, 52, 109]. 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

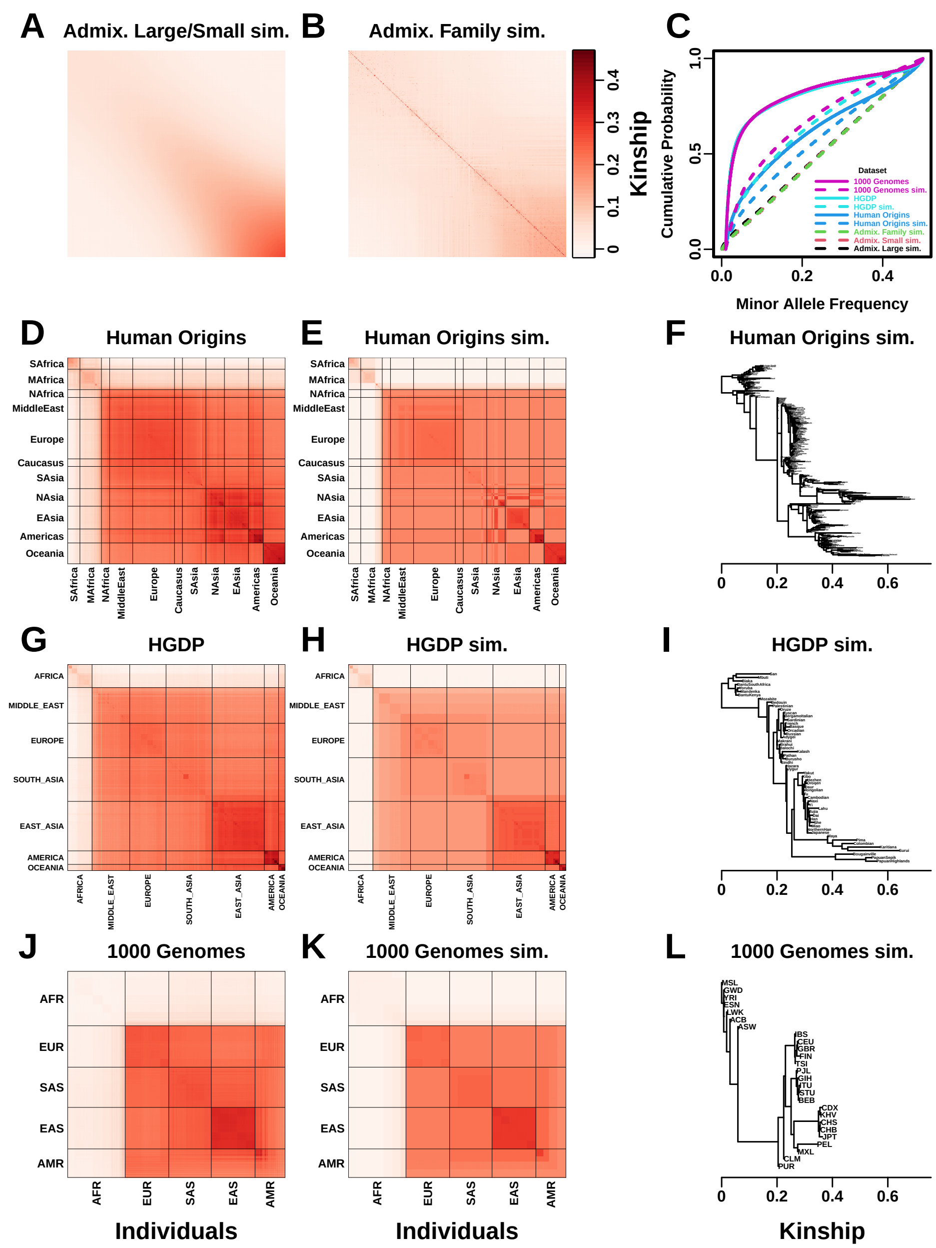

Figure 1: **Population structures of simulated and real human genotype datasets.** First two columns are population kinship matrices as heatmaps: individuals along x- and y-axis, kinship as color. Diagonal shows inbreeding values. **A.** Admixture scenario for both Large and Small simulations. **B.** 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 subpopulation 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.

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 subpopulation 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 [71, 89–91], so it is the focus of our results. The RC model draws coefficients independent of allele frequency, corresponding to neutral traits [71, 91], 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) 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 control measures; there is a correspondence between λ and SRMSD_p, with $\text{SRMSD}_p > 0.01$ giving $\lambda > 1.06$ (Fig. S1) and thus evidence of miscalibration close to the rule of thumb of $\lambda > 1.05$ [44]. There is also a monotonic correspondence between SRMSD_p and type I error rate (Fig. S2). AUC_{PR} has been used to evaluate association models [110], and reflects calibrated statistical power (Fig. S3) while being robust to miscalibrated models (Appendix B).

Both PCA and LMM are 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 $|\text{SRMSD}_p|$ over replicates) is typically large across all datasets (Table 4), although much smaller r values often performed as well (shown in following sections). Most cases have a mean $|\text{SRMSD}_p| < 0.01$, whose p-values are effectively calibrated. However, PCA is often miscalibrated on the family simulation and real datasets (Table 4). In contrast, for LMM, $r = 0$ (no PCs) is always best, and is always calibrated. Comparing LMM with $r = 0$ to PCA with its best

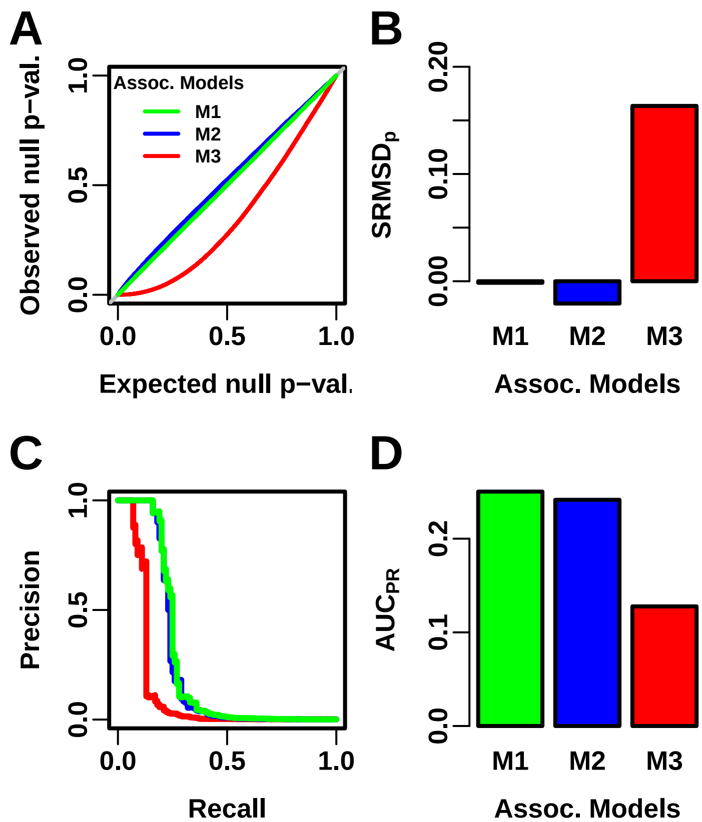

Figure 2: **Illustration of evaluation measures.** Three archetypal models illustrate our complementary measures: M1 is ideal, M2 overfits slightly, M3 is naive. **A.** QQ plot of p-values of "null" (non-causal) loci. M1 has desired uniform p-values, M2/M3 are miscalibrated. **B.** SRMSD$_p$ (p-value 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.** AUC$_{\mathrm{PR}}$ (Area Under the PR Curve) reflects power (higher is better).

Table 4: **Overview of PCA and LMM evaluations for high heritability simulations**

			LMM $r = 0$ vs best r			PCA vs LMM $r = 0$			
Dataset	Metric	Trait[a]	Cal.[b]	Best r[c]	P-value[d]	Best r[c]	Cal.[b]	P-value[d]	Best model[e]
Admix. Large sim.	$\lvert\mathrm{SRMSD}_p\rvert$	FES	True	0	1	12	True	0.036	Tie
Admix. Small sim.	$\lvert\mathrm{SRMSD}_p\rvert$	FES	True	0	1	4	True	0.055	Tie
Admix. Family sim.	$\lvert\mathrm{SRMSD}_p\rvert$	FES	True	0	1	90	False	3.9e-10*	LMM
Human Origins	$\lvert\mathrm{SRMSD}_p\rvert$	FES	True	0	1	89	False	3.9e-10*	LMM
HGDP	$\lvert\mathrm{SRMSD}_p\rvert$	FES	True	0	1	87	True	4.4e-10*	LMM
1000 Genomes	$\lvert\mathrm{SRMSD}_p\rvert$	FES	True	0	1	90	False	3.9e-10*	LMM
Human Origins sim.	$\lvert\mathrm{SRMSD}_p\rvert$	FES	True	0	1	88	True	0.017	Tie
HGDP sim.	$\lvert\mathrm{SRMSD}_p\rvert$	FES	True	0	1	47	True	0.046	Tie
1000 Genomes sim.	$\lvert\mathrm{SRMSD}_p\rvert$	FES	True	0	1	78	True	9.6e-10*	LMM
Admix. Large sim.	$\lvert\mathrm{SRMSD}_p\rvert$	RC	True	0	1	26	True	0.11	Tie
Admix. Small sim.	$\lvert\mathrm{SRMSD}_p\rvert$	RC	True	0	1	4	True	0.00097	Tie
Admix. Family sim.	$\lvert\mathrm{SRMSD}_p\rvert$	RC	True	0	1	90	False	3.9e-10*	LMM
Human Origins	$\lvert\mathrm{SRMSD}_p\rvert$	RC	True	0	1	90	True	0.00065	Tie
HGDP	$\lvert\mathrm{SRMSD}_p\rvert$	RC	True	0	1	37	True	1.5e-05*	LMM
1000 Genomes	$\lvert\mathrm{SRMSD}_p\rvert$	RC	True	0	1	76	True	3.9e-10*	LMM
Human Origins sim.	$\lvert\mathrm{SRMSD}_p\rvert$	RC	True	0	1	85	True	0.14	Tie
HGDP sim.	$\lvert\mathrm{SRMSD}_p\rvert$	RC	True	0	1	44	True	8.8e-07*	LMM
1000 Genomes sim.	$\lvert\mathrm{SRMSD}_p\rvert$	RC	True	0	1	90	True	3.9e-10*	LMM
Admix. Large sim.	$\mathrm{AUC}_{\mathrm{PR}}$	FES		0	1	3		5.9e-06*	LMM
Admix. Small sim.	$\mathrm{AUC}_{\mathrm{PR}}$	FES		0	1	2		0.025	Tie
Admix. Family sim.	$\mathrm{AUC}_{\mathrm{PR}}$	FES		1	0.35	22		3.9e-10*	LMM
Human Origins	$\mathrm{AUC}_{\mathrm{PR}}$	FES		0	1	34		3.9e-10*	LMM
HGDP	$\mathrm{AUC}_{\mathrm{PR}}$	FES		1	0.33	16		4.4e-10*	LMM
1000 Genomes	$\mathrm{AUC}_{\mathrm{PR}}$	FES		1	0.11	8		3.9e-10*	LMM
Human Origins sim.	$\mathrm{AUC}_{\mathrm{PR}}$	FES		0	1	36		3.9e-10*	LMM
HGDP sim.	$\mathrm{AUC}_{\mathrm{PR}}$	FES		0	1	17		1.7e-05*	LMM
1000 Genomes sim.	$\mathrm{AUC}_{\mathrm{PR}}$	FES		0	1	10		5e-10*	LMM
Admix. Large sim.	$\mathrm{AUC}_{\mathrm{PR}}$	RC		0	1	3		1.4e-05*	LMM
Admix. Small sim.	$\mathrm{AUC}_{\mathrm{PR}}$	RC		0	1	1		0.095	Tie
Admix. Family sim.	$\mathrm{AUC}_{\mathrm{PR}}$	RC		0	1	34		3.9e-10*	LMM
Human Origins	$\mathrm{AUC}_{\mathrm{PR}}$	RC		3	0.4	36		9.6e-10*	LMM
HGDP	$\mathrm{AUC}_{\mathrm{PR}}$	RC		4	0.21	16		0.013	Tie
1000 Genomes	$\mathrm{AUC}_{\mathrm{PR}}$	RC		5	0.004	9		0.00043	Tie
Human Origins sim.	$\mathrm{AUC}_{\mathrm{PR}}$	RC		0	1	37		4.1e-10*	LMM
HGDP sim.	$\mathrm{AUC}_{\mathrm{PR}}$	RC		3	0.087	17		0.0014	Tie
1000 Genomes sim.	$\mathrm{AUC}_{\mathrm{PR}}$	RC		3	0.37	10		8.5e-10*	LMM

[a]FES: Fixed Effect Sizes, RC: Random Coefficients.
[b]Calibrated: whether mean $\lvert\mathrm{SRMSD}_p\rvert < 0.01$.
[c]Value of r (number of PCs) with minimum mean $\lvert\mathrm{SRMSD}_p\rvert$ or maximum mean $\mathrm{AUC}_{\mathrm{PR}}$.
[d]Wilcoxon paired 1-tailed test of distributions ($\lvert\mathrm{SRMSD}_p\rvert$ or $\mathrm{AUC}_{\mathrm{PR}}$) between models in header. Asterisk marks significant value using Bonferroni threshold ($p < \alpha/n_{\mathrm{tests}}$ with $\alpha = 0.01$ and $n_{\mathrm{tests}} = 72$ is the number of tests in this table).
[e]Tie if no significant difference using Bonferroni threshold.

r, LMM always has significantly smaller $|\mathrm{SRMSD}_p|$ than PCA or is statistically tied. For $\mathrm{AUC}_{\mathrm{PR}}$ and PCA, the best r is 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 $r = 0$ often has the best mean $\mathrm{AUC}_{\mathrm{PR}}$, and otherwise is not significantly different from the best r. Lastly, LMM with $r = 0$ always has significantly greater or statistically tied $\mathrm{AUC}_{\mathrm{PR}}$ than PCA with its best r.

3.2 Evaluations in admixture simulations

Now we look more closely at results per dataset. The complete SRMSD_p and $\mathrm{AUC}_{\mathrm{PR}}$ distributions for the admixture simulations and FES traits are in Fig. 3. RC traits gave qualitatively similar results (Fig. S4).

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 as r increases (p-values are conservative). The $\mathrm{AUC}_{\mathrm{PR}}$ distribution of PCA is similarly worst at $r = 0$, increases rapidly and peaks at $r = 3$, then decreases slowly for $r > 3$, while the $\mathrm{AUC}_{\mathrm{PR}}$ distribution for LMM starts near its maximum at $r = 0$ and decreases with r. Although the $\mathrm{AUC}_{\mathrm{PR}}$ distributions for LMM and PCA overlap considerably at each r, LMM with $r = 0$ has significantly greater $\mathrm{AUC}_{\mathrm{PR}}$ values than PCA with $r = 3$ (Table 4). 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 (see Materials and Methods), 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,

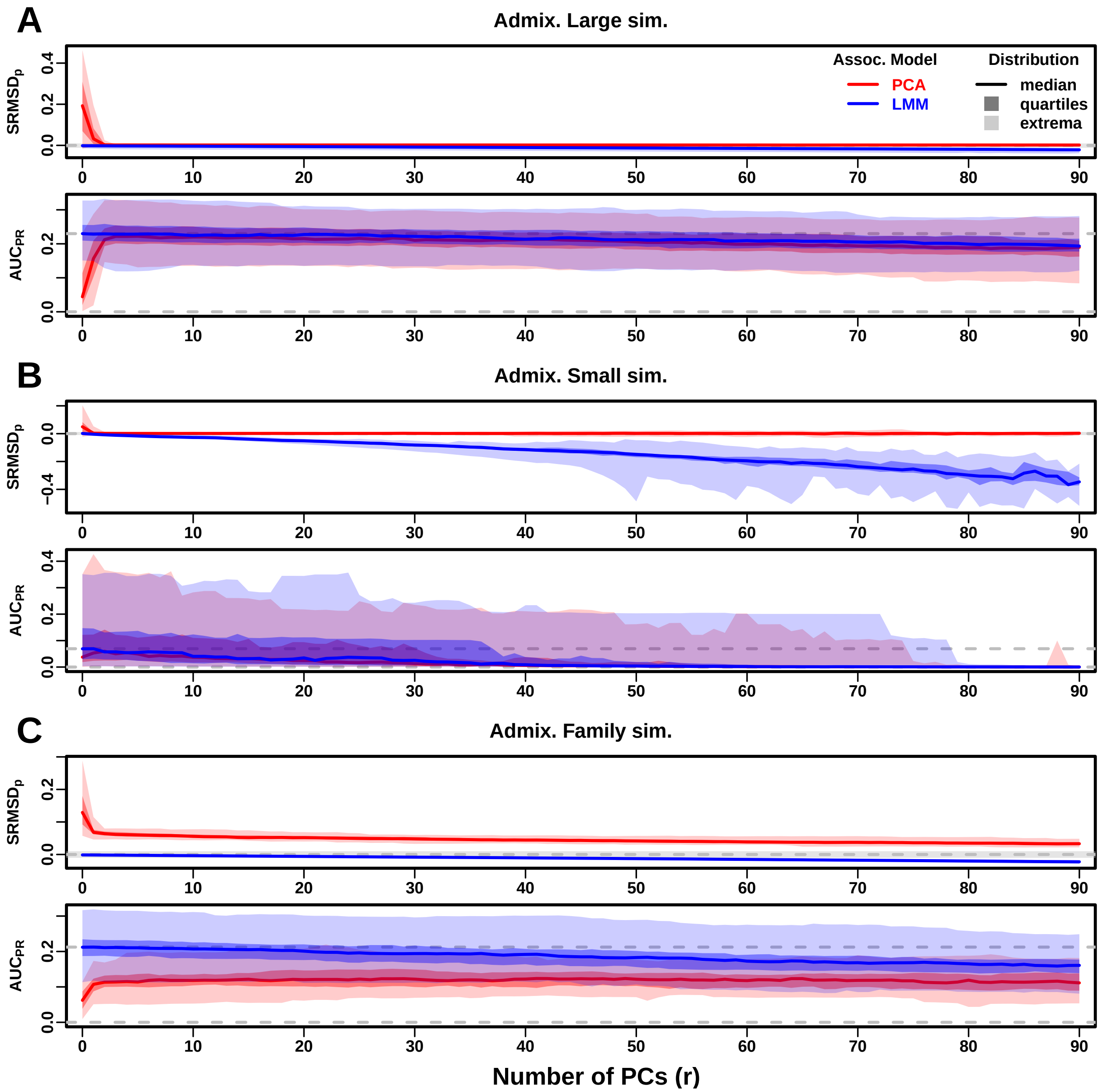

Figure 3: **Evaluations in admixture simulations.** Traits simulated from FES model with high heritability. PCA and LMM models have varying number of PCs ($r \in \{0, ..., 90\}$ on x-axis), with the distributions (y-axis) of SRMSD_p (top subpanel) and AUC_{PR} (bottom subpanel) for 50 replicates. Best 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.

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 4), 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 $\mathrm{AUC}_{\mathrm{PR}}$ 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 F_{ST}, numerous correlated subpopulations, and potential cryptic family relatedness.

Human Origins has the greatest number and diversity of subpopulations. The SRMSD_p and $\mathrm{AUC}_{\mathrm{PR}}$ 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 $|\mathrm{SRMSD}_p| < 0.01$, PCA maintained $\mathrm{SRMSD}_p > 0.01$ for all r and its $\mathrm{AUC}_{\mathrm{PR}}$ were all considerably smaller than the best $\mathrm{AUC}_{\mathrm{PR}}$ 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) again most resembled the family simulations. In particular, LMM with $r = 0$ achieves mean $|\mathrm{SRMSD}_p| < 0.01$ (p-values are calibrated), while PCA does not, and there is a sizable $\mathrm{AUC}_{\mathrm{PR}}$ gap between LMM and PCA. Maximum $\mathrm{AUC}_{\mathrm{PR}}$ 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,

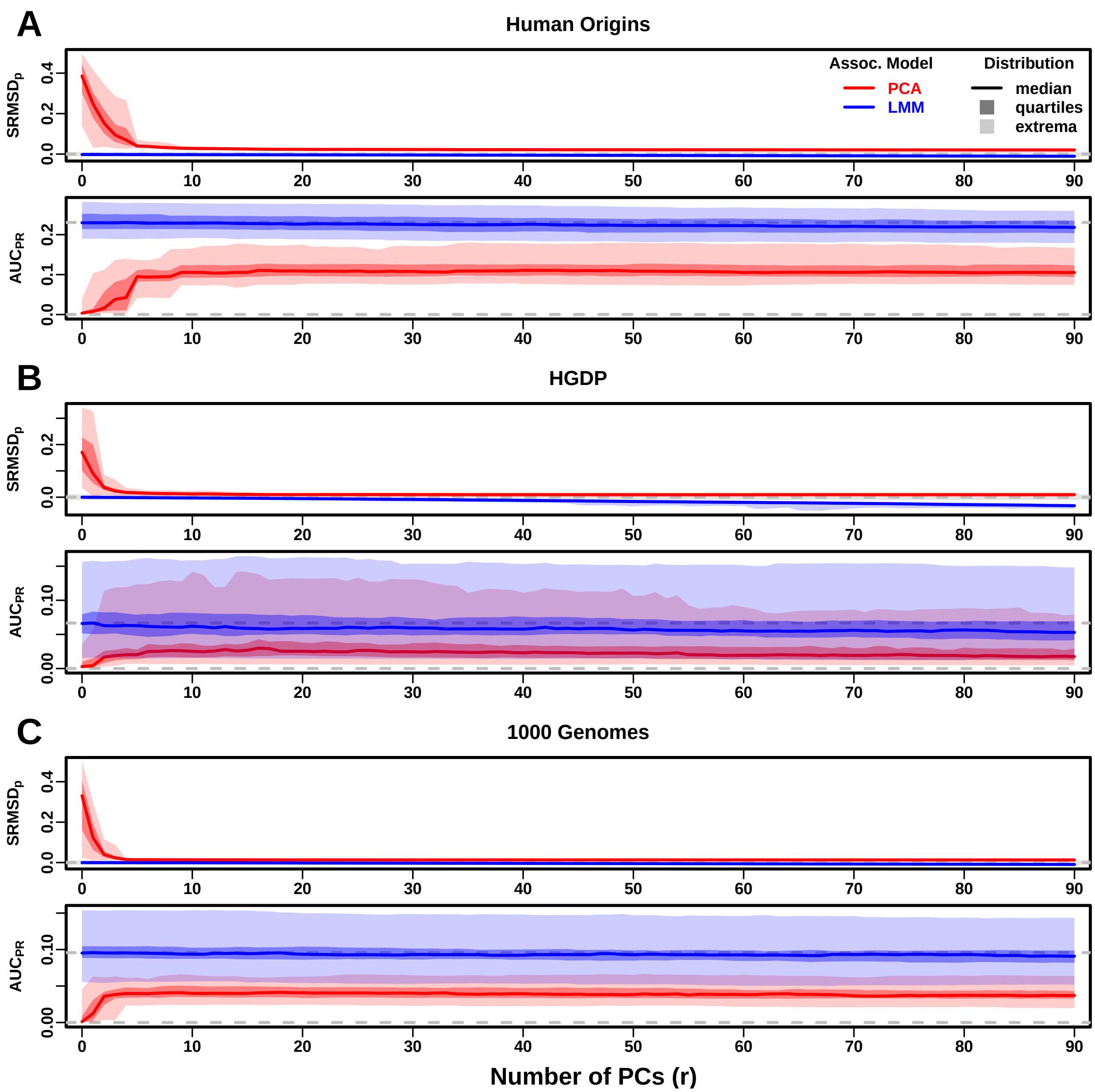

Figure 4: **Evaluations in real human genotype datasets.** Traits simulated from FES model with high heritability. Same setup as Fig. 3, see that for details. These datasets strongly favor LMM with no PCs over PCA, with distributions that most resemble the family simulation. **A.** Human Origins. **B.**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

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. S5). 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 4).

3.4 Evaluations in subpopulation 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 subpopulation tree 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 complex tree structures could recapitulate our 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 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). Both LMM with $r = 0$ and PCA (various r) achieve mean $|\text{SRMSD}_p| < 0.01$ (Table 4). 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 RC traits (Fig. S6 and Table 4). Overall, these subpopulation 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 dimension of relatedness, or kinship matrix rank, since PCA is a low-dimensional model whereas LMM can model high-dimensional relatedness without overfitting. We used the Tracy-Widom test [28] with $p < 0.01$ to estimate

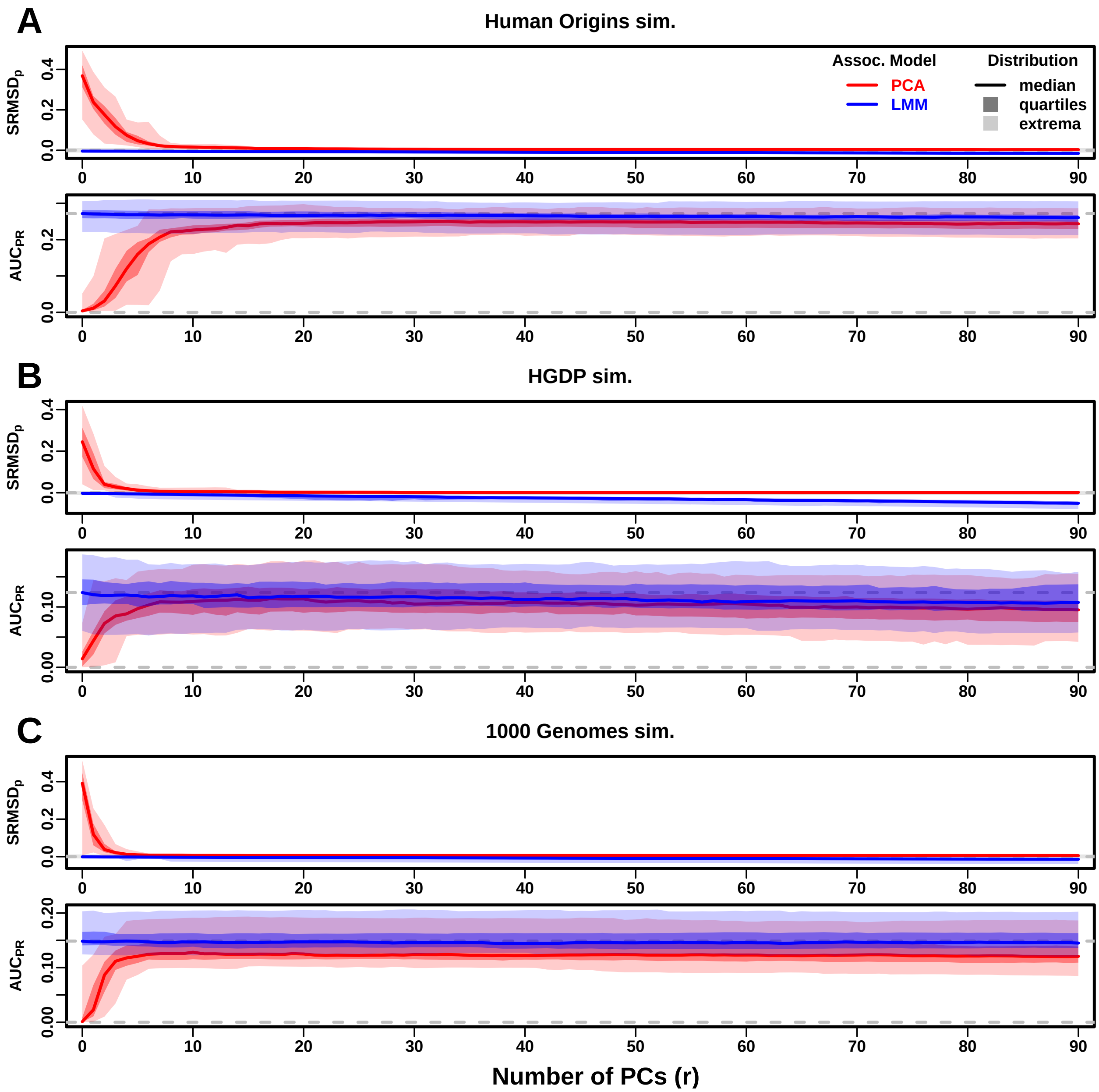

Figure 5: **Evaluations in subpopulation tree simulations fit to human data.** Traits simulated from FES model with high heritability. Same setup as Fig. 3, see that for details. These tree simulations, which exclude family structure by design, do not explain the large gaps in LMM-PCA performance observed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

kinship matrix rank as the number of significant PCs (Fig. S7A). The true rank of our simulations is slightly underestimated (Table 3), but we confirm that the family simulation has the greatest rank, and real datasets have greater estimates than their respective subpopulation tree simulations, which confirms our hypothesis to some extent. However, estimated ranks do not separate real datasets from tree simulations, as required to predict the observed PCA performance. Moreover, the HGDP and 1000 Genomes rank estimates are 45 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 3), 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. S7C). 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. S7B). We also calculated the number of PCs that are significantly associated with the trait, and observed similar results, namely that while the family simulation has more significant PCs than the non-family admixture simulations, the real datasets and their tree simulated counterparts have similar numbers of significant PCs (Fig. S8). Overall, there is no separation between real datasets (where PCA performed poorly) and subpopulation tree simulations (where PCA performed relatively well) in terms of their eigenvalues or kinship matrix rank 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 subpopulation tree simulation counterparts. Instead of inferring local kinship through increased kinship matrix rank, as attempted in the last paragraph, now we measure it directly using the KING-robust estimator [105]. 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 subpopulation 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 [105, 111]. 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.

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 two of the three real datasets ($|\mathrm{SRMSD}_p| > 0.01$), the only marginally calibrated case being HGDP which is also the smallest of these datasets. Otherwise, $\mathrm{AUC}_{\mathrm{PR}}$ 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.

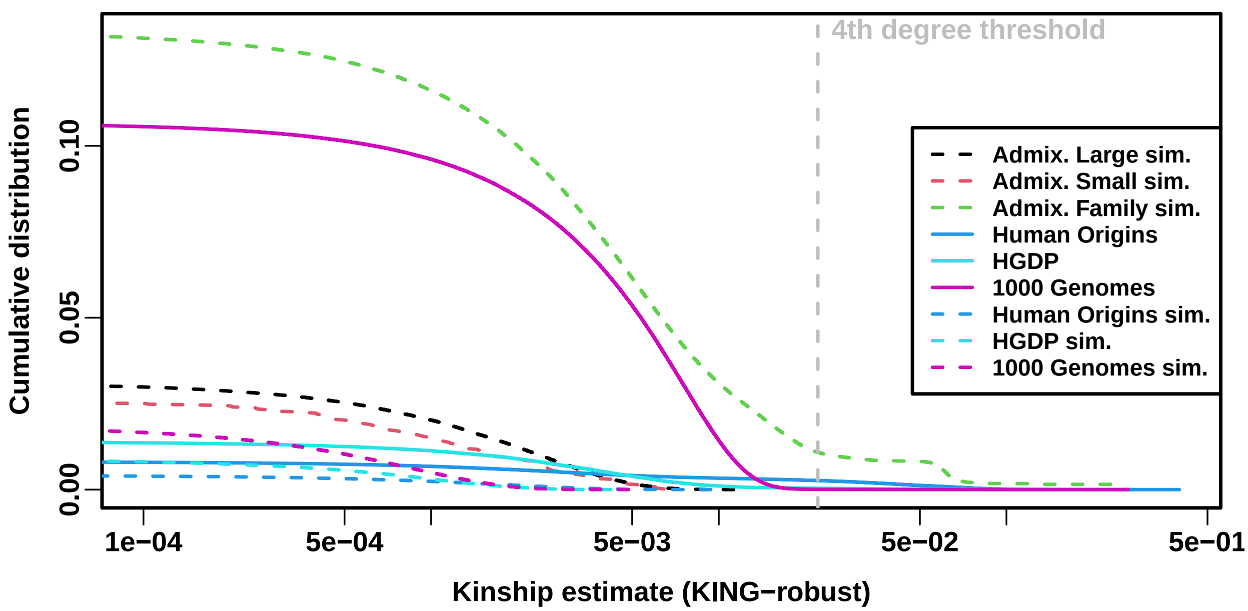

Figure 6: **Local kinship distributions.** Curves are complementary cumulative distribution of lower triangular kinship matrix (self kinship excluded) from KING-robust estimator. Note log x-axis; negative estimates are counted but not shown. Most values are below 4th degree relative threshold. Each real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

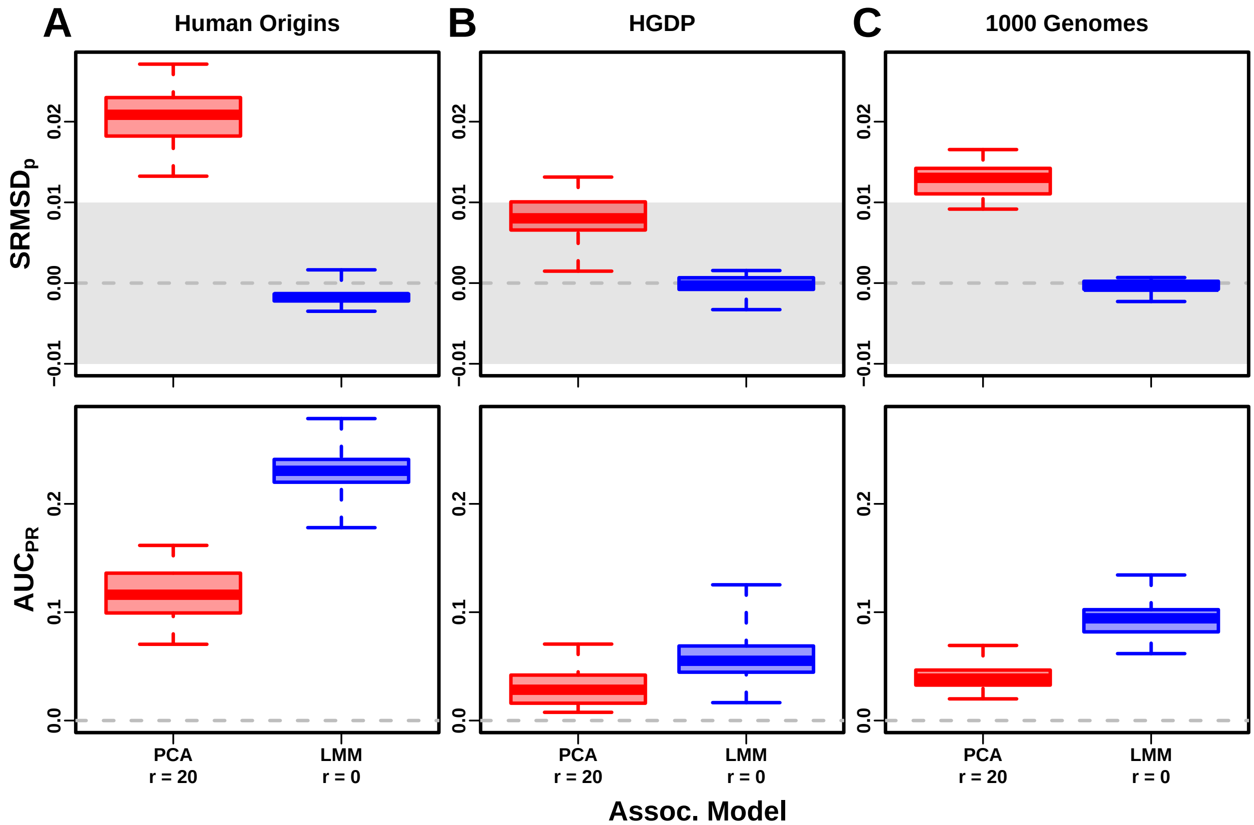

Figure 7: **Evaluation in real datasets excluding 4th degree relatives.** Traits simulated from FES model with high heritability. Each dataset is a column, rows are measures. First row has $|\text{SRMSD}_p| < 0.01$ band marked as gray area.

3.6 Low heritability and environment simulations

Our main evaluations were repeated with traits simulated under a lower heritability value of $h^2 = 0.3$. We reduced the number of causal loci in response to this change in heritability, to result in equal average effect size per locus compared to the previous high heritability evaluations (see Materials and Methods). Despite that, these low heritability evaluations measured lower $\mathrm{AUC}_{\mathrm{PR}}$ values than their high heritability counterparts (Figs. S9 to S13). The gap between LMM and PCA was reduced in these evaluations, but the main conclusion of the high heritability evaluation holds for low heritability as well, namely that LMM with $r = 0$ significantly outperforms or ties LMM with $r > 0$ and PCA in all cases (Table S2).

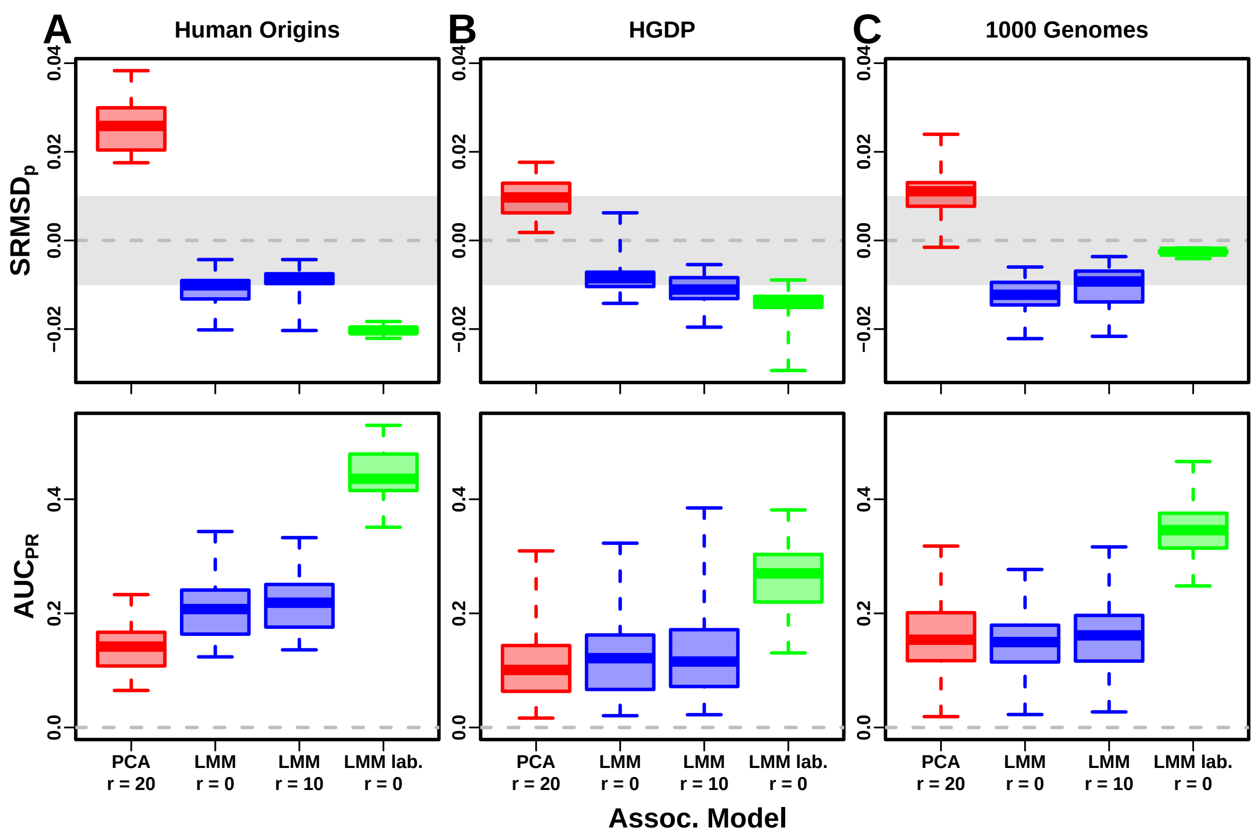

Figure 8: **Evaluation in real datasets excluding 4th degree relatives, FES traits, environment.** Traits simulated with environment effects, otherwise the same as Fig. 7. "LMM lab." includes as fixed effects true groups from which environment was simulated.

Lastly, we simulated traits with both low heritability and large environment effects determined by geography and subpopulation labels, so they are strongly correlated to the low-dimensional population structure (Table 2). For that reason, PCs may be expected to perform better in this setting (in either PCA or LMM). However, we find that both PCA and LMM (even without PCs) increase their $\mathrm{AUC}_{\mathrm{PR}}$ values compared to the low-heritability evaluations (Fig. S14; Fig. 8 also shows representative numbers of PCs, which performed optimally or nearly so in individual simulations shown in Figs. S15 to S18). P-value calibration is comparable with or without environment effects, for LMM for all r and for PCA once r is large enough (Fig. S14). These simulations are the only where we occasionally observed for both metrics a significant, though small, advantage of LMM with PCs versus LMM without PCs (Table S3). Additionally, on RC traits only, PCA significantly outperforms LMM in the three real human datasets (Table S3), the only cases in all of our evaluations where this is observed. For comparison, we also evaluate an "oracle" LMM without PCs but with the finest group labels, the same used to simulate environment, as fixed categorical covariates ("LMM lab."), and see much larger $\mathrm{AUC}_{\mathrm{PR}}$ values than either LMM with PCs or PCA (Figs. 8 and S15 to S18 and Table S3). However, LMM with labels is often more poorly calibrated than LMM or PCA without labels, which may be since these numerous labels are inappropriately modeled as fixed rather than random effects. Overall, we find that association studies with correlated environment and genetic effects remain a challenge for PCA and LMM, that addition of PCs to an LMM improves performance only marginally, and that if the environment effect is driven by geography or ethnicity then use of those labels greatly improves performance compared to using PCs.

4 Discussion

Our evaluations conclusively determined that LMM without PCs performs better than PCA (for any number of PCs) across all scenarios without environment effects, 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 (Tables S2 and 4), which agrees with previous observations [68, 70] but contradicts others [37, 44]. Our findings make sense since PCs are the eigenvectors of the same kinship matrix that parameterized random effects, so including both is redundant.

The presence of environment effects that are correlated to relatedness presents the only scenario where occasionally PCA and LMM with PCs outperform LMM without PCs (Table S3). It is commonly believed that PCs model such environment effects well [20, 39, 40]. However, we observe that LMM without PCs models environment effects nearly as well as with PCs (Fig. 8), consistent with previous findings [53, 54] and with environment inflating heritability estimates using LMM [112]. Moreover, modeling the true environment groups as fixed categorical effects always substantially improved $\mathrm{AUC}_{\mathrm{PR}}$ compared to modeling them with PCs (Fig. 8 and Table S3). Modeling numerous environment groups as fixed effects does result in deflated p-values (Fig. 8 and Table S3), which we expect would be avoided by modeling them as random effects, a strategy we chose not to pursue here as it is both a circular evaluation (the true effects were drawn from that model) and out of scope. Overall, including PCs to model environment effects yields limited power gains if at all, even in an LMM, and is no replacement for more adequate modeling of environment whenever possible.

Previous studies found that PCA was better calibrated than LMM for unusually differentiated markers [44, 55, 57], which as simulated were an artificial scenario not based on a population genetics model, and are otherwise believed to be unusual [58, 78]. Our evaluations on real human data, which contain such loci in relevant proportions if they exist, do not replicate that result. Family 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 $\mathrm{AUC}_{\mathrm{PR}}$ near that of LMM without PCs. Conversely, PCA performed poorly when no number of PCs had either calibrated p-values or acceptably large $\mathrm{AUC}_{\mathrm{PR}}$. There were no cases where high numbers of PCs optimized an acceptable $\mathrm{AUC}_{\mathrm{PR}}$, or cases with miscalibrated p-values but high $\mathrm{AUC}_{\mathrm{PR}}$. PCA performed well in the admixture simulations (without families, both trait models), real human genotypes with RC traits, and the subpopulation 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 restricted to a low-dimensional subspace, whereas

LMM can handle high-dimensional relatedness. Thus, PCA performs well in the admixture simulation, which is explicitly low-dimensional (see Materials and Methods), and our subpopulation tree simulations, which are likely well approximated by a few dimensions despite the large number of subpopulations because there are few long branches. Conversely, PCA performs poorly under family structure because its kinship matrix is high-dimensional (Fig. S7). However, estimating the latent space dimensions of real datasets is challenging because estimated eigenvalues have biased distributions [113]. Kinship matrix rank estimated using the Tracy-Widom test [28] did not fully predict the datasets that PCA performs well on. In contrast, estimated local kinship finds considerable cryptic family 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 cannot tell the full story. There are related tests for numbers of dimensions that consider the trait which we did not consider, including the Bayesian information criterion for the regression with PCs against the trait [38]. Additionally, PCA and LMM goodness of fit could be compared using the coefficient of determination generalized for LMMs [114].

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 subpopulation tree simulations. Admixture is also absent in our tree simulations, but our simulations and theory show that admixture is well handled by PCA. Hundreds of close relative pairs have been identified in 1000 Genomes [115–118], 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 [77, 119, 120]. Our FES trait tests show that family relatedness is more challenging when rarer variants have larger coefficients. Overall, the high relatedness dimensions induced by family 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 [26, 56], 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, which is 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.

There are several variants of the PCA and LMM analyses, most designed for better modeling linkage disequilibrium (LD), that we did not evaluate directly, in which PCs are no longer exactly the top eigenvectors of the kinship matrix (if estimated with different approaches), although this is not a crucial aspect of our arguments. We do not consider the case where samples are projected onto PCs estimated from an external sample [121], which is uncommon in association studies, and whose primary effect is shrinkage, so if all samples are projected then they are all equally affected and larger regression coefficients compensate for the shrinkage, although this will no longer be the case if only a portion of the sample is projected onto the PCs of the rest of the sample. Another approach tests PCs for association against every locus in the genome in order to identify and exclude PCs that capture LD structure (which is localized) instead of ancestry (which should be present across the genome) [121]; a previous proposal removes LD using an autocorrelation model prior to estimating PCs [28]. These improved PCs remain inadequate models of family relatedness, so an LMM will continue to outperform them in that setting. Similarly, the leave-one-chromosome-out (LOCO) approach for estimating kinship matrices for LMMs prevents the test locus and loci in LD with it from being modeled by the random effect as well, which is called "proximal contamination" [55, 62]. While LOCO kinship estimates vary for each chromosome, they continue to model family relatedness, thus maintaining their key advantage over PCA. The LDAK model estimates kinship instead by weighing loci taking LD into account [122]. LD effects must be adjusted for, if present, so in unfiltered data we advise the previous methods be applied. However, in this work, simulated genotypes do not have LD, and the real datasets were filtered to remove LD, so here there is no proximal contamination and LD confounding is minimized if present at all, so these evaluations may be considered the ideal situation where LD effects have been adjusted successfully, and in this setting LMM outperforms PCA. Overall, these alternative PCs or kinship matrices differ from their basic counterparts by either the extent to which LD influences the estimates (which may be a confounder in a small portion of the genome, by definition) or by sampling noise, neither of which are expected to change our key conclusion.

One of the limitations of this work include relatively small sample sizes compared to modern association studies. However, our conclusions are not expected to change with larger sample sizes, as cryptic family relatedness will continue to be abundant in such data, if not increase in abundance, and thus give LMMs an advantage over PCA [77, 119, 120]. One reason PCA has been favored over classic LMMs is because PCA's runtime scales much better with increasing sample size. However, recent approaches not tested in this work have made LMMs more scalable and applicable to biobank-scale data [59, 67, 72], so one clear next step is carefully evaluating these approaches in simulations with larger sample sizes. A different benefit for including PCs were recently reported for BOLT-LMM, which does not result in greater power but rather in reduced runtime, a property that may be specific to its use of scalable algorithms such as conjugate gradient and variational Bayes [77]. Many of these newer LMMs also no longer follow the infinitesimal model of the basic LMM [67, 72], and employ novel approximations, which are features not evaluated in this work and worthy of future study.

Another limitation of this work is ignoring rare variants, a necessity given our smaller sample sizes, where rare variant association is miscalibrated and underpowered. Using simulations mimicking the UK Biobank, recent work has found that rare variants can have a more pronounced structure than common variants, and that modeling this rare variant structure (with either PCA and LMM) may better model environment confounding, reduce inflation in association studies, and ameliorate stratification in polygenic risk scores [123]. Better modeling rare variants and their structure is a key next step in association studies.

The largest limitation of our work is that we only considered quantitative traits. 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 and imbalance, which appears to affect LMMs more severely, although recent work appears to solve this problem [55, 59]. Future evaluations should aim to include our simulations and real datasets, to ensure that previous results were not biased in favor of PCA by not simulating family structure or larger coefficients for rare variants that are 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 [77, 119, 120]. 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 [42, 110]. 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, population structure increases the variance of these sample $\hat{p}_i^T$ relative to the true p_i^T [52]. We present a new algorithm for constructing a new 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

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

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 [52]. 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

$$\mathrm{Var}(p^*) = w^2 W^T + (1 - w)^2 \,\mathrm{Var}(q),$$

which we equate to the desired V^T (Eq. (9)) and solve for w. For simplicity, we 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 $\text{Var}(p^*) = V^T$, 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, AUC_{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, 44]. The inflation factor λ is defined as the median χ^2 association statistic divided by theoretical median under the null hypothesis [23]. To compare p-values from non-χ^2 tests (such as t-statistics), λ can be calculated from p-values using

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

where p_{median} is the median observed p-value (including causal loci), $u_{\text{median}} = \frac{1}{2}$ 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 $\text{SRMSD}_p = 0$. However, non-uniform p-values with the expected median, such as from genomic control [23], result in $\lambda = 1$, but $\text{SRMSD}_p \neq 0$ 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

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 $\text{SRMSD}_p > 0$) 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 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 $a = 0.564$ and $b = 0.619$. The value $\lambda = 1.05$, a common threshold for benign inflation [44], 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), and it requires large simulations to estimate well as it depends on the tail of the distribution. In contrast, SRMSD_p uses the entire distribution so it is easier to estimate, $\text{SRMSD}_p = 0$ guarantees calibrated type I error rates at all t, while large $|\text{SRMSD}_p|$ indicates incorrect type I errors for a range of t. Empirically, we find the expected agreement and monotonic relationship between SRMSD_p and type I error rate (Fig. S2).

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, and it is also harder to estimate than $\mathrm{AUC}_{\mathrm{PR}}$ because power depends on the tail of the distribution.

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

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

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

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

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 $\mathrm{AUC}_{\mathrm{PR}} = \int_0^1 \mathrm{Precision}(\mathrm{Recall}) d\mathrm{Recall}$.

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 $\mathrm{AUC}_{\mathrm{PR}}$ is greater for A than B. Thus, $\mathrm{AUC}_{\mathrm{PR}}$ ranks calibrated models according to power.

Empirically, we find the predicted positive correlation between $\mathrm{AUC}_{\mathrm{PR}}$ and calibrated power (Fig. S3). The correlation is clear when considered separately per dataset, but the slope varies per dataset, which is expected because the proportion of alternative cases π_1 varies per dataset.

Competing interests

The authors declare no competing interests.

Acknowledgments

Thanks to Tiffany Tu, Ratchanon Pornmongkolsuk, and Zhuoran Hou for feedback on this article. This work was funded in part by the Duke University School of Medicine Whitehead Scholars Program, a gift from the Whitehead Charitable Foundation. The 1000 Genomes data were generated at the New York Genome Center with funds provided by NHGRI Grant 3UM1HG008901-03S1.

Web resources

plink2, `https://www.cog-genomics.org/plink/2.0/`

GCTA, `https://yanglab.westlake.edu.cn/software/gcta/`

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] W. Astle and D. J. Balding. "Population Structure and Cryptic Relatedness in Genetic Association Studies". *Statist. Sci.* 24(4) (2009), pp. 451–471. DOI: 10.1214/09-STS307.

[2] N. A. Rosenberg et al. "Genome-wide association studies in diverse populations". *Nat Rev Genet* 11(5) (2010), pp. 356–366. DOI: 10.1038/nrg2760.

[3] G. E. Hoffman. "Correcting for population structure and kinship using the linear mixed model: theory and extensions". *PLoS ONE* 8(10) (2013), e75707. DOI: 10.1371/journal.pone.0075707.

[4] M. A. Coram et al. "Genome-wide Characterization of Shared and Distinct Genetic Components that Influence Blood Lipid Levels in Ethnically Diverse Human Populations". *The American Journal of Human Genetics* 92(6) (6, 2013), pp. 904–916. DOI: 10.1016/j.ajhg.2013.04.025.

[5] C. Medina-Gomez et al. "Challenges in conducting genome-wide association studies in highly admixed multi-ethnic populations: the Generation R Study". *Eur J Epidemiol* 30(4) (1, 2015), pp. 317–330. DOI: 10.1007/s10654-015-9998-4.

[6] M. Conomos et al. "Genetic Diversity and Association Studies in US Hispanic/Latino Populations: Applications in the Hispanic Community Health Study/Study of Latinos". *The American Journal of Human Genetics* 98(1) (7, 2016), pp. 165–184. DOI: 10.1016/j.ajhg.2015.12.001.

[7] C. J. Hodonsky et al. “Genome-wide association study of red blood cell traits in Hispanics/Latinos: The Hispanic Community Health Study/Study of Latinos”. *PLOS Genetics* 13(4) (28, 2017), e1006760. DOI: 10.1371/journal.pgen.1006760.

[8] A. R. Martin et al. “Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations”. *Am. J. Hum. Genet.* 100(4) (6, 2017), pp. 635–649. DOI: 10.1016/j.ajhg.2017.03.004.

[9] A. R. Martin et al. “An Unexpectedly Complex Architecture for Skin Pigmentation in Africans”. *Cell* 171(6) (30, 2017), 1340–1353.e14. DOI: 10.1016/j.cell.2017.11.015.

[10] L. A. Hindorff et al. “Prioritizing diversity in human genomics research”. *Nature Reviews Genetics* 19(3) (2018), pp. 175–185. DOI: 10.1038/nrg.2017.89.

[11] T. J. Hoffmann et al. “A Large Multiethnic Genome-Wide Association Study of Adult Body Mass Index Identifies Novel Loci”. *Genetics* 210(2) (1, 2018), pp. 499–515. DOI: 10.1534/genetics.118.301479.

[12] L. S. Mogil et al. “Genetic architecture of gene expression traits across diverse populations”. *PLOS Genetics* 14(8) (10, 2018), e1007586. DOI: 10.1371/journal.pgen.1007586.

[13] C. Roselli et al. “Multi-ethnic genome-wide association study for atrial fibrillation”. *Nature Genetics* 50(9) (2018), p. 1225. DOI: 10.1038/s41588-018-0133-9.

[14] G. L. Wojcik et al. “Genetic analyses of diverse populations improves discovery for complex traits”. *Nature* 570(7762) (2019), pp. 514–518. DOI: 10.1038/s41586-019-1310-4.

[15] R. E. Peterson et al. “Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations”. *Cell* 179(3) (17, 2019), pp. 589–603. DOI: 10.1016/j.cell.2019.08.051.

[16] Y. Zhong, M. A. Perera, and E. R. Gamazon. “On Using Local Ancestry to Characterize the Genetic Architecture of Human Traits: Genetic Regulation of Gene Expression in Multiethnic or Admixed Populations”. *The American Journal of Human Genetics* 104(6) (6, 2019), pp. 1097–1115. DOI: 10.1016/j.ajhg.2019.04.009.

[17] Y. Hu et al. “Minority-centric meta-analyses of blood lipid levels identify novel loci in the Population Architecture using Genomics and Epidemiology (PAGE) study”. *PLOS Genetics* 16(3) (30, 2020), e1008684. DOI: `10.1371/journal.pgen.1008684`.

[18] I. Simonin-Wilmer et al. “An Overview of Strategies for Detecting Genotype-Phenotype Associations Across Ancestrally Diverse Populations”. *Frontiers in Genetics* 12 (2021).

[19] M. Kamariza et al. “Misuse of the term ‘trans-ethnic’ in genomics research”. *Nat Genet* 53(11) (2021), pp. 1520–1521. DOI: `10.1038/s41588-021-00952-6`.

[20] M. Lin et al. “Admixed Populations Improve Power for Variant Discovery and Portability in Genome-Wide Association Studies”. *Frontiers in Genetics* 12 (2021).

[21] A. Mahajan et al. “Multi-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation”. *Nat Genet* 54(5) (2022), pp. 560–572. DOI: `10.1038/s41588-022-01058-3`.

[22] K. Hou et al. “Causal effects on complex traits are similar for common variants across segments of different continental ancestries within admixed individuals”. *Nat Genet* (20, 2023), pp. 1–10. DOI: `10.1038/s41588-023-01338-6`.

[23] B. Devlin and K. Roeder. “Genomic Control for Association Studies”. *Biometrics* 55(4) (1999), pp. 997–1004. DOI: `10.1111/j.0006-341X.1999.00997.x`.

[24] B. F. Voight and J. K. Pritchard. “Confounding from Cryptic Relatedness in Case-Control Association Studies”. *PLOS Genetics* 1(3) (2005), e32. DOI: `10.1371/journal.pgen.0010032`.

[25] S. Zhang, X. Zhu, and H. Zhao. “On a semiparametric test to detect associations between quantitative traits and candidate genes using unrelated individuals”. *Genetic Epidemiology* 24(1) (2003), pp. 44–56. DOI: `10.1002/gepi.10196`.

[26] A. L. Price et al. “Principal components analysis corrects for stratification in genome-wide association studies”. *Nat. Genet.* 38(8) (2006), pp. 904–909. DOI: `10.1038/ng1847`.

[27] M. Bouaziz, C. Ambroise, and M. Guedj. “Accounting for Population Stratification in Practice: A Comparison of the Main Strategies Dedicated to Genome-Wide Association Studies”. *PLOS ONE* 6(12) (2011), e28845. DOI: `10.1371/journal.pone.0028845`.

[28] N. Patterson, A. L. Price, and D. Reich. "Population Structure and Eigenanalysis". *PLoS Genet* 2(12) (2006), e190. DOI: 10.1371/journal.pgen.0020190.

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

[30] J. K. Pritchard et al. "Association Mapping in Structured Populations". *The American Journal of Human Genetics* 67(1) (2000), pp. 170–181. DOI: 10.1086/302959.

[31] D. H. Alexander, J. Novembre, and K. Lange. "Fast model-based estimation of ancestry in unrelated individuals". *Genome Res.* 19(9) (2009), pp. 1655–1664. DOI: 10.1101/gr.094052.109.

[32] Q. Zhou, L. Zhao, and Y. Guan. "Strong Selection at MHC in Mexicans since Admixture". *PLoS Genet.* 12(2) (2016), e1005847. DOI: 10.1371/journal.pgen.1005847.

[33] G. McVean. "A genealogical interpretation of principal components analysis". *PLoS Genet* 5(10) (2009), e1000686. DOI: 10.1371/journal.pgen.1000686.

[34] X. Zheng and B. S. Weir. "Eigenanalysis of SNP data with an identity by descent interpretation". *Theor Popul Biol* 107 (2016), pp. 65–76. DOI: 10.1016/j.tpb.2015.09.004.

[35] I. Cabreros and J. D. Storey. "A Likelihood-Free Estimator of Population Structure Bridging Admixture Models and Principal Components Analysis". *Genetics* 212(4) (2019), pp. 1009–1029. DOI: 10.1534/genetics.119.302159.

[36] A. M. Chiu et al. "Inferring population structure in biobank-scale genomic data". *The American Journal of Human Genetics* 0(0) (2022). DOI: 10.1016/j.ajhg.2022.02.015.

[37] K. Zhao et al. "An Arabidopsis Example of Association Mapping in Structured Samples". *PLOS Genetics* 3(1) (2007), e4. DOI: 10.1371/journal.pgen.0030004.

[38] C. Zhu and J. Yu. "Nonmetric Multidimensional Scaling Corrects for Population Structure in Association Mapping With Different Sample Types". *Genetics* 182(3) (1, 2009), pp. 875–888. DOI: 10.1534/genetics.108.098863.

[39] J. Novembre et al. "Genes mirror geography within Europe". *Nature* 456(7218) (2008), pp. 98–101. DOI: 10.1038/nature07331.

[40] Y. Zhang and W. Pan. "Principal Component Regression and Linear Mixed Model in Association Analysis of Structured Samples: Competitors or Complements?" *Genetic Epidemiology* 39(3) (2015), pp. 149–155. DOI: 10.1002/gepi.21879.

[41] H. Xu and Y. Guan. "Detecting Local Haplotype Sharing and Haplotype Association". *Genetics* 197(3) (2014), pp. 823–838. DOI: 10.1534/genetics.114.164814.

[42] J. Qian et al. "A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK Biobank". *PLOS Genetics* 16(10) (2020), e1009141. DOI: 10.1371/journal.pgen.1009141.

[43] T. Thornton and M. S. McPeek. "ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure". *Am. J. Hum. Genet.* 86(2) (2010), pp. 172–184. DOI: 10.1016/j.ajhg.2010.01.001.

[44] A. L. Price et al. "New approaches to population stratification in genome-wide association studies". *Nature Reviews Genetics* 11(7) (2010), pp. 459–463. DOI: 10.1038/nrg2813.

[45] S. Lee et al. "Sparse Principal Component Analysis for Identifying Ancestry-Informative Markers in Genome-Wide Association Studies". *Genetic Epidemiology* 36(4) (2012), pp. 293–302. DOI: 10.1002/gepi.21621.

[46] G. Abraham and M. Inouye. "Fast Principal Component Analysis of Large-Scale Genome-Wide Data". *PLOS ONE* 9(4) (2014), e93766. DOI: 10.1371/journal.pone.0093766.

[47] K. Galinsky et al. "Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia". *The American Journal of Human Genetics* 98(3) (2016), pp. 456–472. DOI: 10.1016/j.ajhg.2015.12.022.

[48] G. Abraham, Y. Qiu, and M. Inouye. "FlashPCA2: principal component analysis of Biobank-scale genotype datasets". *Bioinformatics* 33(17) (2017), pp. 2776–2778. DOI: 10.1093/bioinformatics/btx299.

[49] A. Agrawal et al. "Scalable probabilistic PCA for large-scale genetic variation data". *PLOS Genetics* 16(5) (2020), e1008773. DOI: 10.1371/journal.pgen.1008773.

[50] J. Yu et al. "A unified mixed-model method for association mapping that accounts for multiple levels of relatedness". *Nat. Genet.* 38(2) (2006), pp. 203–208. DOI: 10.1038/ng1702.

[51] H. M. Kang et al. "Efficient control of population structure in model organism association mapping". *Genetics* 178(3) (2008), pp. 1709–1723. DOI: 10.1534/genetics.107.080101.

[52] A. Ochoa and J. D. Storey. "Estimating FST and kinship for arbitrary population structures". *PLoS Genet* 17(1) (2021), e1009241. DOI: 10.1371/journal.pgen.1009241.

[53] B. J. Vilhjálmsson and M. Nordborg. "The nature of confounding in genome-wide association studies". *Nat Rev Genet* 14(1) (2013), pp. 1–2. DOI: 10.1038/nrg3382.

[54] H. Wang, B. Aragam, and E. P. Xing. "Trade-offs of Linear Mixed Models in Genome-Wide Association Studies". *Journal of Computational Biology* 29(3) (2022), pp. 233–242. DOI: 10.1089/cmb.2021.0157.

[55] J. Yang et al. "Advantages and pitfalls in the application of mixed-model association methods". *Nat Genet* 46(2) (2014), pp. 100–106. DOI: 10.1038/ng.2876.

[56] H. M. Kang et al. "Variance component model to account for sample structure in genome-wide association studies". *Nat. Genet.* 42(4) (2010), pp. 348–354. DOI: 10.1038/ng.548.

[57] C. Wu et al. "A Comparison of Association Methods Correcting for Population Stratification in Case–Control Studies". *Annals of Human Genetics* 75(3) (2011), pp. 418–427. DOI: 10.1111/j.1469-1809.2010.00639.x.

[58] J. H. Sul and E. Eskin. "Mixed models can correct for population structure for genomic regions under selection". *Nature Reviews Genetics* 14(4) (2013), p. 300. DOI: 10.1038/nrg2813-c1.

[59] W. Zhou et al. "Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies". *Nat Genet* 50(9) (2018), pp. 1335–1341. DOI: 10.1038/s41588-018-0184-y.

[60] Y. S. Aulchenko, D.-J. de Koning, and C. Haley. "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) (2007), pp. 577–585. DOI: 10.1534/genetics.107.075614.

[61] Z. Zhang et al. "Mixed linear model approach adapted for genome-wide association studies". *Nat Genet* 42(4) (2010), pp. 355–360. DOI: 10.1038/ng.546.

[62] C. Lippert et al. "FaST linear mixed models for genome-wide association studies". *Nat. Methods* 8(10) (2011), pp. 833–835. DOI: 10.1038/nmeth.1681.

[63] J. Yang et al. "GCTA: a tool for genome-wide complex trait analysis". *Am. J. Hum. Genet.* 88(1) (2011), pp. 76–82. DOI: 10.1016/j.ajhg.2010.11.011.

[64] J. Listgarten et al. "Improved linear mixed models for genome-wide association studies". *Nat Methods* 9(6) (2012), pp. 525–526. DOI: 10.1038/nmeth.2037.

[65] X. Zhou and M. Stephens. "Genome-wide efficient mixed-model analysis for association studies". *Nat. Genet.* 44(7) (2012), pp. 821–824. DOI: 10.1038/ng.2310.

[66] G. R. Svishcheva et al. "Rapid variance components–based method for whole-genome association analysis". *Nat Genet* 44(10) (2012), pp. 1166–1170. DOI: 10.1038/ng.2410.

[67] P.-R. Loh et al. "Efficient Bayesian mixed-model analysis increases association power in large cohorts". *Nat. Genet.* 47(3) (2015), pp. 284–290. DOI: 10.1038/ng.3190.

[68] L. Janss et al. "Inferences from Genomic Models in Stratified Populations". *Genetics* 192(2) (1, 2012), pp. 693–704. DOI: 10.1534/genetics.112.141143.

[69] G. Tucker, A. L. Price, and B. Berger. "Improving the Power of GWAS and Avoiding Confounding from Population Stratification with PC-Select". *Genetics* 197(3) (2014), pp. 1045–1049. DOI: 10.1534/genetics.114.164285.

[70] N. Liu et al. "Controlling Population Structure in Human Genetic Association Studies with Samples of Unrelated Individuals". *Stat Interface* 4(3) (2011), pp. 317–326. DOI: 10.4310/sii.2011.v4.n3.a6.

[71] J. Zeng et al. “Signatures of negative selection in the genetic architecture of human complex traits”. *Nature Genetics* 50(5) (2018), pp. 746–753. DOI: 10.1038/s41588-018-0101-4.

[72] J. Mbatchou et al. “Computationally efficient whole-genome regression for quantitative and binary traits”. *Nat Genet* 53(7) (2021), pp. 1097–1103. DOI: 10.1038/s41588-021-00870-7.

[73] N. Matoba et al. “GWAS of 165,084 Japanese individuals identified nine loci associated with dietary habits”. *Nat Hum Behav* 4(3) (2020), pp. 308–316. DOI: 10.1038/s41562-019-0805-1.

[74] M. Song, W. Hao, and J. D. Storey. “Testing for genetic associations in arbitrarily structured populations”. *Nat. Genet.* 47(5) (2015), pp. 550–554. DOI: 10.1038/ng.3244.

[75] X. Liu et al. “Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies”. *PLOS Genet* 12(2) (2016), e1005767. DOI: 10.1371/journal.pgen.1005767.

[76] J. H. Sul, L. S. Martin, and E. Eskin. “Population structure in genetic studies: Confounding factors and mixed models”. *PLoS Genet.* 14(12) (2018), e1007309. DOI: 10.1371/journal.pgen.1007309.

[77] P.-R. Loh et al. “Mixed-model association for biobank-scale datasets”. *Nat Genet* 50(7) (2018), pp. 906–908. DOI: 10.1038/s41588-018-0144-6.

[78] A. L. Price et al. “Response to Sul and Eskin”. *Nature Reviews Genetics* 14(4) (2013), p. 300. DOI: 10.1038/nrg2813-c2.

[79] A. Ochoa and J. D. Storey. *New kinship and FST estimates reveal higher levels of differentiation in the global human population.* 2019. DOI: 10.1101/653279.

[80] T. G. P. Consortium. “A map of human genome variation from population-scale sequencing”. *Nature* 467(7319) (2010), pp. 1061–1073. DOI: 10.1038/nature09534.

[81] 1000 Genomes Project Consortium et al. “An integrated map of genetic variation from 1,092 human genomes”. *Nature* 491(7422) (2012), pp. 56–65. DOI: 10.1038/nature11632.

[82] H. M. Cann et al. “A human genome diversity cell line panel”. *Science* 296(5566) (2002), pp. 261–262. DOI: `10.1126/science.296.5566.261b`.

[83] N. A. Rosenberg et al. “Genetic Structure of Human Populations”. *Science* 298(5602) (2002), pp. 2381–2385. DOI: `10.1126/science.1078311`.

[84] A. Bergström et al. “Insights into human genetic variation and population history from 929 diverse genomes”. *Science* 367(6484) (2020). DOI: `10.1126/science.aay5012`.

[85] N. Patterson et al. “Ancient admixture in human history”. *Genetics* 192(3) (2012), pp. 1065–1093. DOI: `10.1534/genetics.112.145037`.

[86] I. Lazaridis et al. “Ancient human genomes suggest three ancestral populations for present-day Europeans”. *Nature* 513(7518) (2014), pp. 409–413. DOI: `10.1038/nature13673`.

[87] I. Lazaridis et al. “Genomic insights into the origin of farming in the ancient Near East”. *Nature* 536(7617) (2016), pp. 419–424. DOI: `10.1038/nature19310`.

[88] P. Skoglund et al. “Genomic insights into the peopling of the Southwest Pacific”. *Nature* 538(7626) (2016), pp. 510–513. DOI: `10.1038/nature19844`.

[89] J.-H. Park et al. “Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants”. *PNAS* 108(44) (2011), pp. 18026–18031. DOI: `10.1073/pnas.1114759108`.

[90] L. J. O’Connor et al. “Extreme Polygenicity of Complex Traits Is Explained by Negative Selection”. *The American Journal of Human Genetics* 0(0) (2019). DOI: `10.1016/j.ajhg.2019.07.003`.

[91] Y. B. Simons et al. “A population genetic interpretation of GWAS findings for human quantitative traits”. *PLOS Biology* 16(3) (2018), e2002985. DOI: `10.1371/journal.pbio.2002985`.

[92] G. Malécot. *Mathématiques de l’hérédité*. Masson et Cie, 1948.

[93] S. Wright. “The Genetical Structure of Populations”. *Annals of Eugenics* 15(1) (1949), pp. 323–354. DOI: `10.1111/j.1469-1809.1949.tb02451.x`.

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

[95] Z. Hou and A. Ochoa. "Genetic association models are robust to common population kinship estimation biases". *Genetics* (27, 2023), iyad030. DOI: 10.1093/genetics/iyad030.

[96] C. C. Chang et al. "Second-generation PLINK: rising to the challenge of larger and richer datasets". *GigaScience* 4(1) (2015), p. 7. DOI: 10.1186/s13742-015-0047-8.

[97] D. J. Balding and R. A. Nichols. "A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity". *Genetica* 96(1-2) (1995), pp. 3–12. DOI: https://doi.org/10.1007/BF01441146.

[98] E. Paradis and K. Schliep. "ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R". *Bioinformatics* 35(3) (2019), pp. 526–528. DOI: 10.1093/bioinformatics/bty633.

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

[100] C. L. Lawson and R. J. Hanson. *Solving least squares problems*. Englewood Cliffs: Prentice Hall, 1974.

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

[102] J.-H. Park et al. "Estimation of effect size distribution from genome-wide association studies and implications for future discoveries". *Nature Genetics* 42(7) (2010), pp. 570–575. DOI: 10.1038/ng.610.

[103] A. Grueneberg and G. d. l. Campos. "BGData - A Suite of R Packages for Genomic Analysis with Big Data". *G3: Genes, Genomes, Genetics* 9(5) (2019), pp. 1377–1383. DOI: 10.1534/g3.119.400018.

[104] S. Fairley et al. "The International Genome Sample Resource (IGSR) collection of open human genomic variation resources". *Nucleic Acids Research* 48(D1) (2020), pp. D941–D947. DOI: 10.1093/nar/gkz836.

[105] A. Manichaikul et al. "Robust relationship inference in genome-wide association studies". *Bioinformatics* 26(22) (2010), pp. 2867–2873. DOI: 10.1093/bioinformatics/btq559.

[106] J. D. Storey. “The positive false discovery rate: a Bayesian interpretation and the q-value”. *Ann. Statist.* 31(6) (2003), pp. 2013–2035. DOI: 10.1214/aos/1074290335.

[107] J. D. Storey and R. Tibshirani. “Statistical significance for genomewide studies”. *Proceedings of the National Academy of Sciences of the United States of America* 100(16) (2003), pp. 9440–9445. DOI: 10.1073/pnas.1530509100.

[108] J. Grau, I. Grosse, and J. Keilwagen. “PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R”. *Bioinformatics* 31(15) (2015), pp. 2595–2597. DOI: 10.1093/bioinformatics/btv153.

[109] P. Gopalan et al. “Scaling probabilistic models of genetic variation to millions of humans”. *Nat. Genet.* 48(12) (2016), pp. 1587–1590. DOI: 10.1038/ng.3710.

[110] B. Rakitsch et al. “A Lasso multi-marker mixed model for association mapping with population structure correction”. *Bioinformatics* 29(2) (2013), pp. 206–214. DOI: 10.1093/bioinformatics/bts669.

[111] M. Conomos et al. “Model-free Estimation of Recent Genetic Relatedness”. *The American Journal of Human Genetics* 98(1) (2016), pp. 127–148. DOI: 10.1016/j.ajhg.2015.11.022.

[112] D. Heckerman et al. “Linear mixed model for heritability estimation that explicitly addresses environmental variation”. *Proc. Natl. Acad. Sci. U.S.A.* 113(27) (2016), pp. 7377–7382. DOI: 10.1073/pnas.1510497113.

[113] K. Hayashi, K.-H. Yuan, and L. Liang. “On the Bias in Eigenvalues of Sample Covariance Matrix”. *Quantitative Psychology.* Ed. by M. Wiberg et al. Springer Proceedings in Mathematics & Statistics. Cham: Springer International Publishing, 2018, pp. 221–233. DOI: 10.1007/978-3-319-77249-3_19.

[114] G. Sun et al. “Variation explained in mixed-model association mapping”. *Heredity* 105(4) (2010), pp. 333–340. DOI: 10.1038/hdy.2010.11.

[115] S. Gazal et al. “High level of inbreeding in final phase of 1000 Genomes Project”. *Sci Rep* 5(1) (2015), p. 17453. DOI: 10.1038/srep17453.

[116] A. Al-Khudhair et al. “Inference of Distant Genetic Relations in Humans Using “1000 Genomes””. *Genome Biology and Evolution* 7(2) (2015), pp. 481–492. DOI: `10.1093/gbe/evv003`.

[117] L. Fedorova et al. “Atlas of Cryptic Genetic Relatedness Among 1000 Human Genomes”. *Genome Biology and Evolution* 8(3) (2016), pp. 777–790. DOI: `10.1093/gbe/evw034`.

[118] D. Schlauch, H. Fier, and C. Lange. “Identification of genetic outliers due to sub-structure and cryptic relationships”. *Bioinformatics* 33(13) (2017), pp. 1972–1979. DOI: `10.1093/bioinformatics/btx109`.

[119] B. M. Henn et al. “Cryptic Distant Relatives Are Common in Both Isolated and Cosmopolitan Genetic Samples”. *PLOS ONE* 7(4) (2012), e34267. DOI: `10.1371/journal.pone.0034267`.

[120] V. Shchur and R. Nielsen. “On the number of siblings and p-th cousins in a large population sample”. *J Math Biol* 77(5) (2018), pp. 1279–1298. DOI: `10.1007/s00285-018-1252-8`.

[121] F. Privé et al. “Efficient toolkit implementing best practices for principal component analysis of population genetic data”. *Bioinformatics* 36(16) (15, 2020), pp. 4449–4457. DOI: `10.1093/bioinformatics/btaa520`.

[122] D. Speed et al. “Improved heritability estimation from genome-wide SNPs”. *Am. J. Hum. Genet.* 91(6) (7, 2012), pp. 1011–1021. DOI: `10.1016/j.ajhg.2012.10.010`.

[123] A. A. Zaidi and I. Mathieson. “Demographic history mediates the effect of stratification on polygenic scores”. *eLife* 9 (17, 2020). Ed. by G. H. Perry, M. C. Turchin, and A. R. Martin, e61548. DOI: `10.7554/eLife.61548`.

Supplemental figures

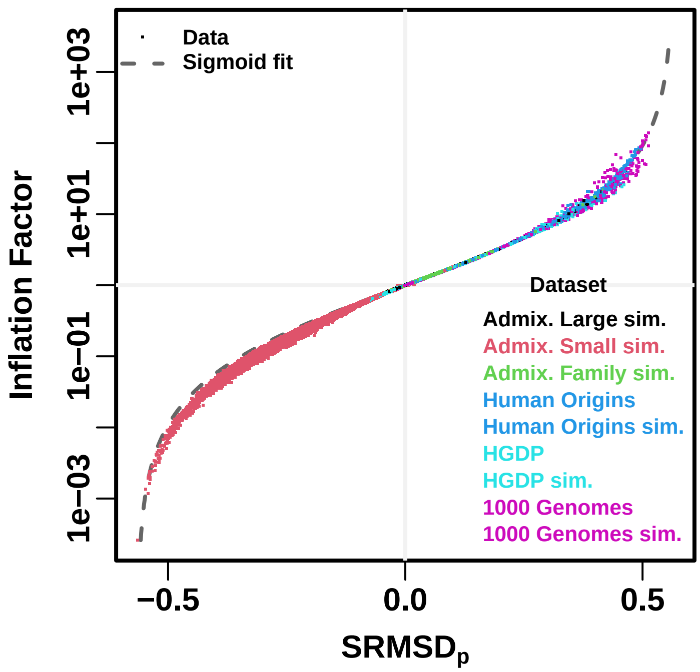

Figure S1: **Comparison between SRMSD$_p$ and inflation factor.** Each point is a pair of statistics for one replicate, one association model (PCA or LMM with some number of PCs r), one trait model (FES vs RC, all heritability/environments tested), and one dataset (color coded by dataset). Note log y-axis.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

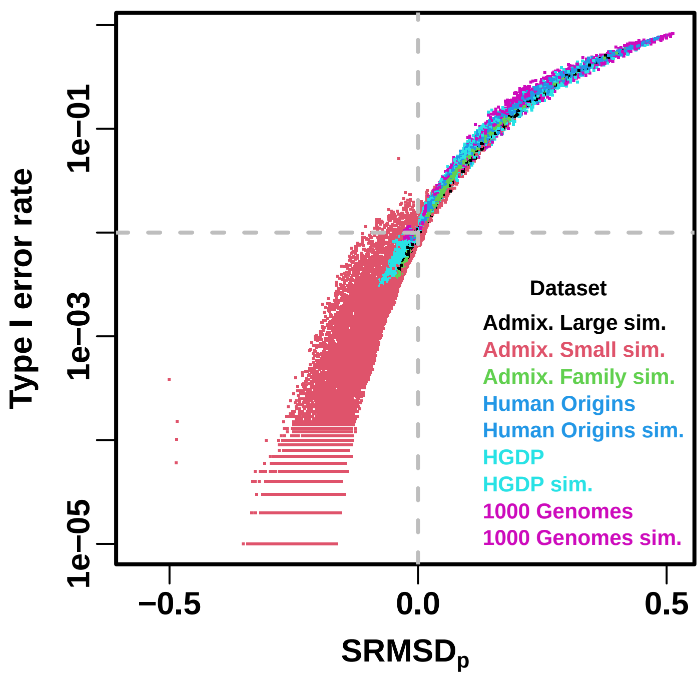

Figure S2: **Comparison between SRMSD$_p$ and type I error rate.** Type I error rate calculated at a p-value threshold of 1e-2 (horizontal dashed gray line). Thus, a calibrated model has a type I error rate of 1e-2 and SRMSD$_p = 0$ (where the dashed lines meet). As expected, increased type I error rates correspond to SRMSD$_p > 0$, while reduced type I error rates correspond to SRMSD$_p < 0$. 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, all heritability/environments tested), and one dataset (color coded by dataset). Note log y-axis.

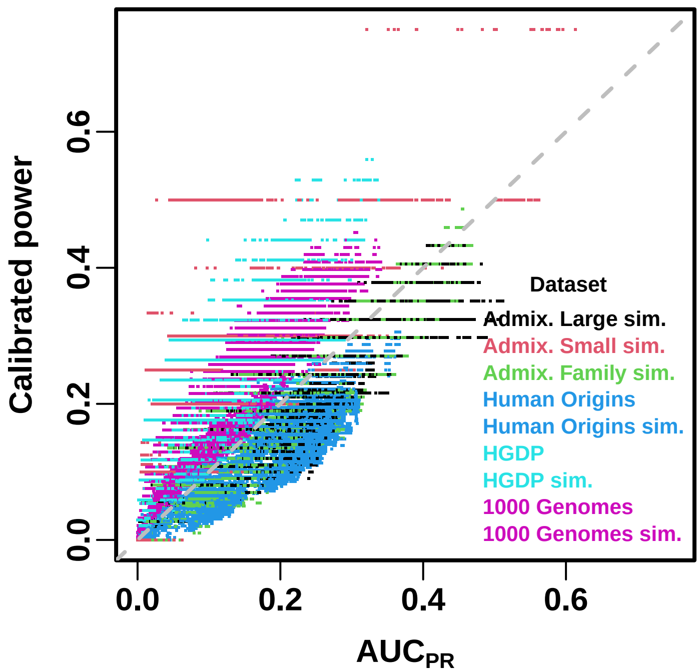

Figure S3: **Comparison between AUC$_{\mathbf{PR}}$ and calibrated power.** Calibrated power is power calculated at an empirical type I error threshold of 1e-4. Each point is a pair of statistics for one replicate, one association model (PCA or LMM with some number of PCs r), one trait model (FES vs RC, all heritability/environments tested), and one dataset (color coded by dataset). Gray dashed line is $y = x$ line.

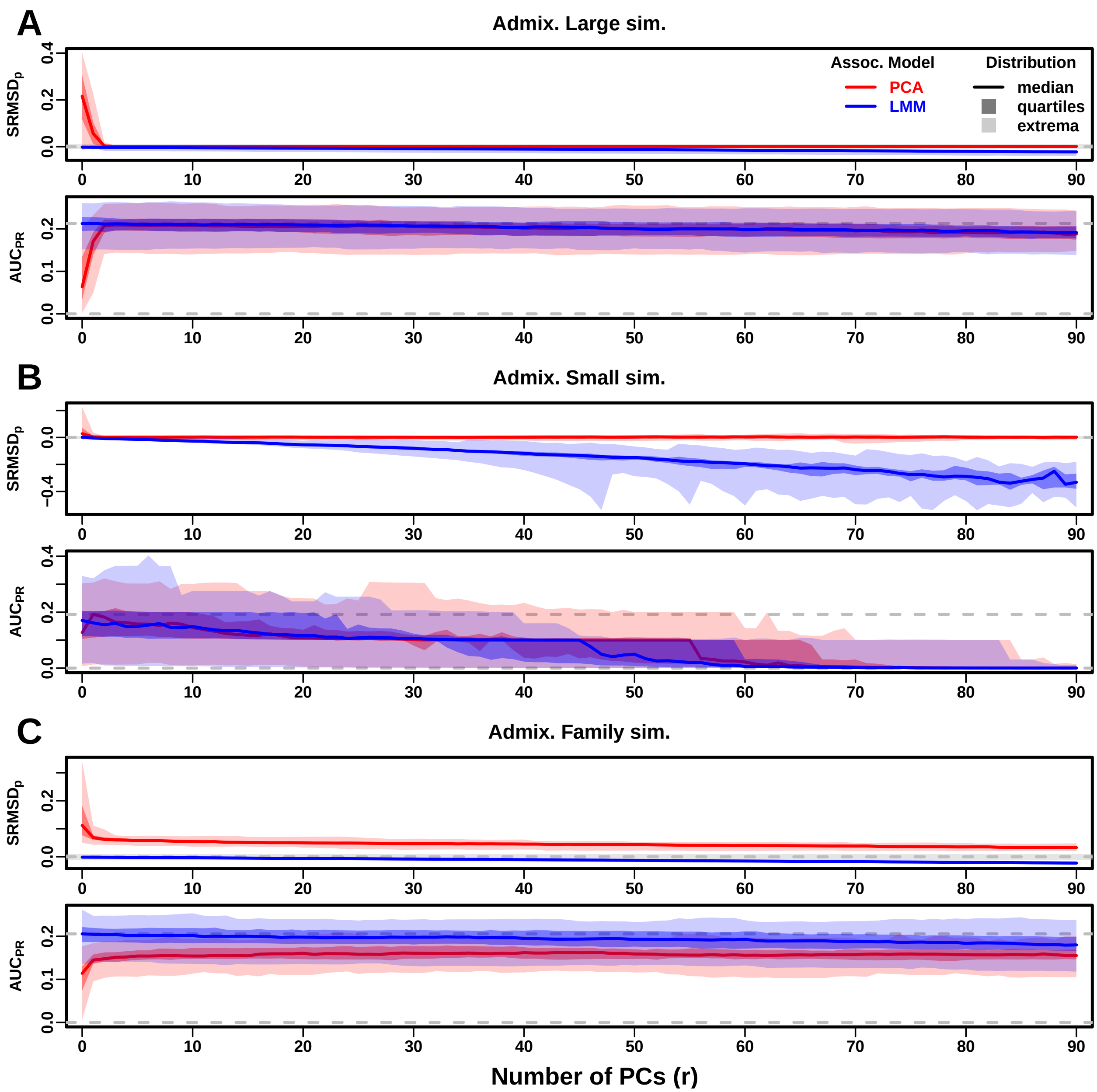

Figure S4: **Evaluations in admixture simulations with RC traits.** Traits simulated from RC model, otherwise the same as Fig. 3.

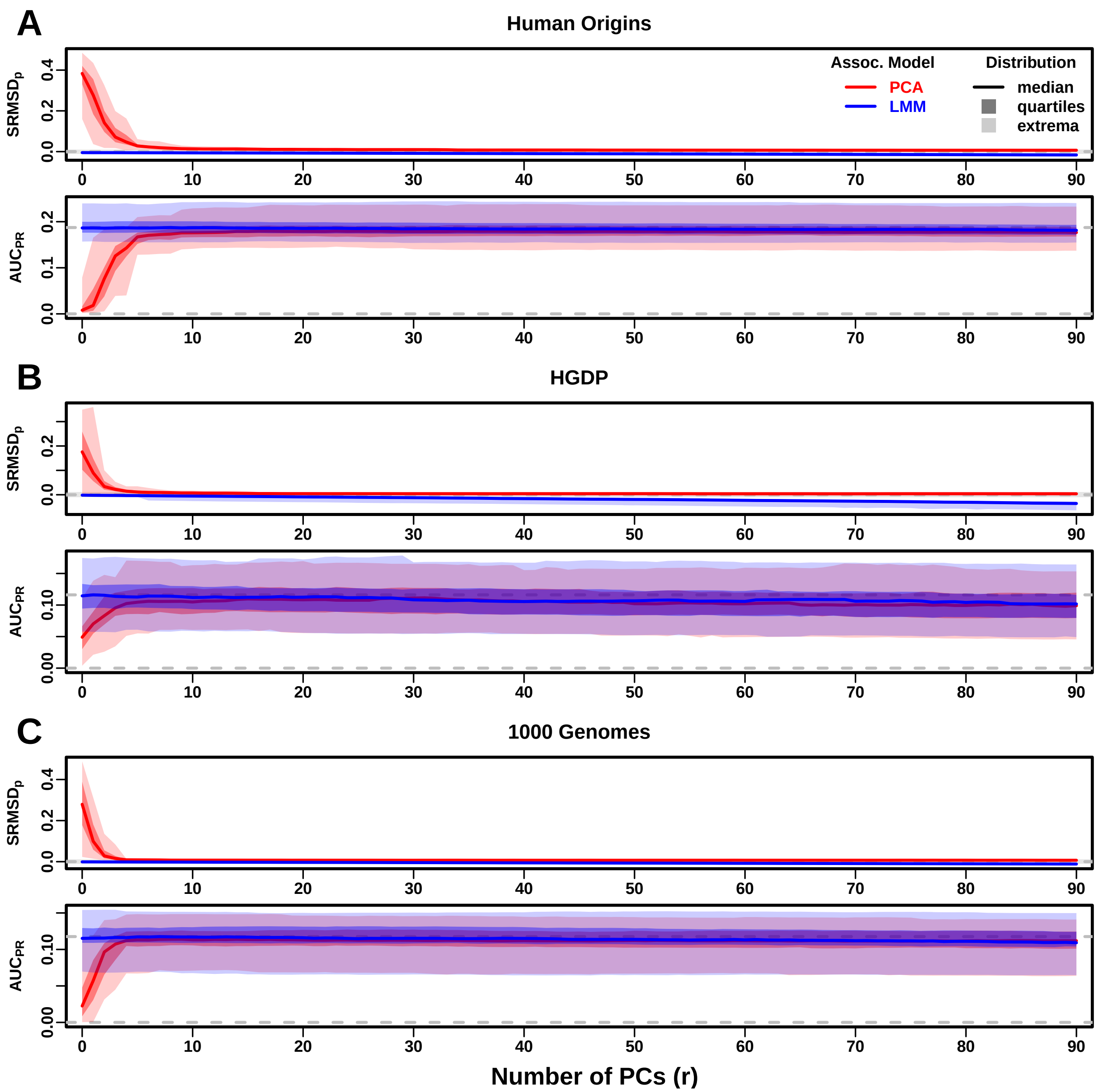

Figure S5: **Evaluations in real human genotype datasets with RC traits.** Traits simulated from RC model, otherwise the same as Fig. 4.

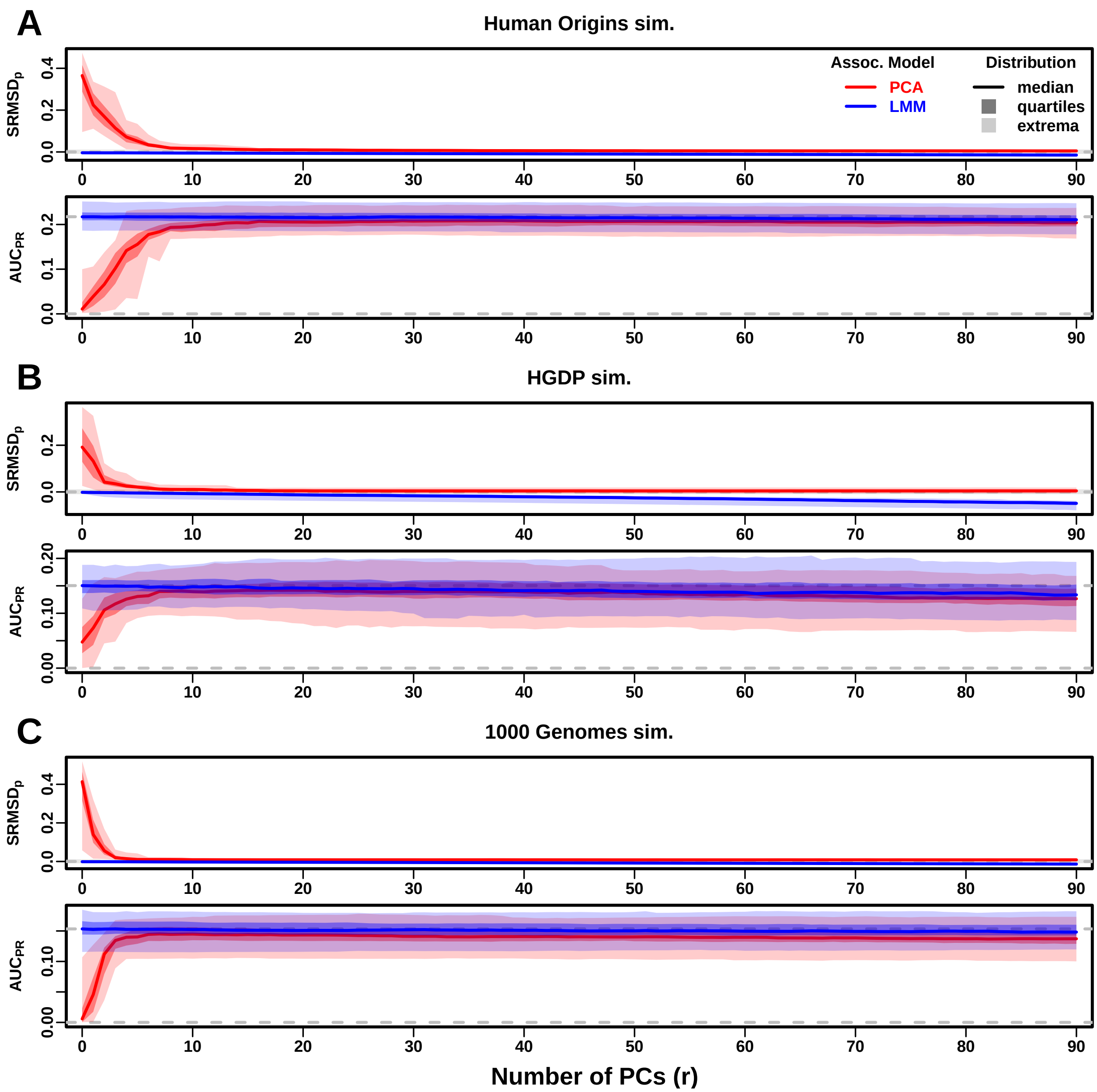

Figure S6: **Evaluations in subpopulation tree simulations fit to human data with RC traits.** Traits simulated from RC model, otherwise the same as Fig. 5.

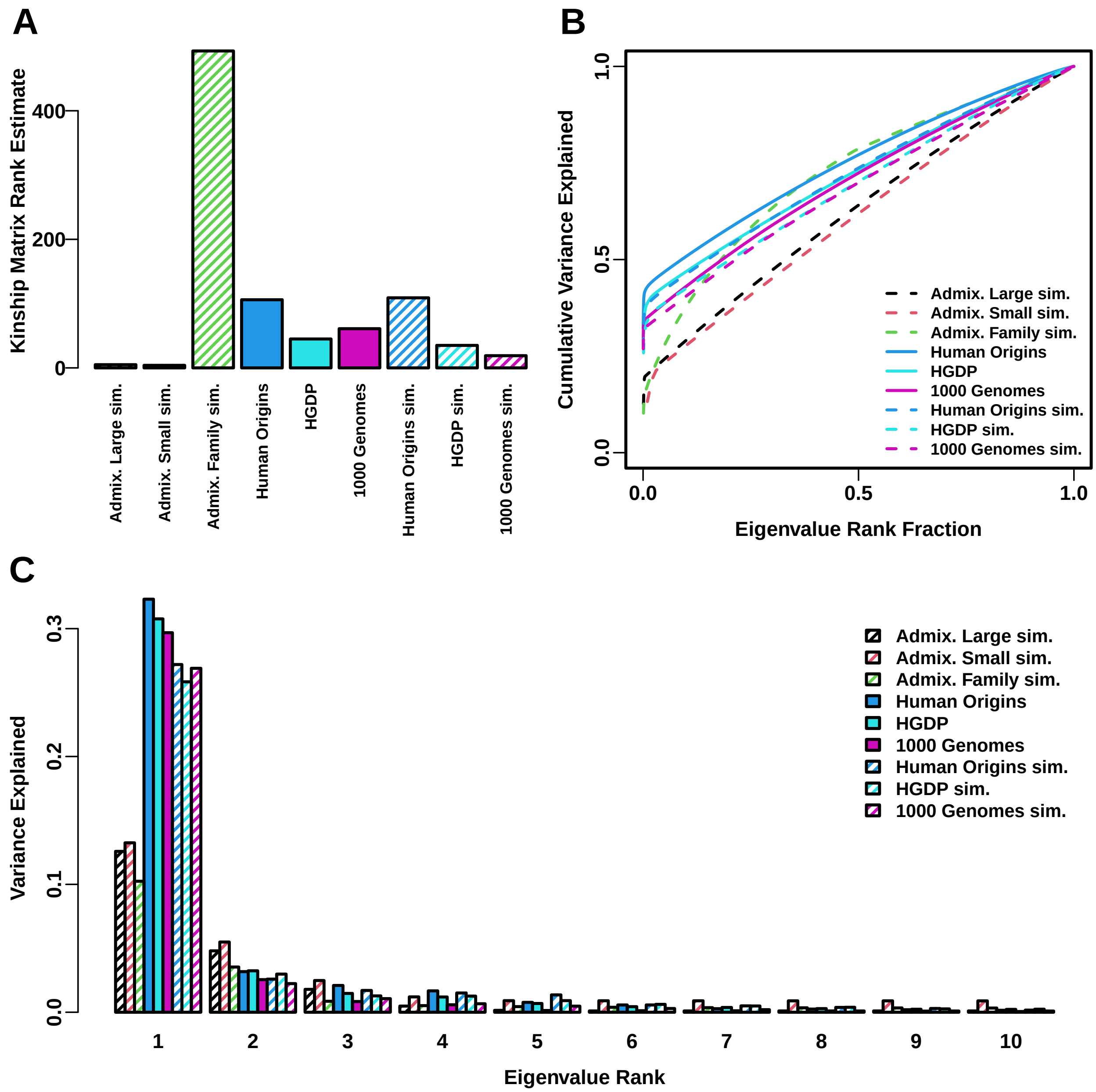

Figure S7: **Estimated relatedness dimensions of datasets. A.** Kinship matrix rank estimated with the Tracy-Widom test with $p < 0.01$. **B.** Cumulative variance explained versus eigenvalue rank fraction. **C.** Variance explained by first 10 eigenvalues.

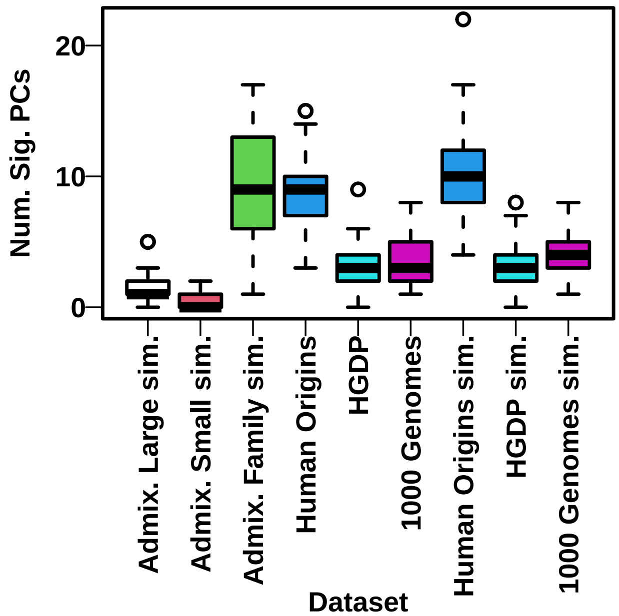

Figure S8: **Number of PCs significantly associated with traits.** PCs are tested using an ordinary linear regression sequentially, with the kth PC tested conditionally on the previous $k-1$ PCs and the intercept. Q-values are estimated from the 90 p-values (one for each PC in a given dataset and replicate) using the R package `qvalue` assuming $\pi_0 = 1$ (necessary since the default π_0 estimates were unreliable for such small numbers of p-values and occasionally produced errors), and an FDR threshold of 0.05 is used to determine the number of significant PCs. Distribution per dataset is over its 50 replicates. Shown are results for FES traits with $h^2 = 0.8$ (the results for RC were very similar, not shown).

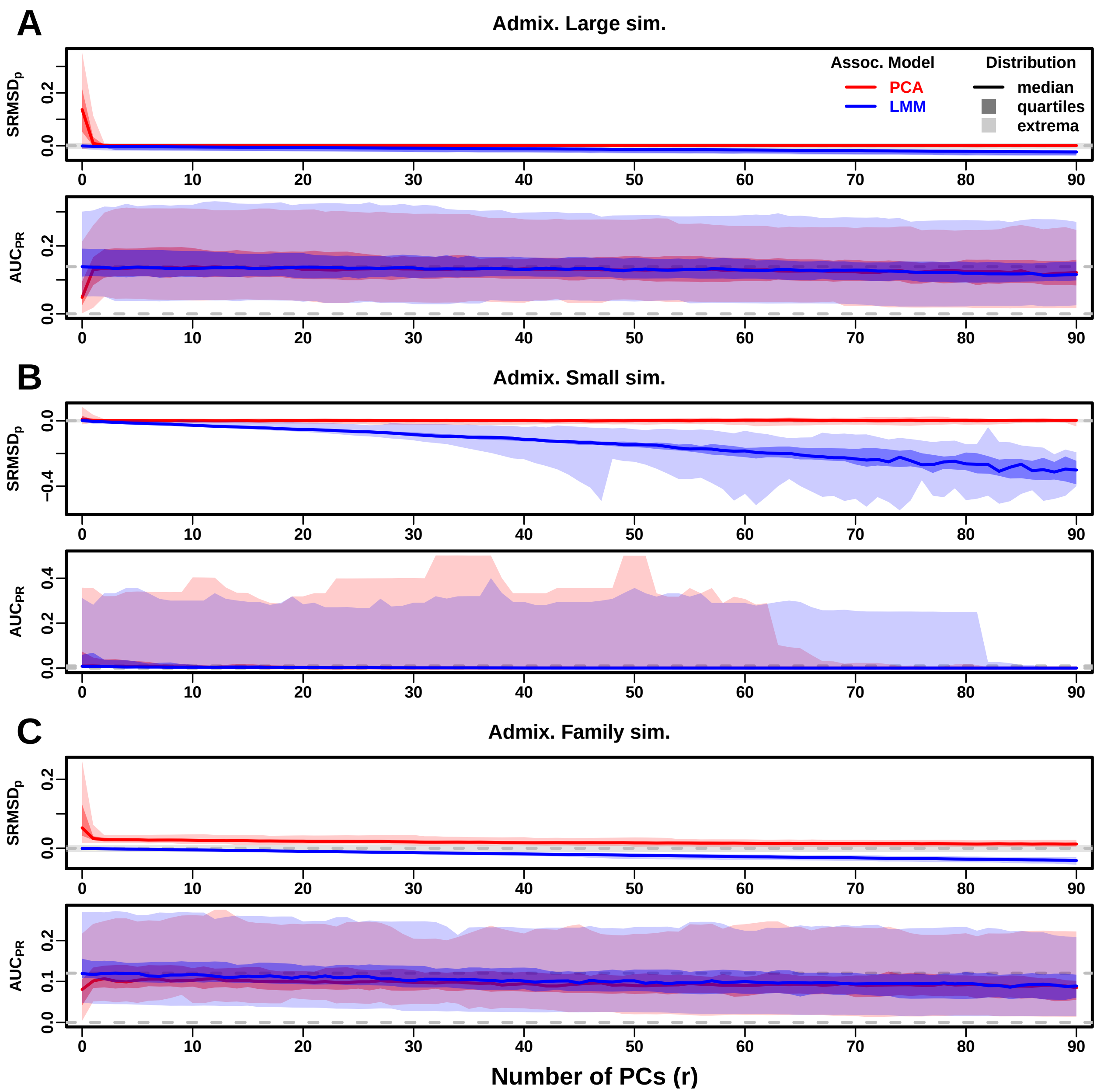

Figure S9: **Evaluations in admixture simulations with FES traits, low heritability.** Traits simulated using $h^2 = 0.3$, otherwise the same as Fig. 3.

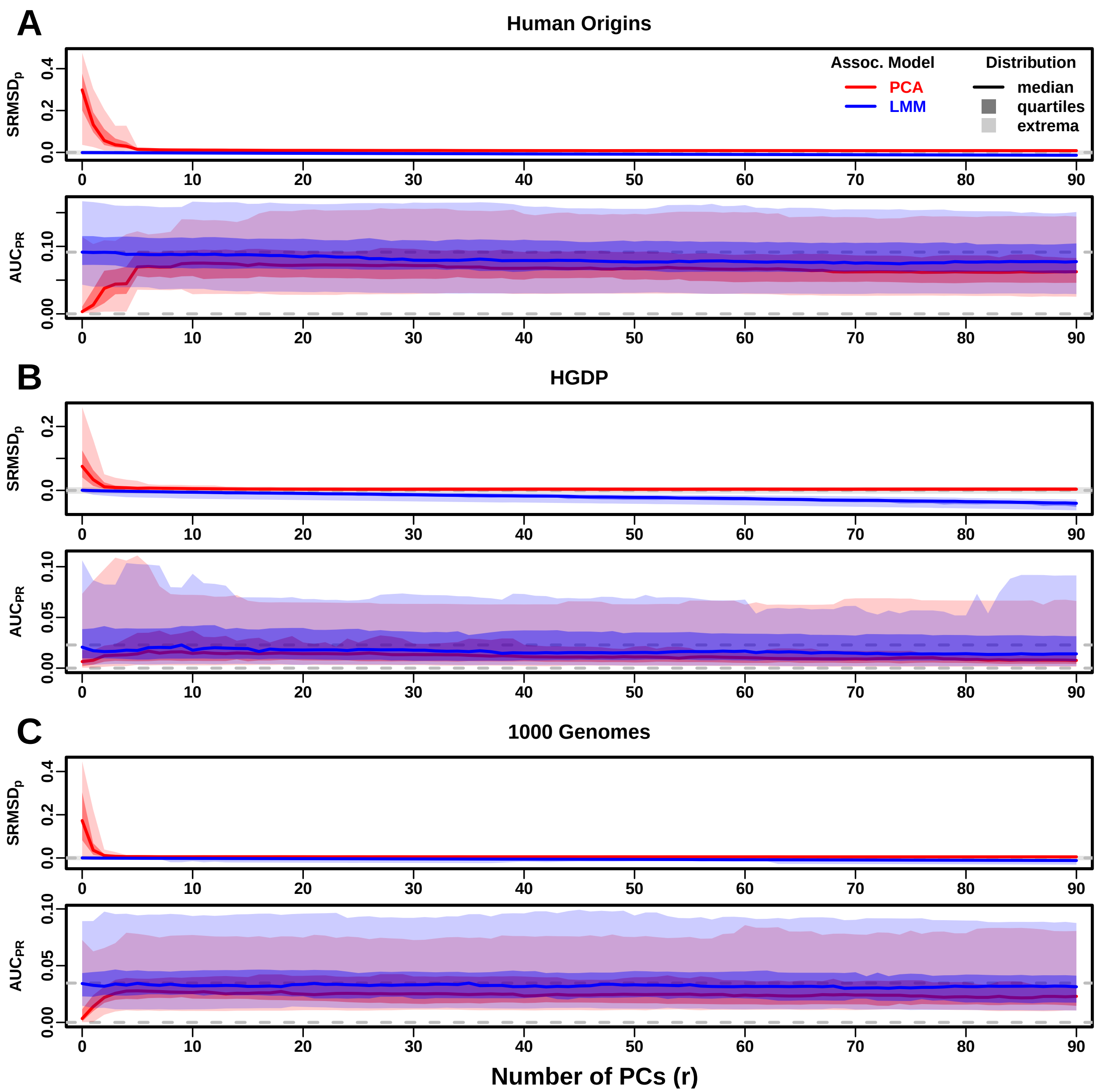

Figure S10: **Evaluations in real human genotype datasets with FES traits, low heritability.** Traits simulated using $h^2 = 0.3$, otherwise the same as Fig. 4.

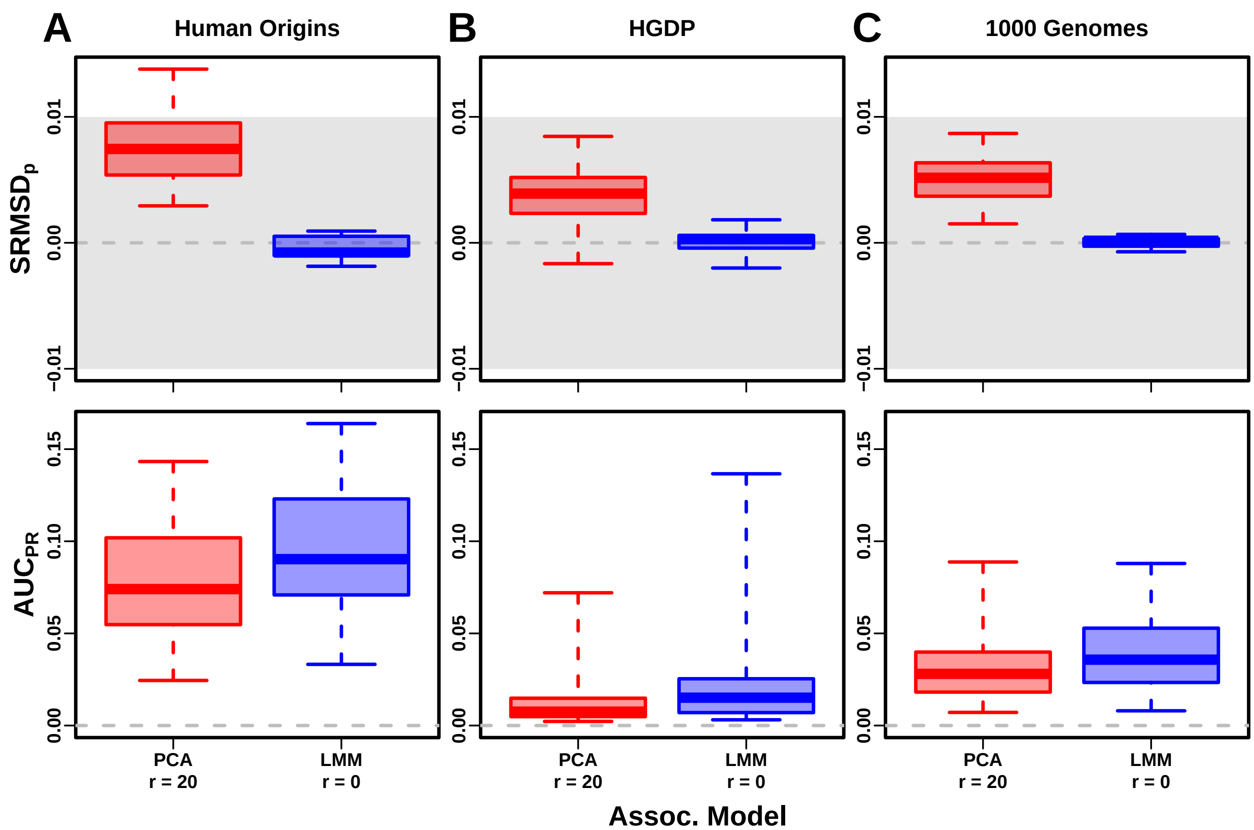

Figure S11: **Evaluation in real datasets excluding 4th degree relatives, FES traits, low heritability.** Traits simulated using $h^2 = 0.3$, otherwise the same as Fig. 7.

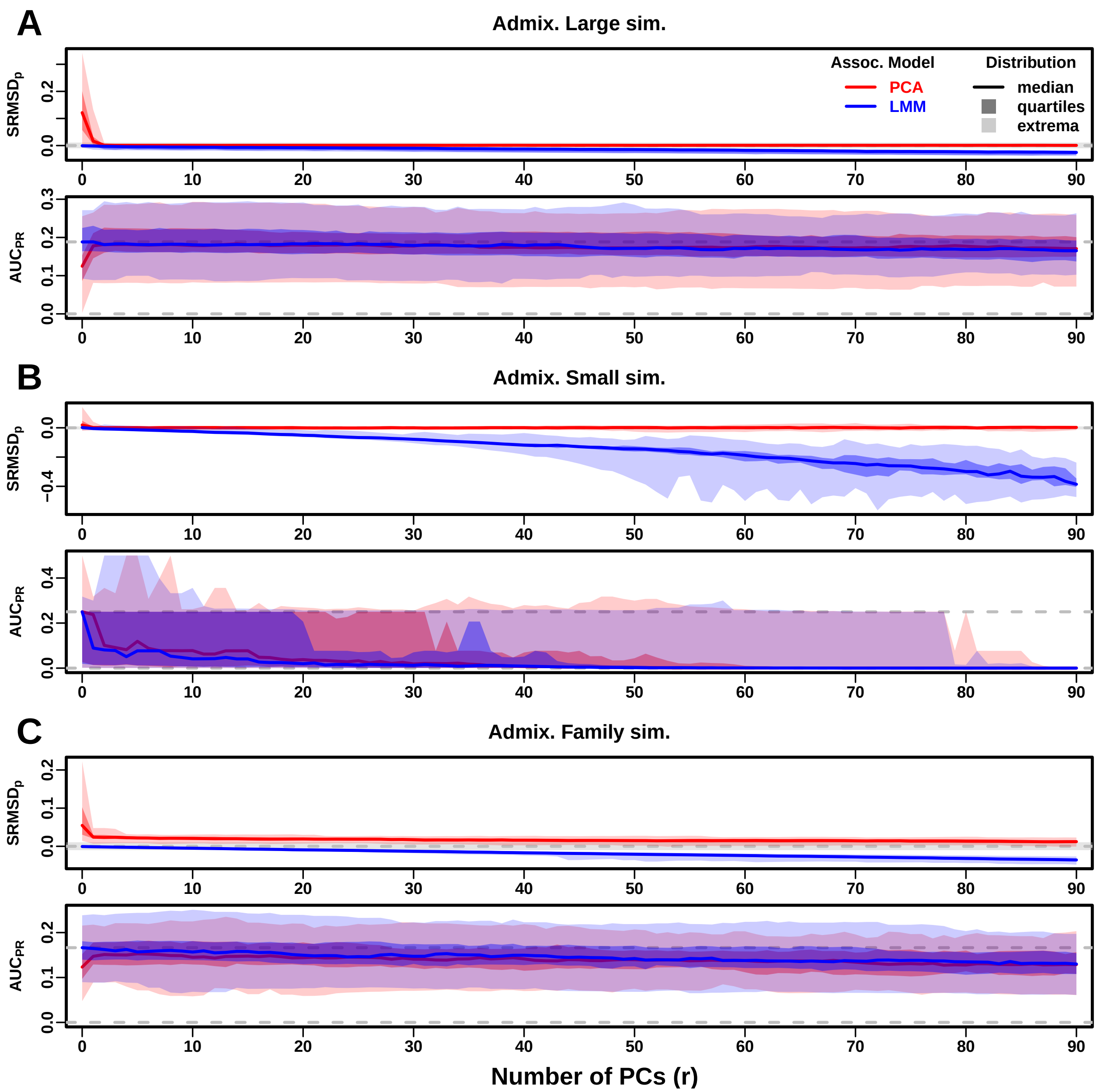

Figure S12: **Evaluations in admixture simulations with RC traits, low heritability.** Traits simulated using $h^2 = 0.3$, otherwise the same as Fig. S4.

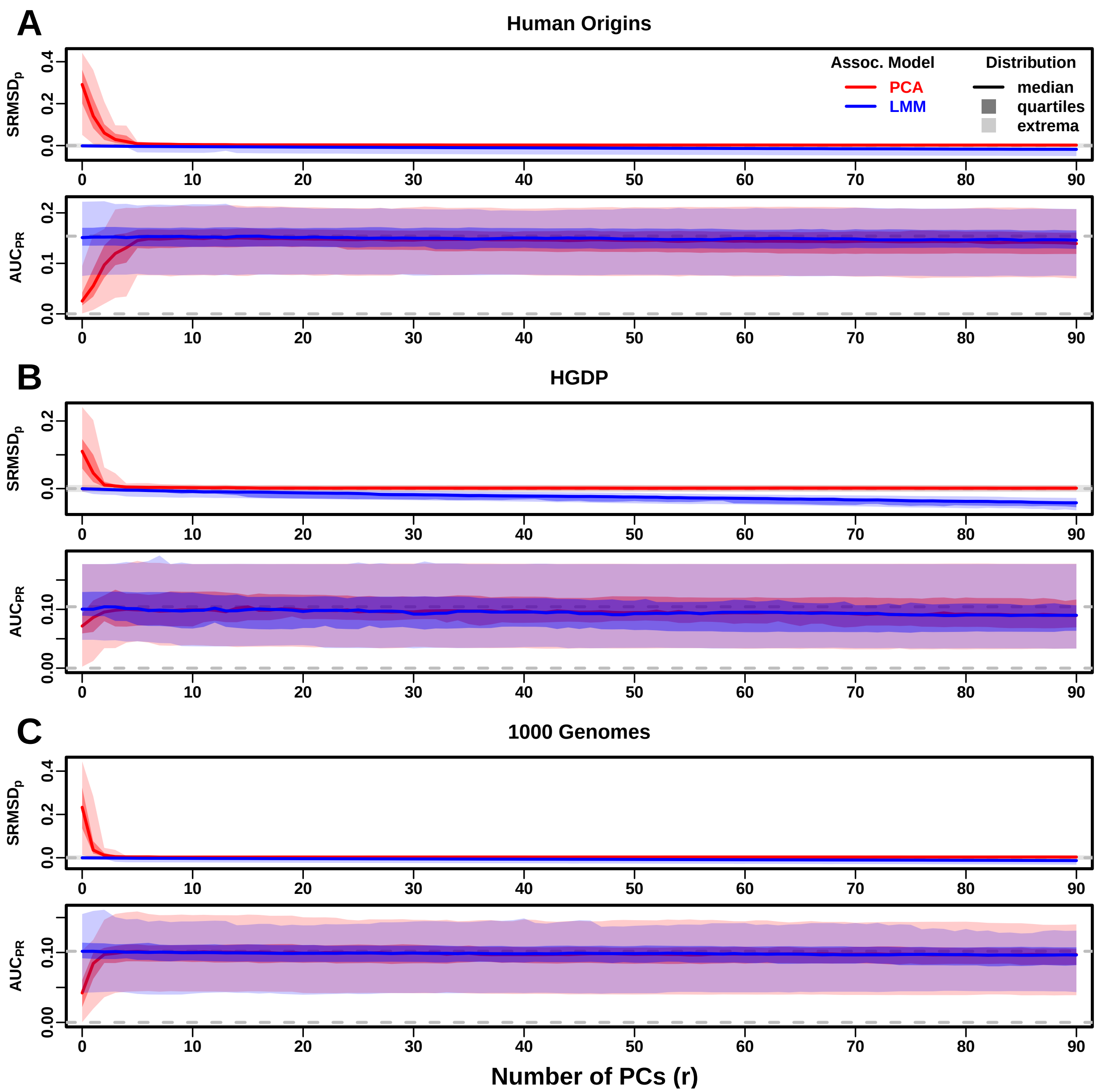

Figure S13: **Evaluations in real human genotype datasets with RC traits, low heritability.** Traits simulated using $h^2 = 0.3$, otherwise the same as Fig. S5.

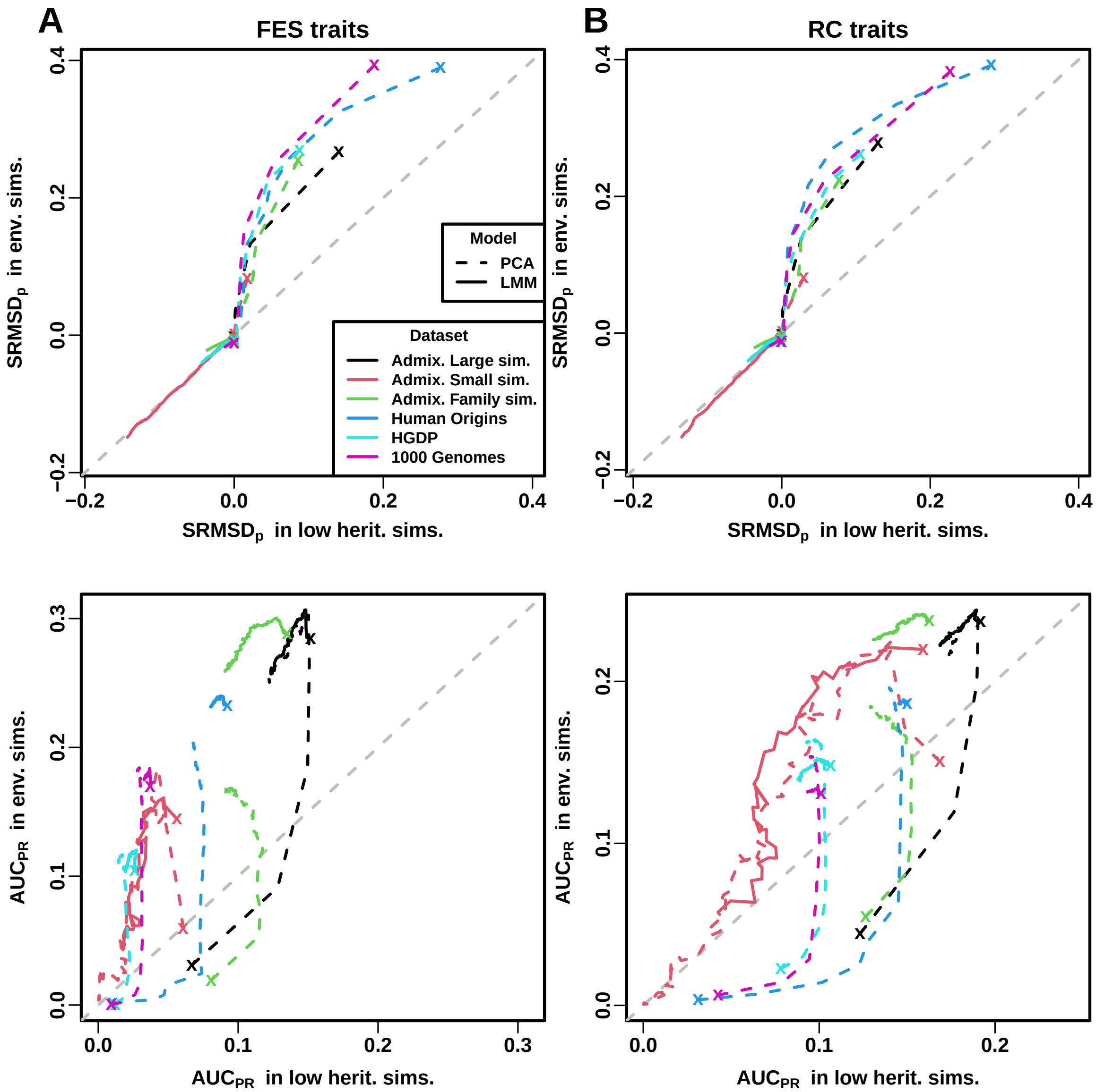

Figure S14: **Comparison of performance in low heritability vs environment simulations.** Each curve traces as the number of PCs r is increased from $r = 0$ (marked with an "x") until $r = 90$ (unmarked end), on one axis is the mean value over replicates of either SRMSD$_p$ or AUC$_{\mathrm{PR}}$, for low heritability simulations on the x-axis and environment simulations on the y-axis. Each curve corresponds to one dataset (color) and association model (solid or dashed line type). Columns: **A.** FES and **B.** RC traits show similar results. First row shows that for PCA curves (dashed), SRMSD$_p$ is higher (worse) in environment simulations for low r, but becomes equal in both simulations once r is sufficiently large; for LMM curves (solid), SRMSD$_p$ is equal in both simulations for all r, all datasets. Second row shows that for PCA, AUC$_{\mathrm{PR}}$ is higher (better) in low heritability simulations for low r, but becomes higher in environment simulations once r is sufficiently large; for LMM, performance is better in environment simulations for all r, all datasets.

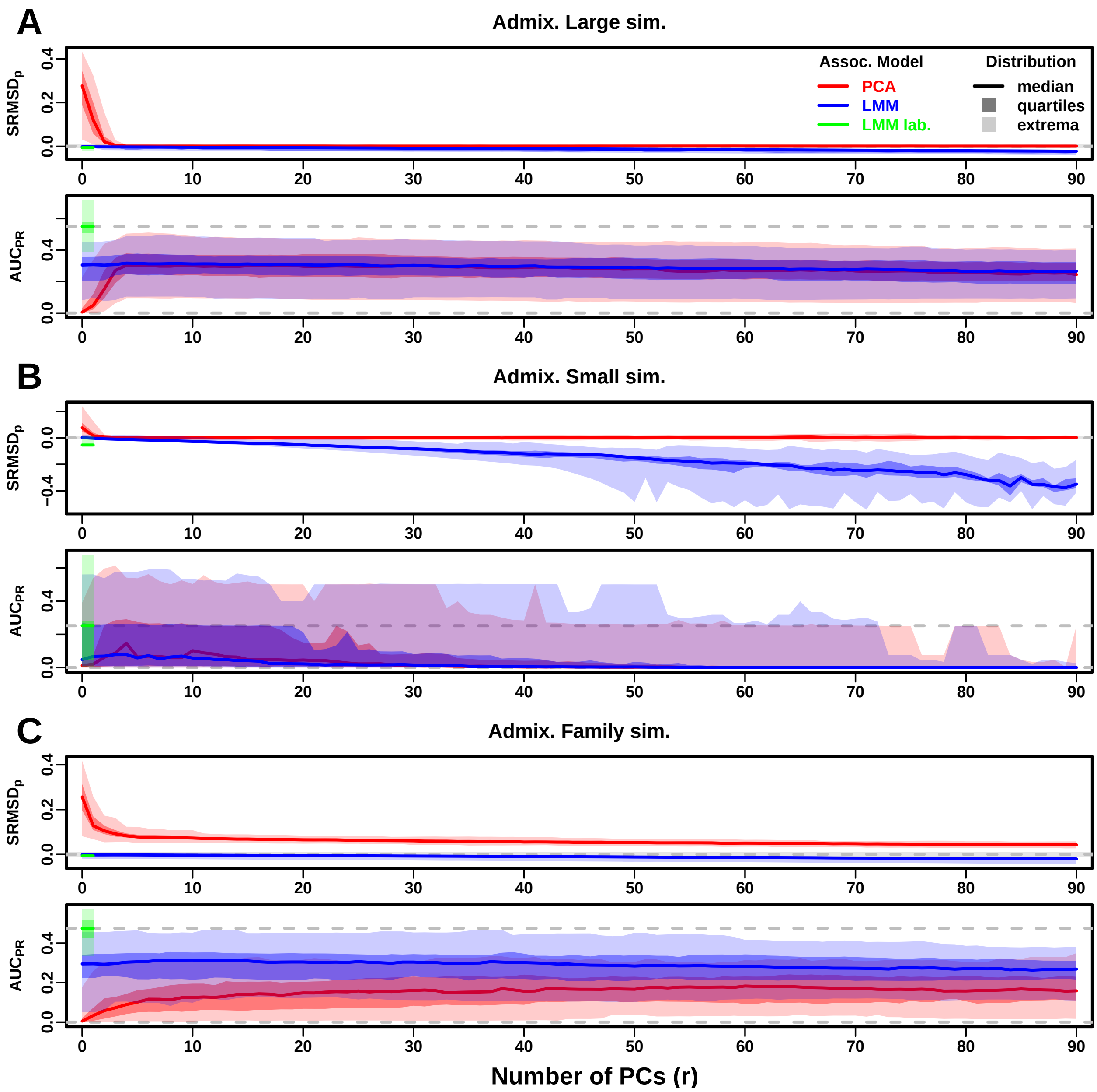

Figure S15: **Evaluations in admixture simulations with FES traits, environment.** Traits simulated with environment effects, otherwise the same as Fig. S9. "LMM lab." was only tested with $r = 0$.

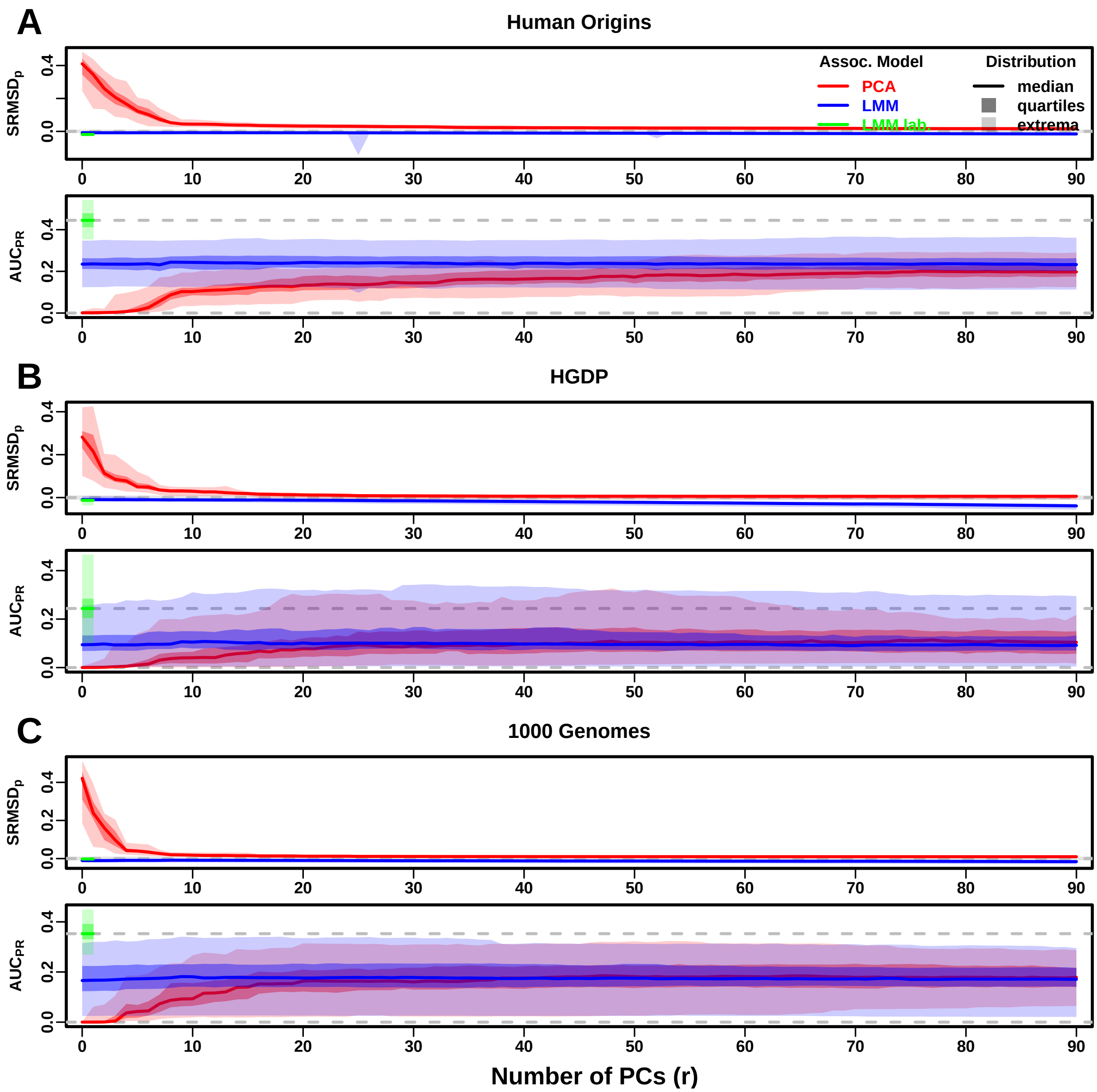

Figure S16: **Evaluations in real human genotype datasets with FES traits, environment.** Traits simulated with environment effects, otherwise the same as Fig. S10. "LMM lab." was only tested with $r = 0$.

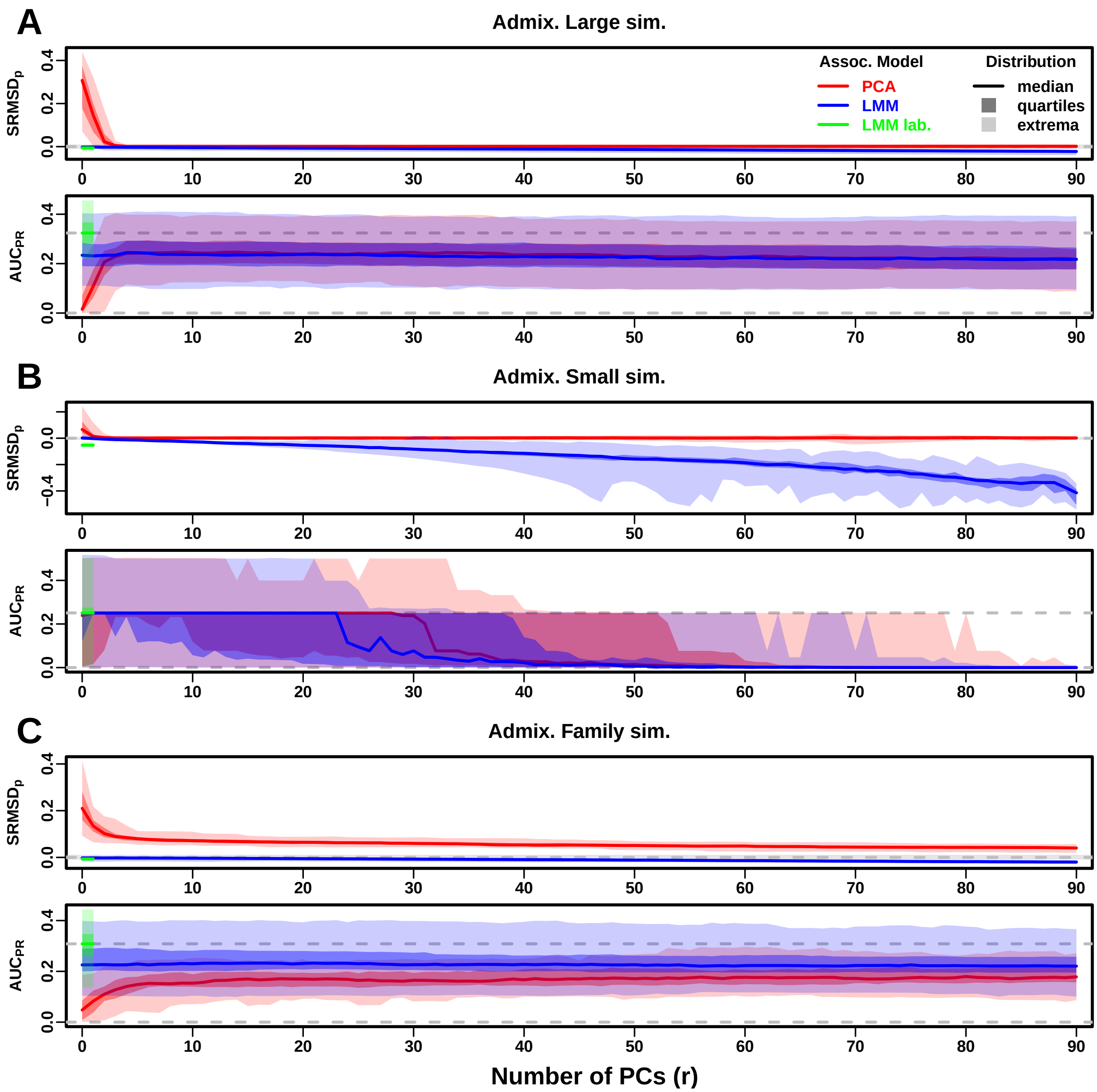

Figure S17: **Evaluations in admixture simulations with RC traits, environment.** Traits simulated with environment effects, otherwise the same as Fig. S12. "LMM lab." was only tested with $r = 0$.

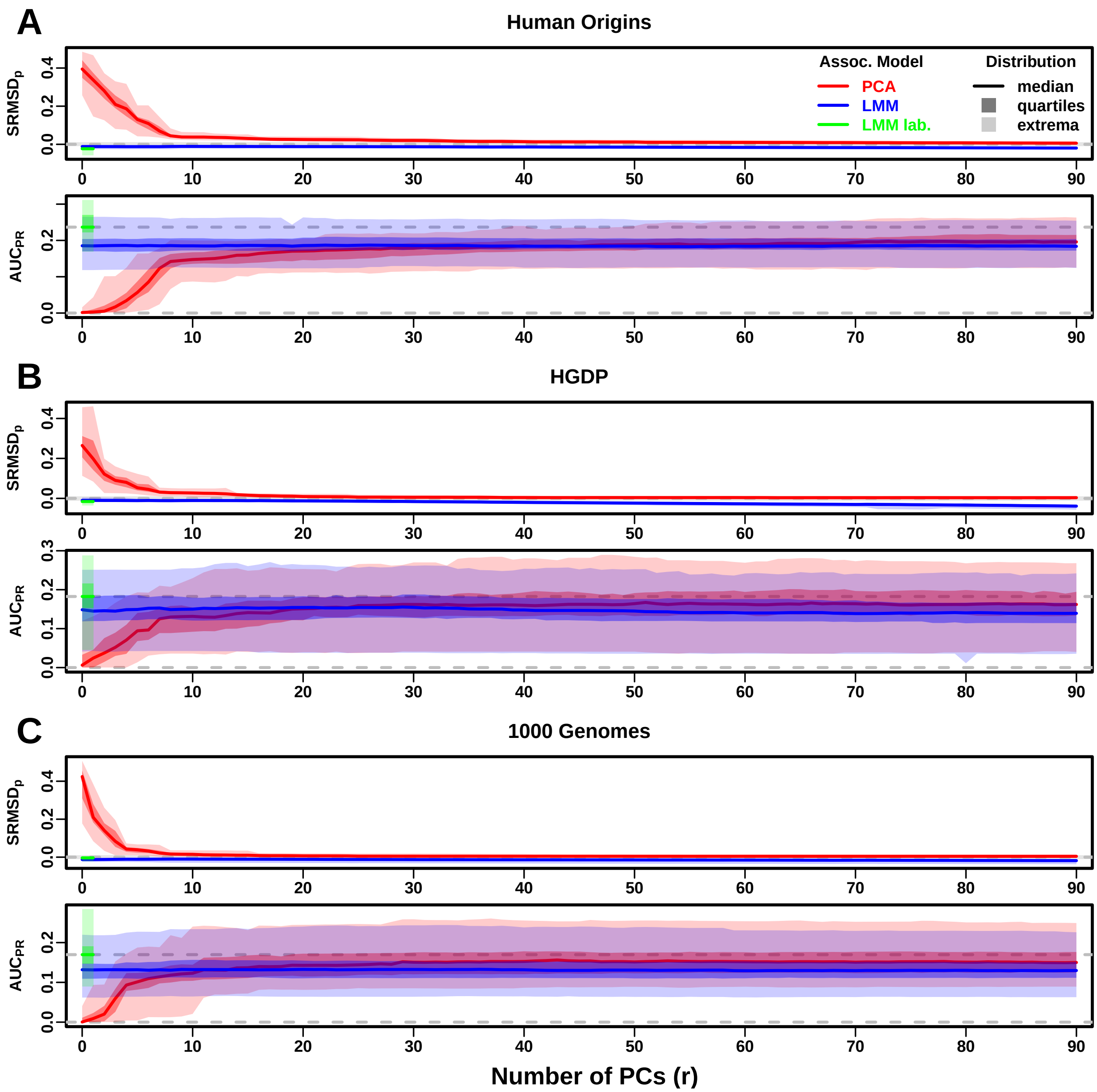

Figure S18: **Evaluations in real human genotype datasets with RC traits, environment.** Traits simulated with environment effects, otherwise the same as Fig. S13. "LMM lab." was only tested with $r = 0$.

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	758,009	847	8.8
1000 Genomes	1,097,415	2390	4.6

Table S2: **Overview of PCA and LMM evaluations for low heritability simulations**

			LMM $r = 0$ vs best r			PCA vs LMM $r = 0$			
Dataset	Metric	Trait[a]	Cal.[b]	Best r[c]	P-value[d]	Best r[c]	Cal.[b]	P-value[d]	Best model[e]
Admix. Large sim.	$\|SRMSD_p\|$	FES	True	0	1	62	True	0.00012*	LMM
Admix. Small sim.	$\|SRMSD_p\|$	FES	True	0	1	3	True	0.27	Tie
Admix. Family sim.	$\|SRMSD_p\|$	FES	True	0	1	90	False	3.9e-10*	LMM
Human Origins	$\|SRMSD_p\|$	FES	True	0	1	81	True	3.9e-10*	LMM
HGDP	$\|SRMSD_p\|$	FES	True	0	1	37	True	6.2e-09*	LMM
1000 Genomes	$\|SRMSD_p\|$	FES	True	0	1	84	True	3.9e-10*	LMM
Admix. Large sim.	$\|SRMSD_p\|$	RC	True	0	1	35	True	0.00094	Tie
Admix. Small sim.	$\|SRMSD_p\|$	RC	True	0	1	3	True	0.087	Tie
Admix. Family sim.	$\|SRMSD_p\|$	RC	True	0	1	90	False	4.1e-10*	LMM
Human Origins	$\|SRMSD_p\|$	RC	True	0	1	75	True	0.00016*	LMM
HGDP	$\|SRMSD_p\|$	RC	True	0	1	23	True	1.7e-05*	LMM
1000 Genomes	$\|SRMSD_p\|$	RC	True	0	1	41	True	6.7e-10*	LMM
Admix. Large sim.	AUC_{PR}	FES		0	1	3		0.11	Tie
Admix. Small sim.	AUC_{PR}	FES		0	1	0		0.58	Tie
Admix. Family sim.	AUC_{PR}	FES		0	1	7		2.2e-06*	LMM
Human Origins	AUC_{PR}	FES		0	1	16		8e-10*	LMM
HGDP	AUC_{PR}	FES		11	0.68	6		0.0043	Tie
1000 Genomes	AUC_{PR}	FES		6	0.34	4		2.3e-07*	LMM
Admix. Large sim.	AUC_{PR}	RC		0	1	3		0.14	Tie
Admix. Small sim.	AUC_{PR}	RC		0	1	0		0.1	Tie
Admix. Family sim.	AUC_{PR}	RC		0	1	5		1.9e-06*	LMM
Human Origins	AUC_{PR}	RC		4	0.16	12		0.003	Tie
HGDP	AUC_{PR}	RC		2	0.14	5		0.14	Tie
1000 Genomes	AUC_{PR}	RC		0	1	4		0.078	Tie

[a]FES: Fixed Effect Sizes, RC: Random Coefficients.
[b]Calibrated: whether mean $|SRMSD_p| < 0.01$.
[c]Value of r (number of PCs) with minimum mean $|SRMSD_p|$ or maximum mean AUC_{PR}.
[d]Wilcoxon paired 1-tailed test of distributions ($|SRMSD_p|$ or AUC_{PR}) between models in header. Asterisk marks significant value using Bonferroni threshold ($p < \alpha/n_{tests}$ with $\alpha = 0.01$ and $n_{tests} = 48$ is the number of tests in this table).
[e]Tie if no significant difference using Bonferroni threshold.

Table S3: **Overview of PCA and LMM evaluations for environment simulations**

			LMM $r = 0$ vs best r			PCA vs LMM $r = 0$				LMM lab. $r = 0$ vs PCA/LMM		
Dataset	Metric	Trait[a]	Cal.[b]	r[c]	P-value[d]	r[c]	Cal.[b]	P-value[d]	Best[e]	Cal.[b]	P-value[d]	Best[e]
Admix. Large sim.	$\lvert\mathrm{SRMSD}_p\rvert$	FES	True	0	1	83	True	0.38	Tie	True	1.8e-14*	PCA/LMM
Admix. Small sim.	$\lvert\mathrm{SRMSD}_p\rvert$	FES	True	0	1	90	True	0.001	Tie	False	1.4e-14*	PCA/LMM
Admix. Family sim.	$\lvert\mathrm{SRMSD}_p\rvert$	FES	True	4	0.18	90	False	3.9e-10*	LMM	True	0.066	LMM/LMM lab.
Human Origins	$\lvert\mathrm{SRMSD}_p\rvert$	FES	True	9	3.9e-05*	90	False	1.4e-08*	LMM	False	3.9e-10*	LMM
HGDP	$\lvert\mathrm{SRMSD}_p\rvert$	FES	True	0	1	90	True	0.0037	Tie	False	2.1e-09*	PCA/LMM
1000 Genomes	$\lvert\mathrm{SRMSD}_p\rvert$	FES	False	8	8.8e-08*	85	True	0.053	Tie	True	3.9e-10*	LMM lab.
Admix. Large sim.	$\lvert\mathrm{SRMSD}_p\rvert$	RC	True	0	1	60	True	0.033	Tie	True	6.3e-10*	PCA/LMM
Admix. Small sim.	$\lvert\mathrm{SRMSD}_p\rvert$	RC	True	0	1	9	True	0.85	Tie	False	1.4e-14*	PCA/LMM
Admix. Family sim.	$\lvert\mathrm{SRMSD}_p\rvert$	RC	True	5	0.14	90	False	3.9e-10*	LMM	True	0.011	LMM/LMM lab.
Human Origins	$\lvert\mathrm{SRMSD}_p\rvert$	RC	False	9	1.1e-08*	90	True	2.3e-07*	PCA	False	3.9e-10*	PCA
HGDP	$\lvert\mathrm{SRMSD}_p\rvert$	RC	True	0	1	89	True	6.5e-09*	PCA	False	3.9e-10*	PCA
1000 Genomes	$\lvert\mathrm{SRMSD}_p\rvert$	RC	False	8	1.6e-08*	88	True	4.9e-09*	PCA	True	0.09	PCA/LMM lab.
Admix. Large sim.	$\mathrm{AUC}_{\mathrm{PR}}$	FES		4	2.4e-06*	6		0.0021	Tie		1.8e-15*	LMM lab.
Admix. Small sim.	$\mathrm{AUC}_{\mathrm{PR}}$	FES		3	0.055	4		0.033	Tie		0.28	Tie
Admix. Family sim.	$\mathrm{AUC}_{\mathrm{PR}}$	FES		12	7e-04	63		3.9e-10*	LMM		3.9e-10*	LMM lab.
Human Origins	$\mathrm{AUC}_{\mathrm{PR}}$	FES		20	3.7e-06*	90		1.4e-05*	LMM		3.9e-10*	LMM lab.
HGDP	$\mathrm{AUC}_{\mathrm{PR}}$	FES		12	4.3e-06*	45		0.0044	Tie		3.9e-10*	LMM lab.
1000 Genomes	$\mathrm{AUC}_{\mathrm{PR}}$	FES		9	1.9e-08*	55		0.028	Tie		3.9e-10*	LMM lab.
Admix. Large sim.	$\mathrm{AUC}_{\mathrm{PR}}$	RC		4	0.00085	5		0.0018	Tie		5e-10*	LMM lab.
Admix. Small sim.	$\mathrm{AUC}_{\mathrm{PR}}$	RC		2	0.13	5		0.093	Tie		0.0028	Tie
Admix. Family sim.	$\mathrm{AUC}_{\mathrm{PR}}$	RC		9	0.01	86		1.7e-09*	LMM		3.9e-10*	LMM lab.
Human Origins	$\mathrm{AUC}_{\mathrm{PR}}$	RC		22	0.0039	90		1e-06*	PCA		3.9e-10*	LMM lab.
HGDP	$\mathrm{AUC}_{\mathrm{PR}}$	RC		19	0.0057	64		2.8e-05*	PCA		3e-07*	LMM lab.
1000 Genomes	$\mathrm{AUC}_{\mathrm{PR}}$	RC		9	8.7e-05*	87		1.2e-09*	PCA		4.4e-10*	LMM lab.

[a]FES: Fixed Effect Sizes, RC: Random Coefficients.
[b]Calibrated: whether mean $|\mathrm{SRMSD}_p| < 0.01$.
[c]Value of r (number of PCs) with minimum mean $|\mathrm{SRMSD}_p|$ or maximum mean $\mathrm{AUC}_{\mathrm{PR}}$.
[d]Wilcoxon paired 1-tailed test of distributions ($|\mathrm{SRMSD}_p|$ or $\mathrm{AUC}_{\mathrm{PR}}$) between models in header. Asterisk marks significant value using Bonferroni threshold ($p < \alpha/n_{\mathrm{tests}}$ with $\alpha = 0.01$ and $n_{\mathrm{tests}} = 72$ is the number of tests in this table).
[e]Tie if no significant difference using Bonferroni threshold; in last column, pairwise ties are specified and "Tie" is three-way tie.