

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

3 Yiqi Yao,^{1,3} Alejandro Ochoa^{1,2,*}

4 ¹ Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27705, USA

5 ² Duke Center for Statistical Genetics and Genomics, Duke University, Durham, NC 27705, USA

6 ³ Present address: BenHealth Consulting, Shanghai, Shanghai, 200023, China

7 * Correspondence: alejandro.ochoa@duke.edu

8 **Abstract**

9 Principal Component Analysis (PCA) and the Linear Mixed-effects Model (LMM), some-
10 times in combination, are the most common genetic association models. Previous PCA-LMM
11 comparisons give mixed results, unclear guidance, and have several limitations, including not
12 varying the number of principal components (PCs), simulating simple population structures,
13 and inconsistent use of real data and power evaluations. We evaluate PCA and LMM both
14 varying number of PCs in realistic genotype and complex trait simulations including admixed
15 families, subpopulation trees, and real multiethnic human datasets with simulated traits. We
16 find that LMM without PCs usually performs best, with the largest effects in family simulations
17 and real human datasets and traits without environment effects. Poor PCA performance on
18 human datasets is driven by large numbers of distant relatives more than the smaller number
19 of closer relatives. While PCA was known to fail on family data, we report strong effects of
20 family relatedness in genetically diverse human datasets, not avoided by pruning close relatives.
21 Environment effects driven by geography and ethnicity are better modeled with LMM includ-
22 ing those labels instead of PCs. This work better characterizes the severe limitations of PCA
23 compared to LMM in modeling the complex relatedness structures of multiethnic human data
24 for association studies.

25 **Abbreviations:** PCA: principal component analysis; PCs: principal components; LMM: linear

26 mixed-effects model; FES: fixed effect sizes (trait model); RC: random coefficients (trait model);
27 MAF: minor allele frequency; WGS: whole genome sequencing; LD: linkage disequilibrium.

28 1 Introduction

29 The goal of a genetic association study is to identify loci whose genotype variation is significantly
30 correlated to given trait. Naive association tests assume that genotypes are drawn independently
31 from a common allele frequency. This assumption does not hold for structured populations, which
32 includes multiethnic cohorts and admixed individuals (ancient relatedness), and for family data
33 (recent relatedness) [1]. When insufficient approaches are applied to data with relatedness, their
34 association statistics are miscalibrated, resulting in excess false positives and loss of power [1–
35 3]. Therefore, many specialized approaches have been developed for genetic association under
36 relatedness, of which PCA and LMM are the most popular.

37 Genetic association with PCA consists of including the top eigenvectors of the population kinship
38 matrix as covariates in a generalized linear model [4–6]. These top eigenvectors are a new set of
39 coordinates for individuals that are commonly referred to as PCs in genetics [7], the convention
40 adopted here, but in other fields PCs instead denote what in genetics would be the projections of loci
41 onto eigenvectors, which are new independent coordinates for loci [8]. The direct ancestor of PCA
42 association is structured association, in which inferred ancestry (genetic cluster membership, often
43 corresponding with labels such as “European”, “African”, “Asian”, etc.) or admixture proportions of
44 these ancestries are used as regression covariates [9]. These models are deeply connected because
45 PCs map to ancestry empirically [10, 11] and theoretically [12–15], and they work as well as global
46 ancestry in association studies but are estimated more easily [6, 7, 10, 16]. Another approach closely
47 related to PCA is nonmetric multidimensional scaling [17]. PCs are also proposed for modeling
48 environment effects that are correlated to ancestry, for example, through geography [18–20]. The
49 strength of PCA is its simplicity, which as covariates can be readily included in more complex
50 models, such as haplotype association [21] and polygenic models [22]. However, PCA assumes that
51 the underlying relatedness space is low dimensional (or low rank), so it can be well modeled with a
52 small number of PCs, which may limit its applicability. PCA is known to be inadequate for family

53 data [7, 17, 23, 24], which is called “cryptic relatedness” when it is unknown to the researchers, but
54 no other troublesome cases have been confidently identified. Recent work has focused on developing
55 more scalable versions of the PCA algorithm [25–29]. PCA remains a popular and powerful approach
56 for association studies.

57 The other dominant association model under relatedness is the LMM, which includes a random
58 effect parameterized by the kinship matrix. Unlike PCA, LMM does not assume that relatedness is
59 low-dimensional, and explicitly models families via the kinship matrix. Early LMMs used kinship
60 matrices estimated from known pedigrees or using methods that captured recent relatedness only,
61 and modeled population structure (ancestry) as fixed effects [16, 17, 30]. Modern LMMs estimate
62 kinship from genotypes using a non-parametric estimator, often referred to as a genetic relationship
63 matrix, that captures the combined covariance due to family relatedness and ancestry [1, 31, 32].
64 Like PCA, LMM has also been proposed for modeling environment correlated to genetics [33, 34].
65 The classic LMM assumes a quantitative (continuous) complex trait, the focus of our work. Although
66 case-control (binary) traits and their underlying ascertainment are theoretically a challenge [35],
67 LMMs have been applied successfully to balanced case-control studies [1, 36] and simulations [24,
68 37, 38], and have been adapted for unbalanced case-control studies [39]. However, LMMs tend to
69 be considerably slower than PCA and other models, so much effort has focused on improving their
70 runtime and scalability [31, 36, 39–47].

71 An LMM variant that incorporates PCs as fixed covariates is tested thoroughly in our work.
72 Since PCs are the top eigenvectors of the same kinship matrix estimate used in modern LMMs [1, 19,
73 48, 49], then population structure is modeled twice in an LMM with PCs. However, some previous
74 work has found the apparent redundancy of an LMM with PCs beneficial [19, 24, 50], while others
75 did not [48, 51], and the approach continues to be used [52, 53] though not always [54]. (Recall
76 that early LMMs used kinship to model family relatedness only, so population structure had to be
77 modeled separately in those models, in practice as admixture fractions instead of PCs [16, 17, 30].)
78 The LMM with PCs (vs no PCs) is also believed to help better model loci that have experienced
79 selection [24, 33] and environment effects correlated with genetics [19].

80 LMM and PCA are closely related models [1, 19, 48, 49], so similar performance is expected

81 particularly under low-dimensional relatedness. Direct comparisons have yielded mixed results, with
 82 several studies finding superior performance for LMM, notably from papers promoting advances in
 83 LMMs, while many others report comparable performance (Table 1). No papers find that PCA
 84 outperforms LMM decisively, although PCA occasionally performs better in isolated and artificial
 85 cases or individual measures, often with unknown significance. Previous studies generally used either
 86 only simulated or only real genotypes, with only two studies using both. The simulated genotype
 87 studies, which tended to have low model dimensions and differentiation (F_{ST}), were more likely to
 88 report ties or mixed results (6/8), whereas real genotypes tended to clearly favor LMMs (9/11).
 89 Similarly, 10/12 papers with quantitative traits favor LMMs, whereas 6/9 papers with case-control
 90 traits gave ties or mixed results—the only factor we do not explore in this work. Additionally,
 91 although all previous evaluations measured type I error (or proxies such as genomic inflation factors

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

Publication	Sim. Genotypes			Real ^d	Trait ^e	Power	PCs (r)	Best
	Type ^a	K ^b	F_{ST} ^c					
Zhao et al. [16]				✓	Q	✓	8	LMM
Zhu and Yu [17]	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. [36]				✓	Both		2-100	LMM
Price et al. [24]	I, F	2	0.01		CC		1	Mixed
Wu et al. [37]	I, A	2-4	0.01		CC	✓	10	Mixed
Liu et al. [51]	S, A	2-3	R		Q	✓	10	Tie
Sul and Eskin [38]	I	2	0.01		CC		1	Tie
Tucker, Price, and Berger [50]	I	2	0.05	✓	Both	✓	5	Tie
Yang et al. [35]				✓	CC	✓	5	Tie
Song, Hao, and Storey [55]	S, A	2-3	R		Q		3	LMM
Loh et al. [47]				✓	Q	✓	10	LMM
Zhang and Pan [19]				✓	Q	✓	20-100	LMM
Liu et al. [56]				✓	Q	✓	3-6	LMM
Sul, Martin, and Eskin [57]				✓	Q		100	LMM
Loh et al. [58]				✓	Both	✓	20	LMM
Mbatchou et al. [53]				✓	Both		1	LMM
This work	A, T, F	10-243	≤ 0.25	✓	Q	✓	0-90	LMM

^aGenotype simulation types. I: Independent subpopulations; S: subpopulations (with parameters drawn from real data); A: Admixture; T: Subpopulation Tree; F: Family.

^bModel dimension (number of subpopulations or ancestries)

^cR: simulated parameters based on real data, F_{ST} not reported.

^dEvaluations using unmodified real genotypes.

^eQ: quantitative; CC: case-control.

[2] 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 [24, 38, 49, 59], 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 differentiation matching that of multiethnic human cohorts, 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 [60, 61], the Human Genome Diversity Panel (HGDP) [62–64], and Human Origins [65–68]. We simulate quantitative traits from two models: fixed effect sizes (FES) construct coefficients inverse to allele frequency, which matches real data [52, 69, 70] and corresponds to high pleiotropy and strong balancing selection [71] and strong negative selection [52, 70], which are appropriate assumptions for diseases; and random coefficients (RC), which are drawn independent of allele frequency, and corresponds to neutral traits [52, 71]. 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.

119 **2 Materials and Methods**

120 **2.1 The complex trait model and PCA and LMM approximations**

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

125
$$\mathbf{y} = \mathbf{1}\alpha + \mathbf{X}'\boldsymbol{\beta} + \mathbf{Z}'\boldsymbol{\eta} + \boldsymbol{\epsilon}, \quad (1)$$

126 where $\mathbf{1}$ is a length- n vector of ones, α is the scalar intercept coefficient, $\boldsymbol{\beta}$ is the length- m vector of
127 locus coefficients, \mathbf{Z} is a design matrix of environment effects and other covariates, $\boldsymbol{\eta}$ is the vector
128 of environment coefficients, $\boldsymbol{\epsilon}$ is a length- n vector of residuals, and the prime symbol ('') denotes
129 matrix transposition. The residuals follow $\epsilon_j \sim \text{Normal}(0, \sigma_\epsilon^2)$ independently per individual j , for
130 some σ_ϵ^2 .

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

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

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

134 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
135 $n \times r$ matrix of PCs, $\boldsymbol{\gamma}_r$ is the length- r vector of PC coefficients, \mathbf{s} is a length- n vector of random
136 effects, $\boldsymbol{\Phi}^T = (\varphi_{jk}^T)$ is the $n \times n$ kinship matrix conditioned on the ancestral population T , and σ_s^2
137 is a variance factor. Both models condition the regression of the focal locus i on an approximation
138 of the total polygenic effect $\mathbf{X}'\boldsymbol{\beta}$ with the same covariance structure, which is parameterized by the

139 kinship matrix. Under the kinship model, genotypes are random variables obeying

140
$$E[\mathbf{x}_i|T] = 2p_i^T \mathbf{1}, \quad \text{Cov}(\mathbf{x}_i|T) = 4p_i^T(1-p_i^T)\Phi^T, \quad (4)$$

141 where p_i^T is the ancestral allele frequency of locus i [1, 72–74]. Assuming independent loci, the

142 covariance of the polygenic effect is

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

143 which is readily modeled by the LMM random effect \mathbf{s} , where the difference in mean is absorbed by

144 the intercept. Alternatively, consider the eigendecomposition of the kinship matrix $\Phi^T = \mathbf{U}\Lambda\mathbf{U}'$

145 where \mathbf{U} is the $n \times n$ eigenvector matrix and Λ is the $n \times n$ diagonal matrix of eigenvalues. The

146 random effect can be written as

$$\mathbf{s} = \mathbf{U}\gamma_{\text{LMM}}, \quad \gamma_{\text{LMM}} \sim \text{Normal}(\mathbf{0}, 2\sigma_s^2\Lambda),$$

147 which follows from the affine transformation property of multivariate normal distributions. There-

148 fore, the PCA term $\mathbf{U}_r\gamma_r$ can be derived from the above equation under the additional assumption

149 that the kinship matrix has approximate rank r and the coefficients γ_r are fit without constraints.

150 In contrast, the LMM uses all eigenvectors, while effectively shrinking their coefficients γ_{LMM} as

151 all random effects models do, although these parameters are marginalized [1, 19, 48, 49]. PCA has

152 more parameters than LMM, so it may overfit more: ignoring the shared terms in Eqs. (2) and (3),

153 PCA fits r parameters (length of γ), whereas LMMs fit only one (σ_s^2).

154 In practice, the kinship matrix used for PCA and LMM is estimated with variations of a method-

155 of-moments formula applied to standardized genotypes \mathbf{X}_S , which is derived from Eq. (4):

156
$$\mathbf{X}_S = \left(\frac{x_{ij} - 2\hat{p}_i^T}{\sqrt{4\hat{p}_i^T(1-\hat{p}_i^T)}} \right), \quad \hat{\Phi}^T = \frac{1}{m}\mathbf{X}_S'\mathbf{X}_S, \quad (5)$$

157 where the unknown p_i^T is estimated by $\hat{p}_i^T = \frac{1}{2n} \sum_{j=1}^n x_{ij}$ [5, 31, 35, 36, 39, 43, 45, 47, 57]. However,

158 this kinship estimator has a complex bias that differs for every individual pair, which arises due
159 to the use of this estimated \hat{p}_i^T [32, 75]. Nevertheless, in PCA and LMM these biased estimates
160 perform as well as unbiased ones [76].

161 We selected fast and robust software implementing the basic PCA and LMM models. PCA
162 association was performed with `plink2` [77]. The quantitative trait association model is a linear
163 regression with covariates, evaluated using the t-test. PCs were calculated with `plink2`, which equal
164 the top eigenvectors of Eq. (5) after removing loci with minor allele frequency MAF < 0.1.

165 LMM association was performed using GCTA [35, 43]. Its kinship estimator equals Eq. (5).
166 PCs were calculated using GCTA from its kinship estimate. Association significance is evaluated
167 with a score test. In the small simulation only, GCTA with large numbers of PCs had convergence
168 and singularity errors in some replicates, which were treated as missing data.

169 2.2 Simulations

170 Every simulation was replicated 50 times, drawing anew all genotypes (except for real datasets)
171 and traits. Below we use the notation f_A^B for the inbreeding coefficient of a subpopulation A from
172 another subpopulation B ancestral to A . In the special case of the *total* inbreeding of A , f_A^T , T is
173 an overall ancestral population, which is ancestral to every individual under consideration, such as
174 the most recent common ancestor (MRCA) population.

175 2.2.1 Genotype simulation from the admixture model

176 The basic admixture model is as described previously [32] and is implemented in the R package
177 `bnpstd`. Both Large and Family simulations have $n = 1,000$ individuals, while Small has $n =$
178 100. The number of loci is $m = 100,000$. Individuals are admixed from $K = 10$ intermediate
179 subpopulations, or ancestries. Each subpopulation S_u ($u \in \{1, \dots, K\}$) is at coordinate u and has an
180 inbreeding coefficient $f_{S_u}^T = u\tau$ for some τ . Ancestry proportions q_{ju} for individual j and S_u arise
181 from a random walk with spread σ on the 1D geography, and τ and σ are fit to give $F_{ST} = 0.1$ and
182 mean kinship $\bar{\theta}^T = 0.5F_{ST}$ for the admixed individuals [32]. Random ancestral allele frequencies
183 p_i^T , subpopulation allele frequencies $p_i^{S_u}$, individual-specific allele frequencies π_{ij} , and genotypes x_{ij}

184 are drawn from this hierarchical model:

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

185 where this Beta is the Balding-Nichols distribution [78] with mean p_i^T and variance $p_i^T (1 - p_i^T) f_{S_u}^T$.
186 Fixed loci (i where $x_{ij} = 0$ for all j , or $x_{ij} = 2$ for all j) are drawn again from the model, starting
187 from p_i^T , iterating until no loci are fixed. Each replicate draws a genotypes starting from p_i^T .

188 As a brief aside, we prove that global ancestry proportions as covariates is equivalent in expec-
189 tation to using PCs under the admixture model. Note that the latent space of \mathbf{X} , which is the
190 subspace to which the data is constrained by the admixture model, is given by (π_{ij}) , which has K
191 dimensions (number of columns of $\mathbf{Q} = (q_{ju})$), so the top K PCs span this space. Since associations
192 include an intercept term ($\mathbf{1}\alpha$ in Eq. (2)), estimated PCs are orthogonal to $\mathbf{1}$ (note $\hat{\Phi}^T \mathbf{1} = \mathbf{0}$ because
193 $\mathbf{X}_S \mathbf{1} = \mathbf{0}$), and the sum of rows of \mathbf{Q} sums to one, then only $K - 1$ PCs plus the intercept are
194 needed to span the latent space of this admixture model.

195 2.2.2 Genotype simulation from random admixed families

196 We simulated a pedigree with admixed founders, no close relative pairings, assortative mating based
197 on a 1D geography (to preserve admixture structure), random family sizes, and arbitrary numbers
198 of generations (20 here). This simulation is implemented in the R package `simfam`. Generations
199 are drawn iteratively. Generation 1 has $n = 1000$ individuals from the above admixture simulation
200 ordered by their 1D geography. Local kinship measures pedigree relatedness; in the first generation,
201 everybody is locally unrelated and outbred. Individuals are randomly assigned sex. In the next
202 generation, individuals are paired iteratively, removing random males from the pool of available
203 males and pairing them with the nearest available female with local kinship $< 1/4^3$ (stay unpaired

204 if there are no matches), until there are no more available males or females. Let $n = 1000$ be the
 205 desired population size, $n_m = 1$ the minimum number of children per family and n_f the number of
 206 families (paired parents) in the current generation, then the number of additional children (beyond
 207 the minimum) is drawn from $\text{Poisson}(n/n_f - n_m)$. Let δ be the difference between desired and
 208 current population sizes. If $\delta > 0$, then δ random families are incremented by 1. If $\delta < 0$, then $|\delta|$
 209 random families with at least $n_m + 1$ children are decremented by 1. If $|\delta|$ exceeds the number of
 210 families, all families are incremented or decremented as needed and the process is iterated. Children
 211 are assigned sex randomly, and are reordered by the average coordinate of their parents. Children
 212 draw alleles from their parents independently per locus. A new random pedigree is drawn for each
 213 replicate, as well as new founder genotypes from the admixture model.

214 2.2.3 Genotype simulation from a subpopulation tree model

215 This model draws subpopulations allele frequencies from a hierarchical model parameterized by a
 216 tree, which is also implemented in `bnpst` and relies on the R package `ape` for general tree data struc-
 217 tures and methods [79]. The ancestral population T is the root, and each node is a subpopulation
 218 S_w indexed arbitrarily. Each edge between S_w and its parent population P_w has an inbreeding coef-
 219 ficient $f_{S_w}^{P_w}$. p_i^T are drawn from a given distribution, which is constructed to mimic each real dataset
 220 in Appendix A. Given the allele frequencies $p_i^{P_w}$ of the parent population, S_w 's allele frequencies
 221 are drawn from:

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

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

225 **2.2.4 Fitting subpopulation tree to real data**

226 We developed new methods to fit trees to real data based on unbiased kinship estimates from
227 `popkin`, implemented in `bnpstd`. A tree with given inbreeding coefficients $f_{S_w}^{P_w}$ for its edges (between
228 subpopulation S_w and its parent P_w) gives rise to a coancestry matrix ϑ_{uv}^T for a subpopulation pair
229 (S_u, S_v) , and the goal is to recover these edge inbreeding coefficients from coancestry estimates.
230 Coancestry values are total inbreeding coefficients of the MRCA population of each subpopulation
231 pair. Therefore, we calculate $f_{S_w}^T$ for every S_w recursively from the root as follows. Nodes with
232 parent $P_w = T$ are already as desired. Given $f_{P_w}^T$, the desired $f_{S_w}^T$ is calculated via the “additive
233 edge” δ_w [32]:

$$\text{234} \quad f_{S_w}^T = f_{P_w}^T + \delta_w, \quad \delta_w = f_{S_w}^{P_w} (1 - f_{P_w}^T). \quad (6)$$

235 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
236 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
237 ancestor nodes,

$$\text{238} \quad \vartheta_{uv}^T = \sum_w \delta_w I_w(u, v), \quad (7)$$

239 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.

240 Note that $I_w(u, v)$ reflects tree topology and δ_w edge values.

241 To estimate population-level coancestry, first kinship ($\hat{\varphi}_{jk}^T$) is estimated using `popkin` [32]. In-
242 dividual coancestry ($\hat{\theta}_{jk}^T$) is estimated from kinship using

$$\text{243} \quad \hat{\theta}_{jk}^T = \begin{cases} \hat{\varphi}_{jk}^T & \text{if } k \neq j, \\ \hat{f}_j^T = 2\hat{\varphi}_{jj}^T - 1 & \text{if } k = j. \end{cases} \quad (8)$$

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

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

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

246 arithmetic mean [80], with distance function $d(S_u, S_v) = \max \left\{ \hat{\vartheta}_{uv}^T \right\} - \hat{\vartheta}_{uv}^T$, which succeeds due to
247 the monotonic relationship between node depth and coancestry (Eq. (7)). This algorithm recovers
248 the true topology from the true coancestry values, and performs well for estimates from genotypes.

249 To estimate tree edge lengths, first δ_w are estimated from $\hat{\vartheta}_{uv}^T$ and the topology using Eq. (7) and
250 non-negative least squares linear regression [81] (implemented in `nnls` [82]) to yield non-negative
251 δ_w , and $f_{S_w}^{P_w}$ are calculated from δ_w by reversing Eq. (6). To account for small biases in coancestry
252 estimation, an intercept term δ_0 is included ($I_0(u, v) = 1$ for all u, v), and when converting δ_w to
253 $f_{S_w}^{P_w}$, δ_0 is treated as an additional edge to the root, but is ignored when drawing allele frequencies
254 from the tree.

255 2.2.5 Trait Simulation

256 Traits are simulated from the quantitative trait model of Eq. (1), with novel bias corrections for
257 simulating the desired heritability from real data relying on the unbiased kinship estimator `popkin`
258 [32]. This simulation is implemented in the R package `simtrait`. All simulations have a fixed
259 narrow-sense heritability of h^2 , a variance proportion due to environment effects σ_η^2 , and residuals
260 are drawn from $\epsilon_j \sim \text{Normal}(0, \sigma_\epsilon^2)$ with $\sigma_\epsilon^2 = 1 - h^2 - \sigma_\eta^2$. The number of causal loci m_1 , which
261 determines the average coefficient size, is chosen with the heuristic formula $m_1 = \text{round}(nh^2/8)$,
262 which empirically balances power well with varying n and h^2 . The set of causal loci C is drawn
263 anew for each replicate, from loci with MAF ≥ 0.01 to avoid rare causal variants, which are not
264 discoverable by PCA or LMM at the sample sizes we considered. Letting $v_i^T = p_i^T (1 - p_i^T)$, the
265 effect size of locus i equals $2v_i^T \beta_i^2$, its contribution of the trait variance [83]. Under the *fixed effect*
266 *sizes* (FES) model, initial causal coefficients are

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

267 for known p_i^T ; otherwise v_i^T is replaced by the unbiased estimator [32] $\hat{v}_i^T = \hat{p}_i^T (1 - \hat{p}_i^T) / (1 - \bar{\varphi}^T)$,
268 where $\bar{\varphi}^T$ is the mean kinship estimated with `popkin`. Each causal locus is multiplied by -1 with
269 probability 0.5. Alternatively, under the *random coefficients* (RC) model, initial causal coefficients

270 are drawn independently from $\beta_i \sim \text{Normal}(0, 1)$. For both models, the initial genetic variance is
 271 $\sigma_0^2 = \sum_{i \in C} 2v_i^T \beta_i^2$, replacing v_i^T with \hat{v}_i^T for unknown p_i^T (so σ_0^2 is an unbiased estimate), so we
 272 multiply every initial β_i by $\frac{h}{\sigma_0}$ to have the desired heritability. Lastly, for known p_i^T , the intercept
 273 coefficient is $\alpha = -\sum_{i \in C} 2p_i^T \beta_i$. When p_i^T are unknown, \hat{p}_i^T should not replace p_i^T since that distorts
 274 the trait covariance (for the same reason the standard kinship estimator in Eq. (5) is biased), which
 275 is avoided with

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

276 Simulations optionally included multiple environment group effects, similarly to previous models
 277 [19, 34], as follows. Each independent environment i has predefined groups, and each group g has
 278 random coefficients drawn independent from $\eta_{gi} \sim \text{Normal}(0, \sigma_{\eta i}^2)$ where $\sigma_{\eta i}^2$ is a specified variance
 279 proportion for environment i . \mathbf{Z} has individuals along columns and environment-groups along rows,
 280 and it contains indicator variables: 1 if the individual belongs to the environment-group, 0 otherwise.

281 We performed trait simulations with the following variance parameters (Table 2): *high heritabil-*
 282 *ity* used $h^2 = 0.8$ and no environment effects; *low heritability* used $h^2 = 0.3$ and no environment
 283 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
 284 genotype datasets, the groups are the continental (environment 1) and fine-grained (environment 2)
 285 subpopulation labels given (see next subsection). For simulated genotypes, we created these labels
 286 by grouping by the index j (geographical coordinate) of each simulated individual, assigning group
 287 $g = \text{ceiling}(jk_i/n)$ where k_i is the number of groups in environment i , and we selected $k_1 = 5$ and
 288 $k_2 = 25$ 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

289 **2.3 Real human genotype datasets**

290 The three datasets were processed as before [75] (summarized below), except with an additional filter
291 so loci are in approximate linkage equilibrium and rare variants are removed. All processing was
292 performed with `plink2` [77], and analysis was uniquely enabled by the R packages `BEDMatrix` [84]
293 and `genio`. Each dataset groups individuals in a two-level hierarchy: continental and fine-grained
294 subpopulations. Final dataset sizes are in Table 3.

295 We obtained the full (including non-public) Human Origins by contacting the authors and
296 agreeing to their usage restrictions. The Pacific data [68] was obtained separately from the rest
297 [66, 67], and datasets were merged using the intersection of loci. We removed ancient individuals,
298 and individuals from singleton and non-native subpopulations. Non-autosomal loci were removed.
299 Our analysis of both the whole-genome sequencing (WGS) version of HGDP [64] and the high-
300 coverage NYGC version of 1000 Genomes [85] was restricted to autosomal biallelic SNP loci with
301 filter “PASS”.

302 Since our evaluations assume uncorrelated loci, we filtered each real dataset with `plink2` using
303 parameters “`--indep-pairwise 1000kb 0.3`”, which iteratively removes loci that have a greater
304 than 0.3 squared correlation coefficient with another locus that is within 1000kb, stopping until no
305 such loci remain. Since all real datasets have numerous rare variants, while PCA and LMM are not
306 able to detect associations involving rare variants, we removed all loci with $\text{MAF} < 0.01$. 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

^aFor 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.

^cModel parameter for simulations, estimated value on real datasets.

307 only HGDP had loci with over 10% missingness removed, as they were otherwise 17% of remaining
 308 loci (for Human Origins and 1000 Genomes they were under 1% of loci so they were not removed).
 309 Kinship matrix rank and eigenvalues were calculated from `popkin` kinship estimates. Eigenvalues
 310 were assigned p-values with `twstats` of the Eigensoft package [7], and kinship matrix rank was
 311 estimated as the largest number of consecutive eigenvalue from the start that all satisfy $p < 0.01$
 312 (p-values did not increase monotonically). For the evaluation with close relatives removed, each
 313 dataset was filtered with `plink2` with option “`--king-cutoff`” with cutoff $0.02209709 (= 2^{-11/2})$
 314 for removing up to 4th degree relatives using KING-robust [86], and $\text{MAF} < 0.01$ filter is reapplied
 315 (Table S1).

316 2.4 Evaluation of performance

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

321 P-values for continuous test statistics have a uniform distribution when the null hypothesis
 322 holds, a crucial assumption for type I error and FDR control [87, 88]. We use the Signed Root
 323 Mean Square Deviation (SRMSD_p) to measure the difference between the observed null p-value
 324 quantiles and the expected uniform quantiles:

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

325 where $m_0 = m - m_1$ is the number of null (non-causal) loci, here i indexes null loci only, $p_{(i)}$ is
 326 the i th ordered null p-value, $u_i = (i - 0.5)/m_0$ is its expectation, p_{median} is the median observed
 327 null p-value, $u_{\text{median}} = \frac{1}{2}$ is its expectation, and sgn is the sign function (1 if $u_{\text{median}} \geq p_{\text{median}}$,
 328 -1 otherwise). Thus, $\text{SRMSD}_p = 0$ corresponds to calibrated p-values, $\text{SRMSD}_p > 0$ indicate anti-
 329 conservative p-values, and $\text{SRMSD}_p < 0$ are conservative p-values. The maximum SRMSD_p is
 330 achieved when all p-values are zero (the limit of anti-conservative p-values), which for infinite loci

331 approaches

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

332 The same value with a negative sign occurs for all p-values of 1.

333 Precision and recall are standard performance measures for binary classifiers that do not require
334 calibrated p-values [89]. Given the total numbers of true positives (TP), false positives (FP) and
335 false negatives (FN) at some threshold or parameter t , precision and recall are

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

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

341 3 Results

342 3.1 Overview of evaluations

343 We use three real genotype datasets and simulated genotypes from six population structure scenarios
344 to cover various features of interest (Table 3). We introduce them in sets of three, as they appear
345 in the rest of our results. Population kinship matrices, which combine population and family
346 relatedness, are estimated without bias using `popkin` [32] (Fig. 1). The first set of three simulated
347 genotypes are based on an admixture model with 10 ancestries (Fig. 1A) [14, 32, 90]. The “large”
348 version (1000 individuals) illustrates asymptotic performance, while the “small” simulation (100
349 individuals) illustrates model overfitting. The “family” simulation has admixed founders and draws
350 a 20-generation random pedigree with assortative mating, resulting in a complex joint family and
351 ancestry structure in the last generation (Fig. 1B). The second set of three are the real human

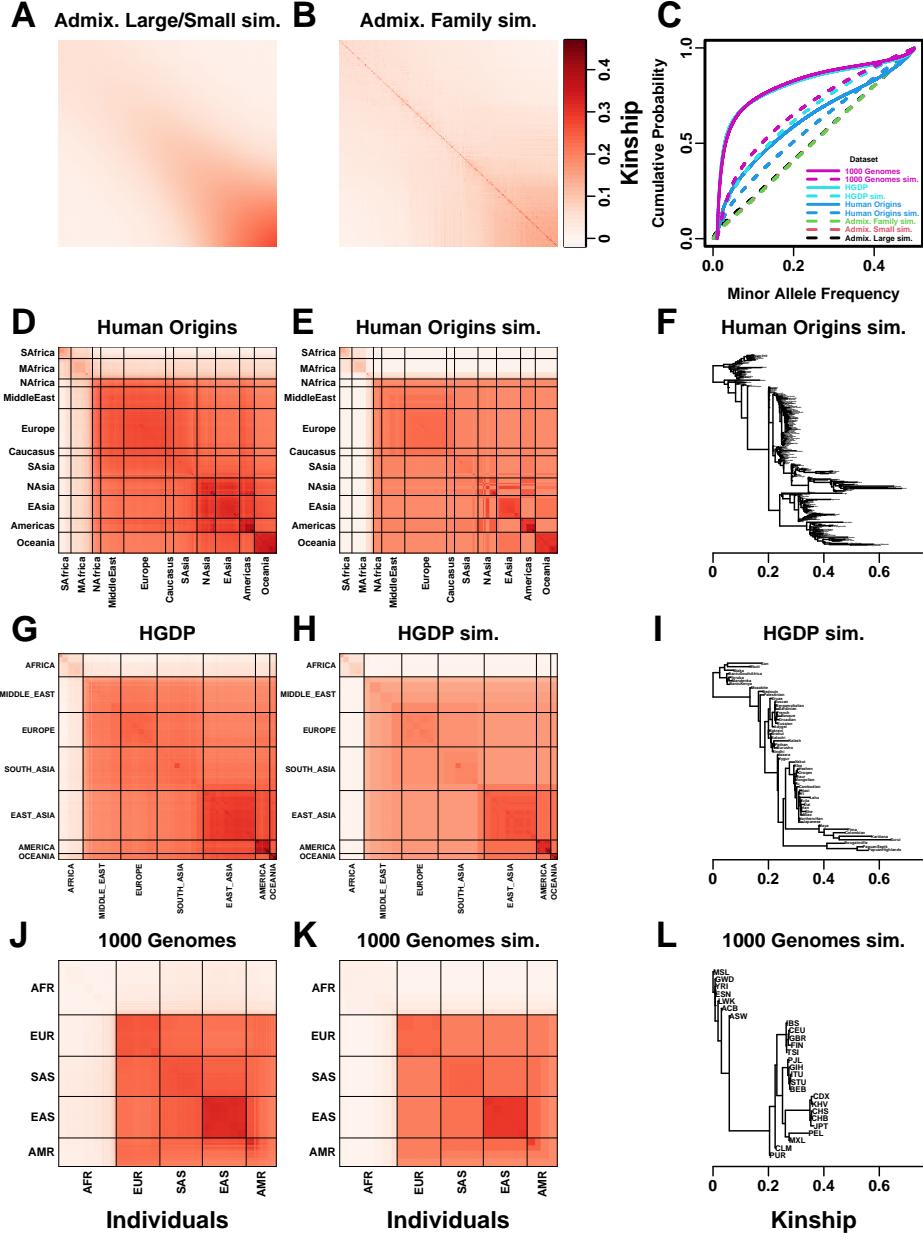


Figure 1: **Population structures of simulated and real human genotype datasets.** First two columns are population kinship matrices as heatmaps: individuals along x- and y-axis, kinship as color. Diagonal shows inbreeding values. **A.** Admixture scenario for both Large and Small simulations. **B.**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

352 datasets representing global human diversity: Human Origins (Fig. 1D), HGDP (Fig. 1G), and
353 1000 Genomes (Fig. 1J), which are enriched for small minor allele frequencies even after $\text{MAF} < 1\%$
354 filter (Fig. 1C). Last are subpopulation tree simulations (Fig. 1F,I,L) fit to the kinship (Fig. 1E,H,K)
355 and MAF (Fig. 1C) of each real human dataset, which by design do not have family structure.

356 All traits in this work are simulated. We repeated all evaluations on two additive quantitative
357 trait models, *fixed effect sizes* (FES) and *random coefficients* (RC), which differ in how causal coef-
358 ficients are constructed. The FES model captures the rough inverse relationship between coefficient
359 and minor allele frequency that arises under strong negative and balancing selection and has been
360 observed in numerous diseases and other traits [52, 69–71], so it is the focus of our results. The
361 RC model draws coefficients independent of allele frequency, corresponding to neutral traits [52,
362 71], which results in a wider effect size distribution that reduces association power and effective
363 polygenicity compared to FES.

364 We evaluate using two complementary measures: (1) SRMSD_p (p-value signed root mean square
365 deviation) measures p-value calibration (closer to zero is better), and (2) AUC_{PR} (precision-recall
366 area under the curve) measures causal locus classification performance (higher is better; Fig. 2).
367 SRMSD_p is a more robust alternative to the common inflation factor λ and type I error control
368 measures; there is a correspondence between λ and SRMSD_p , with $\text{SRMSD}_p > 0.01$ giving $\lambda > 1.06$
369 (Fig. S1) and thus evidence of miscalibration close to the rule of thumb of $\lambda > 1.05$ [24]. There
370 is also a monotonic correspondence between SRMSD_p and type I error rate (Fig. S2). AUC_{PR} has
371 been used to evaluate association models [91], and reflects calibrated statistical power (Fig. S3)
372 while being robust to miscalibrated models (Appendix B).

373 Both PCA and LMM are evaluated in each replicate dataset including a number of PCs r
374 between 0 and 90 as fixed covariates. In terms of p-value calibration, for PCA the best number of
375 PCs r (minimizing mean $|\text{SRMSD}_p|$ over replicates) is typically large across all datasets (Table 4),
376 although much smaller r values often performed as well (shown in following sections). Most cases
377 have a mean $|\text{SRMSD}_p| < 0.01$, whose p-values are effectively calibrated. However, PCA is often
378 miscalibrated on the family simulation and real datasets (Table 4). In contrast, for LMM, $r = 0$ (no
379 PCs) is always best, and is always calibrated. Comparing LMM with $r = 0$ to PCA with its best

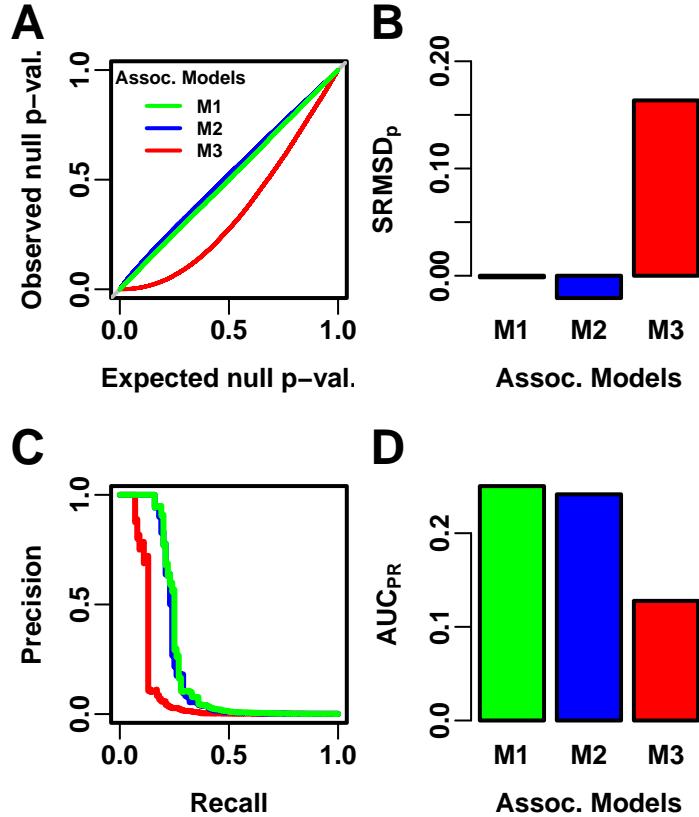


Figure 2: **Illustration of evaluation measures.** Three archetypal models illustrate our complementary measures: M1 is ideal, M2 overfits slightly, M3 is naive. **A.** QQ plot of p-values of “null” (non-causal) loci. M1 has desired uniform p-values, M2/M3 are miscalibrated. **B.** SRMSD_p (p-value Signed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

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

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

^aFES: Fixed Effect Sizes, RC: Random Coefficients.

^bCalibrated: whether mean $|\text{SRMSD}_P| < 0.01$.

^cValue of r (number of PCs) with minimum mean $|\text{SRMSD}_P|$ or maximum mean AUC_{PR}.

^dWilcoxon paired 1-tailed test of distributions ($|\text{SRMSD}_P|$ or AUC_{PR}) between models in header. Asterisk marks significant value using Bonferroni threshold ($p < \alpha/n_{\text{tests}}$ with $\alpha = 0.01$ and $n_{\text{tests}} = 72$ is the number of tests in this table).

^eTie if no significant difference using Bonferroni threshold.

380 r , LMM always has significantly smaller $|\text{SRMSD}_p|$ than PCA or is statistically tied. For AUC_{PR}
381 and PCA, the best r is always smaller than the best r for $|\text{SRMSD}_p|$, so there is often a tradeoff
382 between calibrated p-values versus classification performance. For LMM there is no tradeoff, as
383 $r = 0$ often has the best mean AUC_{PR} , and otherwise is not significantly different from the best
384 r . Lastly, LMM with $r = 0$ always has significantly greater or statistically tied AUC_{PR} than PCA
385 with its best r .

386 3.2 Evaluations in admixture simulations

387 Now we look more closely at results per dataset. The complete SRMSD_p and AUC_{PR} distributions
388 for the admixture simulations and FES traits are in Fig. 3. RC traits gave qualitatively similar
389 results (Fig. S4).

390 In the large admixture simulation, the SRMSD_p of PCA is largest when $r = 0$ (no PCs) and
391 decreases rapidly to near zero at $r = 3$, where it stays for up to $r = 90$ (Fig. 3A). Thus, PCA
392 has calibrated p-values for $r \geq 3$, smaller than the theoretical optimum for this simulation of
393 $r = K - 1 = 9$. In contrast, the SRMSD_p for LMM starts near zero for $r = 0$, but becomes negative
394 as r increases (p-values are conservative). The AUC_{PR} distribution of PCA is similarly worst at
395 $r = 0$, increases rapidly and peaks at $r = 3$, then decreases slowly for $r > 3$, while the AUC_{PR}
396 distribution for LMM starts near its maximum at $r = 0$ and decreases with r . Although the AUC_{PR}
397 distributions for LMM and PCA overlap considerably at each r , LMM with $r = 0$ has significantly
398 greater AUC_{PR} values than PCA with $r = 3$ (Table 4). However, qualitatively PCA performs nearly
399 as well as LMM in this simulation.

400 The observed robustness to large r led us to consider smaller sample sizes. A model with large
401 numbers of parameters r should overfit more as r approaches the sample size n . Rather than increase
402 r beyond 90, we reduce individuals to $n = 100$, which is small for typical association studies but
403 may occur in studies of rare diseases, pilot studies, or other constraints. To compensate for the loss
404 of power due to reducing n , we also reduce the number of causal loci (see Materials and Methods),
405 which increases per-locus effect sizes. We found a large decrease in performance for both models as
406 r increases, and best performance for $r = 1$ for PCA and $r = 0$ for LMM (Fig. 3B). Remarkably,

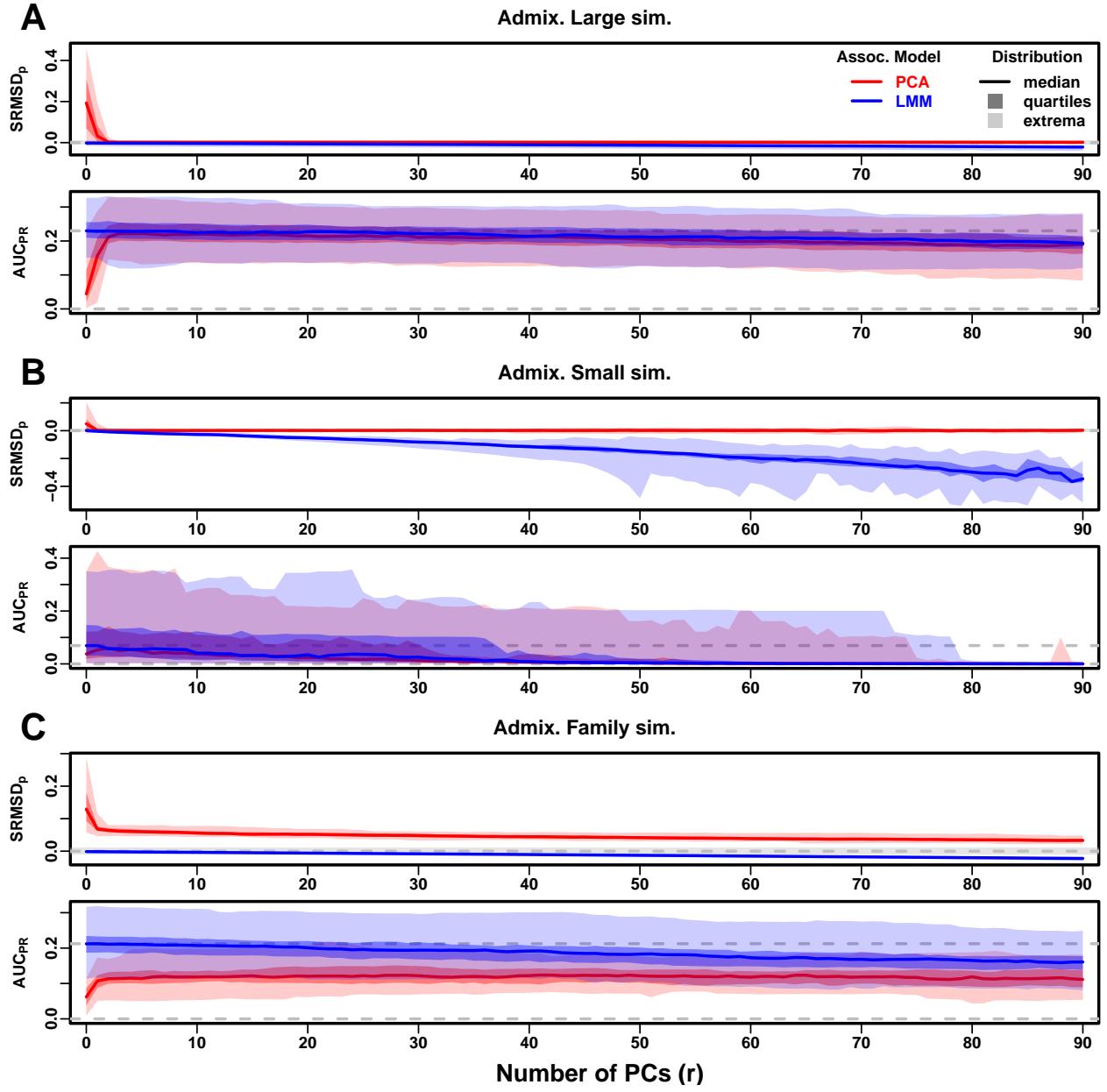


Figure 3: **Evaluations in admixture simulations.** Traits simulated from FES model with high heritability. PCA and LMM models have varying number of PCs ($r \in \{0, \dots, 90\}$ on x-axis), with the distributions (y-axis) of SRMSD_p (top subpanel) and AUC_{PR} (bottom subpanel) for 50 replicates. Best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

407 LMM attains much larger negative SRMSD_p values than in our other evaluations. LMM with $r = 0$
408 is significantly better than PCA ($r = 1$ to 4) in both measures (Table 4), but qualitatively the
409 difference is negligible.

410 The family simulation adds a 20-generation random family to our large admixture simulation.
411 Only the last generation is studied for association, which contains numerous siblings, first cousins,
412 etc., with the initial admixture structure preserved by geographically biased mating. Our evaluation
413 reveals a sizable gap in both measures between LMM and PCA across all r (Fig. 3C). LMM again
414 performs best with $r = 0$ and achieves mean $|\text{SRMSD}_p| < 0.01$. However, PCA does not achieve
415 mean $|\text{SRMSD}_p| < 0.01$ at any r , and its best mean AUC_{PR} is considerably worse than that of
416 LMM. Thus, LMM is conclusively superior to PCA, and the only calibrated model, when there is
417 family structure.

418 3.3 Evaluations in real human genotype datasets

419 Next we repeat our evaluations with real human genotype data, which differs from our simulations in
420 allele frequency distributions and more complex population structures with greater differentiation,
421 numerous correlated subpopulations, and potential cryptic family relatedness.

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

427 HGDP has the fewest individuals among real datasets, but compared to Human Origins contains
428 more loci and low-frequency variants. Performance (Fig. 4B) again most resembled the family sim-
429 ulations. In particular, LMM with $r = 0$ achieves mean $|\text{SRMSD}_p| < 0.01$ (p-values are calibrated),
430 while PCA does not, and there is a sizable AUC_{PR} gap between LMM and PCA. Maximum AUC_{PR}
431 values were lowest in HGDP compared to the two other real datasets.

432 1000 Genomes has the fewest subpopulations but largest number of individuals per subpopula-
433 tion. Thus, although this dataset has the simplest subpopulation structure among the real datasets,

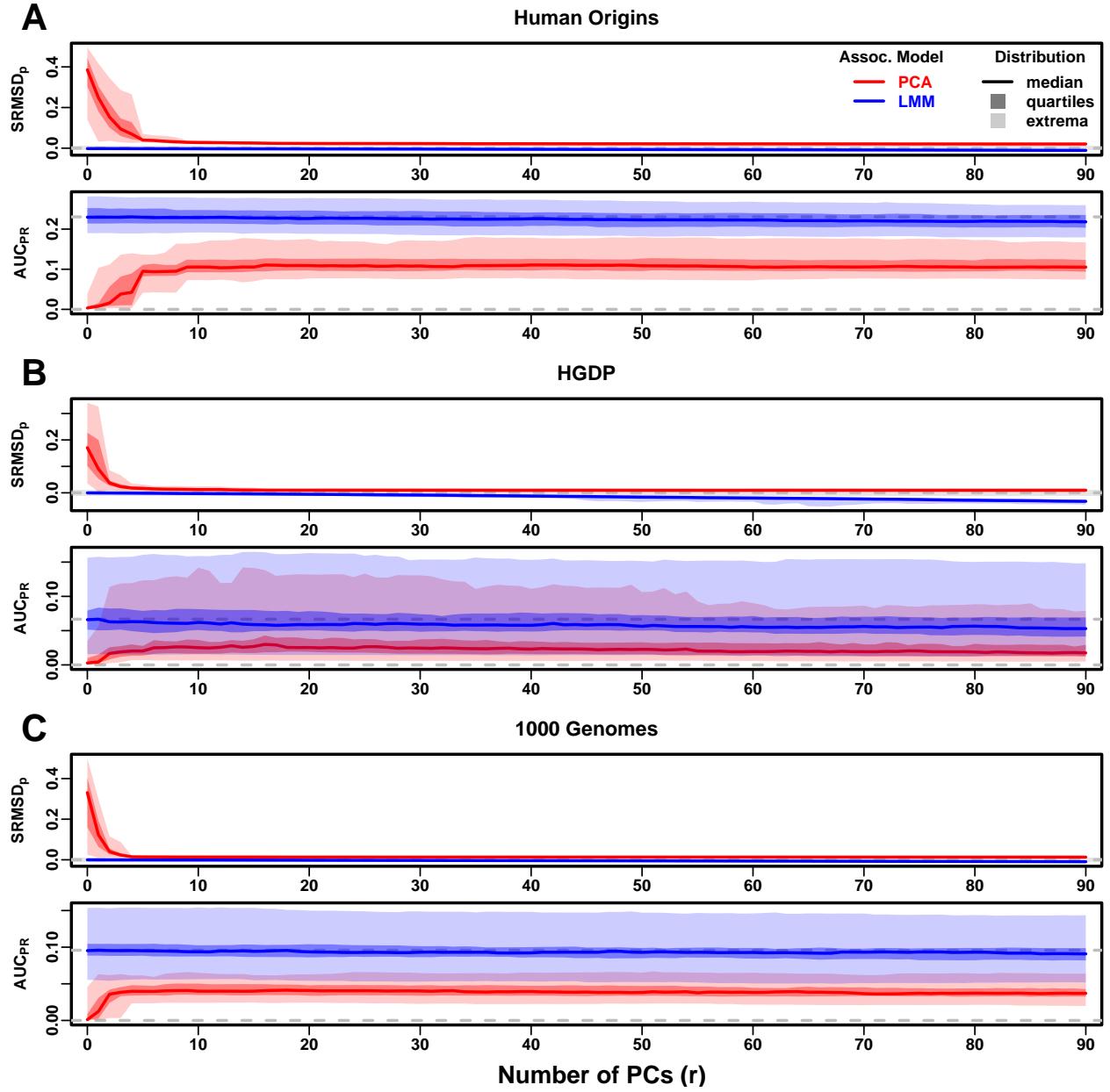


Figure 4: **Evaluations in real human genotype datasets.** Traits simulated from FES model with high heritability. Same setup as Fig. 3, see that for details. These datasets strongly favor LMM with no PCs over PCA, with distributions that most resemble the family simulation. **A.** Human Origins. **B.**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

434 we find SRMSD_p and AUC_{PR} distributions (Fig. 4C) that again most resemble our earlier family
435 simulation, with mean |SRMSD_p| < 0.01 for LMM only and large AUC_{PR} gaps between LMM and
436 PCA.

437 Our results are qualitatively different for RC traits, which had smaller AUC_{PR} gaps between
438 LMM and PCA (Fig. S5). Maximum AUC_{PR} were smaller in RC compared to FES in Human Origins
439 and 1000 Genomes, suggesting lower power for RC traits across association models. Nevertheless,
440 LMM with $r = 0$ was significantly better than PCA for all measures in the real datasets and RC
441 traits (Table 4).

442 3.4 Evaluations in subpopulation tree simulations fit to human data

443 To better understand which features of the real datasets lead to the large differences in performance
444 between LMM and PCA, we carried out subpopulation tree simulations. Human subpopulations
445 are related roughly by trees, which induce the strongest correlations, so we fit trees to each real
446 dataset and tested if data simulated from these complex tree structures could recapitulate our
447 previous results (Fig. 1). These tree simulations also feature non-uniform ancestral allele frequency
448 distributions, which recapitulated some of the skew for smaller minor allele frequencies of the real
449 datasets (Fig. 1C). The SRMSD_p and AUC_{PR} distributions for these tree simulations (Fig. 5)
450 resembled our admixture simulation more than either the family simulation (Fig. 3) or real data
451 results (Fig. 4). Both LMM with $r = 0$ and PCA (various r) achieve mean |SRMSD_p| < 0.01
452 (Table 4). The AUC_{PR} distributions of both LMM and PCA track closely as r is varied, although
453 there is a small gap resulting in LMM ($r = 0$) besting PCA in all three simulations. The results
454 are qualitatively similar for RC traits (Fig. S6 and Table 4). Overall, these subpopulation tree
455 simulations do not recapitulate the large LMM advantage over PCA observed on the real data.

456 3.5 Numerous distant relatives explain poor PCA performance in real data

457 In principle, PCA performance should be determined by the dimension of relatedness, or kinship
458 matrix rank, since PCA is a low-dimensional model whereas LMM can model high-dimensional
459 relatedness without overfitting. We used the Tracy-Widom test [7] with $p < 0.01$ to estimate

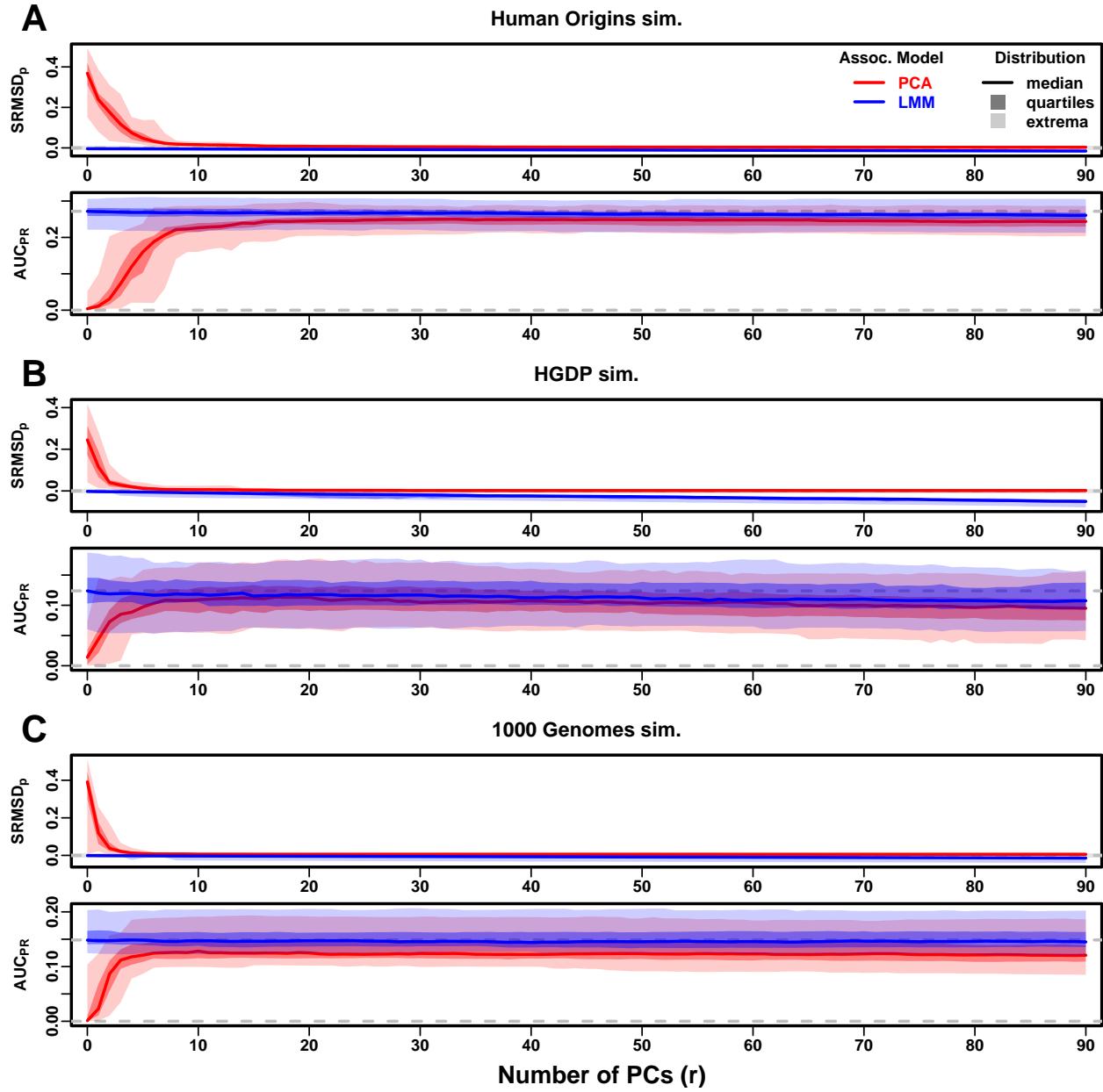


Figure 5: **Evaluations in subpopulation tree simulations fit to human data.** Traits simulated from FES model with high heritability. Same setup as Fig. 3, see that for details. These tree simulations, which exclude family structure by design, do not explain the large gaps in LMM-PCA performance observed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

460 kinship matrix rank as the number of significant PCs (Fig. S7A). The true rank of our simulations
461 is slightly underestimated (Table 3), but we confirm that the family simulation has the greatest rank,
462 and real datasets have greater estimates than their respective subpopulation tree simulations, which
463 confirms our hypothesis to some extent. However, estimated ranks do not separate real datasets from
464 tree simulations, as required to predict the observed PCA performance. Moreover, the HGDP and
465 1000 Genomes rank estimates are 45 and 61, respectively, yet PCA performed poorly for all $r \leq 90$
466 numbers of PCs (Fig. 4). The top eigenvalue explained a proportion of variance proportional to F_{ST}
467 (Table 3), but the rest of the top 10 eigenvalues show no clear differences between datasets, except
468 the small simulation had larger variances explained per eigenvalue (expected since it has fewer
469 eigenvalues; Fig. S7C). Comparing cumulative variance explained versus rank fraction across all
470 eigenvalues, all datasets increase from their starting point almost linearly until they reach 1, except
471 the family simulation has much greater variance explained by mid-rank eigenvalues (Fig. S7B). We
472 also calculated the number of PCs that are significantly associated with the trait, and observed
473 similar results, namely that while the family simulation has more significant PCs than the non-
474 family admixture simulations, the real datasets and their tree simulated counterparts have similar
475 numbers of significant PCs (Fig. S8). Overall, there is no separation between real datasets (where
476 PCA performed poorly) and subpopulation tree simulations (where PCA performed relatively well)
477 in terms of their eigenvalues or kinship matrix rank estimates.

478 Local kinship, which is recent relatedness due to family structure excluding population structure,
479 is the presumed cause of the LMM to PCA performance gap observed in real datasets but not their
480 subpopulation tree simulation counterparts. Instead of inferring local kinship through increased
481 kinship matrix rank, as attempted in the last paragraph, now we measure it directly using the
482 KING-robust estimator [86]. We observe more large local kinship in the real datasets and the family
483 simulation compared to the other simulations (Fig. 6). However, for real data this distribution
484 depends on the subpopulation structure, since locally related pairs are most likely in the same
485 subpopulation. Therefore, the only comparable curve to each real dataset is their corresponding
486 subpopulation tree simulation, which matches subpopulation structure. In all real datasets we
487 identified highly related individual pairs with kinship above the 4th degree relative threshold of

488 0.022 [86, 92]. However, these highly related pairs are vastly outnumbered by more distant pairs
 489 with evident non-zero local kinship as compared to the extreme tree simulation values.

490 To try to improve PCA performance, we followed the standard practice of removing 4th degree
 491 relatives, which reduced sample sizes between 5% and 10% (Table S1). Only $r = 0$ for LMM
 492 and $r = 20$ for PCA were tested, as these performed well in our earlier evaluation, and only FES
 493 traits were tested because they previously displayed the large PCA-LMM performance gap. LMM
 494 significantly outperforms PCA in all these cases (Wilcoxon paired 1-tailed $p < 0.01$; Fig. 7). Notably,
 495 PCA still had miscalibrated p-values two of the three real datasets ($|SRMSD_p| > 0.01$), the only
 496 marginally calibrated case being HGDP which is also the smallest of these datasets. Otherwise,
 497 AUC_{PR} and SRMSD_p ranges were similar here as in our earlier evaluation. Therefore, the removal
 498 of the small number of highly related individual pairs had a negligible effect in PCA performance,
 499 so the larger number of more distantly related pairs explain the poor PCA performance in the real
 500 datasets.

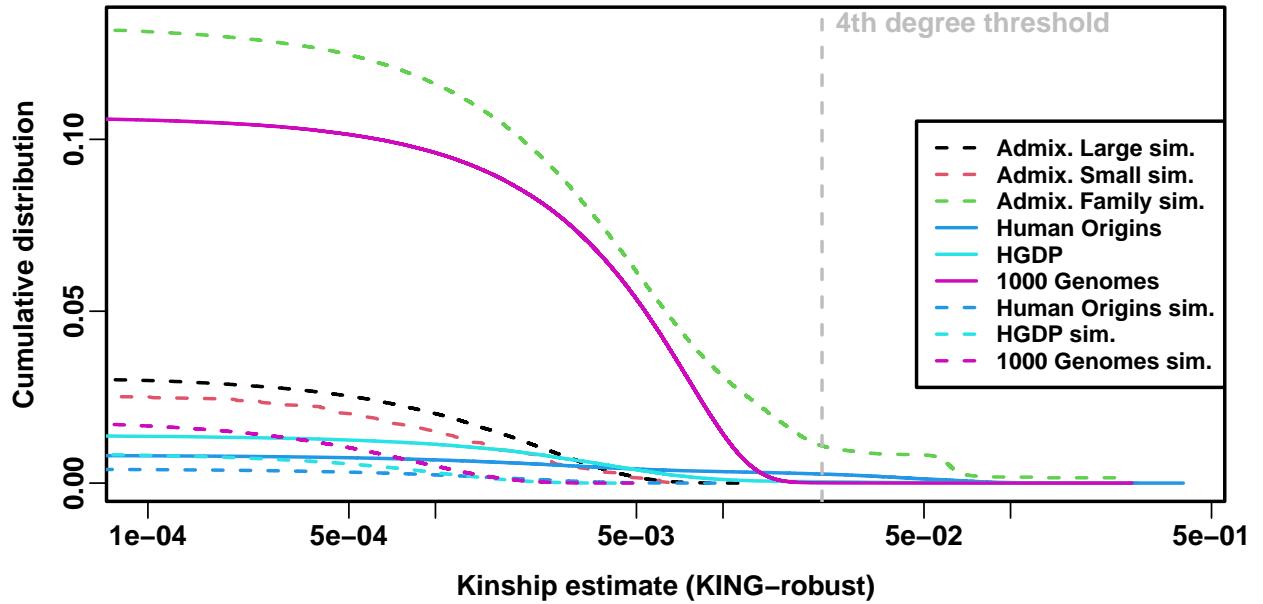


Figure 6: **Local kinship distributions.** Curves are complementary cumulative distribution of lower triangular kinship matrix (self kinship excluded) from KING-robust estimator. Note log x-axis; negative estimates are counted but not shown. Most values are below 4th degree relative threshold. Each real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

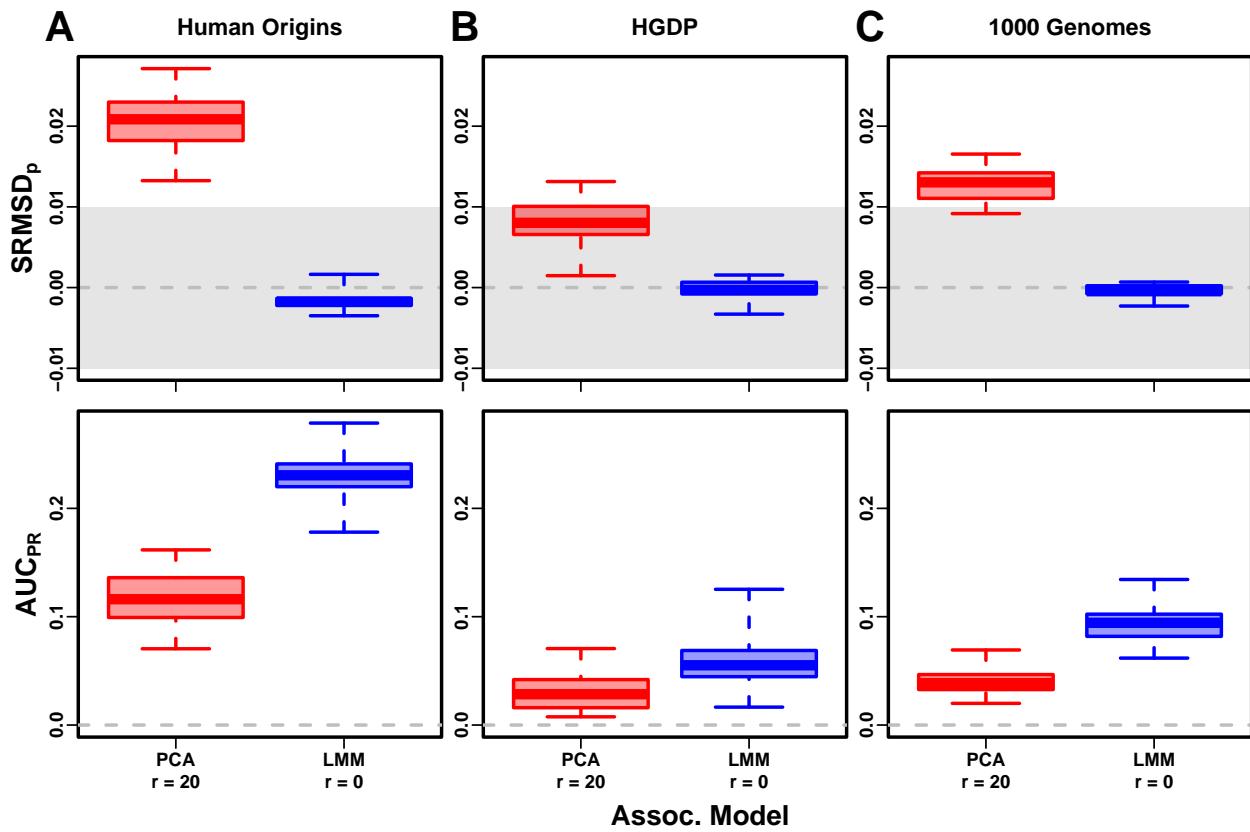


Figure 7: **Evaluation in real datasets excluding 4th degree relatives.** Traits simulated from FES model with high heritability. Each dataset is a column, rows are measures. First row has $|SRMSD_p| < 0.01$ band marked as gray area.

501 **3.6 Low heritability and environment simulations**

502 Our main evaluations were repeated with traits simulated under a lower heritability value of $h^2 =$
 503 0.3. We reduced the number of causal loci in response to this change in heritability, to result
 504 in equal average effect size per locus compared to the previous high heritability evaluations (see
 505 Materials and Methods). Despite that, these low heritability evaluations measured lower AUC_{PR}
 506 values than their high heritability counterparts (Figs. S9 to S13). The gap between LMM and PCA
 507 was reduced in these evaluations, but the main conclusion of the high heritability evaluation holds
 508 for low heritability as well, namely that LMM with $r = 0$ significantly outperforms or ties LMM
 509 with $r > 0$ and PCA in all cases (Table S2).

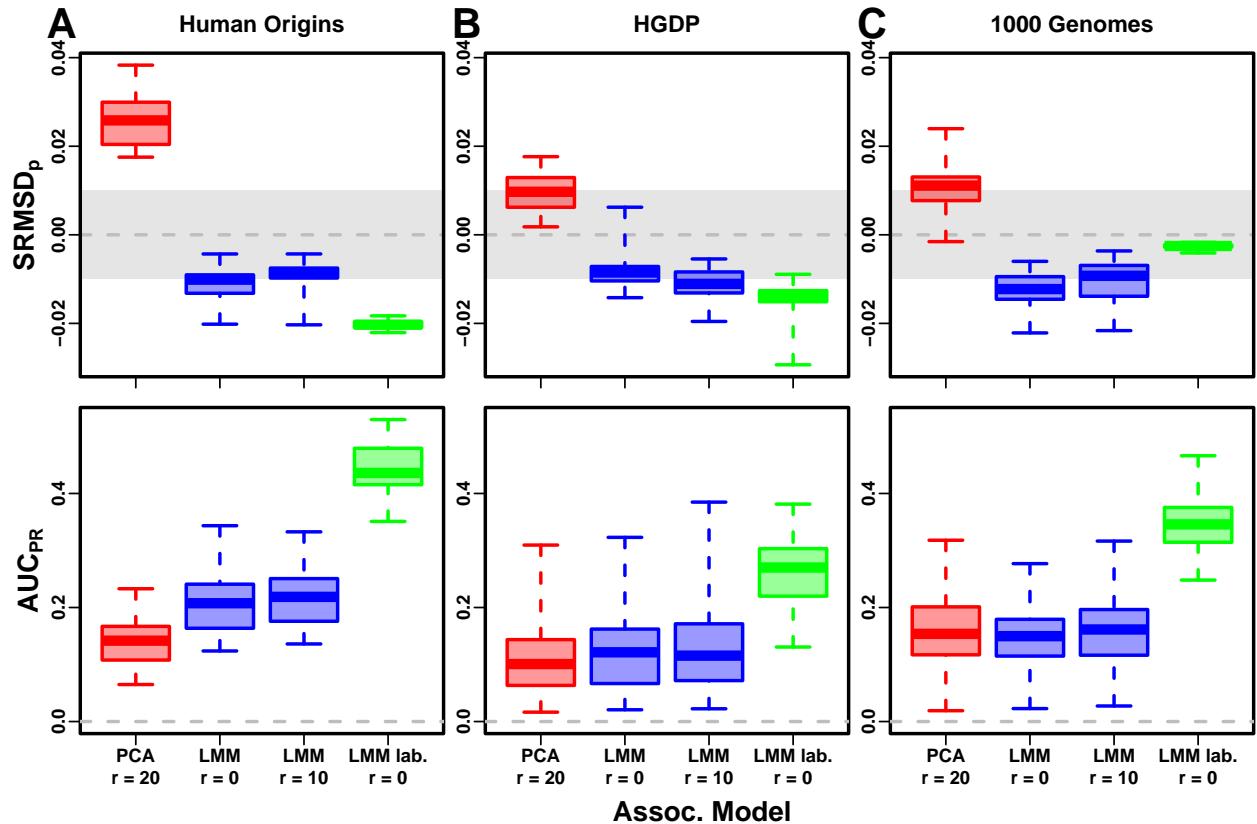


Figure 8: **Evaluation in real datasets excluding 4th degree relatives, FES traits, environment.** Traits simulated with environment effects, otherwise the same as Fig. 7.

510 Lastly, we simulated traits with both low heritability and large environment effects determined

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

530 4 Discussion

531 Our evaluations conclusively determined that LMM without PCs performs better than PCA (for any
532 number of PCs) across all scenarios without environment effects, including all real and simulated
533 genotypes and two trait simulation models. Although the addition of a few PCs to LMM does
534 not greatly hurt its performance (except for small sample sizes), they generally did not improve it
535 either (Tables S2 and 4), which agrees with previous observations [48, 51] but contradicts others
536 [16, 24]. Our findings make sense since PCs are the eigenvectors of the same kinship matrix that
537 parameterized random effects, so including both is redundant.

538 The presence of environment effects that are correlated to relatedness presents the only scenario
539 where occasionally PCA and LMM with PCs outperform LMM without PCs (Table S3). It is
540 commonly believed that PCs model such environment effects well [18–20]. However, we observe
541 that LMM without PCs models environment effects nearly as well as with PCs (Fig. 8), consistent
542 with previous findings [33, 34] and with environment inflating heritability estimates using LMM [93].
543 Moreover, modeling the true environment groups as fixed categorical effects always substantially
544 improved AUC_{PR} compared to modeling them with PCs (Fig. 8 and Table S3). Modeling numerous
545 environment groups as fixed effects does result in deflated p-values (Fig. 8 and Table S3), which we
546 expect would be avoided by modeling them as random effects, a strategy we chose not to pursue
547 here as it is both a circular evaluation (the true effects were drawn from that model) and out of
548 scope. Overall, including PCs to model environment effects yields limited power gains if at all, even
549 in an LMM, and is no replacement for more adequate modeling of environment whenever possible.

550 Previous studies found that PCA was better calibrated than LMM for unusually differentiated
551 markers [24, 35, 37], which as simulated were an artificial scenario not based on a population
552 genetics model, and are otherwise believed to be unusual [38, 59]. Our evaluations on real human
553 data, which contain such loci in relevant proportions if they exist, do not replicate that result.
554 Family relatedness strongly favors LMM, an advantage that probably outweighs this potential PCA
555 benefit in real data.

556 Relative to LMM, the behavior of PCA fell between two extremes. When PCA performed well,
557 there was a small number of PCs with both calibrated p-values and AUC_{PR} near that of LMM
558 without PCs. Conversely, PCA performed poorly when no number of PCs had either calibrated
559 p-values or acceptably large AUC_{PR}. There were no cases where high numbers of PCs optimized
560 an acceptable AUC_{PR}, or cases with miscalibrated p-values but high AUC_{PR}. PCA performed well
561 in the admixture simulations (without families, both trait models), real human genotypes with RC
562 traits, and the subpopulation tree simulations (both trait models). Conversely, PCA performed
563 poorly in the admixed family simulation (both trait models) and the real human genotypes with
564 FES traits.

565 PCA assumes that genetic relatedness is restricted to a low-dimensional subspace, whereas

566 LMM can handle high-dimensional relatedness. Thus, PCA performs well in the admixture sim-
567 ulation, which is explicitly low-dimensional (see Materials and Methods), and our subpopulation
568 tree simulations, which are likely well approximated by a few dimensions despite the large number
569 of subpopulations because there are few long branches. Conversely, PCA performs poorly under
570 family structure because its kinship matrix is high-dimensional (Fig. S7). However, estimating the
571 latent space dimensions of real datasets is challenging because estimated eigenvalues have biased
572 distributions [94]. Kinship matrix rank estimated using the Tracy-Widom test [7] did not fully pre-
573 dict the datasets that PCA performs well on. In contrast, estimated local kinship finds considerable
574 cryptic family relatedness in all real human datasets and better explains why PCA performs poorly
575 there. The trait model also influences the relative performance of PCA, so genotype-only parameters
576 (eigenvalues or local kinship) alone cannot tell the full story. There are related tests for numbers of
577 dimensions that consider the trait which we did not consider, including the Bayesian information
578 criterion for the regression with PCs against the trait [17]. Additionally, PCA and LMM goodness
579 of fit could be compared using the coefficient of determination generalized for LMMs [95].

580 PCA is at best underpowered relative to LMMs, and at worst miscalibrated regardless of the
581 numbers of PCs included, in real human genotype tests. Among our simulations, such poor per-
582 formance occurred only in the admixed family. Local kinship estimates reveal considerable family
583 relatedness in the real datasets absent in the corresponding subpopulation tree simulations. Admix-
584 ture is also absent in our tree simulations, but our simulations and theory show that admixture is
585 well handled by PCA. Hundreds of close relative pairs have been identified in 1000 Genomes [96–99],
586 but their removal does not improve PCA performance sufficiently in our tests, so the larger number
587 of more distantly related pairs are PCA’s most serious obstacle in practice. Distant relatives are
588 expected to be numerous in any large human dataset [58, 100, 101]. Our FES trait tests show that
589 family relatedness is more challenging when rarer variants have larger coefficients. Overall, the high
590 relatedness dimensions induced by family relatedness is the key challenge for PCA association in
591 modern datasets that is readily overcome by LMM.

592 Our tests also found PCA robust to large numbers of PCs, far beyond the optimal choice,
593 agreeing with previous anecdotal observations [5, 36], in contrast to using too few PCs for which

594 there is a large performance penalty. The exception was the small sample size simulation, where only
595 small numbers of PCs performed well. In contrast, LMM is simpler since there is no need to choose
596 the number of PCs. However, an LMM with a large number of covariates may have conservative
597 p-values, as observed for LMM with large numbers of PCs, which is a weakness of the score test
598 used by the LMM we evaluated that may be overcome with other statistical tests. Simulations or
599 post hoc evaluations remain crucial for ensuring that statistics are calibrated.

600 There are several variants of the PCA and LMM analyses, most designed for better modeling
601 linkage disequilibrium (LD), that we did not evaluate directly, in which PCs are no longer exactly
602 the top eigenvectors of the kinship matrix (if estimated with different approaches), although this is
603 not a crucial aspect of our arguments. We do not consider the case where samples are projected
604 onto PCs estimated from an external sample [102], which is uncommon in association studies, and
605 whose primary effect is shrinkage, so if all samples are projected then they are all equally affected
606 and larger regression coefficients compensate for the shrinkage, although this will no longer be the
607 case if only a portion of the sample is projected onto the PCs of the rest of the sample. Another
608 approach tests PCs for association against every locus in the genome in order to identify and exclude
609 PCs that capture LD structure (which is localized) instead of ancestry (which should be present
610 across the genome) [102]; a previous proposal removes LD using an autocorrelation model prior to
611 estimating PCs [7]. These improved PCs remain inadequate models of family relatedness, so an
612 LMM will continue to outperform them in that setting. Similarly, the leave-one-chromosome-out
613 (LOCO) approach for estimating kinship matrices for LMMs prevents the test locus and loci in LD
614 with it from being modeled by the random effect as well, which is called “proximal contamination”
615 [35, 42]. While LOCO kinship estimates vary for each chromosome, they continue to model family
616 relatedness, thus maintaining their key advantage over PCA. The LDAK model estimates kinship
617 instead by weighing loci taking LD into account [103]. LD effects must be adjusted for, if present,
618 so in unfiltered data we advise the previous methods be applied. However, in this work, simulated
619 genotypes do not have LD, and the real datasets were filtered to remove LD, so here there is no
620 proximal contamination and LD confounding is minimized if present at all, so these evaluations
621 may be considered the ideal situation where LD effects have been adjusted successfully, and in

622 this setting LMM outperforms PCA. Overall, these alternative PCs or kinship matrices differ from
623 their basic counterparts by either the extent to which LD influences the estimates (which may be a
624 confounder in a small portion of the genome, by definition) or by sampling noise, neither of which
625 are expected to change our key conclusion.

626 One of the limitations of this work include relatively small sample sizes compared to modern
627 association studies. However, our conclusions are not expected to change with larger sample sizes,
628 as cryptic family relatedness will continue to be abundant in such data, if not increase in abundance,
629 and thus give LMMs an advantage over PCA [58, 100, 101]. One reason PCA has been favored over
630 classic LMMs is because PCA’s runtime scales much better with increasing sample size. However,
631 recent approaches not tested in this work have made LMMs more scalable and applicable to biobank-
632 scale data [39, 47, 53], so one clear next step is carefully evaluating these approaches in simulations
633 with larger sample sizes. A different benefit for including PCs were recently reported for BOLT-
634 LMM, which does not result in greater power but rather in reduced runtime, a property that may
635 be specific to its use of scalable algorithms such as conjugate gradient and variational Bayes [58].
636 Many of these newer LMMs also no longer follow the infinitesimal model of the basic LMM [47,
637 53], and employ novel approximations, which are features not evaluated in this work and worthy of
638 future study.

639 Another limitation of this work is ignoring rare variants, a necessity given our smaller sample
640 sizes, where rare variant association is miscalibrated and underpowered. Using simulations mimick-
641 ing the UK Biobank, recent work has found that rare variants can have a more pronounced structure
642 than common variants, and that modeling this rare variant structure (with either PCA and LMM)
643 may better model environment confounding, reduce inflation in association studies, and ameliorate
644 stratification in polygenic risk scores [104]. Better modeling rare variants and their structure is a
645 key next step in association studies.

646 The largest limitation of our work is that we only considered quantitative traits. Previous evalua-
647 tions involving case-control traits tended to report PCA-LMM ties or mixed results, an observation
648 potentially confounded by the use of low-dimensional simulations without family relatedness (Ta-
649 ble 1). An additional concern is case-control ascertainment bias and imbalance, which appears to

650 affect LMMs more severely, although recent work appears to solve this problem [35, 39]. Future
651 evaluations should aim to include our simulations and real datasets, to ensure that previous results
652 were not biased in favor of PCA by not simulating family structure or larger coefficients for rare
653 variants that are expected for diseases by various selection models.

654 Overall, our results lead us to recommend LMM over PCA for association studies in general.
655 Although PCA offer flexibility and speed compared to LMM, additional work is required to ensure
656 that PCA is adequate, including removal of close relatives (lowering sample size and wasting re-
657 sources) followed by simulations or other evaluations of statistics, and even then PCA may perform
658 poorly in terms of both type I error control and power. The large numbers of distant relatives
659 expected of any real dataset all but ensures that PCA will perform poorly compared to LMM [58,
660 100, 101]. Our findings also suggest that related applications such as polygenic models may enjoy
661 gains in power and accuracy by employing an LMM instead of PCA to model relatedness [22, 91].
662 PCA remains indispensable across population genetics, from visualizing population structure and
663 performing quality control to its deep connection to admixture models, but the time has come to
664 limit its use in association testing in favor of LMM or other, richer models capable of modeling all
665 forms of relatedness.

666 5 Appendices

667 5.1 Appendix A: Fitting ancestral allele frequency distribution to real data

668 We calculated \hat{p}_i^T distributions of each real dataset. However, differentiation increases the variance
669 of these sample \hat{p}_i^T relative to the true p_i^T [32]. We present a new algorithm for constructing an
670 “undifferentiated” distribution based on the input data but with the lower variance of the true
671 ancestral distribution. Suppose the p_i^T distribution over loci i satisfies $E[p_i^T] = \frac{1}{2}$ and $\text{Var}(p_i^T) =$
672 V^T . The sample allele frequency \hat{p}_i^T , conditioned on p_i^T , satisfies

$$E[\hat{p}_i^T | p_i^T] = p_i^T, \quad \text{Var}(\hat{p}_i^T | p_i^T) = p_i^T (1 - p_i^T) \bar{\varphi}^T,$$

673 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 [32]. The unconditional
 674 moments of \hat{p}_i^T follow from the laws of total expectation and variance: $E[\hat{p}_i^T] = \frac{1}{2}$ and

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

675 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
 676 the original, lower variance of

$$677 \quad V^T = \frac{W^T - \frac{1}{4}\bar{\varphi}^T}{1 - \bar{\varphi}^T}. \quad (9)$$

678 We use the unbiased estimator $\hat{W}^T = \frac{1}{m} \sum_{i=1}^m (\hat{p}_i^T - \frac{1}{2})^2$, while $\bar{\varphi}^T$ is calculated from the tree
 679 parameters: the subpopulation coancestry matrix (Eq. (7)), expanded from subpopulations to indi-
 680 viduals, the diagonal converted to kinship (reversing Eq. (8)), and the matrix averaged. However,
 681 since our model ignores the MAF filters imposed in our simulations, $\bar{\varphi}^T$ was adjusted. For Human
 682 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
 683 and 0.124, respectively, but 0.4 for both produced a better fit.

684 Lastly, we construct new allele frequencies,

$$p^* = w\hat{p}_i^T + (1 - w)q,$$

685 by a weighted average of \hat{p}_i^T and $q \in (0, 1)$ drawn independently from a different distribution.
 686 $E[q] = \frac{1}{2}$ is required to have $E[p^*] = \frac{1}{2}$. The resulting variance is

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

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

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

689 Although $w = 0$ yields $\text{Var}(p^*) = V^T$, we use the second root of the quadratic equation to use \hat{p}_i^T :

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

690 **5.2 Appendix B: comparisons between SRMSD_p, AUC_{PR}, and evaluation mea-
691 sures from the literature**

692 **5.2.1 The inflation factor λ**

693 Test statistic inflation has been used to measure model calibration [1, 24]. The inflation factor
694 λ is defined as the median χ^2 association statistic divided by theoretical median under the null
695 hypothesis [2]. To compare p-values from non- χ^2 tests (such as t-statistics), λ can be calculated
696 from p-values using

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

697 where p_{median} is the median observed p-value (including causal loci), $u_{\text{median}} = \frac{1}{2}$ is its null expec-
698 tation, and F is the χ^2 cumulative density function (F^{-1} is the quantile function).

699 To compare λ and SRMSD_p directly, for simplicity assume that all p-values are null. In this
700 case, calibrated p-values give $\lambda = 1$ and SRMSD_p = 0. However, non-uniform p-values with the
701 expected median, such as from genomic control [2], result in $\lambda = 1$, but SRMSD_p ≠ 0 except for
702 uniform p-values, a key flaw of λ that SRMSD_p overcomes. Inflated statistics (anti-conservative
703 p-values) give $\lambda > 1$ and SRMSD_p > 0. Deflated statistics (conservative p-values) give $\lambda < 1$ and
704 SRMSD_p < 0. Thus, $\lambda \neq 1$ always implies SRMSD_p ≠ 0 (where $\lambda - 1$ and SRMSD_p have the
705 same sign), but not the other way around. Overall, λ depends only on the median p-value, while
706 SRMSD_p uses the complete distribution. However, SRMSD_p requires knowing which loci are null,
707 so unlike λ it is only applicable to simulated traits.

708 **5.2.2 Empirical comparison of SRMSD_p and λ**

709 There is a near one-to-one correspondence between λ and SRMSD_p in our data (Fig. S1). PCA
710 tended to be inflated ($\lambda > 1$ and SRMSD_p > 0) whereas LMM tended to be deflated ($\lambda < 1$ and

711 SRMSD_p < 0), otherwise the data for both models fall on the same contiguous curve. We fit a
712 sigmoidal function to this data,

713
$$\text{SRMSD}_p(\lambda) = a \frac{\lambda^b - 1}{\lambda^b + 1}, \quad (10)$$

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

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

715 We fit this model to $\lambda > 1$ only since it was less noisy and of greater interest, and obtained the
716 curve shown in Fig. S1 with $a = 0.564$ and $b = 0.619$. The value $\lambda = 1.05$, a common threshold
717 for benign inflation [24], corresponds to $\text{SRMSD}_p = 0.0085$ according to Eq. (10). Conversely,
718 $\text{SRMSD}_p = 0.01$, serving as a simpler rule of thumb, corresponds to $\lambda = 1.06$.

719 **5.2.3 Type I error rate**

720 The type I error rate is the proportion of null p-values with $p \leq t$. Calibrated p-values have type
721 I error rate near t , which may be evaluated with a binomial test. This measure may give different
722 results for different t , for example be significantly miscalibrated only for large t (due to lack of
723 power for smaller t). In contrast, $\text{SRMSD}_p = 0$ guarantees calibrated type I error rates at all t ,
724 while large $|\text{SRMSD}_p|$ indicates incorrect type I errors for a range of t . Empirically, we find the
725 expected agreement and monotonic relationship between SRMSD_p and type I error rate (Fig. S2).

726 **5.2.4 Statistical power and comparison to AUC_{PR}**

727 Power is the probability that a test is declared significant when the alternative hypothesis H_1 holds.

728 At a p-value threshold t , power equals

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

729 $F(t)$ is a cumulative function, so it is monotonically increasing and has an inverse. Like type I error
730 control, power may rank models differently depending on t .

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

⁷³² AUC_{PR}, assume calibrated (uniform) null p-values: $\Pr(p < t | H_0) = t$. TPs, FPs, and FNs at t are

$$\text{TP}(t) = m\pi_1 F(t),$$

$$\text{FP}(t) = m\pi_0 t,$$

$$\text{FN}(t) = m\pi_1(1 - F(t)),$$

⁷³³ where $\pi_0 = \Pr(H_0)$ is the proportion of null cases and $\pi_1 = 1 - \pi_0$ of alternative cases. Therefore,

$$\text{Precision}(t) = \frac{\pi_1 F(t)}{\pi_1 F(t) + \pi_0 t},$$

$$\text{Recall}(t) = F(t).$$

⁷³⁴ Noting that $t = F^{-1}(\text{Recall})$, precision can be written as a function of recall, the power function,

⁷³⁵ and constants:

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

⁷³⁶ This last form leads most clearly to $\text{AUC}_{\text{PR}} = \int_0^1 \text{Precision}(\text{Recall}) d\text{Recall}$.

⁷³⁷ Lastly, consider a simple yet common case in which model A is uniformly more powerful than

⁷³⁸ model B : $F_A(t) > F_B(t)$ for every t . Therefore $F_A^{-1}(\text{Recall}) < F_B^{-1}(\text{Recall})$ for every recall value.

⁷³⁹ This ensures that the precision of A is greater than that of B at every recall value, so AUC_{PR} is

⁷⁴⁰ greater for A than B . Thus, AUC_{PR} ranks calibrated models according to power.

⁷⁴¹ Empirically, we find the predicted positive correlation between AUC_{PR} and calibrated power

⁷⁴² (Fig. S3). The correlation is clear when considered separately per dataset, but the slope varies per

⁷⁴³ dataset, which is expected because the proportion of alternative cases π_1 varies per dataset.

⁷⁴⁴ Competing interests

⁷⁴⁵ The authors declare no competing interests.

⁷⁴⁶ **Acknowledgments**

⁷⁴⁷ Thanks to Tiffany Tu, Ratchanon Pornmongkolsuk, and Zhuoran Hou for feedback on this article.
⁷⁴⁸ This work was funded in part by the Duke University School of Medicine Whitehead Scholars
⁷⁴⁹ Program, a gift from the Whitehead Charitable Foundation. The 1000 Genomes data were generated
⁷⁵⁰ at the New York Genome Center with funds provided by NHGRI Grant 3UM1HG008901-03S1.

⁷⁵¹ **Web resources**

⁷⁵² plink2, <https://www.cog-genomics.org/plink/2.0/>
⁷⁵³ GCTA, <https://yanglab.westlake.edu.cn/software/gcta/>
⁷⁵⁴ Eigensoft, <https://github.com/DReichLab/EIG>
⁷⁵⁵ bnpsd, <https://cran.r-project.org/package=bnpsd>
⁷⁵⁶ simfam, <https://cran.r-project.org/package=simfam>
⁷⁵⁷ simtrait, <https://cran.r-project.org/package=simtrait>
⁷⁵⁸ genio, <https://cran.r-project.org/package=genio>
⁷⁵⁹ popkin, <https://cran.r-project.org/package=popkin>
⁷⁶⁰ ape, <https://cran.r-project.org/package=ape>
⁷⁶¹ nnls, <https://cran.r-project.org/package=nnls>
⁷⁶² PRROC, <https://cran.r-project.org/package=PRROC>
⁷⁶³ BEDMatrix, <https://cran.r-project.org/package=BEDMatrix>

⁷⁶⁴ **Data and code availability**

⁷⁶⁵ 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
⁷⁶⁶ website at <https://reich.hms.harvard.edu/datasets>; non-public samples have to be requested
⁷⁶⁷ from David Reich. The WGS version of HGDP was downloaded from the Wellcome Sanger In-
⁷⁶⁸ stitute FTP site at ftp://ngs.sanger.ac.uk/production/hgdp/hgdp_wgs.20190516/. The high-
⁷⁶⁹ coverage version of the 1000 Genomes Project was downloaded from <ftp://ftp.1000genomes.ebi.ac.uk>.
⁷⁷⁰

772 References

- 773 [1] W. Astle and D. J. Balding. “Population Structure and Cryptic Relatedness in Genetic
774 Association Studies”. *Statist. Sci.* 24(4) (2009), pp. 451–471. DOI: 10.1214/09-STS307.
- 775 [2] B. Devlin and K. Roeder. “Genomic Control for Association Studies”. *Biometrics* 55(4)
776 (1999), pp. 997–1004. DOI: 10.1111/j.0006-341X.1999.00997.x.
- 777 [3] B. F. Voight and J. K. Pritchard. “Confounding from Cryptic Relatedness in Case-Control As-
778 sociation Studies”. *PLOS Genetics* 1(3) (2005), e32. DOI: 10.1371/journal.pgen.0010032.
- 779 [4] S. Zhang, X. Zhu, and H. Zhao. “On a semiparametric test to detect associations between
780 quantitative traits and candidate genes using unrelated individuals”. *Genetic Epidemiology*
781 24(1) (2003), pp. 44–56. DOI: 10.1002/gepi.10196.
- 782 [5] A. L. Price et al. “Principal components analysis corrects for stratification in genome-wide
783 association studies”. *Nat. Genet.* 38(8) (2006), pp. 904–909. DOI: 10.1038/ng1847.
- 784 [6] M. Bouaziz, C. Ambroise, and M. Guedj. “Accounting for Population Stratification in Prac-
785 tice: A Comparison of the Main Strategies Dedicated to Genome-Wide Association Studies”.
786 *PLOS ONE* 6(12) (2011), e28845. DOI: 10.1371/journal.pone.0028845.
- 787 [7] N. Patterson, A. L. Price, and D. Reich. “Population Structure and Eigenanalysis”. *PLoS
788 Genet* 2(12) (2006), e190. DOI: 10.1371/journal.pgen.0020190.
- 789 [8] I. T. Jolliffe. *Principal Component Analysis*. 2nd ed. New York: Springer-Verlag, 2002.
- 790 [9] J. K. Pritchard et al. “Association Mapping in Structured Populations”. *The American Jour-
791 nal of Human Genetics* 67(1) (2000), pp. 170–181. DOI: 10.1086/302959.
- 792 [10] D. H. Alexander, J. Novembre, and K. Lange. “Fast model-based estimation of ancestry in
793 unrelated individuals”. *Genome Res.* 19(9) (2009), pp. 1655–1664. DOI: 10.1101/gr.094052.
794 109.

- 795 [11] Q. Zhou, L. Zhao, and Y. Guan. “Strong Selection at MHC in Mexicans since Admixture”.
796 *PLoS Genet.* 12(2) (2016), e1005847. DOI: 10.1371/journal.pgen.1005847.
- 797 [12] G. McVean. “A genealogical interpretation of principal components analysis”. *PLoS Genet*
798 5(10) (2009), e1000686. DOI: 10.1371/journal.pgen.1000686.
- 799 [13] X. Zheng and B. S. Weir. “Eigenanalysis of SNP data with an identity by descent interpre-
800 tation”. *Theor Popul Biol* 107 (2016), pp. 65–76. DOI: 10.1016/j.tpb.2015.09.004.
- 801 [14] I. Cabreros and J. D. Storey. “A Likelihood-Free Estimator of Population Structure Bridging
802 Admixture Models and Principal Components Analysis”. *Genetics* 212(4) (2019), pp. 1009–
803 1029. DOI: 10.1534/genetics.119.302159.
- 804 [15] A. M. Chiu et al. “Inferring population structure in biobank-scale genomic data”. *The Amer-
805 ican Journal of Human Genetics* 0(0) (2022). DOI: 10.1016/j.ajhg.2022.02.015.
- 806 [16] K. Zhao et al. “An Arabidopsis Example of Association Mapping in Structured Samples”.
807 *PLOS Genetics* 3(1) (2007), e4. DOI: 10.1371/journal.pgen.0030004.
- 808 [17] C. Zhu and J. Yu. “Nonmetric Multidimensional Scaling Corrects for Population Structure in
809 Association Mapping With Different Sample Types”. *Genetics* 182(3) (1, 2009), pp. 875–888.
810 DOI: 10.1534/genetics.108.098863.
- 811 [18] J. Novembre et al. “Genes mirror geography within Europe”. *Nature* 456(7218) (2008), pp. 98–
812 101. DOI: 10.1038/nature07331.
- 813 [19] Y. Zhang and W. Pan. “Principal Component Regression and Linear Mixed Model in Associ-
814 ation Analysis of Structured Samples: Competitors or Complements?” *Genetic Epidemiology*
815 39(3) (2015), pp. 149–155. DOI: 10.1002/gepi.21879.
- 816 [20] M. Lin et al. “Admixed Populations Improve Power for Variant Discovery and Portability in
817 Genome-Wide Association Studies”. *Frontiers in Genetics* 12 (2021).
- 818 [21] H. Xu and Y. Guan. “Detecting Local Haplotype Sharing and Haplotype Association”. *Ge-
819 netics* 197(3) (2014), pp. 823–838. DOI: 10.1534/genetics.114.164814.

- 820 [22] J. Qian et al. “A fast and scalable framework for large-scale and ultrahigh-dimensional sparse
821 regression with application to the UK Biobank”. *PLOS Genetics* 16(10) (2020), e1009141.
822 DOI: 10.1371/journal.pgen.1009141.
- 823 [23] T. Thornton and M. S. McPeek. “ROADTRIPS: case-control association testing with par-
824 tially or completely unknown population and pedigree structure”. *Am. J. Hum. Genet.* 86(2)
825 (2010), pp. 172–184. DOI: 10.1016/j.ajhg.2010.01.001.
- 826 [24] A. L. Price et al. “New approaches to population stratification in genome-wide association
827 studies”. *Nature Reviews Genetics* 11(7) (2010), pp. 459–463. DOI: 10.1038/nrg2813.
- 828 [25] S. Lee et al. “Sparse Principal Component Analysis for Identifying Ancestry-Informative
829 Markers in Genome-Wide Association Studies”. *Genetic Epidemiology* 36(4) (2012), pp. 293–
830 302. DOI: 10.1002/gepi.21621.
- 831 [26] G. Abraham and M. Inouye. “Fast Principal Component Analysis of Large-Scale Genome-
832 Wide Data”. *PLOS ONE* 9(4) (2014), e93766. DOI: 10.1371/journal.pone.0093766.
- 833 [27] K. Galinsky et al. “Fast Principal-Component Analysis Reveals Convergent Evolution of
834 ADH1B in Europe and East Asia”. *The American Journal of Human Genetics* 98(3) (2016),
835 pp. 456–472. DOI: 10.1016/j.ajhg.2015.12.022.
- 836 [28] G. Abraham, Y. Qiu, and M. Inouye. “FlashPCA2: principal component analysis of Biobank-
837 scale genotype datasets”. *Bioinformatics* 33(17) (2017), pp. 2776–2778. DOI: 10.1093/
838 bioinformatics/btx299.
- 839 [29] A. Agrawal et al. “Scalable probabilistic PCA for large-scale genetic variation data”. *PLOS
840 Genetics* 16(5) (2020), e1008773. DOI: 10.1371/journal.pgen.1008773.
- 841 [30] J. Yu et al. “A unified mixed-model method for association mapping that accounts for mul-
842 tiple levels of relatedness”. *Nat. Genet.* 38(2) (2006), pp. 203–208. DOI: 10.1038/ng1702.
- 843 [31] H. M. Kang et al. “Efficient control of population structure in model organism association
844 mapping”. *Genetics* 178(3) (2008), pp. 1709–1723. DOI: 10.1534/genetics.107.080101.
- 845 [32] A. Ochoa and J. D. Storey. “Estimating FST and kinship for arbitrary population structures”.
846 *PLoS Genet* 17(1) (2021), e1009241. DOI: 10.1371/journal.pgen.1009241.

- 847 [33] B. J. Vilhjálmsdóttir and M. Nordborg. “The nature of confounding in genome-wide association
848 studies”. *Nat Rev Genet* 14(1) (2013), pp. 1–2. DOI: 10.1038/nrg3382.
- 849 [34] H. Wang, B. Aragam, and E. P. Xing. “Trade-offs of Linear Mixed Models in Genome-
850 Wide Association Studies”. *Journal of Computational Biology* 29(3) (2022), pp. 233–242.
851 DOI: 10.1089/cmb.2021.0157.
- 852 [35] J. Yang et al. “Advantages and pitfalls in the application of mixed-model association meth-
853 ods”. *Nat Genet* 46(2) (2014), pp. 100–106. DOI: 10.1038/ng.2876.
- 854 [36] H. M. Kang et al. “Variance component model to account for sample structure in genome-
855 wide association studies”. *Nat. Genet.* 42(4) (2010), pp. 348–354. DOI: 10.1038/ng.548.
- 856 [37] C. Wu et al. “A Comparison of Association Methods Correcting for Population Stratification
857 in Case–Control Studies”. *Annals of Human Genetics* 75(3) (2011), pp. 418–427. DOI: 10.
858 1111/j.1469-1809.2010.00639.x.
- 859 [38] J. H. Sul and E. Eskin. “Mixed models can correct for population structure for genomic
860 regions under selection”. *Nature Reviews Genetics* 14(4) (2013), p. 300. DOI: 10.1038/
861 nrg2813-c1.
- 862 [39] W. Zhou et al. “Efficiently controlling for case-control imbalance and sample relatedness in
863 large-scale genetic association studies”. *Nat Genet* 50(9) (2018), pp. 1335–1341. DOI: 10.
864 1038/s41588-018-0184-y.
- 865 [40] Y. S. Aulchenko, D.-J. de Koning, and C. Haley. “Genomewide rapid association using mixed
866 model and regression: a fast and simple method for genomewide pedigree-based quantitative
867 trait loci association analysis”. *Genetics* 177(1) (2007), pp. 577–585. DOI: 10.1534/genetics.
868 107.075614.
- 869 [41] Z. Zhang et al. “Mixed linear model approach adapted for genome-wide association studies”.
870 *Nat Genet* 42(4) (2010), pp. 355–360. DOI: 10.1038/ng.546.
- 871 [42] C. Lippert et al. “FaST linear mixed models for genome-wide association studies”. *Nat.*
872 *Methods* 8(10) (2011), pp. 833–835. DOI: 10.1038/nmeth.1681.

- 873 [43] J. Yang et al. “GCTA: a tool for genome-wide complex trait analysis”. *Am. J. Hum. Genet.*
874 88(1) (2011), pp. 76–82. DOI: [10.1016/j.ajhg.2010.11.011](https://doi.org/10.1016/j.ajhg.2010.11.011).
- 875 [44] J. Listgarten et al. “Improved linear mixed models for genome-wide association studies”. *Nat
876 Methods* 9(6) (2012), pp. 525–526. DOI: [10.1038/nmeth.2037](https://doi.org/10.1038/nmeth.2037).
- 877 [45] X. Zhou and M. Stephens. “Genome-wide efficient mixed-model analysis for association stud-
878 ies”. *Nat. Genet.* 44(7) (2012), pp. 821–824. DOI: [10.1038/ng.2310](https://doi.org/10.1038/ng.2310).
- 879 [46] G. R. Svishcheva et al. “Rapid variance components-based method for whole-genome asso-
880 ciation analysis”. *Nat Genet* 44(10) (2012), pp. 1166–1170. DOI: [10.1038/ng.2410](https://doi.org/10.1038/ng.2410).
- 881 [47] P.-R. Loh et al. “Efficient Bayesian mixed-model analysis increases association power in large
882 cohorts”. *Nat. Genet.* 47(3) (2015), pp. 284–290. DOI: [10.1038/ng.3190](https://doi.org/10.1038/ng.3190).
- 883 [48] L. Janss et al. “Inferences from Genomic Models in Stratified Populations”. *Genetics* 192(2)
884 (1, 2012), pp. 693–704. DOI: [10.1534/genetics.112.141143](https://doi.org/10.1534/genetics.112.141143).
- 885 [49] G. E. Hoffman. “Correcting for population structure and kinship using the linear mixed
886 model: theory and extensions”. *PLoS ONE* 8(10) (2013), e75707. DOI: [10.1371/journal.
887 pone.0075707](https://doi.org/10.1371/journal.pone.0075707).
- 888 [50] G. Tucker, A. L. Price, and B. Berger. “Improving the Power of GWAS and Avoiding Con-
889 founding from Population Stratification with PC-Select”. *Genetics* 197(3) (2014), pp. 1045–
890 1049. DOI: [10.1534/genetics.114.164285](https://doi.org/10.1534/genetics.114.164285).
- 891 [51] N. Liu et al. “Controlling Population Structure in Human Genetic Association Studies with
892 Samples of Unrelated Individuals”. *Stat Interface* 4(3) (2011), pp. 317–326. DOI: [10.4310/
sii.2011.v4.n3.a6](https://doi.org/10.4310/
893 sii.2011.v4.n3.a6).
- 894 [52] J. Zeng et al. “Signatures of negative selection in the genetic architecture of human complex
895 traits”. *Nature Genetics* 50(5) (2018), pp. 746–753. DOI: [10.1038/s41588-018-0101-4](https://doi.org/10.1038/s41588-018-0101-4).
- 896 [53] J. Mbatchou et al. “Computationally efficient whole-genome regression for quantitative and
897 binary traits”. *Nat Genet* 53(7) (2021), pp. 1097–1103. DOI: [10.1038/s41588-021-00870-7](https://doi.org/10.1038/s41588-021-00870-7).

- 898 [54] N. Matoba et al. “GWAS of 165,084 Japanese individuals identified nine loci associated with
899 dietary habits”. *Nat Hum Behav* 4(3) (2020), pp. 308–316. DOI: 10.1038/s41562-019-0805-
900 1.
- 901 [55] M. Song, W. Hao, and J. D. Storey. “Testing for genetic associations in arbitrarily structured
902 populations”. *Nat. Genet.* 47(5) (2015), pp. 550–554. DOI: 10.1038/ng.3244.
- 903 [56] X. Liu et al. “Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient
904 Genome-Wide Association Studies”. *PLOS Genet.* 12(2) (2016), e1005767. DOI: 10.1371/journal.
905 pgen.1005767.
- 906 [57] J. H. Sul, L. S. Martin, and E. Eskin. “Population structure in genetic studies: Confounding
907 factors and mixed models”. *PLoS Genet.* 14(12) (2018), e1007309. DOI: 10.1371/journal.
908 pgen.1007309.
- 909 [58] P.-R. Loh et al. “Mixed-model association for biobank-scale datasets”. *Nat Genet* 50(7)
910 (2018), pp. 906–908. DOI: 10.1038/s41588-018-0144-6.
- 911 [59] A. L. Price et al. “Response to Sul and Eskin”. *Nature Reviews Genetics* 14(4) (2013), p. 300.
912 DOI: 10.1038/nrg2813-c2.
- 913 [60] T. G. P. Consortium. “A map of human genome variation from population-scale sequencing”.
914 *Nature* 467(7319) (2010), pp. 1061–1073. DOI: 10.1038/nature09534.
- 915 [61] 1000 Genomes Project Consortium et al. “An integrated map of genetic variation from 1,092
916 human genomes”. *Nature* 491(7422) (2012), pp. 56–65. DOI: 10.1038/nature11632.
- 917 [62] H. M. Cann et al. “A human genome diversity cell line panel”. *Science* 296(5566) (2002),
918 pp. 261–262. DOI: 10.1126/science.296.5566.261b.
- 919 [63] N. A. Rosenberg et al. “Genetic Structure of Human Populations”. *Science* 298(5602) (2002),
920 pp. 2381–2385. DOI: 10.1126/science.1078311.
- 921 [64] A. Bergström et al. “Insights into human genetic variation and population history from 929
922 diverse genomes”. *Science* 367(6484) (2020). DOI: 10.1126/science.aay5012.

- 923 [65] N. Patterson et al. “Ancient admixture in human history”. *Genetics* 192(3) (2012), pp. 1065–
924 1093. DOI: [10.1534/genetics.112.145037](https://doi.org/10.1534/genetics.112.145037).
- 925 [66] I. Lazaridis et al. “Ancient human genomes suggest three ancestral populations for present-
926 day Europeans”. *Nature* 513(7518) (2014), pp. 409–413. DOI: [10.1038/nature13673](https://doi.org/10.1038/nature13673).
- 927 [67] I. Lazaridis et al. “Genomic insights into the origin of farming in the ancient Near East”.
928 *Nature* 536(7617) (2016), pp. 419–424. DOI: [10.1038/nature19310](https://doi.org/10.1038/nature19310).
- 929 [68] P. Skoglund et al. “Genomic insights into the peopling of the Southwest Pacific”. *Nature*
930 538(7626) (2016), pp. 510–513. DOI: [10.1038/nature19844](https://doi.org/10.1038/nature19844).
- 931 [69] J.-H. Park et al. “Distribution of allele frequencies and effect sizes and their interrelationships
932 for common genetic susceptibility variants”. *PNAS* 108(44) (2011), pp. 18026–18031. DOI:
933 [10.1073/pnas.1114759108](https://doi.org/10.1073/pnas.1114759108).
- 934 [70] L. J. O’Connor et al. “Extreme Polygenicity of Complex Traits Is Explained by Negative
935 Selection”. *The American Journal of Human Genetics* 0(0) (2019). DOI: [10.1016/j.ajhg.2019.07.003](https://doi.org/10.1016/j.ajhg.2019.07.003).
- 936 [71] Y. B. Simons et al. “A population genetic interpretation of GWAS findings for human quanti-
937 tative traits”. *PLOS Biology* 16(3) (2018), e2002985. DOI: [10.1371/journal.pbio.2002985](https://doi.org/10.1371/journal.pbio.2002985).
- 938 [72] G. Malécot. *Mathématiques de l'hérédité*. Masson et Cie, 1948.
- 939 [73] S. Wright. “The Genetical Structure of Populations”. *Annals of Eugenics* 15(1) (1949),
940 pp. 323–354. DOI: [10.1111/j.1469-1809.1949.tb02451.x](https://doi.org/10.1111/j.1469-1809.1949.tb02451.x).
- 941 [74] A. Jacquard. *Structures génétiques des populations*. Paris: Masson et Cie, 1970.
- 942 [75] A. Ochoa and J. D. Storey. *New kinship and FST estimates reveal higher levels of differen-
943 tiation in the global human population*. 2019. DOI: [10.1101/653279](https://doi.org/10.1101/653279).
- 944 [76] Z. Hou and A. Ochoa. “Genetic association models are robust to common population kinship
945 estimation biases”. *Genetics* (27, 2023), iyad030. DOI: [10.1093/genetics/iyad030](https://doi.org/10.1093/genetics/iyad030).
- 946 [77] C. C. Chang et al. “Second-generation PLINK: rising to the challenge of larger and richer
947 datasets”. *GigaScience* 4(1) (2015), p. 7. DOI: [10.1186/s13742-015-0047-8](https://doi.org/10.1186/s13742-015-0047-8).

- 949 [78] D. J. Balding and R. A. Nichols. “A method for quantifying differentiation between pop-
950ulations at multi-allelic loci and its implications for investigating identity and paternity”.
951 *Genetica* 96(1-2) (1995), pp. 3–12. DOI: <https://doi.org/10.1007/BF01441146>.
- 952 [79] E. Paradis and K. Schliep. “ape 5.0: an environment for modern phylogenetics and evolution-
953ary analyses in R”. *Bioinformatics* 35(3) (2019), pp. 526–528. DOI: [10.1093/bioinformatics/bty633](https://doi.org/10.1093/bioinformatics/bty633).
- 955 [80] R. R. Sokal and C. D. Michener. “A statistical method for evaluating systematic relation-
956ships.” *Univ. Kansas, Sci. Bull.* 38 (1958), pp. 1409–1438.
- 957 [81] C. L. Lawson and R. J. Hanson. *Solving least squares problems*. Englewood Cliffs: Prentice
958 Hall, 1974.
- 959 [82] K. M. Mullen and I. H. M. v. Stokkum. *nnls: The Lawson-Hanson algorithm for non-negative*
960 *least squares (NNLS)*. 2012.
- 961 [83] J.-H. Park et al. “Estimation of effect size distribution from genome-wide association studies
962 and implications for future discoveries”. *Nature Genetics* 42(7) (2010), pp. 570–575. DOI:
963 [10.1038/ng.610](https://doi.org/10.1038/ng.610).
- 964 [84] A. Grueneberg and G. d. l. Campos. “BGData - A Suite of R Packages for Genomic Analysis
965 with Big Data”. *G3: Genes, Genomes, Genetics* 9(5) (2019), pp. 1377–1383. DOI: [10.1534/g3.119.400018](https://doi.org/10.1534/g3.119.400018).
- 967 [85] S. Fairley et al. “The International Genome Sample Resource (IGSR) collection of open
968 human genomic variation resources”. *Nucleic Acids Research* 48(D1) (2020), pp. D941–D947.
969 DOI: [10.1093/nar/gkz836](https://doi.org/10.1093/nar/gkz836).
- 970 [86] A. Manichaikul et al. “Robust relationship inference in genome-wide association studies”.
971 *Bioinformatics* 26(22) (2010), pp. 2867–2873. DOI: [10.1093/bioinformatics/btq559](https://doi.org/10.1093/bioinformatics/btq559).
- 972 [87] J. D. Storey. “The positive false discovery rate: a Bayesian interpretation and the q-value”.
973 *Ann. Statist.* 31(6) (2003), pp. 2013–2035. DOI: [10.1214/aos/1074290335](https://doi.org/10.1214/aos/1074290335).

- 974 [88] 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),
975 pp. 9440–9445. DOI: 10.1073/pnas.1530509100.
- 976
- 977 [89] J. Grau, I. Grosse, and J. Keilwagen. “PRROC: computing and visualizing precision-recall
978 and receiver operating characteristic curves in R”. *Bioinformatics* 31(15) (2015), pp. 2595–
979 2597. DOI: 10.1093/bioinformatics/btv153.
- 980
- 981 [90] 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.
- 982
- 983 [91] B. Rakitsch et al. “A Lasso multi-marker mixed model for association mapping with pop-
984 ulation structure correction”. *Bioinformatics* 29(2) (2013), pp. 206–214. DOI: 10.1093/
bioinformatics/bts669.
- 985
- 986 [92] 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.
- 987
- 988 [93] D. Heckerman et al. “Linear mixed model for heritability estimation that explicitly addresses
989 environmental variation”. *Proc. Natl. Acad. Sci. U.S.A.* 113(27) (2016), pp. 7377–7382. DOI:
10.1073/pnas.1510497113.
- 990
- 991 [94] K. Hayashi, K.-H. Yuan, and L. Liang. “On the Bias in Eigenvalues of Sample Covariance Ma-
992 trix”. *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-
993 3-319-77249-3_19.
- 994
- 995 [95] G. Sun et al. “Variation explained in mixed-model association mapping”. *Heredity* 105(4)
(2010), pp. 333–340. DOI: 10.1038/hdy.2010.11.
- 996
- 997 [96] 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.
- 998
- 999 [97] 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.

- 1000 [98] L. Fedorova et al. “Atlas of Cryptic Genetic Relatedness Among 1000 Human Genomes”.
1001 *Genome Biology and Evolution* 8(3) (2016), pp. 777–790. DOI: 10.1093/gbe/evw034.
- 1002 [99] D. Schlauch, H. Fier, and C. Lange. “Identification of genetic outliers due to sub-structure
1003 and cryptic relationships”. *Bioinformatics* 33(13) (2017), pp. 1972–1979. DOI: 10.1093/
1004 *bioinformatics/btx109*.
- 1005 [100] B. M. Henn et al. “Cryptic Distant Relatives Are Common in Both Isolated and Cosmopolitan
1006 Genetic Samples”. *PLOS ONE* 7(4) (2012), e34267. DOI: 10.1371/journal.pone.0034267.
- 1007 [101] V. Shchur and R. Nielsen. “On the number of siblings and p-th cousins in a large population
1008 sample”. *J Math Biol* 77(5) (2018), pp. 1279–1298. DOI: 10.1007/s00285-018-1252-8.
- 1009 [102] F. Privé et al. “Efficient toolkit implementing best practices for principal component analysis
1010 of population genetic data”. *Bioinformatics* 36(16) (15, 2020), pp. 4449–4457. DOI: 10.1093/
1011 *bioinformatics/btaa520*.
- 1012 [103] D. Speed et al. “Improved heritability estimation from genome-wide SNPs”. *Am. J. Hum.
1013 Genet.* 91(6) (7, 2012), pp. 1011–1021. DOI: 10.1016/j.ajhg.2012.10.010.
- 1014 [104] A. A. Zaidi and I. Mathieson. “Demographic history mediates the effect of stratification on
1015 polygenic scores”. *eLife* 9 (17, 2020). Ed. by G. H. Perry, M. C. Turchin, and A. R. Martin,
1016 e61548. DOI: 10.7554/eLife.61548.

Supplemental figures

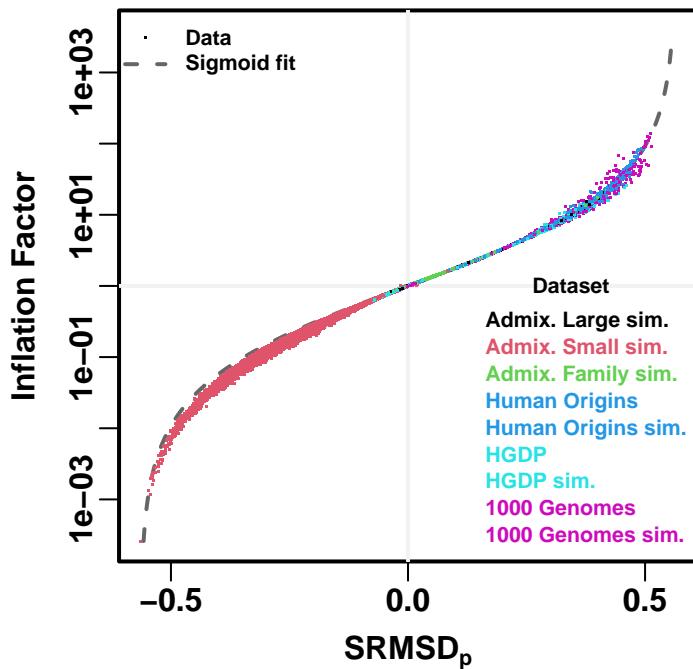


Figure S1: **Comparison between SRMSD_p and inflation factor.** Each point is a pair of statistics for one replicate, one association model (PCA or LMM with some number of PCs r), one trait model (FES vs RC, all heritability/environments tested), and one dataset (color coded by dataset). Note log y-axis.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

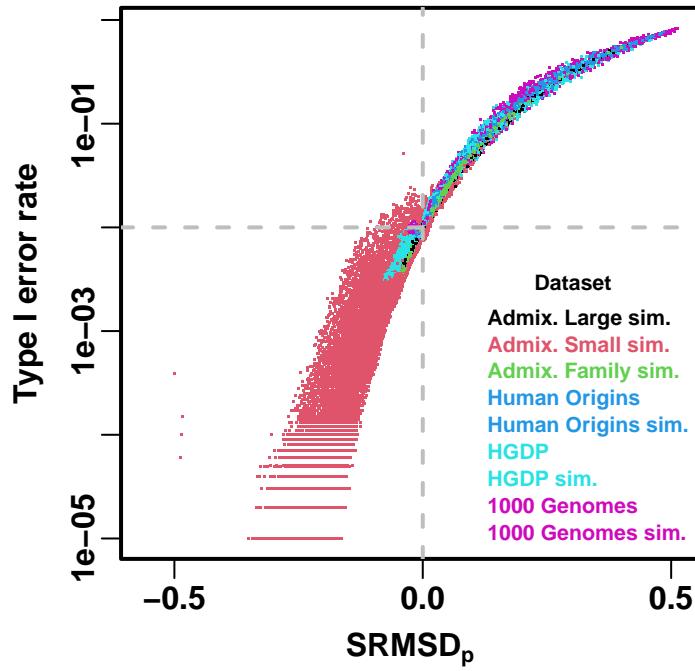


Figure S2: **Comparison between SRMSD_p and type I error rate.** Type I error rate calculated at a p-value threshold of $1\text{e-}2$ (horizontal dashed gray line). Thus, a calibrated model has a type I error rate of $1\text{e-}2$ and $\text{SRMSD}_p = 0$ (where the dashed lines meet). As expected, increased type I error rates correspond to $\text{SRMSD}_p > 0$, while reduced type I error rates correspond to $\text{SRMSD}_p < 0$. 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.

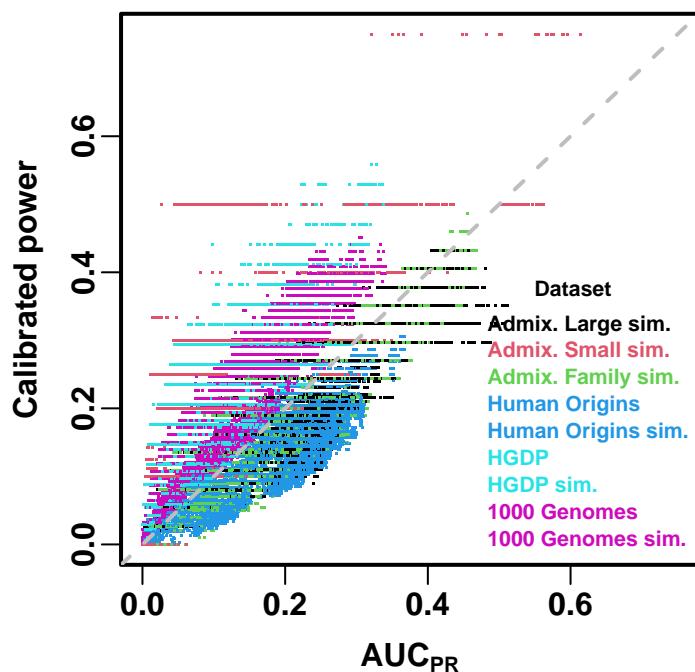


Figure S3: **Comparison between AUC_{PR} and calibrated power.** Calibrated power is power calculated at an empirical type I error threshold of $1\text{e-}4$. Each point is a pair of statistics for one replicate, one association model (PCA or LMM with some number of PCs r), one trait model (FES vs RC,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

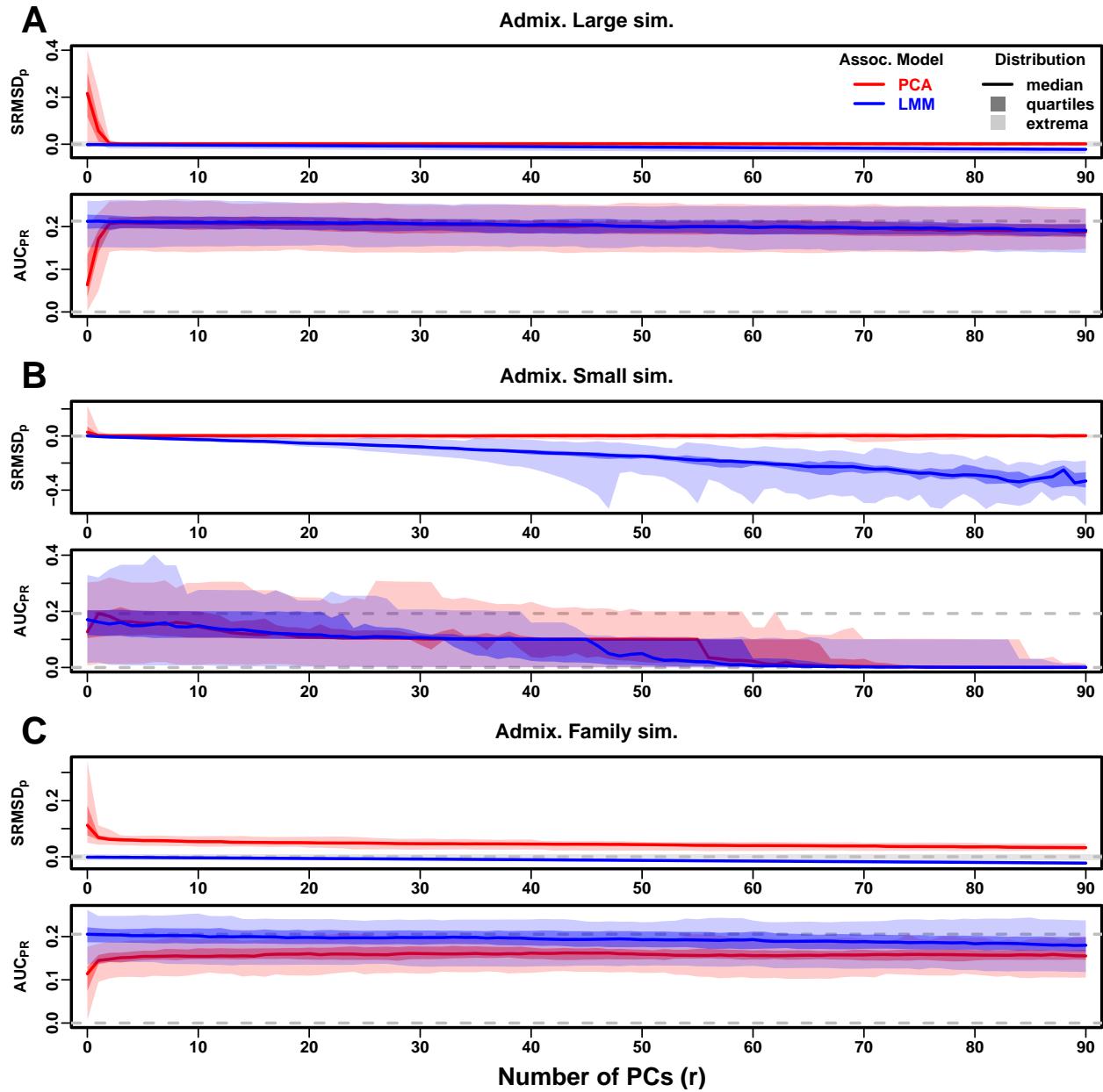


Figure S4: Evaluations in admixture simulations with RC traits. Traits simulated from RC model, otherwise the same as Fig. 3.

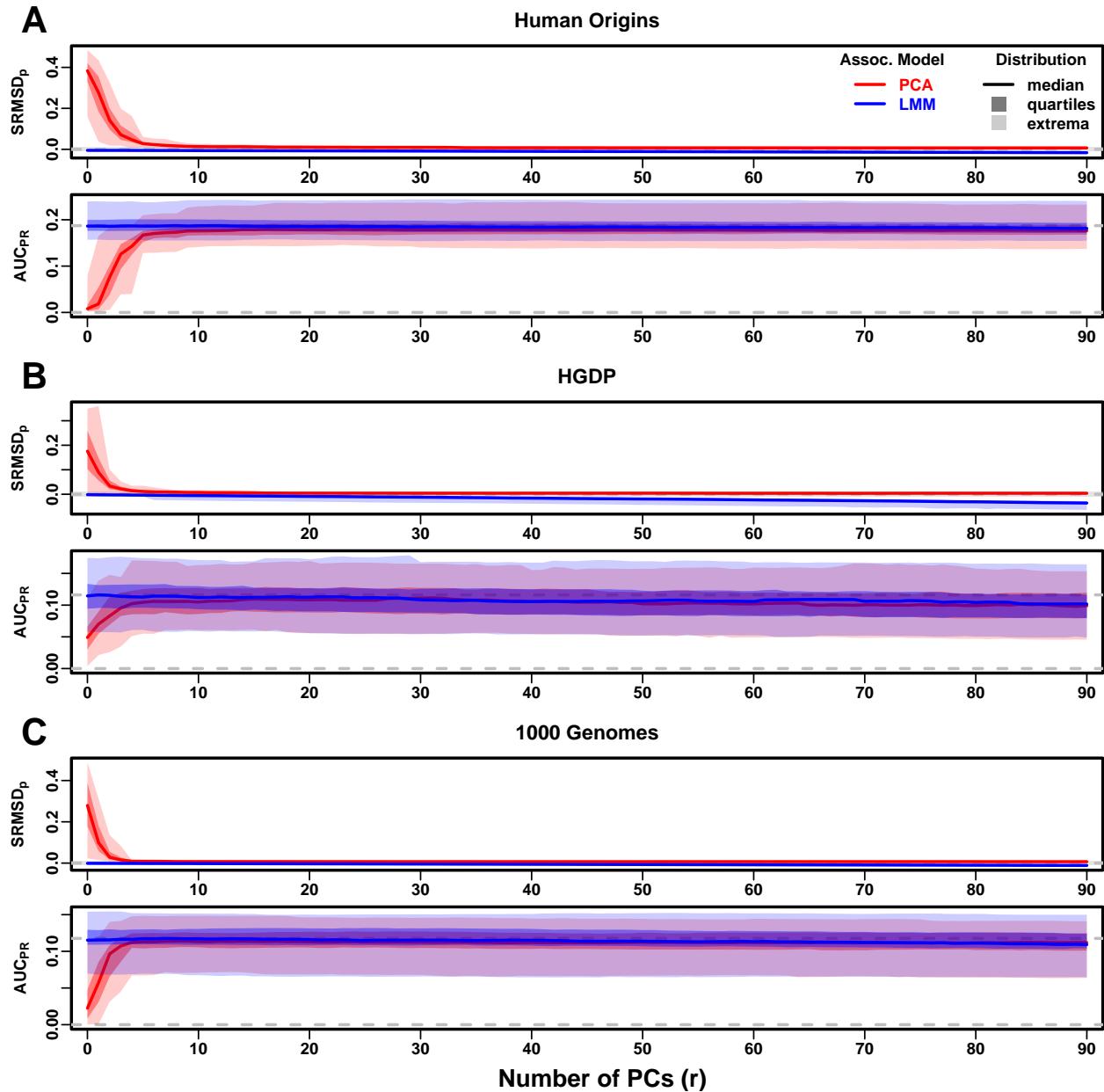


Figure S5: Evaluations in real human genotype datasets with RC traits. Traits simulated from RC model, otherwise the same as Fig. 4.

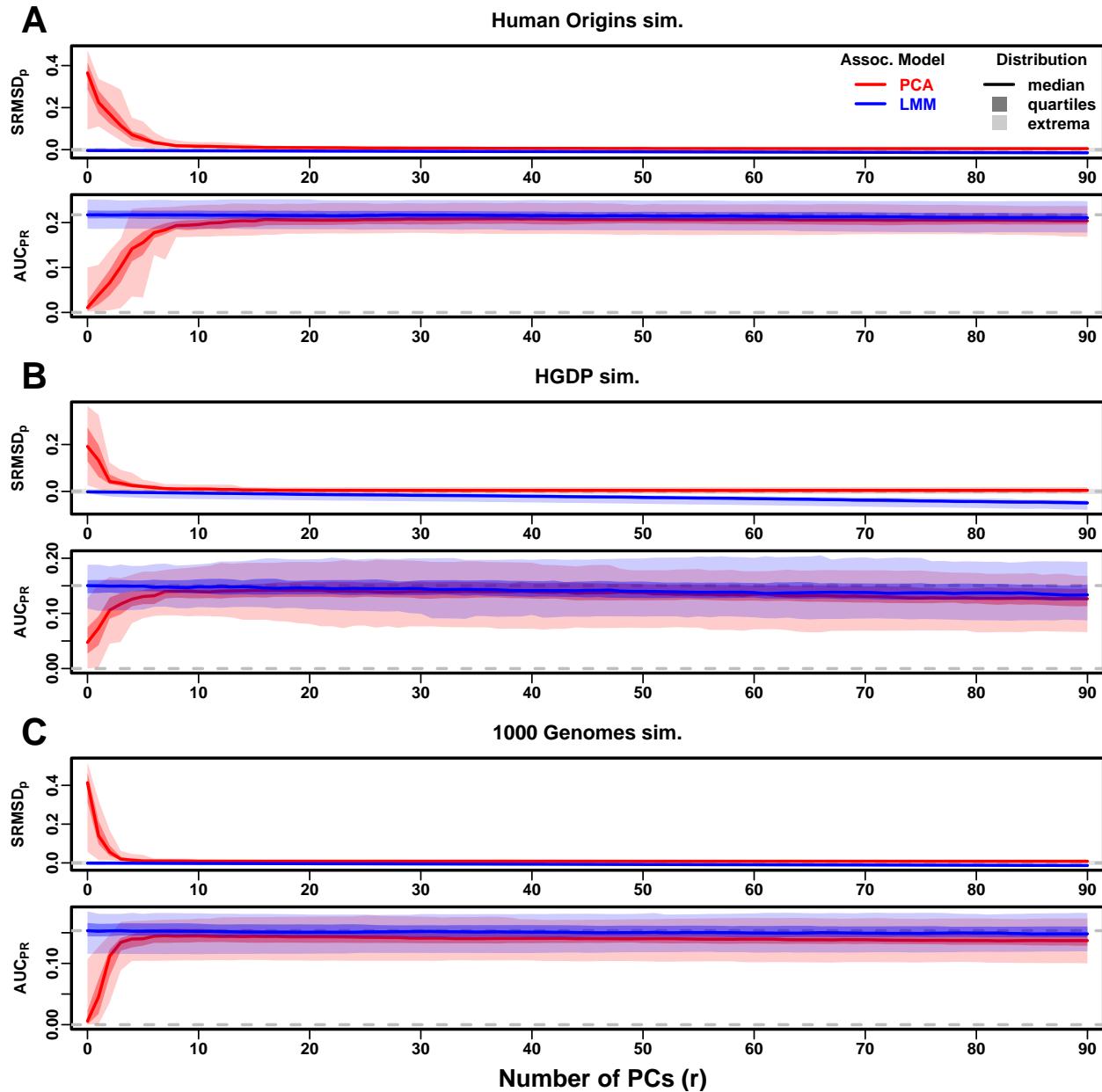


Figure S6: **Evaluations in subpopulation tree simulations fit to human data with RC traits.** Traits simulated from RC model, otherwise the same as Fig. 5.

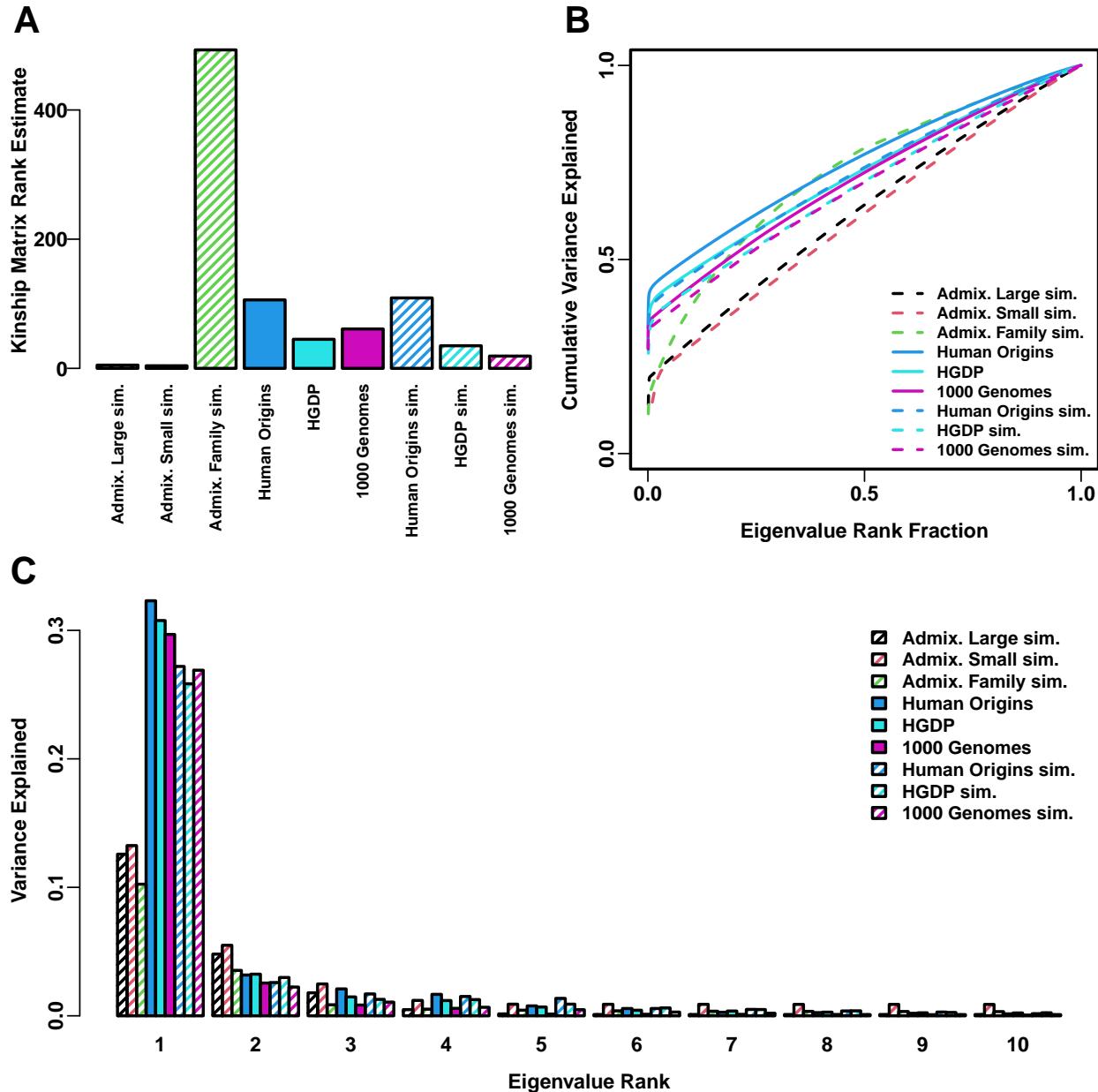


Figure S7: **Estimated relatedness dimensions of datasets.** **A.** Kinship matrix rank estimated with the Tracy-Widom test with $p < 0.01$. **B.** Cumulative variance explained versus eigenvalue rank fraction. **C.** Variance explained by first 10 eigenvalues.

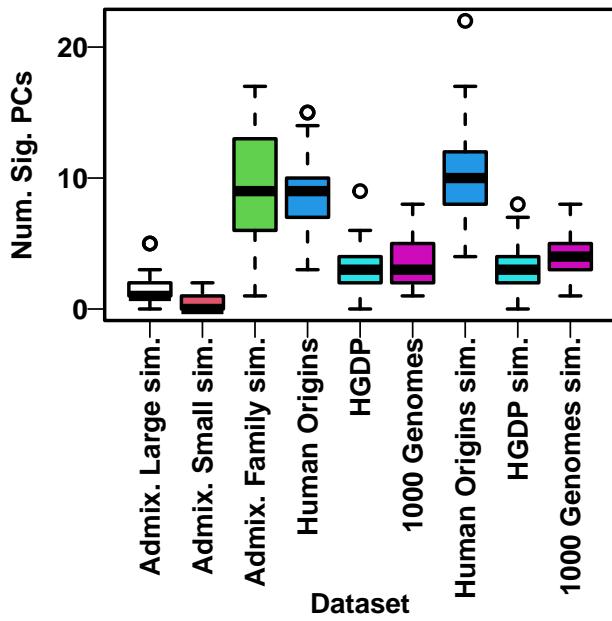


Figure S8: **Number of PCs significantly associated with traits.** PCs are tested using an ordinary linear regression sequentially, with the k th PC tested conditionally on the previous $k - 1$ PCs and the intercept. Q-values are estimated from the 90 p-values (one for each PC in a given dataset and 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).

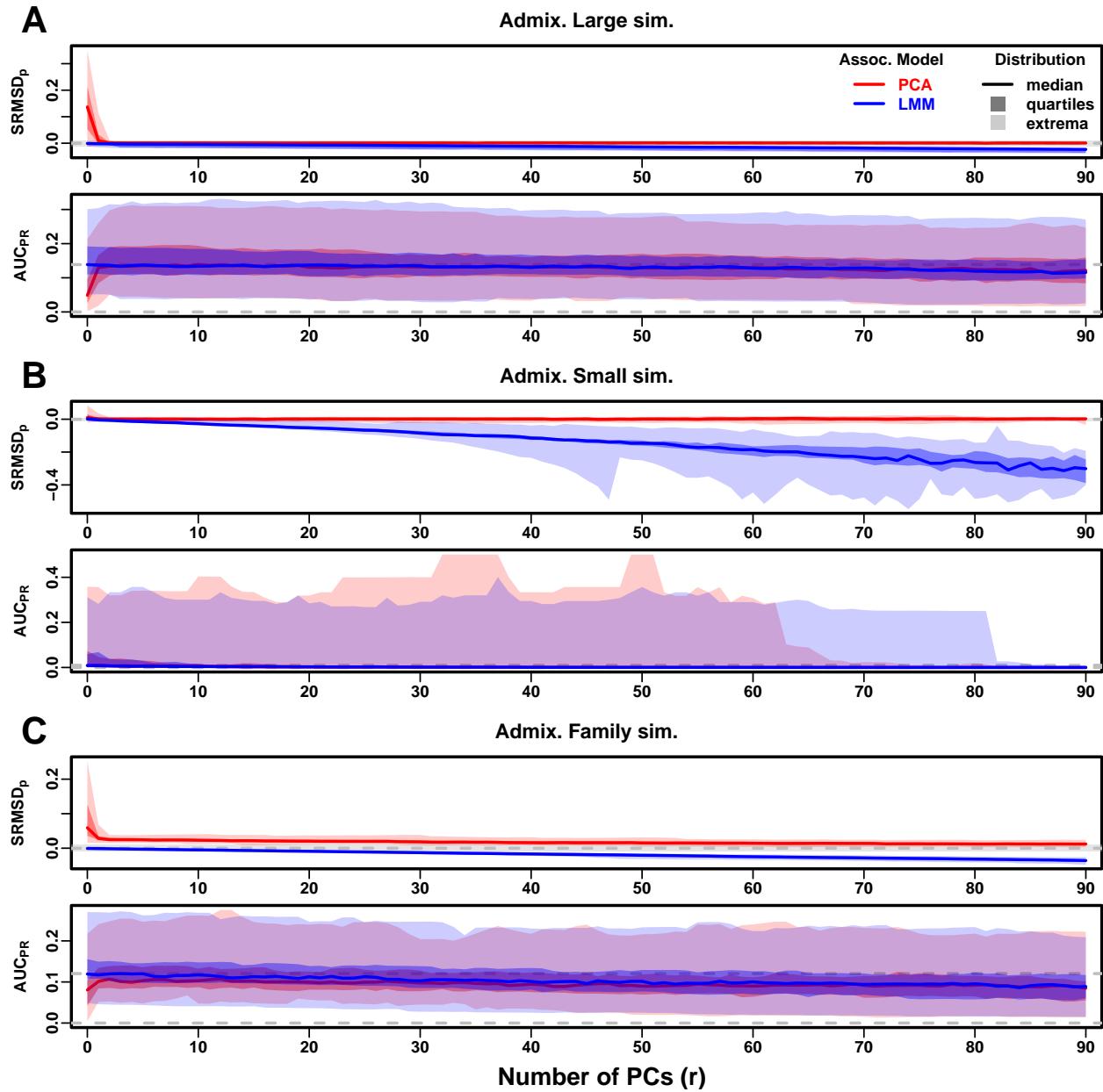


Figure S9: Evaluations in admixture simulations with FES traits, low heritability. Traits simulated using $h^2 = 0.3$, otherwise the same as Fig. 3.

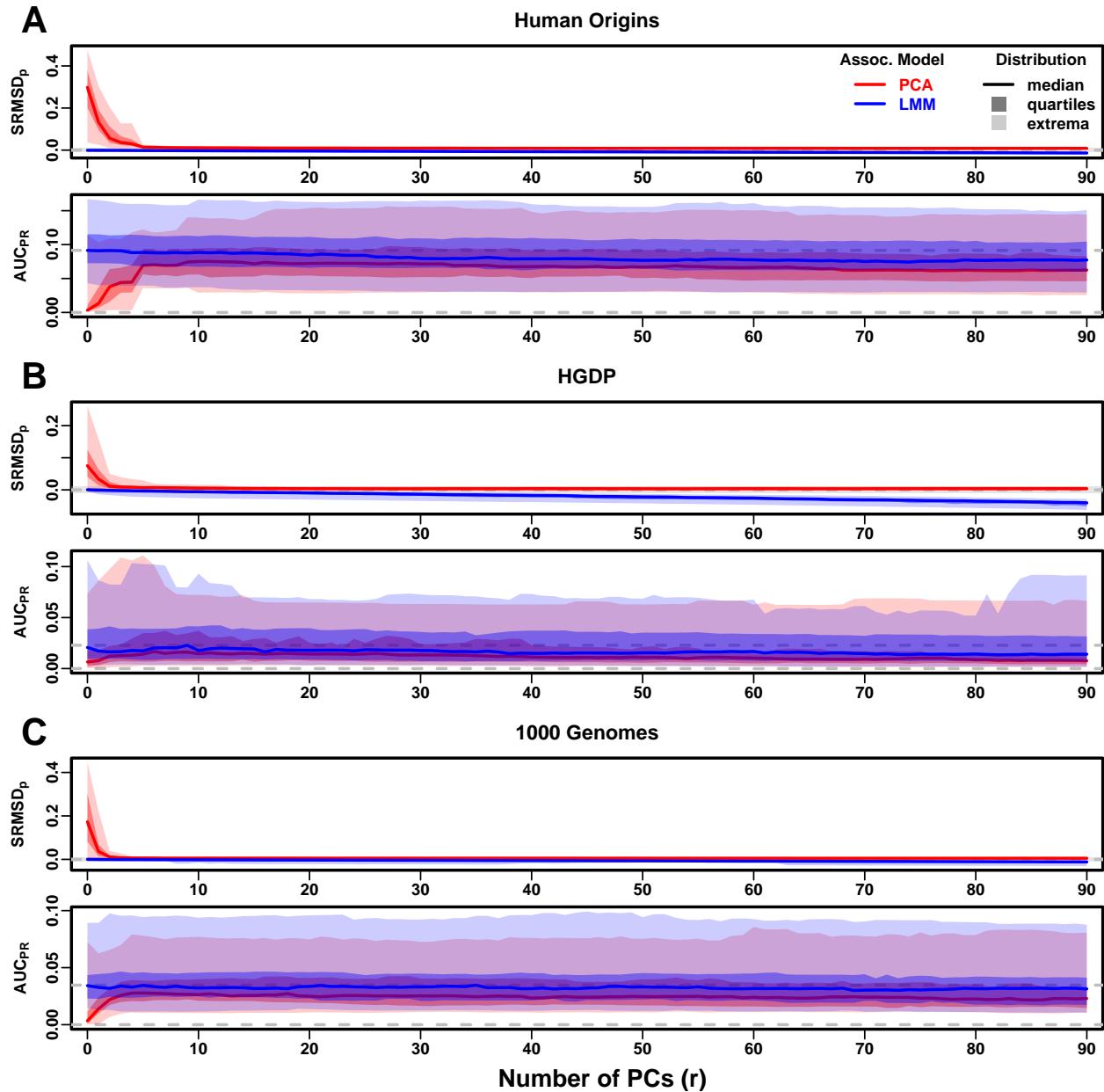


Figure S10: Evaluations in real human genotype datasets with FES traits, low heritability. Traits simulated using $h^2 = 0.3$, otherwise the same as Fig. 4.

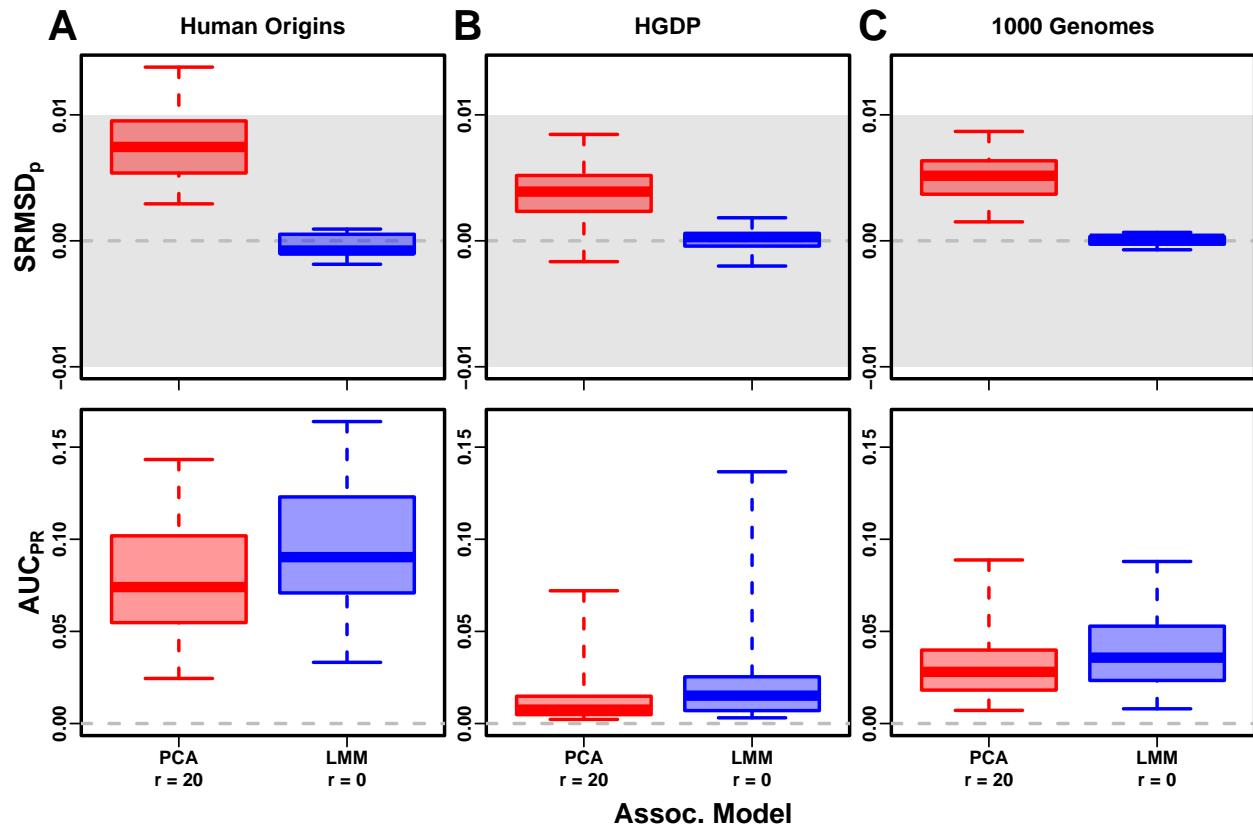


Figure S11: Evaluation in real datasets excluding 4th degree relatives, FES traits, low heritability. Traits simulated using $h^2 = 0.3$, otherwise the same as Fig. 7.

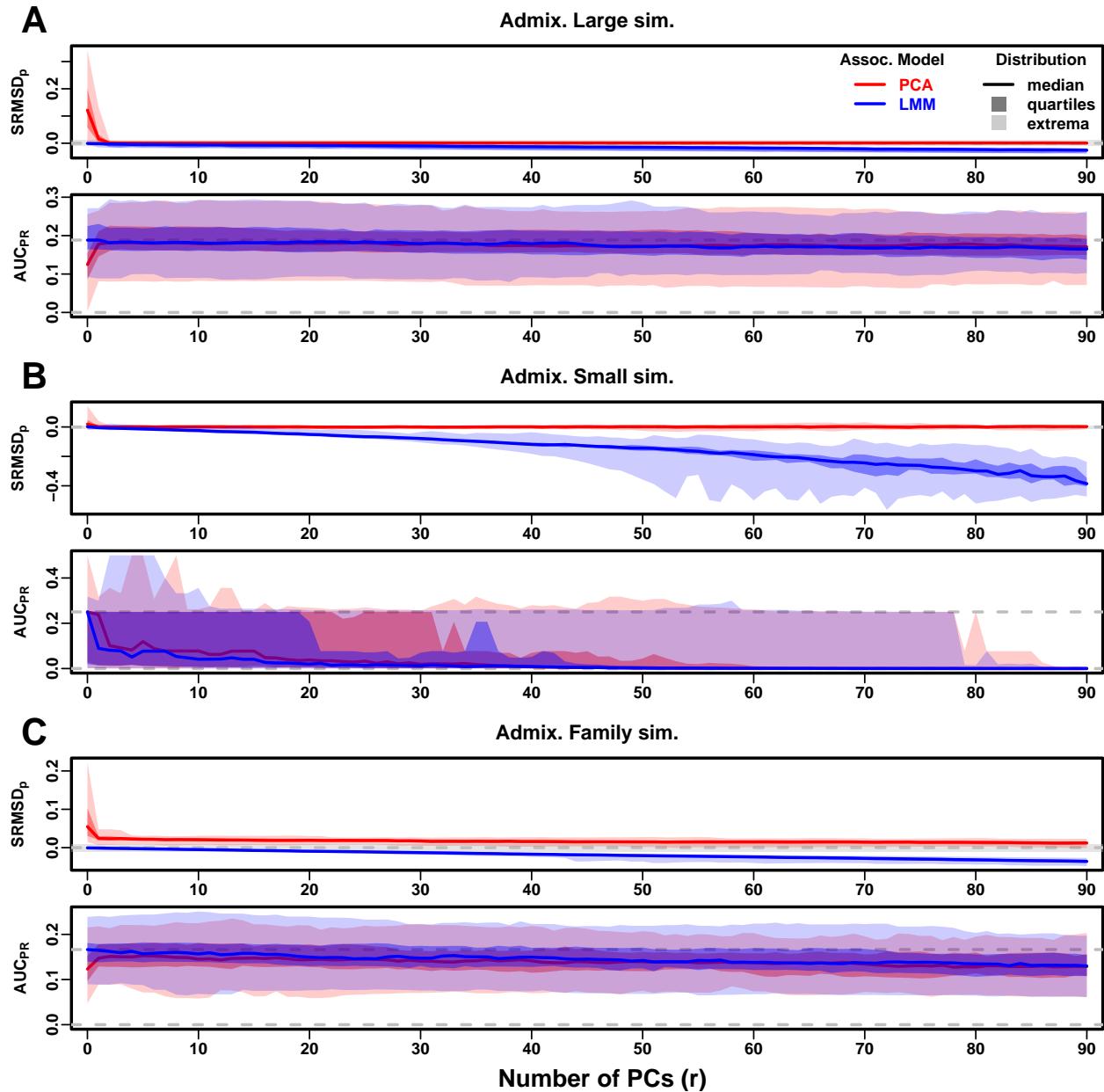


Figure S12: **Evaluations in admixture simulations with RC traits, low heritability.** Traits simulated using $h^2 = 0.3$, otherwise the same as Fig. S4.

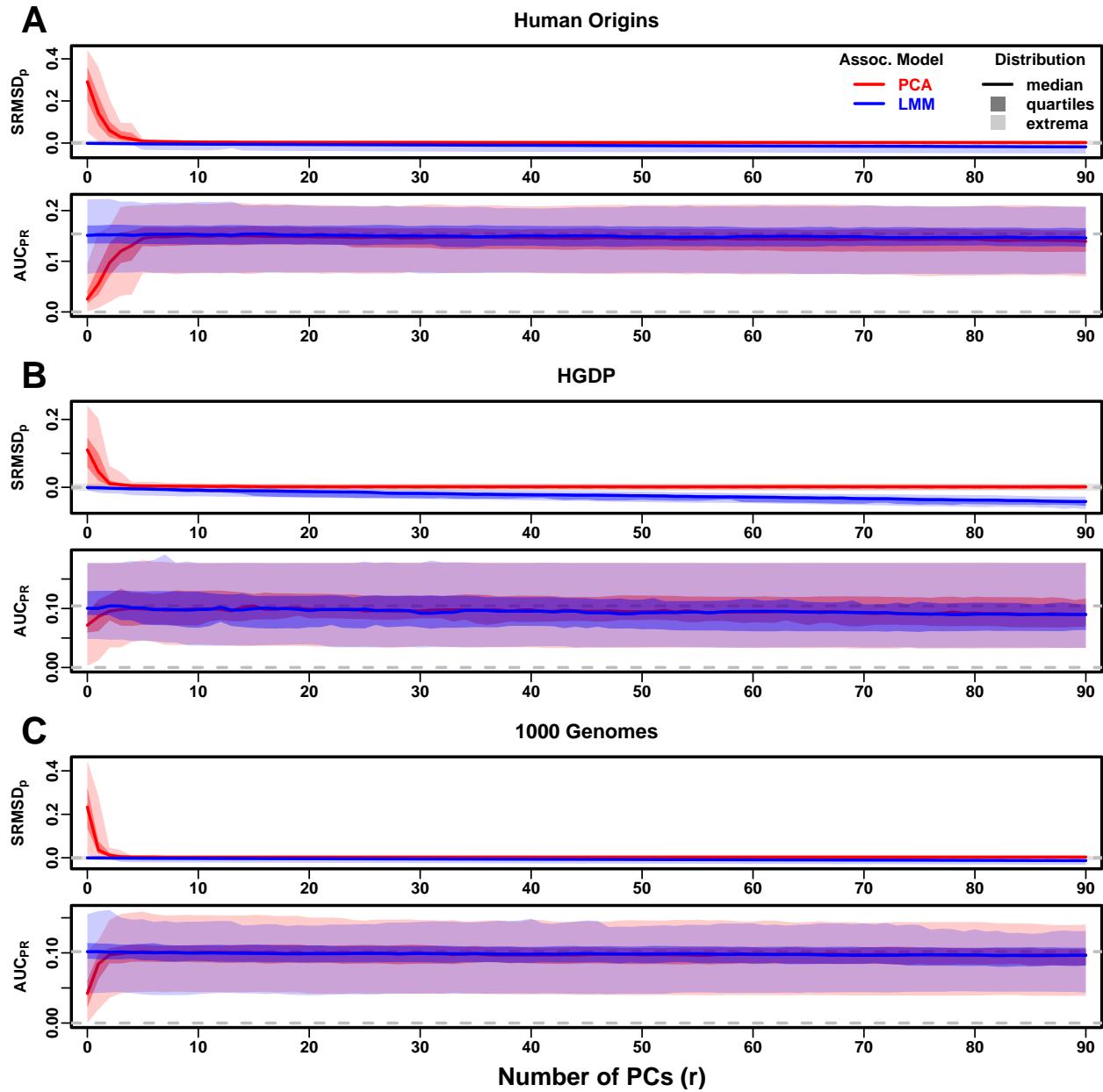


Figure S13: Evaluations in real human genotype datasets with RC traits, low heritability. Traits simulated using $h^2 = 0.3$, otherwise the same as Fig. S5.

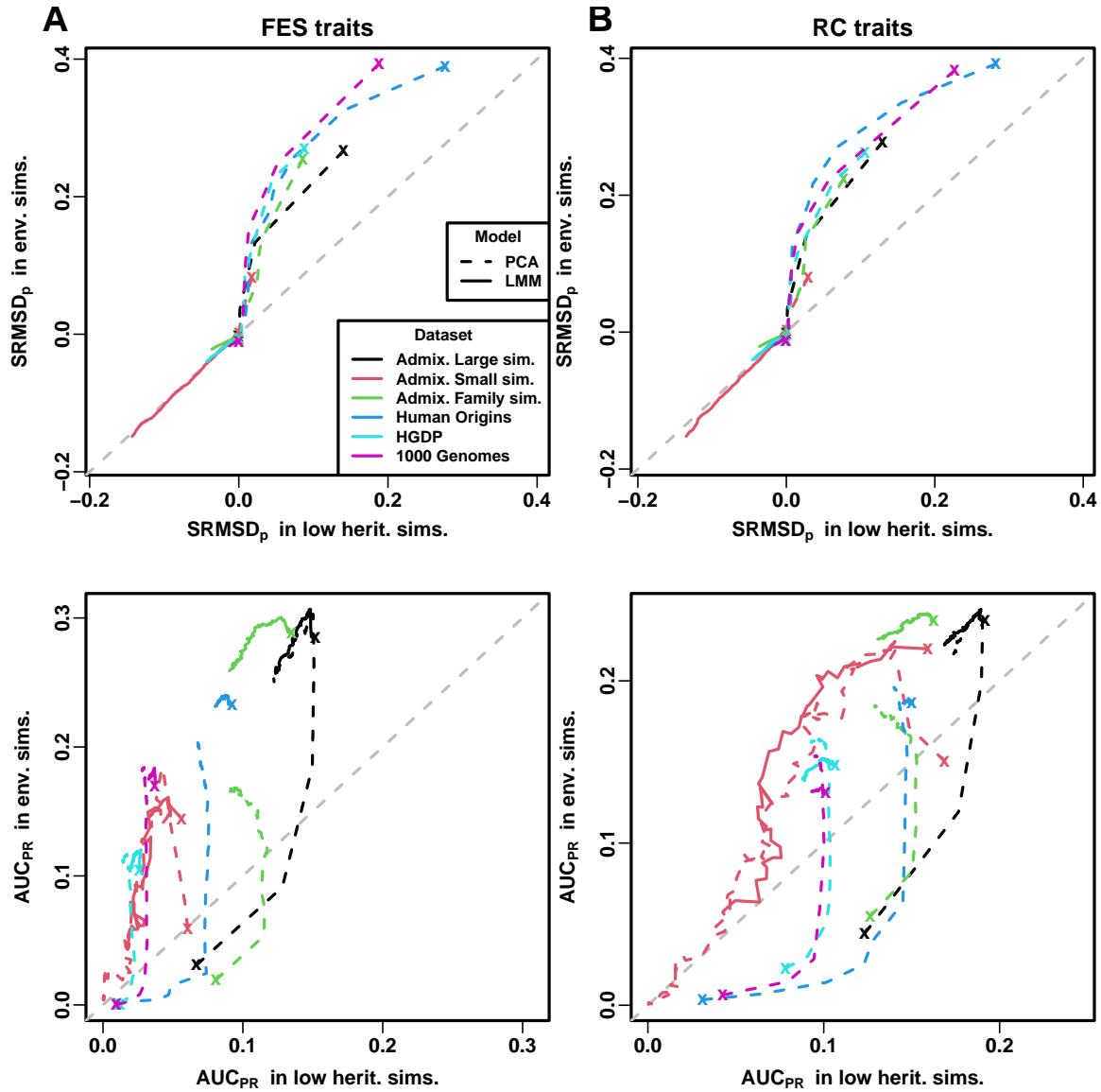


Figure S14: Comparison of performance in low heritability vs environment simulations. Each curve traces as the number of PCs r is increased from $r = 0$ (marked with an “ x ”) until $r = 90$ (unmarked end), on one axis is the mean value over replicates of either SRMSD_p or AUC_{PR} , for low heritability 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_{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.

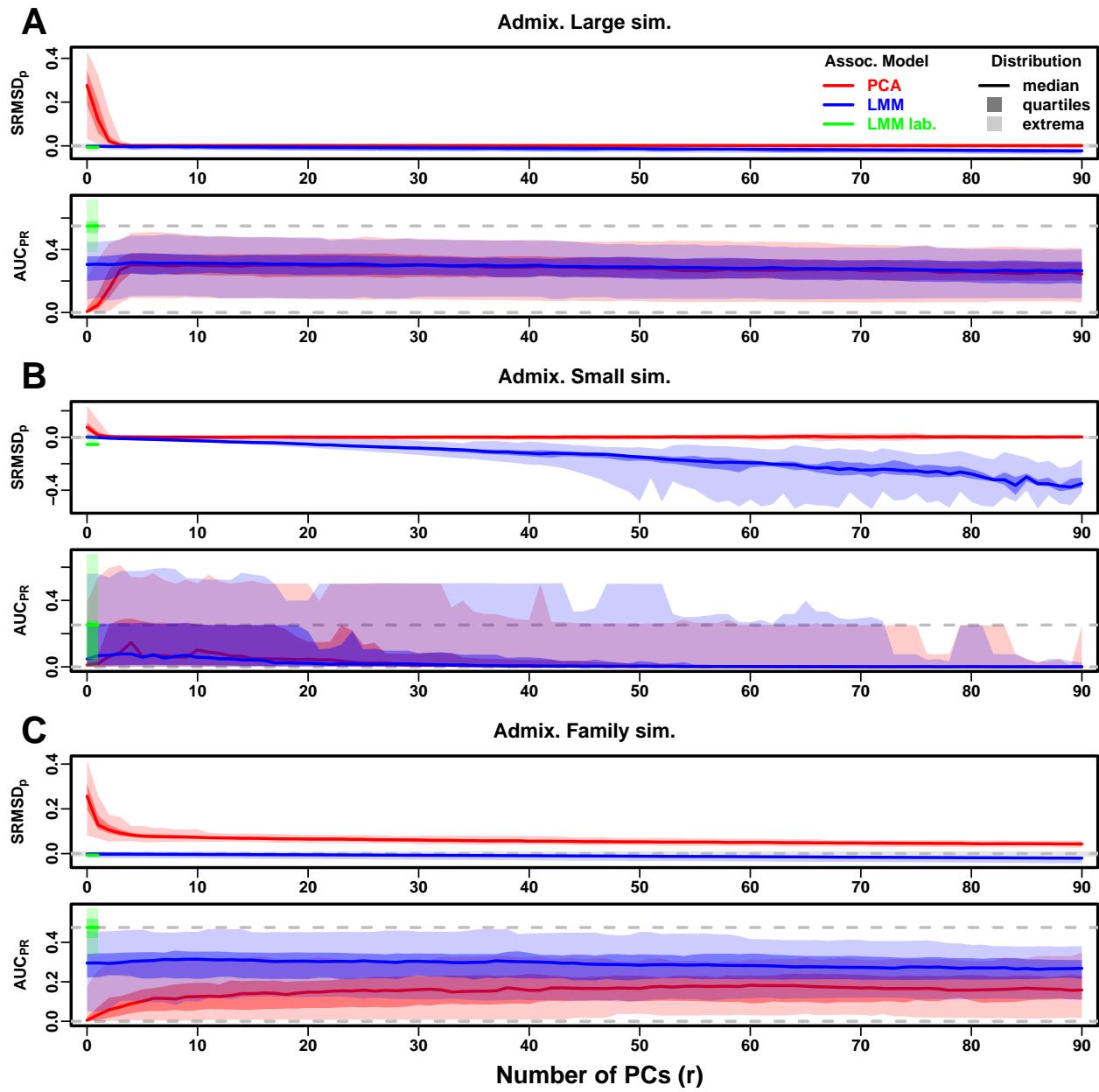


Figure S15: **Evaluations in admixture simulations with FES traits, environment.** Traits simulated with environment effects, otherwise the same as Fig. S9. “LMM lab.” was only tested with $r = 0$.

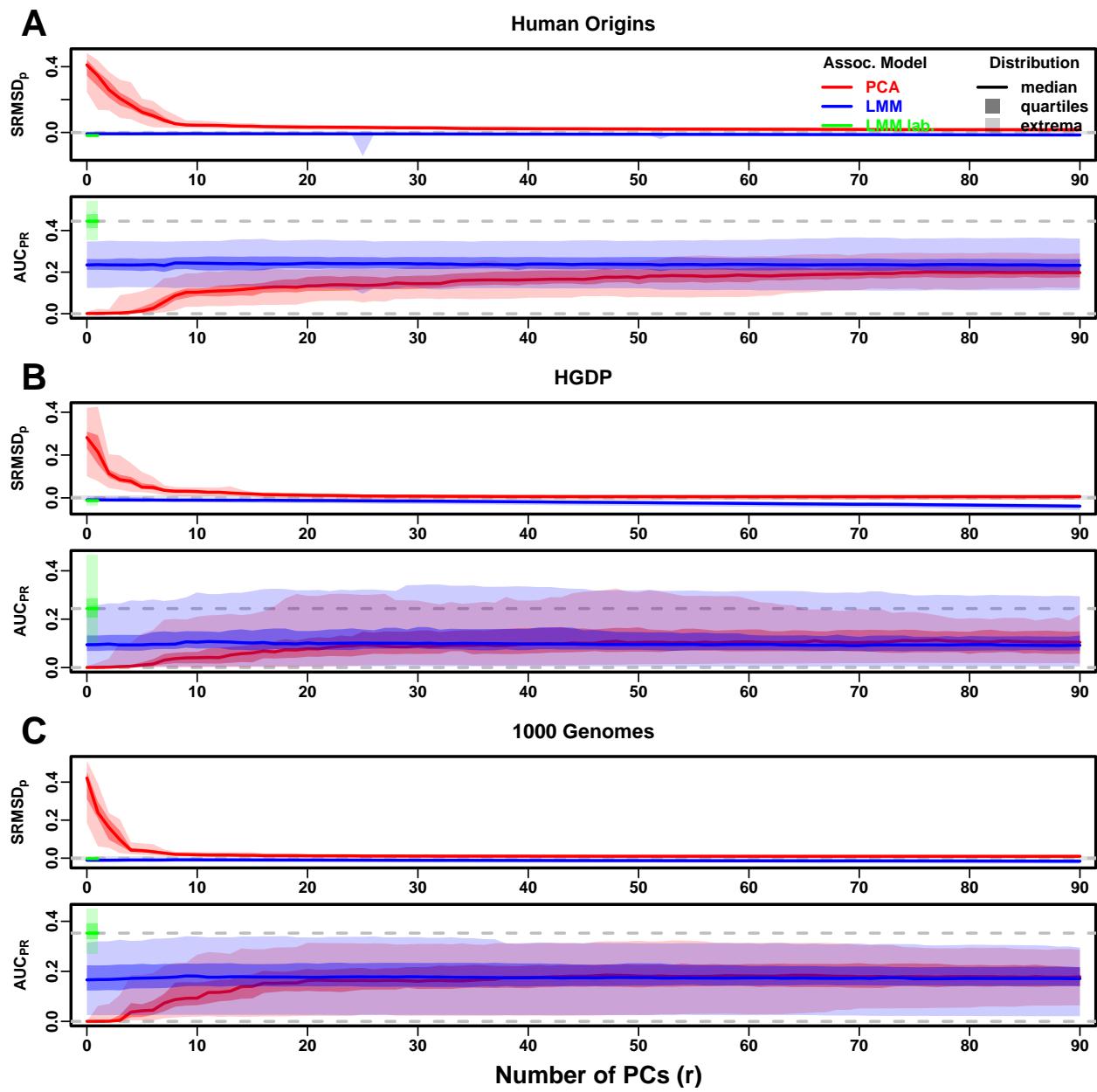


Figure S16: **Evaluations in real human genotype datasets with FES traits, environment.** Traits simulated with environment effects, otherwise the same as Fig. S10. “LMM lab.” was only tested with $r = 0$.

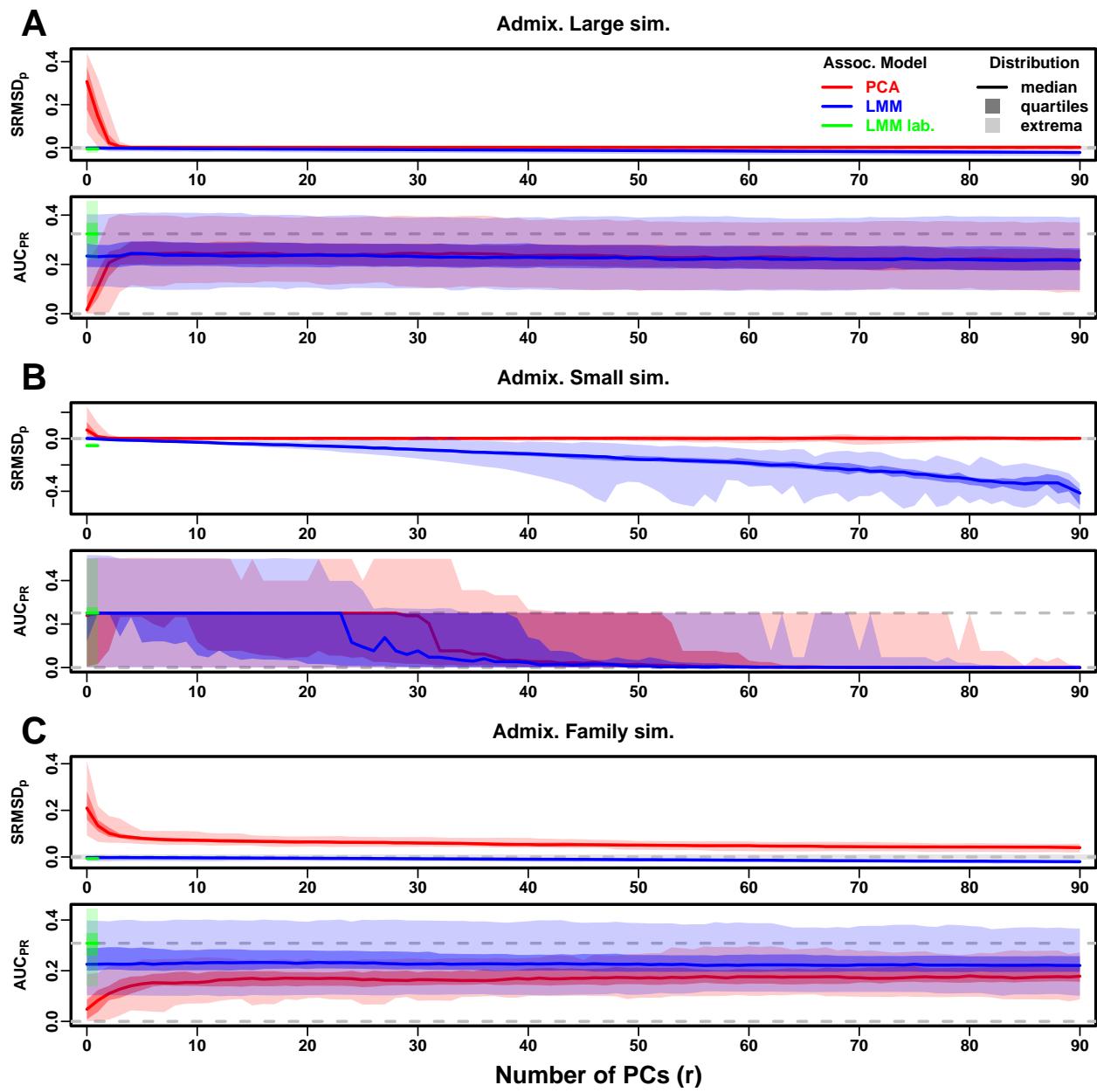


Figure S17: **Evaluations in admixture simulations with RC traits, environment.** Traits simulated with environment effects, otherwise the same as Fig. S12. “LMM lab.” was only tested with $r = 0$.

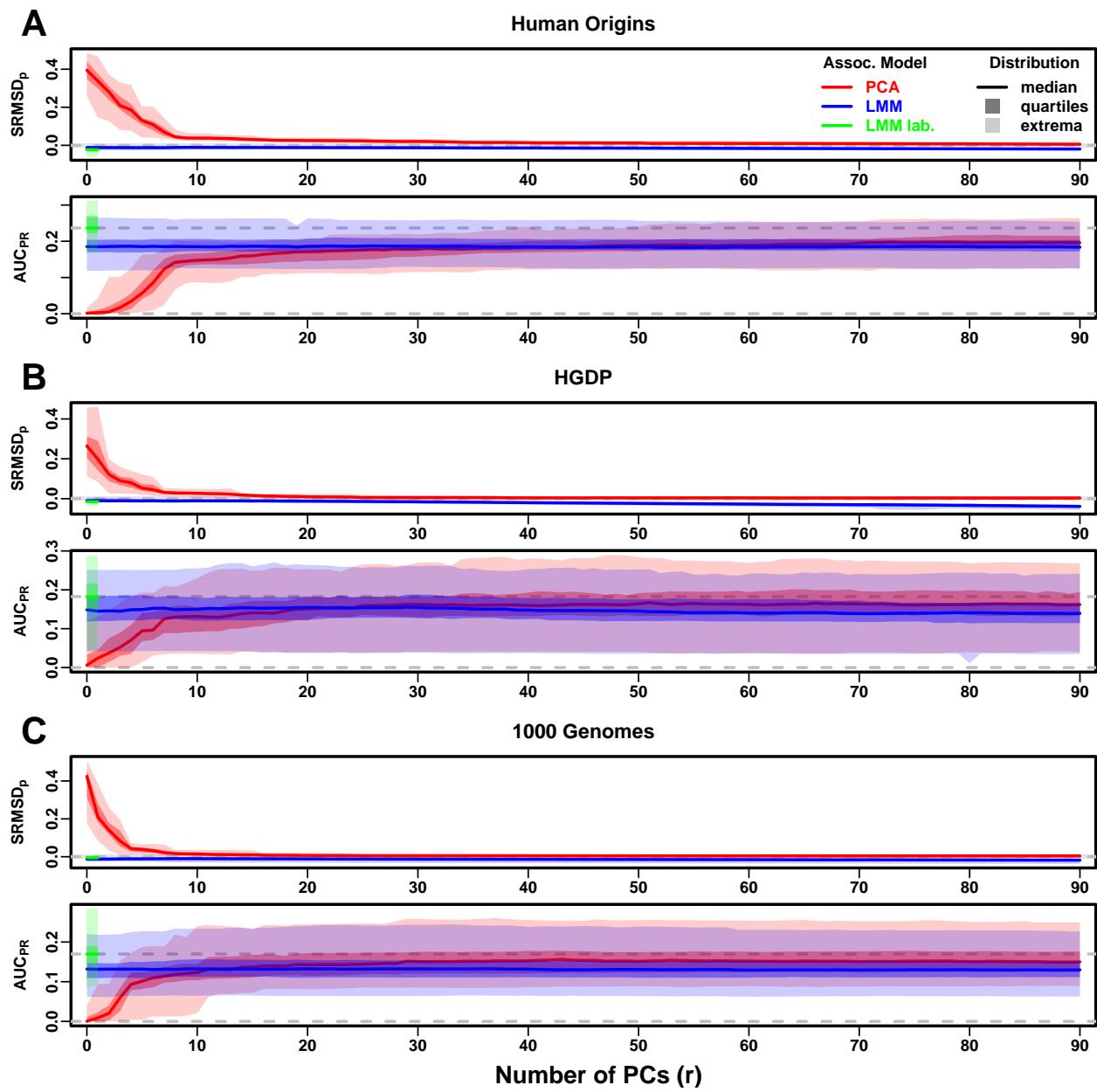


Figure S18: Evaluations in real human genotype datasets with RC traits, environment. Traits simulated with environment effects, otherwise the same as Fig. S13. “LMM lab.” was only tested with $r = 0$.

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

Dataset	Metric	Trait ^a	LMM $r = 0$ vs best r			Best r^c	PCA vs LMM $r = 0$		
			Cal. ^b	Best r^c	P-value ^d		Cal. ^b	P-value ^d	Best model ^e
Admix. Large sim.	$ \text{SRMSD}_p $	FES	True	0	1	62	True	0.00012*	LMM
Admix. Small sim.	$ \text{SRMSD}_p $	FES	True	0	1	3	True	0.27	Tie
Admix. Family sim.	$ \text{SRMSD}_p $	FES	True	0	1	90	False	3.9e-10*	LMM
Human Origins	$ \text{SRMSD}_p $	FES	True	0	1	81	True	3.9e-10*	LMM
HGDP	$ \text{SRMSD}_p $	FES	True	0	1	37	True	6.2e-09*	LMM
1000 Genomes	$ \text{SRMSD}_p $	FES	True	0	1	84	True	3.9e-10*	LMM
Admix. Large sim.	$ \text{SRMSD}_p $	RC	True	0	1	35	True	0.00094	Tie
Admix. Small sim.	$ \text{SRMSD}_p $	RC	True	0	1	3	True	0.087	Tie
Admix. Family sim.	$ \text{SRMSD}_p $	RC	True	0	1	90	False	4.1e-10*	LMM
Human Origins	$ \text{SRMSD}_p $	RC	True	0	1	75	True	0.00016*	LMM
HGDP	$ \text{SRMSD}_p $	RC	True	0	1	23	True	1.7e-05*	LMM
1000 Genomes	$ \text{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

^aFES: Fixed Effect Sizes, RC: Random Coefficients.

^bCalibrated: whether mean $|\text{SRMSD}_p| < 0.01$.

^cValue of r (number of PCs) with minimum mean $|\text{SRMSD}_p|$ or maximum mean AUC_{PR}.

^dWilcoxon paired 1-tailed test of distributions ($|\text{SRMSD}_p|$ or AUC_{PR}) between models in header. Asterisk marks significant value using Bonferroni threshold ($p < \alpha/n_{\text{tests}}$ with $\alpha = 0.01$ and $n_{\text{tests}} = 48$ is the number of tests in this table).

^eTie if no significant difference using Bonferroni threshold.

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

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

^aFES: Fixed Effect Sizes, RC: Random Coefficients.

^bCalibrated: whether mean $|\text{SRMSD}_p| < 0.01$.

^cValue of r (number of PCs) with minimum mean $|\text{SRMSD}_p|$ or maximum mean AUC_{PR}.

^dWilcoxon paired 1-tailed test of distributions ($|\text{SRMSD}_p|$ or AUC_{PR}) between models in header. Asterisk marks significant value using Bonferroni threshold ($p < \alpha/n_{\text{tests}}$ with $\alpha = 0.01$ and $n_{\text{tests}} = 72$ is the number of tests in this table).

^eTie if no significant difference using Bonferroni threshold; in last column, pairwise ties are specified and “Tie” is three-way tie.