

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, trees, and real multiethnic human datasets with simulated traits. We find that LMM
16 without PCs usually performs best, with the largest effects in family simulations and real human
17 datasets and traits without environment effects. Poor PCA performance on human datasets is
18 driven by large numbers of distant relatives more than the smaller number of closer relatives.
19 While PCA was known to fail on family data, we report strong effects of family relatedness in
20 genetically diverse human datasets, not avoided by pruning close relatives. Environment effects
21 driven by geography and ethnicity are better modeled with LMM including those labels instead
22 of PCs. This work better characterizes the severe limitations of PCA compared to LMM in
23 modeling the complex relatedness structures of multiethnic human data for association studies.

24 **Abbreviations:** PCA: principal component analysis; PCs: principal components; LMM: linear
25 mixed-effects model; FES: fixed effect sizes (trait model); RC: random coefficients (trait model);

26 MAF: minor allele frequency; WGS: whole genome sequencing.

27 1 Introduction

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

36 Genetic association with PCA consists of including the top eigenvectors of the population kin-
37 ship matrix as covariates in a generalized linear model [4–6]. These top eigenvectors are commonly
38 referred to as PCs in genetics [7], the convention adopted here, but in other fields PCs denote
39 the projections of loci onto eigenvectors [8]. The direct ancestor of PCA association is structured
40 association, in which inferred ancestry or admixture proportions are used as regression covariates
41 [9]. These models are deeply connected because PCs map to ancestry empirically [10, 11] and the-
42oretically [12–15], and they work as well as global ancestry in association studies but are estimated
43 more easily [6, 7, 10, 16]. Another closely related approach to PCA is nonmetric multidimensional
44 scaling [17]. PCs are also proposed for modeling environment effects that are correlated to ancestry,
45 for example, through geography [18–20]. The strength of PCA is its simplicity, which as covariates
46 can be readily included in more complex models, such as haplotype association [21] and polygenic
47 models [22]. However, PCA assumes that relatedness is low-dimensional (or low-rank), which may
48 limit its applicability. PCA is known to be inadequate for family data [7, 17, 23, 24], which is called
49 “cryptic relatedness” when it is unknown to the researchers, but no other troublesome cases have
50 been confidently identified. Recent work has focused on developing more scalable versions of the
51 PCA algorithm [25–29]. PCA remains a popular and powerful approach for association studies.

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

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

66 An LMM variant that incorporates PCs as fixed covariates is tested thoroughly in our work.
67 Since PCs are the top eigenvectors of the same kinship matrix estimate used in modern LMMs
68 [1, 19, 48, 49], then population structure is modeled twice in an LMM with PCs. However, some
69 previous work has found the apparent redundancy of an LMM with PCs beneficial [19, 24, 50],
70 while others did not [48, 51], and the approach continues to be used [52, 53] though not always [54].
71 (Recall that early LMMs used kinship to model family relatedness only, so population structure had
72 to be modeled separately, in practice as admixture fractions instead of PCs [16, 17, 30].) The LMM
73 with PCs (vs no PCs) is believed to help better model loci that have experienced selection [24, 33]
74 and environment effects correlated with genetics [19].

75 LMM and PCA are closely related models [1, 19, 48, 49], so similar performance is expected
76 particularly under low-dimensional relatedness. Direct comparisons have yielded mixed results, with
77 several studies finding superior performance for LMM (notably from papers promoting advances in
78 LMMs) while many others report comparable performance (Table 1). No papers find that PCA
79 outperforms LMM decisively, although PCA occasionally performs better in isolated and artificial
80 cases or individual measures (often with unknown significance). Previous studies were generally

81 divided into those that employed simulated versus real genotypes (only two studies used both). The
 82 simulated genotype studies, which tended to have low dimensionalities and differentiation (F_{ST}),
 83 were more likely to report ties or mixed results (6/8), whereas real genotypes tended to clearly favor
 84 LMMs (9/11). Similarly, 10/12 papers with quantitative traits favor LMMs, whereas 6/9 papers
 85 with case-control traits gave ties or mixed results (the only factor we do not explore). Additionally,
 86 although all previous evaluations measured type I error (or proxies such as inflation factors or QQ
 87 plots), a large fraction (6/17) did not measure power (including proxies such as ROC curves), and
 88 only four used more than one number of PCs for PCA. Lastly, no consensus has emerged as to why
 89 LMM might outperform PCA or vice versa [24, 38, 49, 59], or which features of the real datasets
 90 are critical for the LMM advantage other than cryptic relatedness, resulting in unclear guidance
 91 for using PCA. Hence, our work includes real and simulated genotypes with higher dimensionalities

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: Tree; F: Family.

^bModel dimensionality (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.

92 and differentiation matching that of multiethnic human cohorts, we vary the number of PCs, and
93 measure robust proxies for type I error control and calibrated power.

94 In this work, we evaluate the PCA and LMM association models under various numbers of
95 PCs (included in LMM too). We use genotype simulations (admixture, family, and tree models)
96 and three real datasets: the 1000 Genomes Project [60, 61], the Human Genome Diversity Panel
97 (HGDP) [62–64], and Human Origins [65–68]. We simulate quantitative traits from two models:
98 fixed effect sizes (FES; coefficients inverse to allele frequency) that matches real data [52, 69, 70] and
99 corresponds to high pleiotropy and strong balancing selection [71] and strong negative selection [52,
100 70], which are appropriate assumptions for diseases; and random coefficients (RC; independent of
101 allele frequency) that corresponds to neutral traits [52, 71]. LMM without PCs consistently performs
102 best in simulations without environment, and greatly outperforms PCA in the family simulation
103 and in all real datasets. The tree simulations do not recapitulate the real data results, suggesting
104 that family relatedness in real data is the reason for poor PCA performance. Lastly, removing up
105 to 4th degree relatives in the real datasets recapitulates poor PCA performance, showing that the
106 more numerous distant relatives explain the result, and suggesting that PCA is generally not an
107 appropriate model for real data. We find that both LMM and PCA are able to model environment
108 effects correlated with genetics, and LMM with PCs gains a small advantage in this setting only, but
109 direct modeling of environment performs much better. All together, we find that LMMs without PCs
110 are generally a preferable association model, and present novel simulation and evaluation approaches
111 to measure the performance of these and other genetic association approaches.

112 2 Materials and Methods

113 2.1 The complex trait model and PCA and LMM approximations

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

117 for a quantitative (continuous) trait is:

$$118 \quad \mathbf{y} = \mathbf{1}\alpha + \mathbf{X}'\boldsymbol{\beta} + \mathbf{Z}'\boldsymbol{\eta} + \boldsymbol{\epsilon}, \quad (1)$$

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

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

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

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

127 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
 128 $n \times r$ matrix of PCs, $\boldsymbol{\gamma}_r$ is the length- r vector of PC coefficients, \mathbf{s} is a length- n vector of random
 129 effects, $\boldsymbol{\Phi}^T = (\varphi_{jk}^T)$ is the $n \times n$ kinship matrix conditioned on the ancestral population T , and σ_s^2
 130 is a variance factor. Both models condition the regression of the focal locus i on an approximation
 131 of the total polygenic effect $\mathbf{X}'\boldsymbol{\beta}$ with the same covariance structure, which is parametrized by the
 132 kinship matrix. Under the kinship model, genotypes are random variables obeying

$$133 \quad \mathbb{E}[\mathbf{x}_i|T] = 2p_i^T \mathbf{1}, \quad \text{Cov}(\mathbf{x}_i|T) = 4p_i^T(1 - p_i^T)\boldsymbol{\Phi}^T, \quad (4)$$

134 where p_i^T is the ancestral allele frequency of locus i [1, 72–74]. Assuming independent loci, the
 135 covariance of the polygenic effect is

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

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

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

140 which follows from the affine transformation property of multivariate normal distributions. There-
 141 fore, the PCA term $\mathbf{U}_r\gamma_r$ can be derived from the above equation under the additional assumption
 142 that the kinship matrix has dimensionality r and the coefficients γ_r are fit without constraints. In
 143 contrast, the LMM uses all eigenvectors, while effectively shrinking their coefficients γ_{LMM} as all
 144 random effects models do, although these parameters are marginalized [1, 19, 48, 49]. PCA has
 145 more parameters than LMM, so it may overfit more: ignoring the shared terms in Eqs. (2) and (3),
 146 PCA fits r parameters (length of γ), whereas LMMs fit only one (σ_s^2).

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

$$149 \mathbf{X}_S = \left(\frac{x_{ij} - 2\hat{p}_i^T}{\sqrt{4\hat{p}_i^T(1-\hat{p}_i^T)}} \right), \quad \hat{\Phi}^T = \frac{1}{m} \mathbf{X}_S' \mathbf{X}_S, \quad (5)$$

150 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,
 151 this kinship estimator has a complex bias that differs for every individual pair, which arises due
 152 to the use of this estimated \hat{p}_i^T [32, 75]. Nevertheless, in PCA and LMM these biased estimates
 153 perform as well as unbiased ones [76].

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

158 LMM association was performed using GCTA [35, 43]. Its kinship estimator equals Eq. (5).

159 PCs were calculated using GCTA from its kinship estimate. Association significance is evaluated
160 with a score test. GCTA with large numbers of PCs (small simulation only) had convergence and
161 singularity errors in some replicates, which were treated as missing data.

162 **2.2 Simulations**

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

168 **2.2.1 Genotype simulation from the admixture model**

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

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

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

181 As a brief aside, we prove that global ancestry proportions as covariates is equivalent in expec-
182 tation to using PCs under the admixture model. Note that the latent space of \mathbf{X} , given by (π_{ij}) ,
183 has K dimensions (number of columns of $\mathbf{Q} = (q_{ju})$), so the top K PCs span this space. Since
184 associations include an intercept term ($\mathbf{1}\alpha$ in Eq. (2)), estimated PCs are orthogonal to $\mathbf{1}$ (note
185 $\hat{\Phi}^T \mathbf{1} = \mathbf{0}$ because $\mathbf{X}_S \mathbf{1} = \mathbf{0}$), and the sum of rows of \mathbf{Q} sums to one, then only $K - 1$ PCs (plus
186 intercept) are needed to span the latent space of this admixture model.

187 2.2.2 Genotype simulation from random admixed families

188 We simulated a pedigree with admixed founders, no close relative pairings, assortative mating based
189 on a 1D geography (to preserve admixture structure), random family sizes, and arbitrary numbers
190 of generations (20 here). This simulation is implemented in the R package `simfam`. Generations
191 are drawn iteratively. Generation 1 has $n = 1000$ individuals from the above admixture simulation
192 ordered by their 1D geography. Local kinship measures pedigree relatedness; in the first generation,
193 everybody is locally unrelated and outbred. Individuals are randomly assigned sex. In the next
194 generation, individuals are paired iteratively, removing random males from the pool of available
195 males and pairing them with the nearest available female with local kinship $< 1/4^3$ (stay unpaired
196 if there are no matches), until there are no more available males or females. Let $n = 1000$ be the
197 desired population size, $n_m = 1$ the minimum number of children and n_f the number of families
198 (paired parents) in the current generation, then the number of additional children (beyond the
199 minimum) is drawn from $\text{Poisson}(n/n_f - n_m)$. Let δ be the difference between desired and current
200 population sizes. If $\delta > 0$, then δ random families are incremented by 1. If $\delta < 0$, then $|\delta|$ random
201 families with at least $n_m + 1$ children are decremented by 1. If $|\delta|$ exceeds the number of families, all
202 families are incremented or decremented as needed and the process is iterated. Children are assigned
203 sex randomly, and are reordered by the average coordinate of their parents. Children draw alleles
204 from their parents independently per locus. A new random pedigree is drawn for each replicate, as

205 well as new founder genotypes from the admixture model.

206 **2.2.3 Genotype simulation from a tree model**

207 This model draws subpopulations allele frequencies from a hierarchical model parametrized by a
208 tree, which is also implemented in `bnpsd` and relies on `ape` for general tree data structures and
209 methods [79]. The ancestral population T is the root, and each node is a subpopulation S_w indexed
210 arbitrarily. Each edge between S_w and its parent population P_w has an inbreeding coefficient $f_{S_w}^{P_w}$.
211 p_i^T are drawn from a given distribution (constructed to mimic each real dataset in Appendix A).
212 Given the allele frequencies $p_i^{P_w}$ of the parent population, S_w 's allele frequencies are drawn from:

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

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

216 **2.2.4 Fitting tree to real data**

217 We developed new methods to fit trees to real data based on unbiased kinship estimates from
218 `popkin`, implemented in `bnpsd`. A tree with given inbreeding edges $f_{S_w}^{P_w}$ gives rise to a coancestry
219 matrix ϑ_{uv}^T for a subpopulation pair (S_u, S_v) , and the goal is to recover the inbreeding edges from
220 coancestry estimates. Coancestry values are total inbreeding coefficients of the MRCA population
221 of each subpopulation pair. Therefore, we calculate $f_{S_w}^T$ for every S_w recursively from the root as
222 follows. Nodes with parent $P_w = T$ are already as desired. Given $f_{P_w}^T$, the desired $f_{S_w}^T$ is calculated
223 via the additive edge δ_w [32]:

$$224 f_{S_w}^T = f_{P_w}^T + \delta_w, \quad \delta_w = f_{S_w}^{P_w} (1 - f_{P_w}^T). \quad (6)$$

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

227

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

228 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.

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

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

232

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

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

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

234 Topology is estimated with hierarchical clustering using the weighted pair group method with
 235 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
 236 the monotonic relationship between node depth and coancestry (Eq. (7)). This algorithm recovers
 237 the true topology from the true coancestry values, and performs well for estimates from genotypes.

238 To estimate tree edge lengths, first δ_w are estimated from $\hat{\vartheta}_{uv}^T$ and the topology using Eq. (7) and
 239 non-negative least squares linear regression [81] (implemented in `nnls` [82]) to yield non-negative
 240 δ_w , and $f_{S_w}^{P_w}$ are calculated from δ_w by reversing Eq. (6). To account for small biases in coancestry
 241 estimation, an intercept term δ_0 is included ($I_0(u, v) = 1$ for all u, v), and when converting δ_w to
 242 $f_{S_w}^{P_w}$, δ_0 is treated as an additional edge to the root, but is ignored when drawing allele frequencies
 243 from the tree.

²⁴⁴ **2.2.5 Trait Simulation**

²⁴⁵ Traits are simulated from the quantitative trait model of Eq. (1), with novel bias corrections for
²⁴⁶ simulating the desired heritability from real data relying on the unbiased kinship estimator `popkin`
²⁴⁷ [32]. This simulation is implemented in the R package `simtrait`. All simulations have a fixed
²⁴⁸ narrow-sense heritability of h^2 , a variance proportion due to environment effects σ_η^2 , and residuals
²⁴⁹ 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
²⁵⁰ determines the average coefficient size, is chosen with the formula $m_1 = \text{round}(nh^2/8)$, which
²⁵¹ empirically balances power well with varying n and h^2 . The set of causal loci C is drawn anew for
²⁵² each replicate, from loci with MAF ≥ 0.01 to avoid rare causal variants (inappropriate for PCA
²⁵³ and LMM). Letting $v_i^T = p_i^T (1 - p_i^T)$, the effect size of locus i equals $2v_i^T \beta_i^2$, its contribution of the
²⁵⁴ trait variance [83]. Under the *fixed effect sizes* (FES) model, initial causal coefficients are

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

²⁵⁵ for known p_i^T ; otherwise v_i^T is replaced by the unbiased estimator [32] $\hat{v}_i^T = \hat{p}_i^T (1 - \hat{p}_i^T) / (1 - \bar{\varphi}^T)$,
²⁵⁶ where $\bar{\varphi}^T$ is the mean kinship estimated with `popkin`. Each causal locus is multiplied by -1 with
²⁵⁷ probability 0.5. Alternatively, under the *random coefficients* (RC) model, initial causal coefficients
²⁵⁸ are drawn independently from $\beta_i \sim \text{Normal}(0, 1)$. For both models, the initial genetic variance is
²⁵⁹ $\sigma_0^2 = \sum_{i \in C} 2v_i^T \beta_i^2$, replacing v_i^T with \hat{v}_i^T for unknown p_i^T (so σ_0^2 is an unbiased estimate), so we
²⁶⁰ multiply every initial β_i by $\frac{h}{\sigma_0}$ to have the desired heritability. Lastly, for known p_i^T , the intercept
²⁶¹ 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
²⁶² the trait covariance (for the same reason the standard kinship estimator in Eq. (5) is biased), which
²⁶³ is avoided with

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

²⁶⁴ Simulations optionally included multiple environment group effects, similarly to previous models
²⁶⁵ [19, 34], as follows. Each independent environment i has predefined groups, and each group g has
²⁶⁶ random coefficients drawn independent from $\eta_{gi} \sim \text{Normal}(0, \sigma_{\eta i}^2)$ where $\sigma_{\eta i}^2$ is a specified variance

267 proportion for environment i . \mathbf{Z} has individuals along columns and environment-groups along rows,
 268 and it contains indicator variables: 1 if the individual belongs to the environment-group, 0 otherwise.

269 We performed trait simulations with the following variance parameters (Table 2): *high heritability*
 270 used $h^2 = 0.8$ and no environment effects; *low heritability* used $h^2 = 0.3$ and no environment
 271 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
 272 real genotype datasets, the groups are the subpopulation (environment 1) and sub-subpopulation
 273 (environment 2) labels given (see next subsection). For simulated genotypes, we created these labels
 274 by grouping by the index j (geographical coordinate) of each simulated individual, assigning group
 275 $g = \text{ceiling}(jk_i/n)$ where k_i is the number of groups in environment i , and we selected $k_1 = 5$ and
 276 $k_2 = 25$ 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

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 dimensionality of 20 generation pedigree structure. For real datasets, lower range is continental subpopulations, upper range is number of fine-grained subpopulations.

^b $m_1 = \text{round}(nh^2/8)$ to balance power across datasets, shown for $h^2 = 0.8$ only.

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

277 **2.3 Real human genotype datasets**

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

283 We obtained the full (including non-public) Human Origins by contacting the authors and
284 agreeing to their usage restrictions. The Pacific data [68] was obtained separately from the rest [66,
285 67], and datasets were merged using the intersection of loci. We removed ancient individuals, and
286 individuals from singleton and non-native subpopulations. Non-autosomal loci were removed. Our
287 analysis of the whole-genome sequencing (WGS) version of HGDP [64] was restricted to autosomal
288 biallelic SNP loci with filter “PASS”. Our analysis of the high-coverage NYGC version of 1000
289 Genomes [85] was restricted to autosomal biallelic SNP loci with filter “PASS”.

290 Since our evaluations assume uncorrelated loci, we filtered each real dataset with `plink2` using
291 parameters “`--indep-pairwise 1000kb 0.3`”, which iteratively removes loci that have a greater
292 than 0.3 squared correlation coefficient with another locus that is within 1000kb, stopping until no
293 such loci remain. Since all real datasets have numerous rare variants, while PCA and LMM are not
294 able to detect associations involving rare variants, we removed all loci with $MAF < 0.01$. Lastly,
295 only HGDP had loci with over 10% missingness removed, as they were otherwise 17% of remaining
296 loci (for Human Origins and 1000 Genomes they were under 1% of loci so they were not removed).
297 Kinship dimensionality and eigenvalues were calculated from `popkin` kinship estimates. Eigenvalues
298 were assigned p-values with `twstats` of the Eigensoft package [7], and dimensionality was estimated
299 as the largest number of consecutive eigenvalue from the start that all satisfy $p < 0.01$ (p-values
300 did not increase monotonically). For the evaluation with close relatives removed, each dataset was
301 filtered with `plink2` with option “`--king-cutoff`” with cutoff $0.02209709 (= 2^{-11/2})$ for removing
302 up to 4th degree relatives using KING-robust [86], and $MAF < 0.01$ filter is reapplied (Table S1).

303 **2.4 Evaluation of performance**

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

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

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

312 where $m_0 = m - m_1$ is the number of null (non-causal) loci, here i indexes null loci only, $p_{(i)}$ is
 313 the i th ordered null p-value, $u_i = (i - 0.5)/m_0$ is its expectation, p_{median} is the median observed
 314 null p-value, $u_{\text{median}} = \frac{1}{2}$ is its expectation, and sgn is the sign function (1 if $u_{\text{median}} \geq p_{\text{median}}$,
 315 -1 otherwise). Thus, $\text{SRMSD}_p = 0$ corresponds to calibrated p-values, $\text{SRMSD}_p > 0$ indicate anti-
 316 conservative p-values, and $\text{SRMSD}_p < 0$ are conservative p-values. The maximum SRMSD_p is
 317 achieved when all p-values are zero (the limit of anti-conservative p-values), which for infinite loci
 318 approaches

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

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

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

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

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

328 3 Results

329 3.1 Overview of evaluations

330 We use three real genotype datasets and simulated genotypes from six population structure scenarios
331 to cover various features of interest (Table 3). We introduce them in sets of three, as they appear
332 in the rest of our results. Population kinship matrices, which combine population and family
333 relatedness, are estimated without bias using `popkin` [32] (Fig. 1). The first set of three simulated
334 genotypes are based on an admixture model with 10 ancestries (Fig. 1A) [14, 32, 90]. The “large”
335 version (1000 individuals) illustrates asymptotic performance, while the “small” simulation (100
336 individuals) illustrates model overfitting. The “family” simulation has admixed founders and draws
337 a 20-generation random pedigree with assortative mating, resulting in a complex joint family and
338 ancestry structure in the last generation (Fig. 1B). The second set of three are the real human
339 datasets representing global human diversity: Human Origins (Fig. 1D), HGDP (Fig. 1G), and
340 1000 Genomes (Fig. 1J), which are enriched for small minor allele frequencies even after $MAF < 1\%$
341 filter (Fig. 1C). Last are tree simulations (Fig. 1F,I,L) fit to the kinship (Fig. 1E,H,K) and MAF
342 (Fig. 1C) of each real human dataset, which by design do not have family structure.

343 All traits in this work are simulated. We repeated all evaluations on two additive quantitative
344 trait models, *fixed effect sizes* (FES) and *random coefficients* (RC), which differ in how causal coef-
345 ficients are constructed. The FES model captures the rough inverse relationship between coefficient
346 and minor allele frequency that arises under strong negative and balancing selection and has been
347 observed in numerous diseases and other traits [52, 69–71], so it is the focus of our results. The
348 RC model draws coefficients independent of allele frequency, corresponding to neutral traits [52,

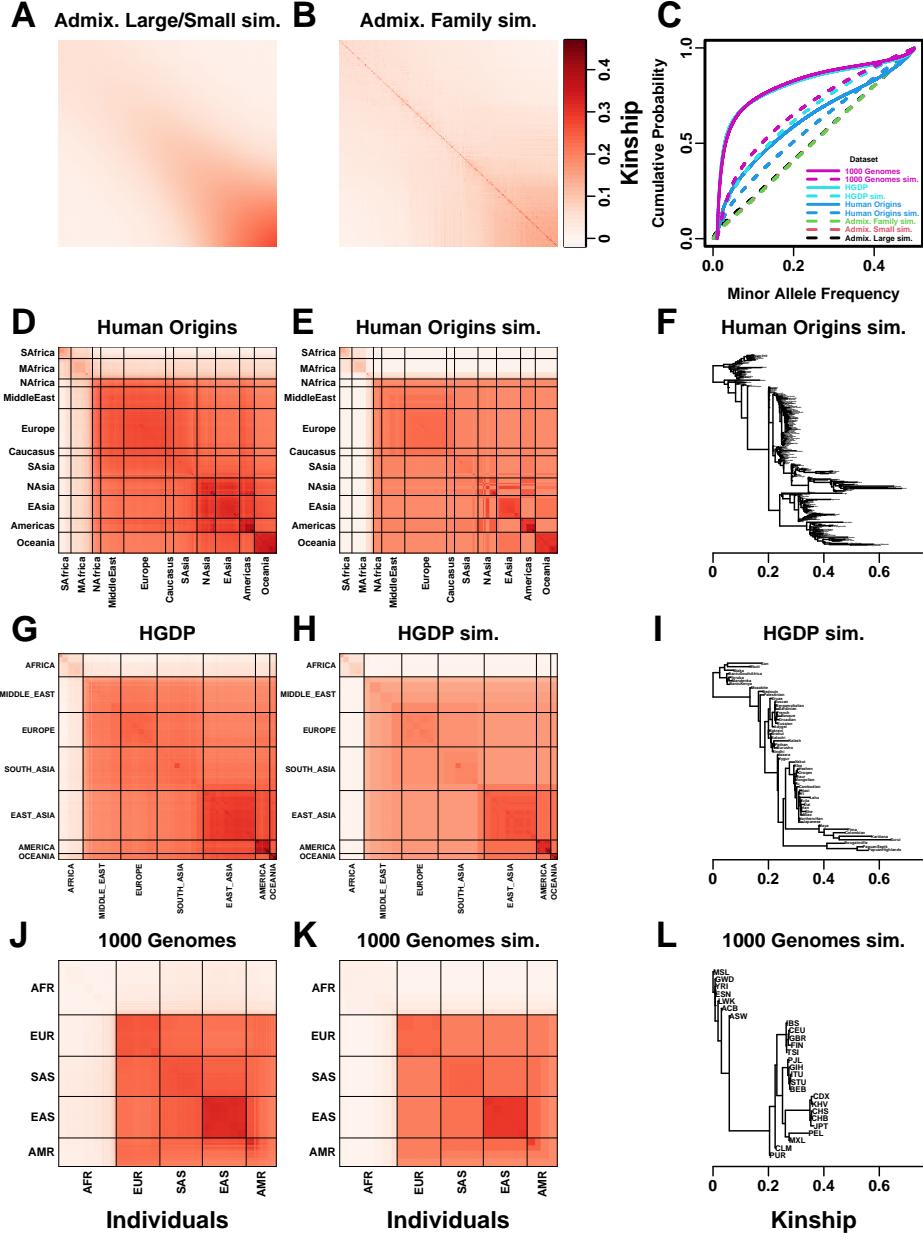


Figure 1: Population structures of simulated and real human genotype datasets. First two columns are population kinship matrices as heatmaps: individuals along x- and y-axis, kinship as color. Diagonal shows inbreeding values. **A.** Admixture scenario for both Large and Small simulations. **B.** Last 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 tree simulations had $\text{MAF} \geq 0.01$ 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.

349 71], which results in a wider effect size distribution that reduces association power and effective
 350 polygenicity compared to FES.

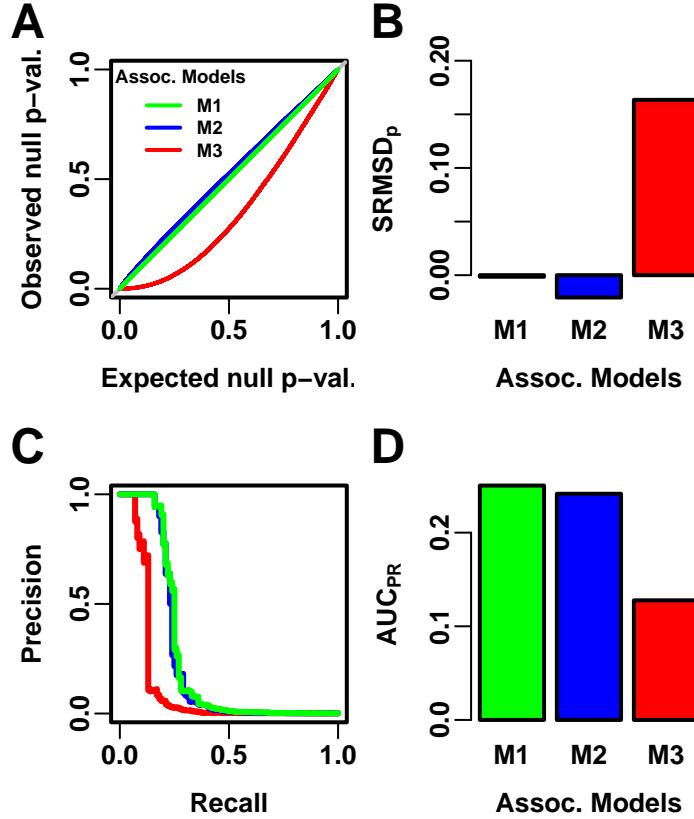


Figure 2: **Illustration of evaluation measures.** Three archetypal models illustrate our complementary measures: M1 is ideal, M2 overfits slightly, M3 is naive. **A.** QQ plot of p-values of “null” (non-causal) loci. M1 has desired uniform p-values, M2/M3 are miscalibrated. **B.** SRMSD_p (p-value Signed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

351 We evaluate using two complementary measures: (1) SRMSD_p (p-value signed root mean square
 352 deviation) measures p-value calibration (closer to zero is better), and (2) AUC_{PR} (precision-recall
 353 area under the curve) measures causal locus classification performance (higher is better; Fig. 2).
 354 SRMSD_p is a more robust alternative to the common inflation factor λ and type I error control
 355 measures; there is a correspondence between λ and SRMSD_p, with SRMSD_p > 0.01 giving $\lambda > 1.06$
 356 (Fig. S1) and thus evidence of miscalibration close to the rule of thumb of $\lambda > 1.05$ [24]. There

357 is also a monotonic correspondence between SRMSD_p and type I error rate (Fig. S2). AUC_{PR} has
 358 been used to evaluate association models [91], and reflects calibrated statistical power (Fig. S3)
 359 while being robust to miscalibrated models (Appendix B).

360 Both PCA and LMM are evaluated in each replicate dataset including a number of PCs r
 361 between 0 and 90 as fixed covariates. In terms of p-value calibration, for PCA the best number of
 362 PCs r (minimizing mean $|\text{SRMSD}_p|$ over replicates) is typically large across all datasets (Table 4),

Table 4: Overview of PCA and LMM evaluations for high 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	12	True	0.036	Tie
Admix. Small sim.	$ \text{SRMSD}_p $	FES	True	0	1	4	True	0.055	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	89	False	3.9e-10*	LMM
HGDP	$ \text{SRMSD}_p $	FES	True	0	1	87	True	4.4e-10*	LMM
1000 Genomes	$ \text{SRMSD}_p $	FES	True	0	1	90	False	3.9e-10*	LMM
Human Origins sim.	$ \text{SRMSD}_p $	FES	True	0	1	88	True	0.017	Tie
HGDP sim.	$ \text{SRMSD}_p $	FES	True	0	1	47	True	0.046	Tie
1000 Genomes sim.	$ \text{SRMSD}_p $	FES	True	0	1	78	True	9.6e-10*	LMM
Admix. Large sim.	$ \text{SRMSD}_p $	RC	True	0	1	26	True	0.11	Tie
Admix. Small sim.	$ \text{SRMSD}_p $	RC	True	0	1	4	True	0.00097	Tie
Admix. Family sim.	$ \text{SRMSD}_p $	RC	True	0	1	90	False	3.9e-10*	LMM
Human Origins	$ \text{SRMSD}_p $	RC	True	0	1	90	True	0.00065	Tie
HGDP	$ \text{SRMSD}_p $	RC	True	0	1	37	True	1.5e-05*	LMM
1000 Genomes	$ \text{SRMSD}_p $	RC	True	0	1	76	True	3.9e-10*	LMM
Human Origins sim.	$ \text{SRMSD}_p $	RC	True	0	1	85	True	0.14	Tie
HGDP sim.	$ \text{SRMSD}_p $	RC	True	0	1	44	True	8.8e-07*	LMM
1000 Genomes sim.	$ \text{SRMSD}_p $	RC	True	0	1	90	True	3.9e-10*	LMM
Admix. Large sim.	AUC_{PR}	FES		0	1	3		5.9e-06*	LMM
Admix. Small sim.	AUC_{PR}	FES		0	1	2		0.025	Tie
Admix. Family sim.	AUC_{PR}	FES		1	0.35	22		3.9e-10*	LMM
Human Origins	AUC_{PR}	FES		0	1	34		3.9e-10*	LMM
HGDP	AUC_{PR}	FES		1	0.33	16		4.4e-10*	LMM
1000 Genomes	AUC_{PR}	FES		1	0.11	8		3.9e-10*	LMM
Human Origins sim.	AUC_{PR}	FES		0	1	36		3.9e-10*	LMM
HGDP sim.	AUC_{PR}	FES		0	1	17		1.7e-05*	LMM
1000 Genomes sim.	AUC_{PR}	FES		0	1	10		5e-10*	LMM
Admix. Large sim.	AUC_{PR}	RC		0	1	3		1.4e-05*	LMM
Admix. Small sim.	AUC_{PR}	RC		0	1	1		0.095	Tie
Admix. Family sim.	AUC_{PR}	RC		0	1	34		3.9e-10*	LMM
Human Origins	AUC_{PR}	RC		3	0.4	36		9.6e-10*	LMM
HGDP	AUC_{PR}	RC		4	0.21	16		0.013	Tie
1000 Genomes	AUC_{PR}	RC		5	0.004	9		0.00043	Tie
Human Origins sim.	AUC_{PR}	RC		0	1	37		4.1e-10*	LMM
HGDP sim.	AUC_{PR}	RC		3	0.087	17		0.0014	Tie
1000 Genomes sim.	AUC_{PR}	RC		3	0.37	10		8.5e-10*	LMM

^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.

363 although much smaller r values often performed as well (shown in following sections). Most cases
364 have a mean $|\text{SRMSD}_p| < 0.01$, whose p-values are effectively calibrated. However, PCA is often
365 miscalibrated on the family simulation and real datasets (Table 4). In contrast, for LMM, $r = 0$ (no
366 PCs) is always best, and is always calibrated. Comparing LMM with $r = 0$ to PCA with its best
367 r , LMM always has significantly smaller $|\text{SRMSD}_p|$ than PCA or is statistically tied. For AUC_{PR}
368 and PCA, the best r is always smaller than the best r for $|\text{SRMSD}_p|$, so there is often a tradeoff
369 between calibrated p-values versus classification performance. For LMM there is no tradeoff, as
370 $r = 0$ often has the best mean AUC_{PR} , and otherwise is not significantly different from the best
371 r . Lastly, LMM with $r = 0$ always has significantly greater or statistically tied AUC_{PR} than PCA
372 with its best r .

373 3.2 Evaluations in admixture simulations

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

377 In the large admixture simulation, the SRMSD_p of PCA is largest when $r = 0$ (no PCs) and
378 decreases rapidly to near zero at $r = 3$, where it stays for up to $r = 90$ (Fig. 3A). Thus, PCA
379 has calibrated p-values for $r \geq 3$, smaller than the theoretical optimum for this simulation of
380 $r = K - 1 = 9$. In contrast, the SRMSD_p for LMM starts near zero for $r = 0$, but becomes negative
381 as r increases (p-values are conservative). The AUC_{PR} distribution of PCA is similarly worst at
382 $r = 0$, increases rapidly and peaks at $r = 3$, then decreases slowly for $r > 3$, while the AUC_{PR}
383 distribution for LMM starts near its maximum at $r = 0$ and decreases with r . Although the AUC_{PR}
384 distributions for LMM and PCA overlap considerably at each r , LMM with $r = 0$ has significantly
385 greater AUC_{PR} values than PCA with $r = 3$ (Table 4). However, qualitatively PCA performs nearly
386 as well as LMM in this simulation.

387 The observed robustness to large r led us to consider smaller sample sizes. A model with large
388 numbers of parameters r should overfit more as r approaches the sample size n . Rather than increase
389 r beyond 90, we reduce individuals to $n = 100$, which is small for typical association studies but

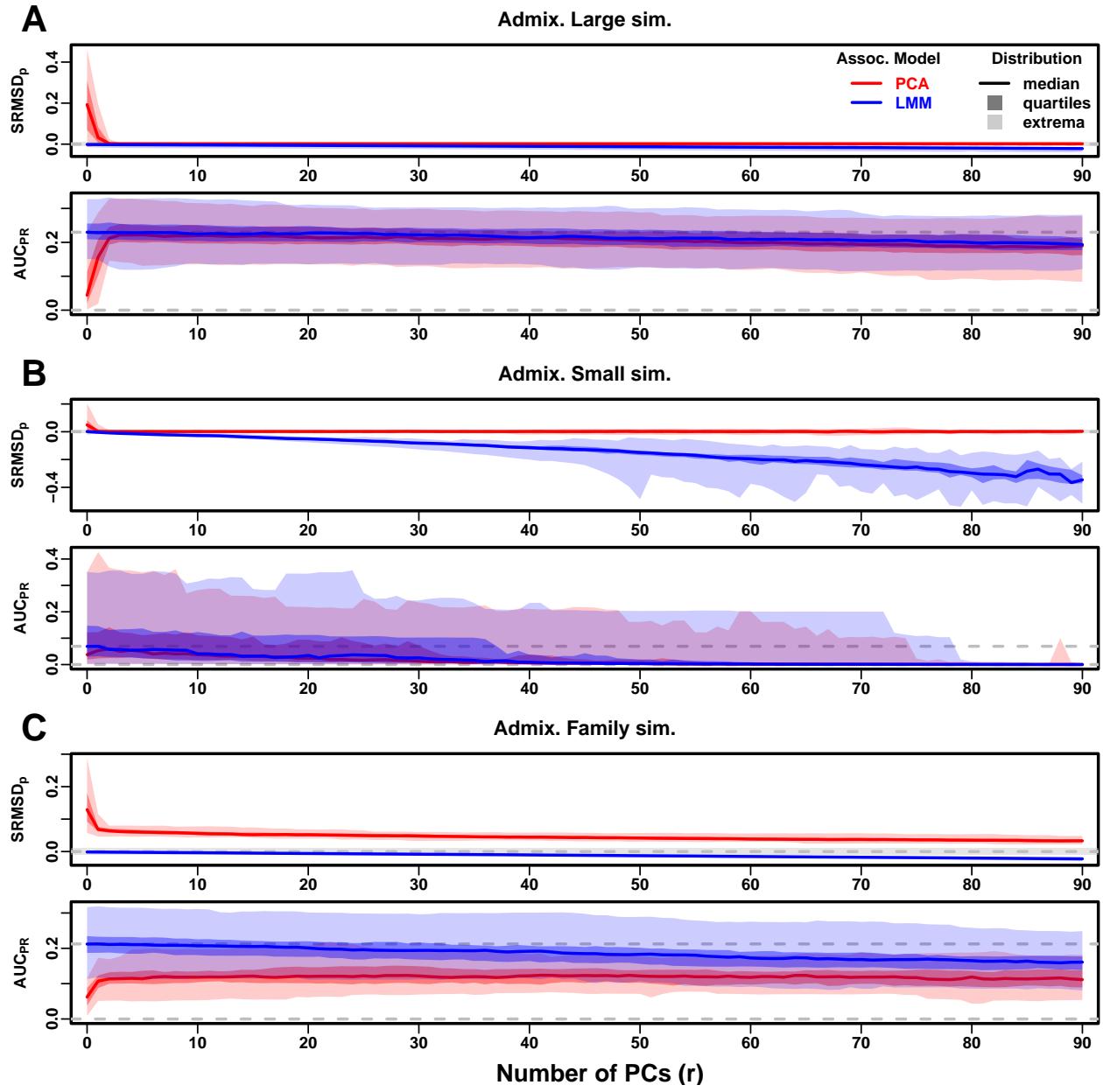


Figure 3: **Evaluations in admixture simulations.** Traits simulated from FES model with high heritability. PCA and LMM models have varying number of PCs ($r \in \{0, \dots, 90\}$ on x-axis), with the distributions (y-axis) of SRMSD_p (top subpanel) and AUC_{PR} (bottom subpanel) for 50 replicates. Best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

may occur in studies of rare diseases, pilot studies, or other constraints. To compensate for the loss of power due to reducing n , we also reduce the number of causal loci (fixed ratio $n/m_1 = 10$), which increases per-locus effect sizes. We found a large decrease in performance for both models as r increases, and best performance for $r = 1$ for PCA and $r = 0$ for LMM (Fig. 3B). Remarkably, LMM attains much larger negative SRMSD_p values than in our other evaluations. LMM with $r = 0$ is significantly better than PCA ($r = 1$ to 4) in both measures (Table 4), but qualitatively the difference is negligible.

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

3.3 Evaluations in real human genotype datasets

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

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

HGDP has the fewest individuals among real datasets, but compared to Human Origins contains more loci and low-frequency variants. Performance (Fig. 4B) again most resembled the family simulations. In particular, LMM with $r = 0$ achieves mean $|\text{SRMSD}_p| < 0.01$ (p-values are calibrated),

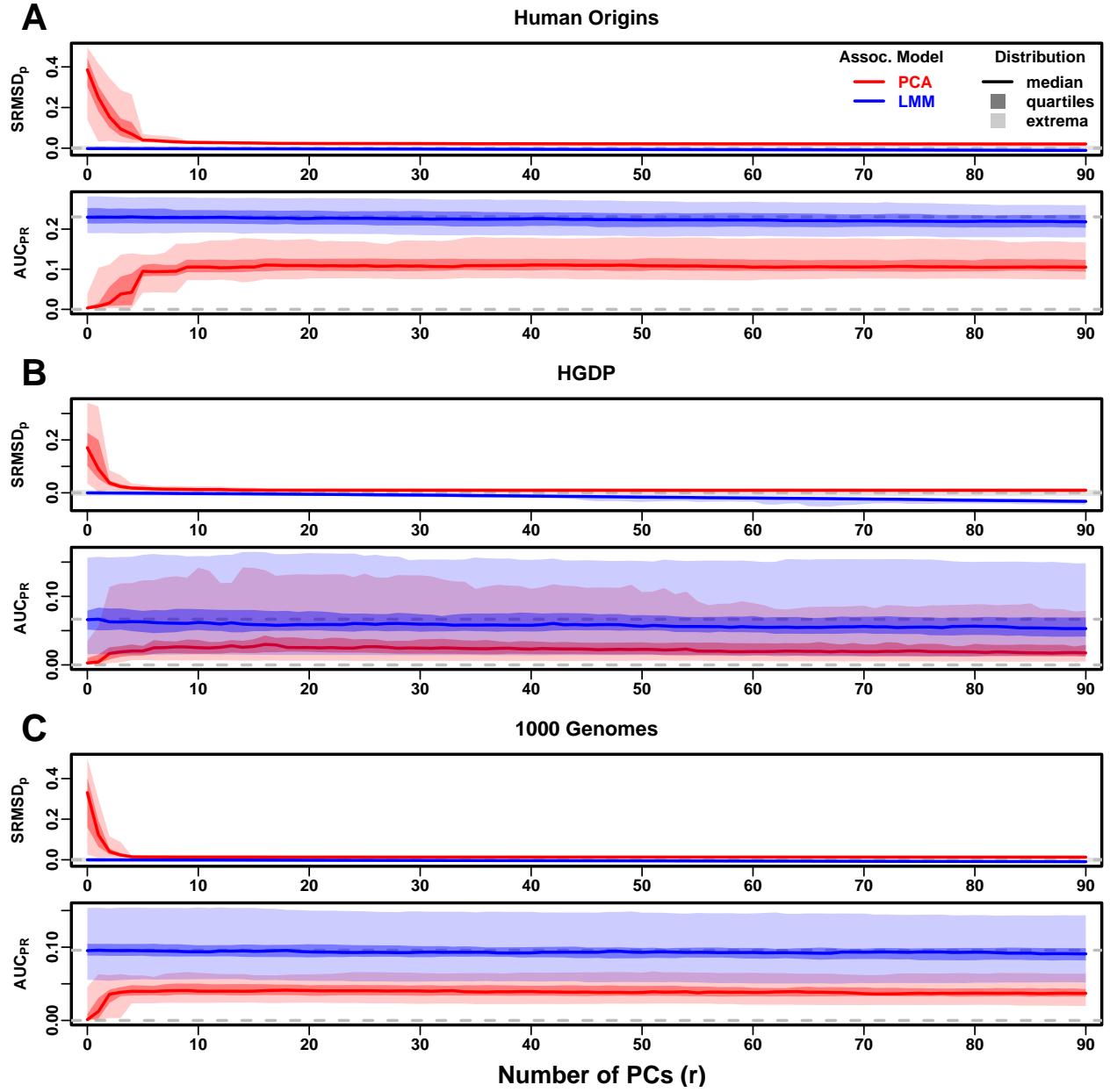


Figure 4: **Evaluations in real human genotype datasets.** Traits simulated from FES model with high heritability. Same setup as Fig. 3, see that for details. These datasets strongly favor LMM with no PCs over PCA, with distributions that most resemble the family simulation. **A.** Human Origins. **B.**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

417 while PCA does not, and there is a sizable AUC_{PR} gap between LMM and PCA. Maximum AUC_{PR}
418 values were lowest in HGDP compared to the two other real datasets.

419 1000 Genomes has the fewest subpopulations but largest number of individuals per subpopula-
420 tion. Thus, although this dataset has the simplest subpopulation structure among the real datasets,
421 we find $SRMSD_p$ and AUC_{PR} distributions (Fig. 4C) that again most resemble our earlier family
422 simulation, with mean $|SRMSD_p| < 0.01$ for LMM only and large AUC_{PR} gaps between LMM and
423 PCA.

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

429 3.4 Evaluations in tree simulations fit to human data

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

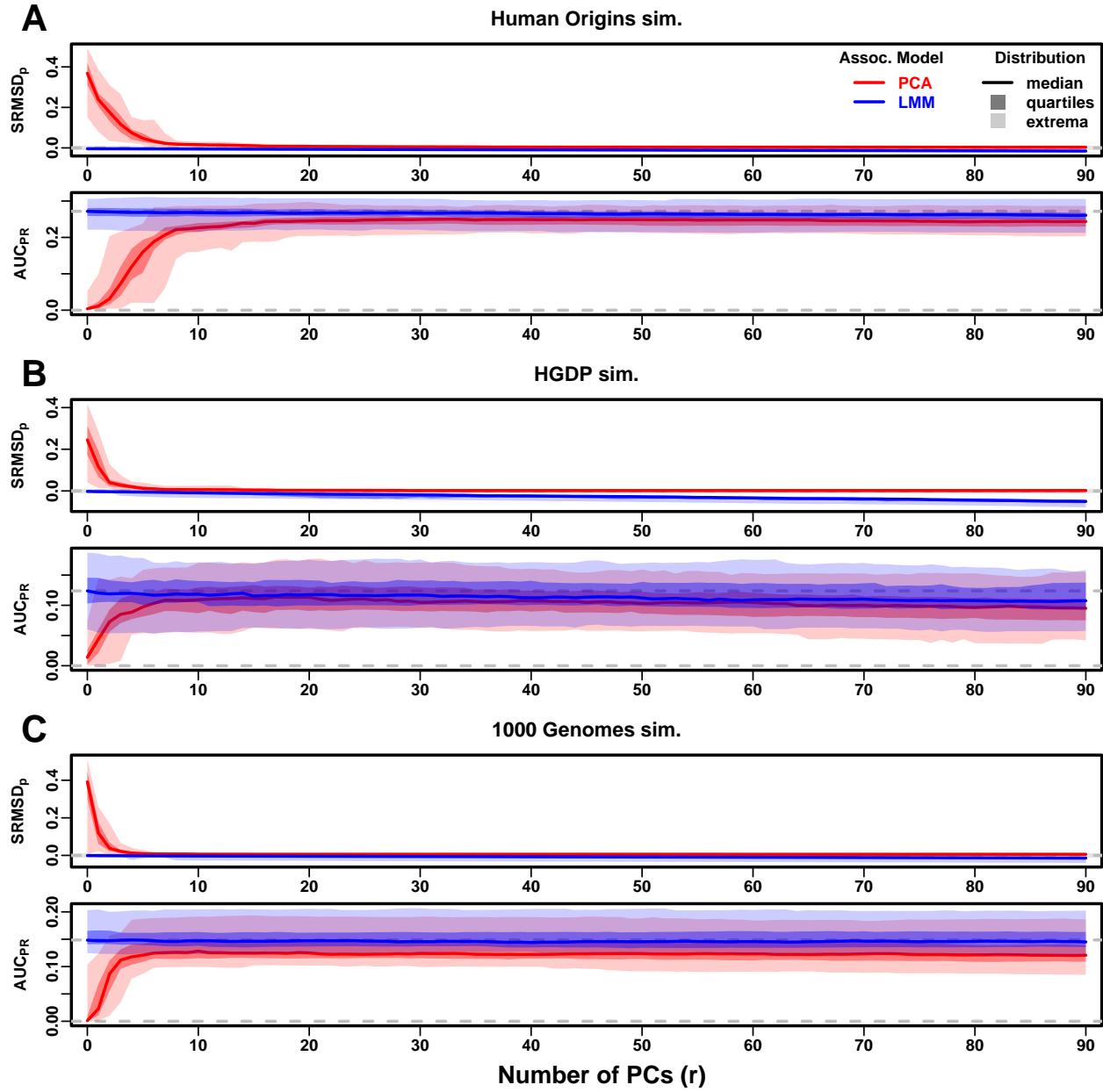


Figure 5: **Evaluations in tree simulations fit to human data.** Traits simulated from FES model with high heritability. Same setup as Fig. 3, see that for details. These tree simulations, which exclude family structure by design, do not explain the large gaps in LMM-PCA performance observed in the real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

443 **3.5 Numerous distant relatives explain poor PCA performance in real data**

444 In principle, PCA performance should be determined by the dimensionality of relatedness, since
445 PCA is a low-dimensional model whereas LMM can model high-dimensional relatedness without
446 overfitting. We used the Tracy-Widom test [7] with $p < 0.01$ to estimate dimensionality as the
447 number of significant PCs (Fig. S7A). The true dimensionality of our simulations is slightly un-
448 derestimated (Table 3), but we confirm that the family simulation has the greatest dimensionality,
449 and real datasets have greater estimates than their respective tree simulations, which confirms our
450 hypothesis to some extent. However, estimated dimensionalities do not separate real datasets from
451 tree simulations, as required to predict the observed PCA performance. Moreover, the HGDP and
452 1000 Genomes dimensionality estimates are 45 and 61, respectively, yet PCA performed poorly for
453 all $r \leq 90$ numbers of PCs (Fig. 4). The top eigenvalue explained a proportion of variance propor-
454 tional to F_{ST} (Table 3), but the rest of the top 10 eigenvalues show no clear differences between
455 datasets, except the small simulation had larger variances explained per eigenvalue (expected since
456 it has fewer eigenvalues; Fig. S7C). Comparing cumulative variance explained versus rank frac-
457 tion across all eigenvalues, all datasets increase from their starting point almost linearly until they
458 reach 1, except the family simulation has much greater variance explained by mid-rank eigenvalues
459 (Fig. S7B). We also calculated the number of PCs that are significantly associated with the trait,
460 and observed similar results, namely that while the family simulation has more significant PCs than
461 the non-family admixture simulations, the real datasets and their tree simulated counterparts have
462 similar numbers of significant PCs (Fig. S8). Overall, there is no separation between real datasets
463 (where PCA performed poorly) and tree simulations (where PCA performed relatively well) in terms
464 of their eigenvalues or dimensionality estimates.

465 Local kinship, which is recent relatedness due to family structure excluding population structure,
466 is the presumed cause of the LMM to PCA performance gap observed in real datasets but not their
467 tree simulation counterparts. Instead of inferring local kinship through increased dimensionality, as
468 attempted in the last paragraph, now we measure it directly using the KING-robust estimator [86].
469 We observe more large local kinship in the real datasets and the family simulation compared to the
470 other simulations (Fig. 6). However, for real data this distribution depends on the subpopulation

471 structure, since locally related pairs are most likely in the same subpopulation. Therefore, the
 472 only comparable curve to each real dataset is their corresponding tree simulation, which matches
 473 subpopulation structure. In all real datasets we identified highly related individual pairs with
 474 kinship above the 4th degree relative threshold of 0.022 [86, 92]. However, these highly related pairs
 475 are vastly outnumbered by more distant pairs with evident non-zero local kinship as compared to
 476 the extreme tree simulation values.

477 To try to improve PCA performance, we followed the standard practice of removing 4th degree
 478 relatives, which reduced sample sizes between 5% and 10% (Table S1). Only $r = 0$ for LMM
 479 and $r = 20$ for PCA were tested, as these performed well in our earlier evaluation, and only
 480 FES traits were tested because they previously displayed the large PCA-LMM performance gap.
 481 LMM significantly outperforms PCA in all these cases (Wilcoxon paired 1-tailed $p < 0.01$; Fig. 7).
 482 Notably, PCA still had miscalibrated p-values in all real datasets ($|\text{SRMSD}_p| > 0.01$). Otherwise,
 483 AUC_{PR} and SRMSD_p ranges were similar here as in our earlier evaluation. Therefore, the removal

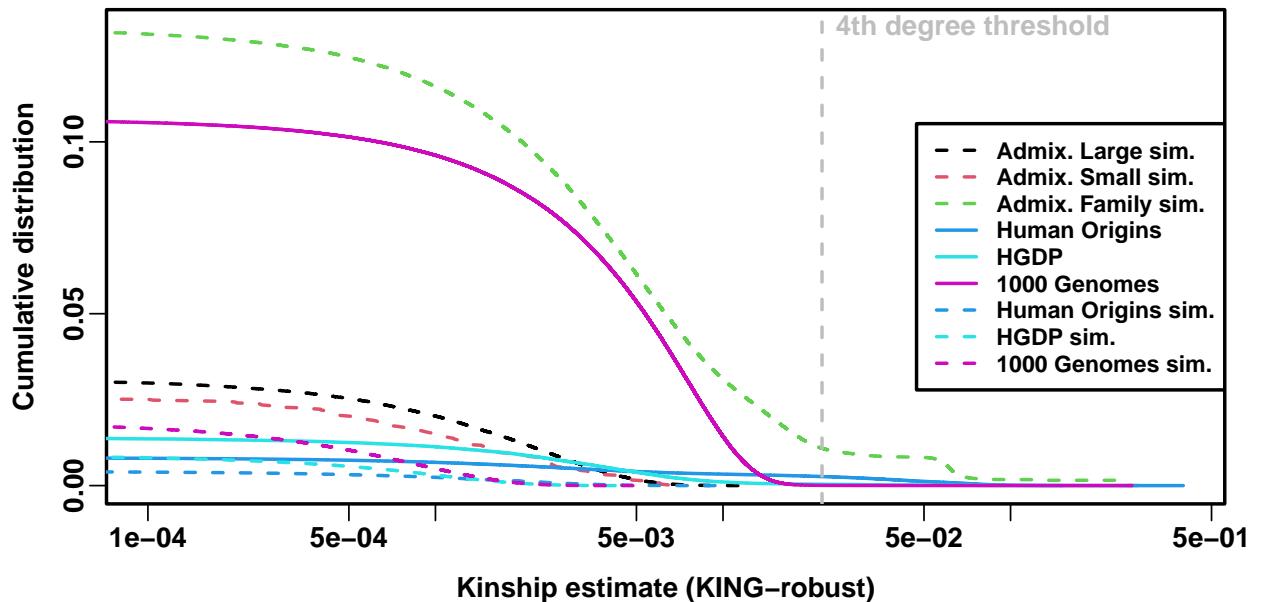


Figure 6: **Local kinship distributions.** Curves are complementary cumulative distribution of lower triangular kinship matrix (self kinship excluded) from KING-robust estimator. Note log x-axis; negative estimates are counted but not shown. Most values are below 4th degree relative threshold. Each real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

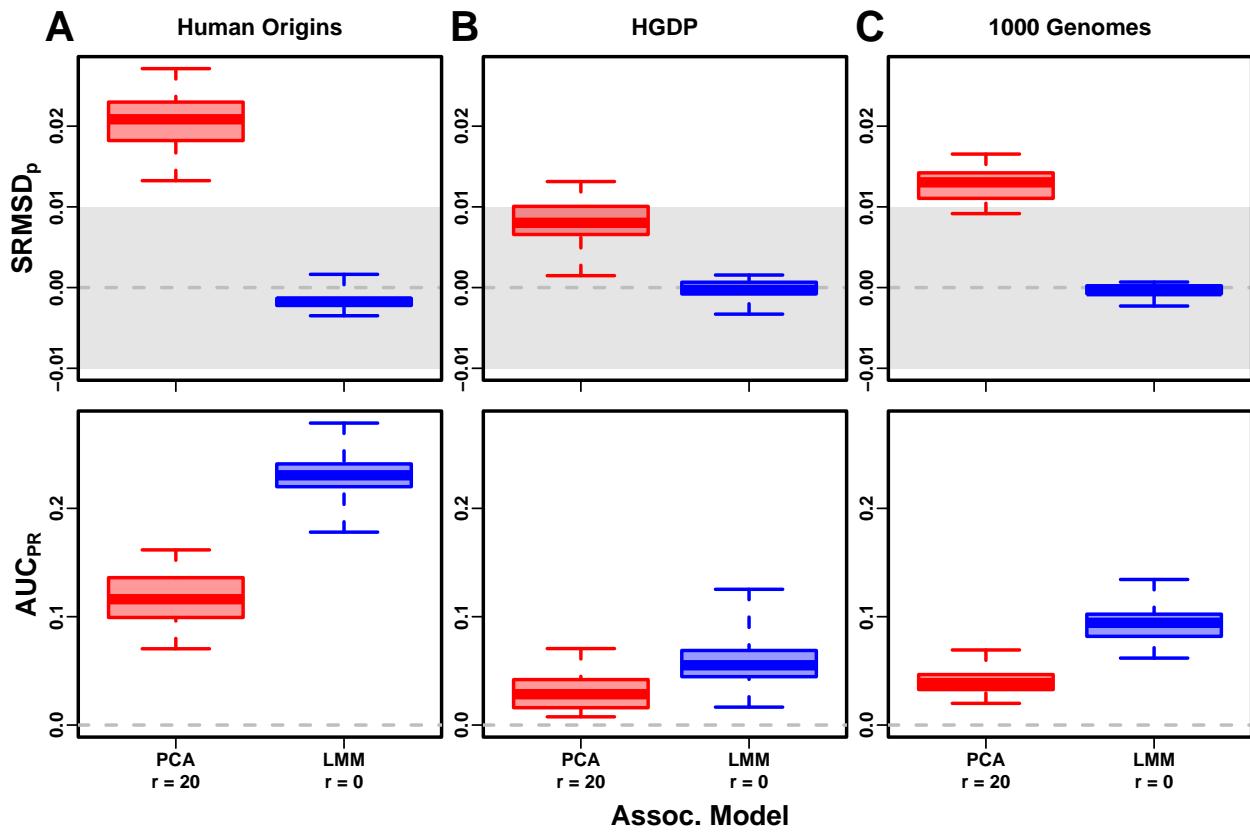


Figure 7: **Evaluation in real datasets excluding 4th degree relatives.** Traits simulated from FES model with high heritability. Each dataset is a column, rows are measures. First row has $|SRMSD_p| < 0.01$ band marked as gray area.

484 of the small number of highly related individual pairs had a negligible effect in PCA performance,
485 so the larger number of more distantly related pairs explain the poor PCA performance in the real
486 datasets.

487 3.6 Low heritability and environment simulations

488 Our main evaluations were repeated with traits simulated under a lower heritability value of $h^2 =$
489 0.3. We reduced the number of causal loci in response to this change in heritability, to result in equal
490 average effect size per locus compared to the previous high heritability evaluations (see Methods).
491 Despite that, these low heritability evaluations measured lower AUC_{PR} values than their high
492 heritability counterparts (Figs. S9 to S13). The gap between LMM and PCA was reduced in these
493 evaluations, but the main conclusion of the high heritability evaluation holds for low heritability as
494 well, namely that LMM with $r = 0$ significantly outperforms or ties LMM with $r > 0$ and PCA in
495 all cases (Table S2).

496 Lastly, we simulated traits with both low heritability and large environment effects determined
497 by geography and race/ethnicity labels, so they are strongly correlated to the low-dimensional pop-
498 ulation structure (Table 2). For that reason, PCs may be expected to perform better in this setting
499 (in either PCA or LMM). However, we find that both PCA and LMM (even without PCs) increase
500 their AUC_{PR} values compared to the low-heritability evaluations (Fig. S14; Fig. 8 also shows repre-
501 sentative numbers of PCs, which performed optimally or nearly so in individual simulations shown
502 in Figs. S15 to S18). P-value calibration (SRMSD_p) is comparable with or without environment
503 effects, for LMM for all r and for PCA once r is large enough (Fig. S14). These simulations are
504 the only where we occasionally observed for both metrics a significant, though small, advantage
505 of LMM with PCs versus LMM without PCs (Table S3). Additionally, on RC traits only, PCA
506 significantly outperforms LMM in the three real human datasets (Table S3), the only cases in all of
507 our evaluations where this is observed. For comparison, we also evaluate an “oracle” LMM without
508 PCs but with the finest group labels, the same used to simulate environment, as fixed categorical
509 covariates (“LMM lab.”), and see much larger AUC_{PR} values than either LMM with PCs or PCA
510 (Figs. 8 and S15 to S18 and Table S3). However, LMM with labels is often more poorly calibrated

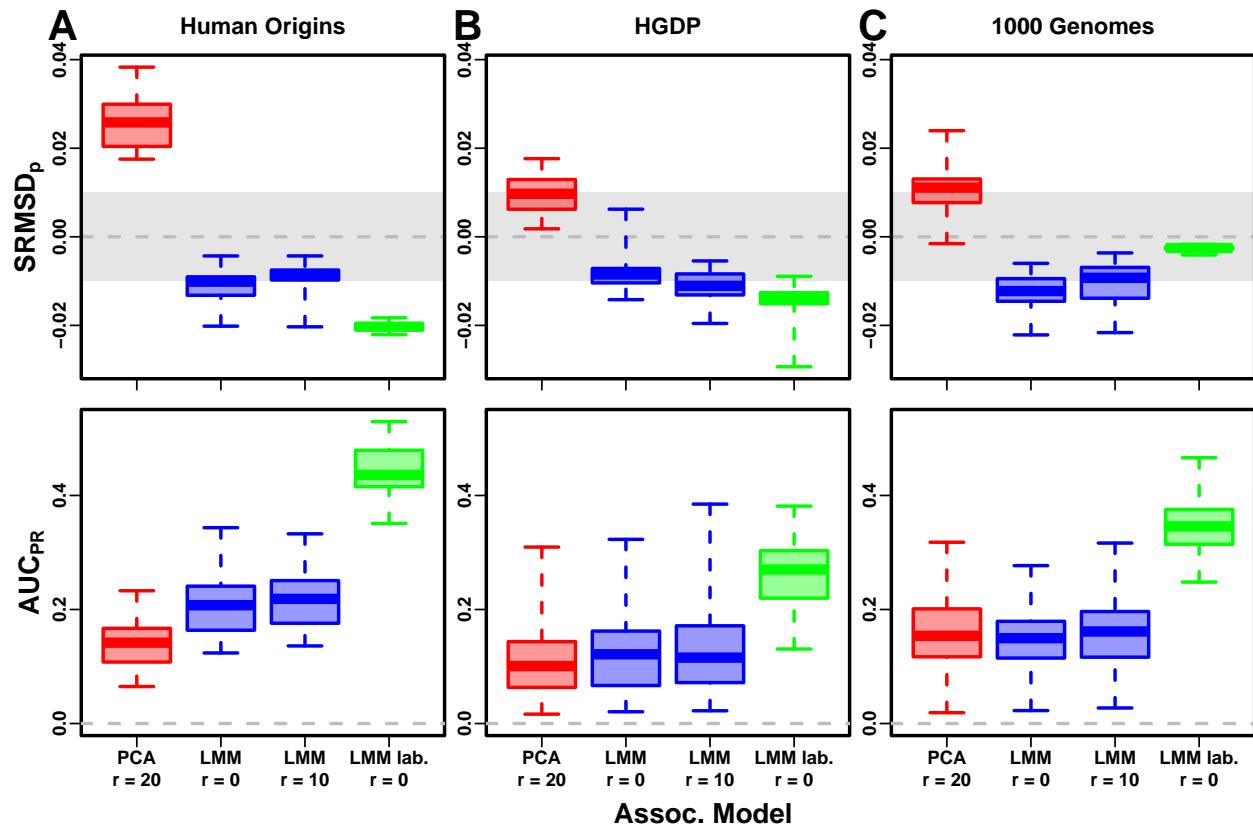


Figure 8: Evaluation in real datasets excluding 4th degree relatives, FES traits, environment. Traits simulated with environment effects, otherwise the same as Fig. 7.

than LMM or PCA without labels, which may be since these numerous labels are inappropriately modeled as fixed rather than random effects. Overall, we find that association studies with correlated environment and genetic effects remain a challenge for PCA and LMM, that addition of PCs to an LMM improves performance only marginally, and that if the environment effect is driven by geography or ethnicity then use of those labels greatly improves performance compared to using PCs.

4 Discussion

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

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

Previous studies found that PCA was better calibrated than LMM for unusually differentiated

538 markers [24, 35, 37], which as simulated were an artificial scenario not based on a population genetics
539 model, and are otherwise believed to be unusual [38, 59]. Our evaluations on real human data,
540 which contain such loci in relevant proportions if they exist, do not replicate that result. Cryptic
541 relatedness strongly favors LMM, an advantage that probably outweighs this potential PCA benefit
542 in real data.

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

551 PCA assumes that genetic relatedness is low-dimensional, whereas LMM can handle high-
552 dimensional relatedness. Thus, PCA performs well in the admixture simulation, which is explicitly
553 low-dimensional (see Materials and Methods), and our tree simulations, which, although complex in
554 principle due to the large number of nodes, had few long branches so a low-dimensional approxima-
555 tion suffices. Conversely, PCA performs poorly under family structure because its kinship matrix is
556 high-dimensional (Fig. S7). However, estimating the dimensionality of real datasets is challenging
557 because estimated eigenvalues have biased distributions. Dimensionality estimated using the Tracy-
558 Widom test [7] did not fully predict the datasets that PCA performs well on. In contrast, estimated
559 local kinship finds considerable cryptic relatedness in all real human datasets and better explains
560 why PCA performs poorly there. The trait model also influences the relative performance of PCA,
561 so genotype-only parameters (eigenvalues or local kinship) alone do not tell the full story. There are
562 related tests for numbers of dimensions that consider the trait which we did not consider, including
563 the Bayesian information criterion for the regression with PCs against the trait [17]. Additionally,
564 PCA and LMM goodness of fit could be compared using the coefficient of determination generalized
565 for mixed models [94].

566 PCA is at best underpowered relative to LMMs, and at worst miscalibrated regardless of the
567 numbers of PCs included, in real human genotype tests. Among our simulations, such poor per-
568 formance occurred only in the admixed family. Local kinship estimates reveal considerable family
569 relatedness in the real datasets absent in the corresponding tree simulations. Admixture is also
570 absent in our tree simulations, but our simulations and theory show that admixture is handled well
571 by PCA. Hundreds of close relative pairs have been identified in 1000 Genomes [95–98], but their
572 removal does not improve PCA performance sufficiently in our tests, so the larger number of more
573 distantly related pairs are PCA’s most serious obstacle in practice. Distant relatives are expected
574 to be numerous in any large human dataset [58, 99, 100]. Our FES trait tests show that cryp-
575 tic relatedness is more challenging when rarer variants have larger coefficients. Overall, the high
576 dimensionality induced by cryptic relatedness is the key challenge for PCA association in modern
577 datasets that is readily overcome by LMM.

578 Our tests also found PCA robust to large numbers of PCs, far beyond the optimal choice,
579 agreeing with previous anecdotal observations [5, 36], in contrast to using too few PCs for which
580 there is a large performance penalty. The exception was the small sample size simulation, where
581 only small numbers of PCs performed well. In contrast, LMM is simpler since there is no need
582 to choose the number of PCs. However, an LMM with a large number of covariates may have
583 conservative p-values (as observed for LMM with large numbers of PCs), a weakness of the score
584 test used by the LMM we evaluated that may be overcome with other statistical tests. Simulations
585 or post hoc evaluations remain crucial for ensuring that statistics are calibrated.

586 There are several variants of the PCA and LMM analyses, most designed for better modeling
587 linkage disequilibrium (LD), that we did not evaluate directly, in which PCs are no longer exactly
588 the top eigenvectors of the kinship matrix (if estimated with different approaches), although this is
589 not a crucial aspect of our arguments. We do not consider the case where samples are projected onto
590 PCs estimated from an external sample [101], which is uncommon in association studies, and whose
591 primary effect is shrinkage, so if all samples are projected then they are all equally affected and
592 larger regression coefficients compensate for the shrinkage, although this will no longer be the case if
593 only a portion of the sample is projected onto the PCs of the rest of the sample. Another approach

594 tests PCs for association against every locus in the genome in order to identify and exclude PCs that
595 capture LD structure (which is localized) instead of ancestry (which should be present across the
596 genome) [101]; a previous proposal removes LD using an autocorrelation model prior to estimating
597 PCs [7]. These improved PCs remain inadequate models of family or cryptic relatedness, so an
598 LMM will continue to outperform them in that setting. Similarly, the leave-one-chromosome-out
599 (LOCO) approach for estimating kinship matrices for LMMs prevents the test locus and loci in LD
600 with it from being modeled by the random effect as well, which is called “proximal contamination”
601 [35, 42]. While LOCO kinship estimates vary for each chromosome, they continue to model family
602 or cryptic relatedness, thus maintaining their key advantage over PCA. The LDAK model estimates
603 kinship instead by weighing loci taking LD into account [102]. LD effects must be adjusted for, if
604 present, so in unfiltered data we advise the previous methods be applied. However, in this work,
605 simulated genotypes do not have LD, and the real datasets were filtered to remove LD, so here
606 there is no proximal contamination and LD confounding is minimized if present at all, so these
607 evaluations may be considered the ideal situation where LD effects have been adjusted successfully,
608 and in this setting LMM outperforms PCA. Overall, these alternative PCs or kinship matrices differ
609 from their basic counterparts by either the extent to which LD influences the estimates (which may
610 be a confounder in a small portion of the genome, by definition) or by sampling noise, neither of
611 which are expected to change our key conclusion.

612 One of the limitations of this work include relatively small sample sizes compared to modern
613 association studies. However, our conclusions are not expected to change with larger sample sizes,
614 as cryptic relatedness will continue to be abundant in such data, if not increase in abundance, and
615 thus give LMMs an advantage over PCA [58, 99, 100]. One reason PCA has been favored over classic
616 LMMs is because PCA’s runtime scales much better with increasing sample size. However, recent
617 approaches not tested in this work have made LMMs more scalable and applicable to biobank-scale
618 data [39, 47, 53], so one clear next step is carefully evaluating these approaches in simulations
619 with larger sample sizes. A different benefit for including PCs were recently reported for BOLT-
620 LMM, which does not result in greater power but rather in reduced runtime, a property that may
621 be specific to its use of scalable algorithms such as conjugate gradient and variational Bayes [58].

622 Many of these newer LMMs also no longer follow the infinitesimal model of the basic LMM [47, 53],
623 and employ additional approximations, which are features not evaluated in this work and worthy
624 of future study.

625 Another limitation of this work is ignoring rare variants, a necessity given our smaller sample
626 sizes, where rare variant association is miscalibrated and underpowered. Using simulations mimick-
627 ing the UK Biobank, recent work has found that rare variants can have a more pronounced structure
628 than common variants, and that modeling this rare variant structure (with either PCA and LMM)
629 may better model environment confounding, improve inflation in association studies, and ameliorate
630 stratification in polygenic risk scores [103]. Better modeling rare variants and their structure is a
631 key next step in association studies.

632 The largest limitation of our work is that we only considered quantitative traits. We noted that
633 previous evaluations involving case-control traits tended to report PCA-LMM ties or mixed results,
634 an observation potentially confounded by the use of low-dimensional simulations without family
635 relatedness (Table 1). An additional concern is case-control ascertainment bias and imbalance,
636 which appears to affect LMMs more severely, although recent work appears to solve this problem
637 [35, 39]. Future evaluations should aim to include our simulations and real datasets, to ensure that
638 previous results were not biased in favor of PCA by employing unrealistic low-dimensional genotype
639 simulations, or by not simulating large coefficients for rare variants expected for diseases by various
640 selection models.

641 Overall, our results lead us to recommend LMM over PCA for association studies in general.
642 Although PCA offer flexibility and speed compared to LMM, additional work is required to ensure
643 that PCA is adequate, including removal of close relatives (lowering sample size and wasting re-
644 sources) followed by simulations or other evaluations of statistics, and even then PCA may perform
645 poorly in terms of both type I error control and power. The large numbers of distant relatives
646 expected of any real dataset all but ensures that PCA will perform poorly compared to LMM [58,
647 99, 100]. Our findings also suggest that related applications such as polygenic models may enjoy
648 gains in power and accuracy by employing an LMM instead of PCA to model relatedness [22, 91].
649 PCA remains indispensable across population genetics, from visualizing population structure and

650 performing quality control to its deep connection to admixture models, but the time has come to
 651 limit its use in association testing in favor of LMM or other, richer models capable of modeling all
 652 forms of relatedness.

653 5 Appendices

654 5.1 Appendix A: Fitting ancestral allele frequency distribution to real data

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

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

660 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
 661 moments of \hat{p}_i^T follow from the laws of total expectation and variance: $E[\hat{p}_i^T] = \frac{1}{2}$ and

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

662 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
 663 the original, lower variance of

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

664 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
 665 parameters: the subpopulation coancestry matrix (Eq. (7)), expanded from subpopulations to indi-
 666 viduals, the diagonal converted to kinship (reversing Eq. (8)), and the matrix averaged. However,
 667 since our model ignores the MAF filters imposed in our simulations, $\bar{\varphi}^T$ was adjusted. For Human
 668 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
 669 and 0.124, respectively, but 0.4 for both produced a better fit.
 670

671 Lastly, we construct new allele frequencies,

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

672 by a weighted average of \hat{p}_i^T and $q \in (0, 1)$ drawn independently from a different distribution.

673 $E[q] = \frac{1}{2}$ is required to have $E[p^*] = \frac{1}{2}$. The resulting variance is

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

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

675 which is achieved with:

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

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

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

677 **5.2 Appendix B: comparisons between SRMSD_p, AUC_{PR}, and evaluation mea-
678 sures from the literature**

679 **5.2.1 The inflation factor λ**

680 Test statistic inflation has been used to measure model calibration [1, 24]. The inflation factor

681 λ is defined as the median χ^2 association statistic divided by theoretical median under the null

682 hypothesis [2]. To compare p-values from non- χ^2 tests (such as t-statistics), λ can be calculated

683 from p-values using

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

684 where p_{median} is the median observed p-value (including causal loci), $u_{\text{median}} = \frac{1}{2}$ is its null expec-

685 tation, and F is the χ^2 cumulative density function (F^{-1} is the quantile function).

686 To compare λ and SRMSD_p directly, for simplicity assume that all p-values are null. In this

687 case, calibrated p-values give $\lambda = 1$ and $\text{SRMSD}_p = 0$. However, non-uniform p-values with the
 688 expected median, such as from genomic control [2], result in $\lambda = 1$, but $\text{SRMSD}_p \neq 0$ except for
 689 uniform p-values, a key flaw of λ that SRMSD_p overcomes. Inflated statistics (anti-conservative
 690 p-values) give $\lambda > 1$ and $\text{SRMSD}_p > 0$. Deflated statistics (conservative p-values) give $\lambda < 1$ and
 691 $\text{SRMSD}_p < 0$. Thus, $\lambda \neq 1$ always implies $\text{SRMSD}_p \neq 0$ (where $\lambda - 1$ and SRMSD_p have the
 692 same sign), but not the other way around. Overall, λ depends only on the median p-value, while
 693 SRMSD_p uses the complete distribution. However, SRMSD_p requires knowing which loci are null,
 694 so unlike λ it is only applicable to simulated traits.

695 5.2.2 Empirical comparison of SRMSD_p and λ

696 There is a near one-to-one correspondence between λ and SRMSD_p in our data (Fig. S1). PCA
 697 tended to be inflated ($\lambda > 1$ and $\text{SRMSD}_p > 0$) whereas LMM tended to be deflated ($\lambda < 1$ and
 698 $\text{SRMSD}_p < 0$), otherwise the data for both models fall on the same contiguous curve. We fit a
 699 sigmoidal function to this data,

$$700 \quad \text{SRMSD}_p(\lambda) = a \frac{\lambda^b - 1}{\lambda^b + 1}, \quad (10)$$

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

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

702 We fit this model to $\lambda > 1$ only since it was less noisy and of greater interest, and obtained the
 703 curve shown in Fig. S1 with $a = 0.564$ and $b = 0.619$. The value $\lambda = 1.05$, a common threshold
 704 for benign inflation [24], corresponds to $\text{SRMSD}_p = 0.0085$ according to Eq. (10). Conversely,
 705 $\text{SRMSD}_p = 0.01$, serving as a simpler rule of thumb, corresponds to $\lambda = 1.06$.

706 5.2.3 Type I error rate

707 The type I error rate is the proportion of null p-values with $p \leq t$. Calibrated p-values have type
 708 I error rate near t , which may be evaluated with a binomial test. This measure may give different
 709 results for different t , for example be significantly miscalibrated only for large t (due to lack of

710 power for smaller t). In contrast, $\text{SRMSD}_p = 0$ guarantees calibrated type I error rates at all t ,
711 while large $|\text{SRMSD}_p|$ indicates incorrect type I errors for a range of t . Empirically, we find the
712 expected agreement and monotonic relationship between SRMSD_p and type I error rate (Fig. S2).

713 **5.2.4 Statistical power and comparison to AUC_{PR}**

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

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

716 $F(t)$ is a cumulative function, so it is monotonically increasing and has an inverse. Like type I error
717 control, power may rank models differently depending on t .

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

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

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

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

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

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

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

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

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

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

724 Lastly, consider a simple yet common case in which model A is uniformly more powerful than
725 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.
726 This ensures that the precision of A is greater than that of B at every recall value, so AUC_{PR} is
727 greater for A than B . Thus, AUC_{PR} ranks calibrated models according to power.

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

731 Competing interests

732 The authors declare no competing interests.

733 Acknowledgments

734 This work was funded in part by the Duke University School of Medicine Whitehead Scholars
735 Program, a gift from the Whitehead Charitable Foundation. The 1000 Genomes data were generated
736 at the New York Genome Center with funds provided by NHGRI Grant 3UM1HG008901-03S1.

737 Web resources

738 plink2, <https://www.cog-genomics.org/plink/2.0/>

739 GCTA, <https://yanglab.westlake.edu.cn/software/gcta/>

740 Eigensoft, <https://github.com/DReichLab/EIG>

741 bnpsd, <https://cran.r-project.org/package=bnpsd>

742 simfam, <https://cran.r-project.org/package=simfam>

743 simtrait, <https://cran.r-project.org/package=simtrait>

744 genio, <https://cran.r-project.org/package=genio>

745 popkin, <https://cran.r-project.org/package=popkin>

746 ape, <https://cran.r-project.org/package=ape>

747 nnls, <https://cran.r-project.org/package=nnls>
748 PRROC, <https://cran.r-project.org/package=PRROC>
749 BEDMatrix, <https://cran.r-project.org/package=BEDMatrix>

750 Data and code availability

751 The data and code generated during this study are available on GitHub at <https://github.com/>
752 OchoaLab/pca-assoc-paper. The public subset of Human Origins is available on the Reich Lab
753 website at <https://reich.hms.harvard.edu/datasets>; non-public samples have to be requested
754 from David Reich. The WGS version of HGDP was downloaded from the Wellcome Sanger In-
755 stitute FTP site at ftp://ngs.sanger.ac.uk/production/hgdp/hgdp_wgs.20190516/. The high-
756 coverage version of the 1000 Genomes Project was downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20190425_NYGC_GATK/.
757

758 References

- 759 [1] W. Astle and D. J. Balding. “Population Structure and Cryptic Relatedness in Genetic
760 Association Studies”. *Statist. Sci.* 24(4) (2009), pp. 451–471. DOI: [10.1214/09-STS307](https://doi.org/10.1214/09-STS307).
- 761 [2] B. Devlin and K. Roeder. “Genomic Control for Association Studies”. *Biometrics* 55(4)
762 (1999), pp. 997–1004. DOI: [10.1111/j.0006-341X.1999.00997.x](https://doi.org/10.1111/j.0006-341X.1999.00997.x).
- 763 [3] B. F. Voight and J. K. Pritchard. “Confounding from Cryptic Relatedness in Case-Control As-
764 sociation Studies”. *PLOS Genetics* 1(3) (2005), e32. DOI: [10.1371/journal.pgen.0010032](https://doi.org/10.1371/journal.pgen.0010032).
- 765 [4] S. Zhang, X. Zhu, and H. Zhao. “On a semiparametric test to detect associations between
766 quantitative traits and candidate genes using unrelated individuals”. *Genetic Epidemiology*
767 24(1) (2003), pp. 44–56. DOI: [10.1002/gepi.10196](https://doi.org/10.1002/gepi.10196).
- 768 [5] A. L. Price et al. “Principal components analysis corrects for stratification in genome-wide
769 association studies”. *Nat. Genet.* 38(8) (2006), pp. 904–909. DOI: [10.1038/ng1847](https://doi.org/10.1038/ng1847).

- 770 [6] M. Bouaziz, C. Ambroise, and M. Guedj. “Accounting for Population Stratification in Prac-
771 tice: A Comparison of the Main Strategies Dedicated to Genome-Wide Association Studies”.
772 *PLOS ONE* 6(12) (2011), e28845. DOI: 10.1371/journal.pone.0028845.
- 773 [7] N. Patterson, A. L. Price, and D. Reich. “Population Structure and Eigenanalysis”. *PLoS*
774 *Genet* 2(12) (2006), e190. DOI: 10.1371/journal.pgen.0020190.
- 775 [8] I. T. Jolliffe. *Principal Component Analysis*. 2nd ed. New York: Springer-Verlag, 2002.
- 776 [9] J. K. Pritchard et al. “Association Mapping in Structured Populations”. *The American Jour-*
777 *nal of Human Genetics* 67(1) (2000), pp. 170–181. DOI: 10.1086/302959.
- 778 [10] D. H. Alexander, J. Novembre, and K. Lange. “Fast model-based estimation of ancestry in
779 unrelated individuals”. *Genome Res.* 19(9) (2009), pp. 1655–1664. DOI: 10.1101/gr.094052.
780 109.
- 781 [11] Q. Zhou, L. Zhao, and Y. Guan. “Strong Selection at MHC in Mexicans since Admixture”.
782 *PLoS Genet.* 12(2) (2016), e1005847. DOI: 10.1371/journal.pgen.1005847.
- 783 [12] G. McVean. “A genealogical interpretation of principal components analysis”. *PLoS Genet*
784 5(10) (2009), e1000686. DOI: 10.1371/journal.pgen.1000686.
- 785 [13] X. Zheng and B. S. Weir. “Eigenanalysis of SNP data with an identity by descent interpre-
786 tation”. *Theor Popul Biol* 107 (2016), pp. 65–76. DOI: 10.1016/j.tpb.2015.09.004.
- 787 [14] I. Cabreros and J. D. Storey. “A Likelihood-Free Estimator of Population Structure Bridging
788 Admixture Models and Principal Components Analysis”. *Genetics* 212(4) (2019), pp. 1009–
789 1029. DOI: 10.1534/genetics.119.302159.
- 790 [15] A. M. Chiu et al. “Inferring population structure in biobank-scale genomic data”. *The Amer-*
791 *ican Journal of Human Genetics* 0(0) (2022). DOI: 10.1016/j.ajhg.2022.02.015.
- 792 [16] K. Zhao et al. “An Arabidopsis Example of Association Mapping in Structured Samples”.
793 *PLOS Genetics* 3(1) (2007), e4. DOI: 10.1371/journal.pgen.0030004.

- 794 [17] C. Zhu and J. Yu. "Nonmetric Multidimensional Scaling Corrects for Population Structure in
795 Association Mapping With Different Sample Types". *Genetics* 182(3) (1, 2009), pp. 875–888.
796 DOI: [10.1534/genetics.108.098863](https://doi.org/10.1534/genetics.108.098863).
- 797 [18] J. Novembre et al. "Genes mirror geography within Europe". *Nature* 456(7218) (2008), pp. 98–
798 101. DOI: [10.1038/nature07331](https://doi.org/10.1038/nature07331).
- 799 [19] Y. Zhang and W. Pan. "Principal Component Regression and Linear Mixed Model in Associa-
800 tion Analysis of Structured Samples: Competitors or Complements?" *Genetic Epidemiology*
801 39(3) (2015), pp. 149–155. DOI: [10.1002/gepi.21879](https://doi.org/10.1002/gepi.21879).
- 802 [20] M. Lin et al. "Admixed Populations Improve Power for Variant Discovery and Portability in
803 Genome-Wide Association Studies". *Frontiers in Genetics* 12 (2021).
- 804 [21] H. Xu and Y. Guan. "Detecting Local Haplotype Sharing and Haplotype Association". *Ge-
805 netics* 197(3) (2014), pp. 823–838. DOI: [10.1534/genetics.114.164814](https://doi.org/10.1534/genetics.114.164814).
- 806 [22] J. Qian et al. "A fast and scalable framework for large-scale and ultrahigh-dimensional sparse
807 regression with application to the UK Biobank". *PLOS Genetics* 16(10) (2020), e1009141.
808 DOI: [10.1371/journal.pgen.1009141](https://doi.org/10.1371/journal.pgen.1009141).
- 809 [23] T. Thornton and M. S. McPeek. "ROADTRIPS: case-control association testing with par-
810 tially or completely unknown population and pedigree structure". *Am. J. Hum. Genet.* 86(2)
811 (2010), pp. 172–184. DOI: [10.1016/j.ajhg.2010.01.001](https://doi.org/10.1016/j.ajhg.2010.01.001).
- 812 [24] A. L. Price et al. "New approaches to population stratification in genome-wide association
813 studies". *Nature Reviews Genetics* 11(7) (2010), pp. 459–463. DOI: [10.1038/nrg2813](https://doi.org/10.1038/nrg2813).
- 814 [25] S. Lee et al. "Sparse Principal Component Analysis for Identifying Ancestry-Informative
815 Markers in Genome-Wide Association Studies". *Genetic Epidemiology* 36(4) (2012), pp. 293–
816 302. DOI: [10.1002/gepi.21621](https://doi.org/10.1002/gepi.21621).
- 817 [26] G. Abraham and M. Inouye. "Fast Principal Component Analysis of Large-Scale Genome-
818 Wide Data". *PLOS ONE* 9(4) (2014), e93766. DOI: [10.1371/journal.pone.0093766](https://doi.org/10.1371/journal.pone.0093766).

- 819 [27] K. Galinsky et al. “Fast Principal-Component Analysis Reveals Convergent Evolution of
820 ADH1B in Europe and East Asia”. *The American Journal of Human Genetics* 98(3) (2016),
821 pp. 456–472. DOI: 10.1016/j.ajhg.2015.12.022.
- 822 [28] G. Abraham, Y. Qiu, and M. Inouye. “FlashPCA2: principal component analysis of Biobank-
823 scale genotype datasets”. *Bioinformatics* 33(17) (2017), pp. 2776–2778. DOI: 10.1093/
824 bioinformatics/btx299.
- 825 [29] A. Agrawal et al. “Scalable probabilistic PCA for large-scale genetic variation data”. *PLOS
826 Genetics* 16(5) (2020), e1008773. DOI: 10.1371/journal.pgen.1008773.
- 827 [30] J. Yu et al. “A unified mixed-model method for association mapping that accounts for mul-
828 tiple levels of relatedness”. *Nat. Genet.* 38(2) (2006), pp. 203–208. DOI: 10.1038/ng1702.
- 829 [31] H. M. Kang et al. “Efficient control of population structure in model organism association
830 mapping”. *Genetics* 178(3) (2008), pp. 1709–1723. DOI: 10.1534/genetics.107.080101.
- 831 [32] A. Ochoa and J. D. Storey. “Estimating FST and kinship for arbitrary population structures”.
832 *PLoS Genet* 17(1) (2021), e1009241. DOI: 10.1371/journal.pgen.1009241.
- 833 [33] B. J. Vilhjálmsdóttir and M. Nordborg. “The nature of confounding in genome-wide association
834 studies”. *Nat Rev Genet* 14(1) (2013), pp. 1–2. DOI: 10.1038/nrg3382.
- 835 [34] H. Wang, B. Aragam, and E. P. Xing. “Trade-offs of Linear Mixed Models in Genome-
836 Wide Association Studies”. *Journal of Computational Biology* 29(3) (2022), pp. 233–242.
837 DOI: 10.1089/cmb.2021.0157.
- 838 [35] J. Yang et al. “Advantages and pitfalls in the application of mixed-model association meth-
839 ods”. *Nat Genet* 46(2) (2014), pp. 100–106. DOI: 10.1038/ng.2876.
- 840 [36] H. M. Kang et al. “Variance component model to account for sample structure in genome-
841 wide association studies”. *Nat. Genet.* 42(4) (2010), pp. 348–354. DOI: 10.1038/ng.548.
- 842 [37] C. Wu et al. “A Comparison of Association Methods Correcting for Population Stratification
843 in Case–Control Studies”. *Annals of Human Genetics* 75(3) (2011), pp. 418–427. DOI: 10.
844 1111/j.1469-1809.2010.00639.x.

- 845 [38] J. H. Sul and E. Eskin. “Mixed models can correct for population structure for genomic
846 regions under selection”. *Nature Reviews Genetics* 14(4) (2013), p. 300. DOI: 10 . 1038 /
847 nrg2813-c1.
- 848 [39] W. Zhou et al. “Efficiently controlling for case-control imbalance and sample relatedness in
849 large-scale genetic association studies”. *Nat Genet* 50(9) (2018), pp. 1335–1341. DOI: 10 .
850 1038/s41588-018-0184-y.
- 851 [40] Y. S. Aulchenko, D.-J. de Koning, and C. Haley. “Genomewide rapid association using mixed
852 model and regression: a fast and simple method for genomewide pedigree-based quantitative
853 trait loci association analysis”. *Genetics* 177(1) (2007), pp. 577–585. DOI: 10 . 1534/genetics .
854 107 . 075614.
- 855 [41] Z. Zhang et al. “Mixed linear model approach adapted for genome-wide association studies”.
856 *Nat Genet* 42(4) (2010), pp. 355–360. DOI: 10 . 1038/ng . 546.
- 857 [42] C. Lippert et al. “FaST linear mixed models for genome-wide association studies”. *Nat.*
858 *Methods* 8(10) (2011), pp. 833–835. DOI: 10 . 1038/nmeth . 1681.
- 859 [43] J. Yang et al. “GCTA: a tool for genome-wide complex trait analysis”. *Am. J. Hum. Genet.*
860 88(1) (2011), pp. 76–82. DOI: 10 . 1016/j.ajhg . 2010 . 11 . 011.
- 861 [44] J. Listgarten et al. “Improved linear mixed models for genome-wide association studies”. *Nat*
862 *Methods* 9(6) (2012), pp. 525–526. DOI: 10 . 1038/nmeth . 2037.
- 863 [45] X. Zhou and M. Stephens. “Genome-wide efficient mixed-model analysis for association stud-
864 ies”. *Nat. Genet.* 44(7) (2012), pp. 821–824. DOI: 10 . 1038/ng . 2310.
- 865 [46] G. R. Svishcheva et al. “Rapid variance components-based method for whole-genome asso-
866 ciation analysis”. *Nat Genet* 44(10) (2012), pp. 1166–1170. DOI: 10 . 1038/ng . 2410.
- 867 [47] P.-R. Loh et al. “Efficient Bayesian mixed-model analysis increases association power in large
868 cohorts”. *Nat. Genet.* 47(3) (2015), pp. 284–290. DOI: 10 . 1038/ng . 3190.
- 869 [48] L. Janss et al. “Inferences from Genomic Models in Stratified Populations”. *Genetics* 192(2)
870 (1, 2012), pp. 693–704. DOI: 10 . 1534/genetics . 112 . 141143.

- 871 [49] G. E. Hoffman. “Correcting for population structure and kinship using the linear mixed
872 model: theory and extensions”. *PLoS ONE* 8(10) (2013), e75707. DOI: [10.1371/journal.pone.0075707](https://doi.org/10.1371/journal.pone.0075707).
- 874 [50] G. Tucker, A. L. Price, and B. Berger. “Improving the Power of GWAS and Avoiding Con-
875 founding from Population Stratification with PC-Select”. *Genetics* 197(3) (2014), pp. 1045–
876 1049. DOI: [10.1534/genetics.114.164285](https://doi.org/10.1534/genetics.114.164285).
- 877 [51] N. Liu et al. “Controlling Population Structure in Human Genetic Association Studies with
878 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/sii.2011.v4.n3.a6).
- 880 [52] J. Zeng et al. “Signatures of negative selection in the genetic architecture of human complex
881 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).
- 882 [53] J. Mbatchou et al. “Computationally efficient whole-genome regression for quantitative and
883 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).
- 884 [54] N. Matoba et al. “GWAS of 165,084 Japanese individuals identified nine loci associated with
885 dietary habits”. *Nat Hum Behav* 4(3) (2020), pp. 308–316. DOI: [10.1038/s41562-019-0805-1](https://doi.org/10.1038/s41562-019-0805-1).
- 887 [55] M. Song, W. Hao, and J. D. Storey. “Testing for genetic associations in arbitrarily structured
888 populations”. *Nat. Genet.* 47(5) (2015), pp. 550–554. DOI: [10.1038/ng.3244](https://doi.org/10.1038/ng.3244).
- 889 [56] X. Liu et al. “Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient
890 Genome-Wide Association Studies”. *PLOS Genet* 12(2) (2016), e1005767. DOI: [10.1371/journal.pgen.1005767](https://doi.org/10.1371/journal.pgen.1005767).
- 892 [57] J. H. Sul, L. S. Martin, and E. Eskin. “Population structure in genetic studies: Confounding
893 factors and mixed models”. *PLoS Genet.* 14(12) (2018), e1007309. DOI: [10.1371/journal.pgen.1007309](https://doi.org/10.1371/journal.pgen.1007309).
- 895 [58] P.-R. Loh et al. “Mixed-model association for biobank-scale datasets”. *Nat Genet* 50(7)
896 (2018), pp. 906–908. DOI: [10.1038/s41588-018-0144-6](https://doi.org/10.1038/s41588-018-0144-6).

- 897 [59] A. L. Price et al. “Response to Sul and Eskin”. *Nature Reviews Genetics* 14(4) (2013), p. 300.
- 898 DOI: 10.1038/nrg2813-c2.
- 899 [60] T. G. P. Consortium. “A map of human genome variation from population-scale sequencing”.
- 900 *Nature* 467(7319) (2010), pp. 1061–1073. DOI: 10.1038/nature09534.
- 901 [61] 1000 Genomes Project Consortium et al. “An integrated map of genetic variation from 1,092
- 902 human genomes”. *Nature* 491(7422) (2012), pp. 56–65. DOI: 10.1038/nature11632.
- 903 [62] H. M. Cann et al. “A human genome diversity cell line panel”. *Science* 296(5566) (2002),
- 904 pp. 261–262. DOI: 10.1126/science.296.5566.261b.
- 905 [63] N. A. Rosenberg et al. “Genetic Structure of Human Populations”. *Science* 298(5602) (2002),
- 906 pp. 2381–2385. DOI: 10.1126/science.1078311.
- 907 [64] A. Bergström et al. “Insights into human genetic variation and population history from 929
- 908 diverse genomes”. *Science* 367(6484) (2020). DOI: 10.1126/science.aay5012.
- 909 [65] N. Patterson et al. “Ancient admixture in human history”. *Genetics* 192(3) (2012), pp. 1065–
- 910 1093. DOI: 10.1534/genetics.112.145037.
- 911 [66] I. Lazaridis et al. “Ancient human genomes suggest three ancestral populations for present-
- 912 day Europeans”. *Nature* 513(7518) (2014), pp. 409–413. DOI: 10.1038/nature13673.
- 913 [67] I. Lazaridis et al. “Genomic insights into the origin of farming in the ancient Near East”.
- 914 *Nature* 536(7617) (2016), pp. 419–424. DOI: 10.1038/nature19310.
- 915 [68] P. Skoglund et al. “Genomic insights into the peopling of the Southwest Pacific”. *Nature*
- 916 538(7626) (2016), pp. 510–513. DOI: 10.1038/nature19844.
- 917 [69] J.-H. Park et al. “Distribution of allele frequencies and effect sizes and their interrelationships
- 918 for common genetic susceptibility variants”. *PNAS* 108(44) (2011), pp. 18026–18031. DOI:
- 919 10.1073/pnas.1114759108.
- 920 [70] L. J. O’Connor et al. “Extreme Polygenicity of Complex Traits Is Explained by Negative
- 921 Selection”. *The American Journal of Human Genetics* 0(0) (2019). DOI: 10.1016/j.ajhg.
- 922 2019.07.003.

- 923 [71] Y. B. Simons et al. “A population genetic interpretation of GWAS findings for human quantitative traits”. *PLOS Biology* 16(3) (2018), e2002985. DOI: 10.1371/journal.pbio.2002985.
- 924
- 925 [72] G. Malécot. *Mathématiques de l'hérédité*. Masson et Cie, 1948.
- 926 [73] S. Wright. “The Genetical Structure of Populations”. *Annals of Eugenics* 15(1) (1949), pp. 323–354. DOI: 10.1111/j.1469-1809.1949.tb02451.x.
- 927
- 928 [74] A. Jacquard. *Structures génétiques des populations*. Paris: Masson et Cie, 1970.
- 929 [75] A. Ochoa and J. D. Storey. *New kinship and FST estimates reveal higher levels of differentiation in the global human population*. 2019. DOI: 10.1101/653279.
- 930
- 931 [76] Z. Hou and A. Ochoa. “Genetic association models are robust to common population kinship estimation biases”. *Genetics* (27, 2023), iyad030. DOI: 10.1093/genetics/iyad030.
- 932
- 933 [77] C. C. Chang et al. “Second-generation PLINK: rising to the challenge of larger and richer datasets”. *GigaScience* 4(1) (2015), p. 7. DOI: 10.1186/s13742-015-0047-8.
- 934
- 935 [78] D. J. Balding and R. A. Nichols. “A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity”. *Genetica* 96(1-2) (1995), pp. 3–12. DOI: <https://doi.org/10.1007/BF01441146>.
- 936
- 937
- 938 [79] E. Paradis and K. Schliep. “ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R”. *Bioinformatics* 35(3) (2019), pp. 526–528. DOI: 10.1093/bioinformatics/bty633.
- 939
- 940
- 941 [80] R. R. Sokal and C. D. Michener. “A statistical method for evaluating systematic relationships.” *Univ. Kansas, Sci. Bull.* 38 (1958), pp. 1409–1438.
- 942
- 943 [81] C. L. Lawson and R. J. Hanson. *Solving least squares problems*. Englewood Cliffs: Prentice Hall, 1974.
- 944
- 945 [82] K. M. Mullen and I. H. M. v. Stokkum. *nnls: The Lawson-Hanson algorithm for non-negative least squares (NNLS)*. 2012.
- 946

- 947 [83] J.-H. Park et al. “Estimation of effect size distribution from genome-wide association studies
948 and implications for future discoveries”. *Nature Genetics* 42(7) (2010), pp. 570–575. DOI:
949 10.1038/ng.610.
- 950 [84] A. Grueneberg and G. d. l. Campos. “BGData - A Suite of R Packages for Genomic Analysis
951 with Big Data”. *G3: Genes, Genomes, Genetics* 9(5) (2019), pp. 1377–1383. DOI: 10.1534/
952 g3.119.400018.
- 953 [85] S. Fairley et al. “The International Genome Sample Resource (IGSR) collection of open
954 human genomic variation resources”. *Nucleic Acids Research* 48(D1) (2020), pp. D941–D947.
955 DOI: 10.1093/nar/gkz836.
- 956 [86] A. Manichaikul et al. “Robust relationship inference in genome-wide association studies”.
957 *Bioinformatics* 26(22) (2010), pp. 2867–2873. DOI: 10.1093/bioinformatics/btq559.
- 958 [87] J. D. Storey. “The positive false discovery rate: a Bayesian interpretation and the q-value”.
959 *Ann. Statist.* 31(6) (2003), pp. 2013–2035. DOI: 10.1214/aos/1074290335.
- 960 [88] J. D. Storey and R. Tibshirani. “Statistical significance for genomewide studies”. *Proceedings
961 of the National Academy of Sciences of the United States of America* 100(16) (2003),
962 pp. 9440–9445. DOI: 10.1073/pnas.1530509100.
- 963 [89] J. Grau, I. Grosse, and J. Keilwagen. “PRROC: computing and visualizing precision-recall
964 and receiver operating characteristic curves in R”. *Bioinformatics* 31(15) (2015), pp. 2595–
965 2597. DOI: 10.1093/bioinformatics/btv153.
- 966 [90] P. Gopalan et al. “Scaling probabilistic models of genetic variation to millions of humans”.
967 *Nat. Genet.* 48(12) (2016), pp. 1587–1590. DOI: 10.1038/ng.3710.
- 968 [91] B. Rakitsch et al. “A Lasso multi-marker mixed model for association mapping with pop-
969 ulation structure correction”. *Bioinformatics* 29(2) (2013), pp. 206–214. DOI: 10.1093/
970 bioinformatics/bts669.
- 971 [92] M. Conomos et al. “Model-free Estimation of Recent Genetic Relatedness”. *The American
972 Journal of Human Genetics* 98(1) (2016), pp. 127–148. DOI: 10.1016/j.ajhg.2015.11.022.

- 973 [93] D. Heckerman et al. “Linear mixed model for heritability estimation that explicitly addresses
974 environmental variation”. *Proc. Natl. Acad. Sci. U.S.A.* 113(27) (2016), pp. 7377–7382. DOI:
975 10.1073/pnas.1510497113.
- 976 [94] G. Sun et al. “Variation explained in mixed-model association mapping”. *Heredity* 105(4)
977 (2010), pp. 333–340. DOI: 10.1038/hdy.2010.11.
- 978 [95] S. Gazal et al. “High level of inbreeding in final phase of 1000 Genomes Project”. *Sci Rep*
979 5(1) (2015), p. 17453. DOI: 10.1038/srep17453.
- 980 [96] A. Al-Khudhair et al. “Inference of Distant Genetic Relations in Humans Using “1000 Genomes””.
981 *Genome Biology and Evolution* 7(2) (2015), pp. 481–492. DOI: 10.1093/gbe/evv003.
- 982 [97] L. Fedorova et al. “Atlas of Cryptic Genetic Relatedness Among 1000 Human Genomes”.
983 *Genome Biology and Evolution* 8(3) (2016), pp. 777–790. DOI: 10.1093/gbe/evw034.
- 984 [98] D. Schlauch, H. Fier, and C. Lange. “Identification of genetic outliers due to sub-structure
985 and cryptic relationships”. *Bioinformatics* 33(13) (2017), pp. 1972–1979. DOI: 10.1093/
986 bioinformatics/btx109.
- 987 [99] B. M. Henn et al. “Cryptic Distant Relatives Are Common in Both Isolated and Cosmopolitan
988 Genetic Samples”. *PLOS ONE* 7(4) (2012), e34267. DOI: 10.1371/journal.pone.0034267.
- 989 [100] V. Shchur and R. Nielsen. “On the number of siblings and p-th cousins in a large population
990 sample”. *J Math Biol* 77(5) (2018), pp. 1279–1298. DOI: 10.1007/s00285-018-1252-8.
- 991 [101] F. Privé et al. “Efficient toolkit implementing best practices for principal component analysis
992 of population genetic data”. *Bioinformatics* 36(16) (15, 2020), pp. 4449–4457. DOI: 10.1093/
993 bioinformatics/btaa520.
- 994 [102] D. Speed et al. “Improved heritability estimation from genome-wide SNPs”. *Am. J. Hum.
995 Genet.* 91(6) (7, 2012), pp. 1011–1021. DOI: 10.1016/j.ajhg.2012.10.010.
- 996 [103] A. A. Zaidi and I. Mathieson. “Demographic history mediates the effect of stratification on
997 polygenic scores”. *eLife* 9 (17, 2020). Ed. by G. H. Perry, M. C. Turchin, and A. R. Martin,
998 e61548. DOI: 10.7554/eLife.61548.

Supplemental figures

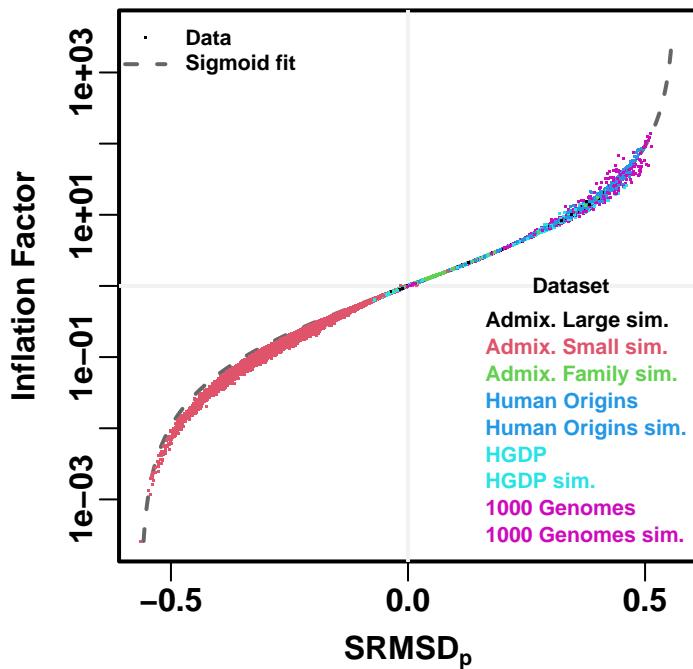


Figure S1: **Comparison between SRMSD_p and inflation factor.** Each point is a pair of statistics for one replicate, one association model (PCA or LMM with some number of PCs r), one trait model (FES vs RC, all heritability/environments tested), and one dataset (color coded by dataset). Note log y-axis.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

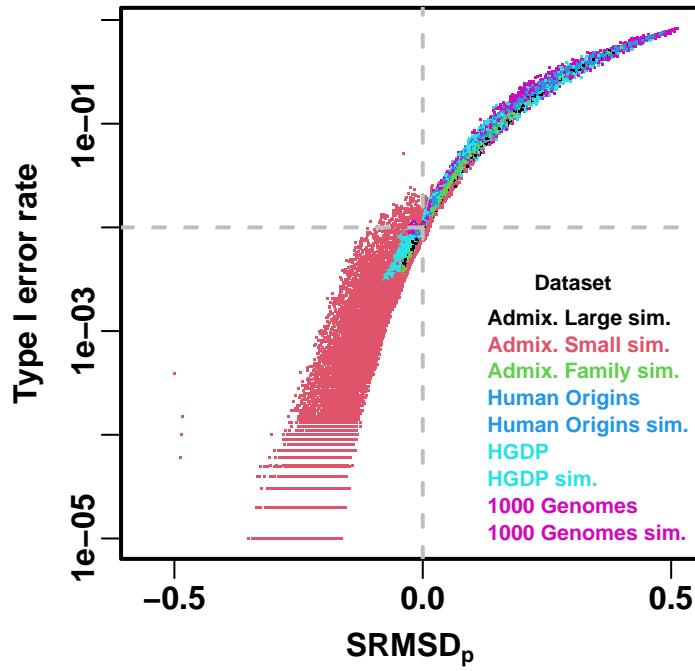


Figure S2: **Comparison between SRMSD_p and type I error rate.** Type I error rate calculated at a p-value threshold of $1\text{e-}2$ (horizontal dashed gray line). Thus, a calibrated model has a type I error rate of $1\text{e-}2$ and $\text{SRMSD}_p = 0$ (where the dashed lines meet). As expected, increased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

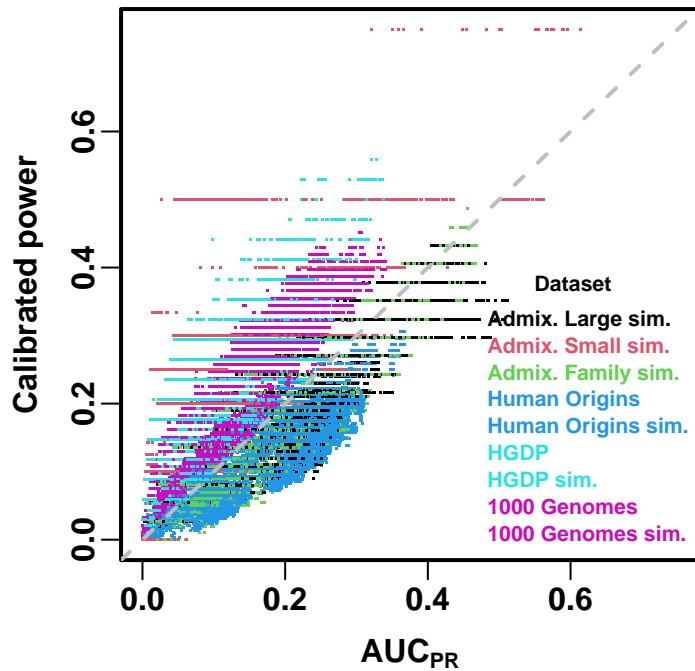


Figure S3: **Comparison between AUC_{PR} and calibrated power.** Calibrated power is power calculated at an empirical type I error threshold of $1e-4$. Each point is a pair of statistics for one replicate, one association model (PCA or LMM with some number of PCs r), one trait model (FES vs RC, all heritability/environments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

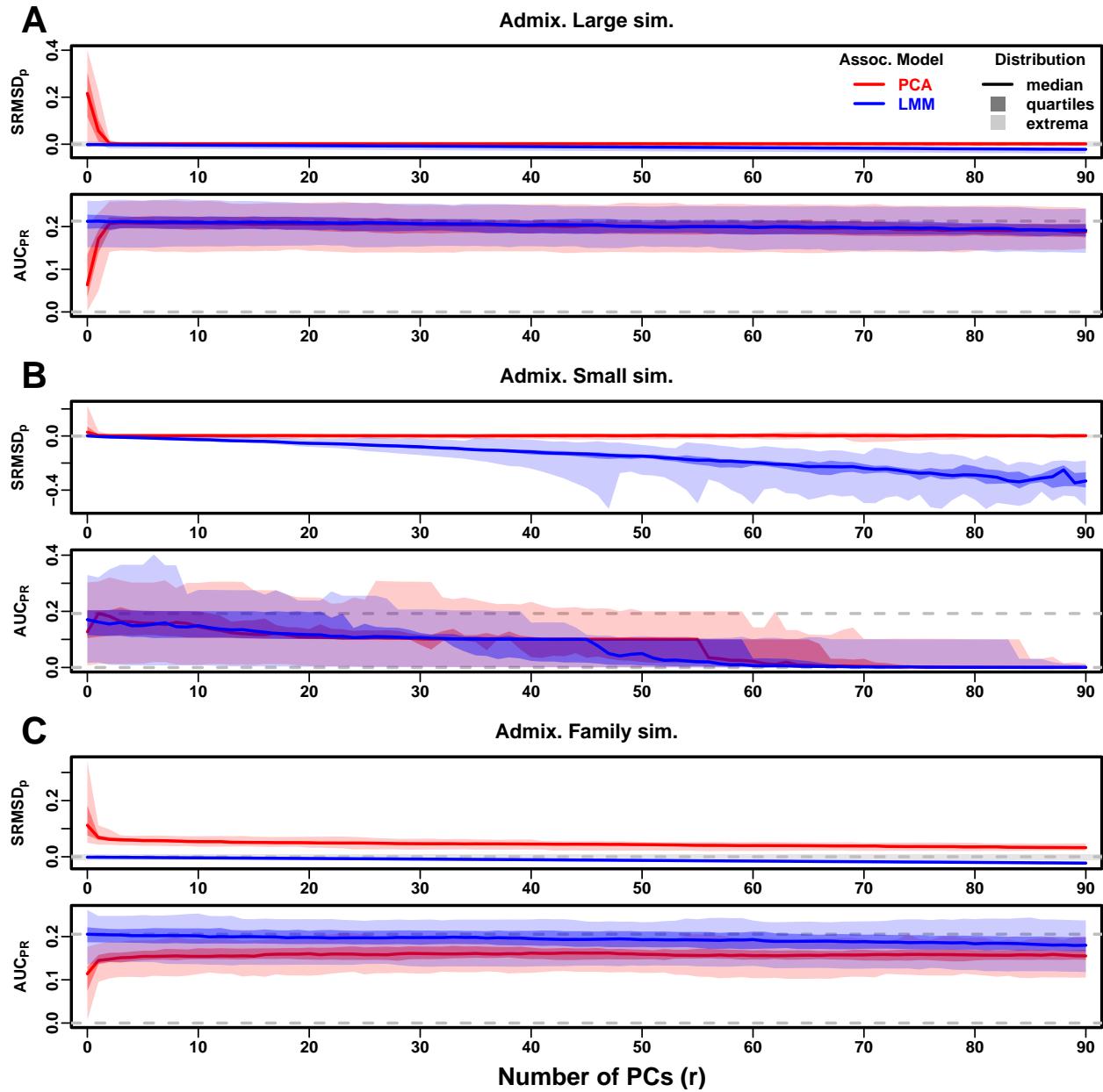


Figure S4: Evaluations in admixture simulations with RC traits. Traits simulated from RC model, otherwise the same as Fig. 3.

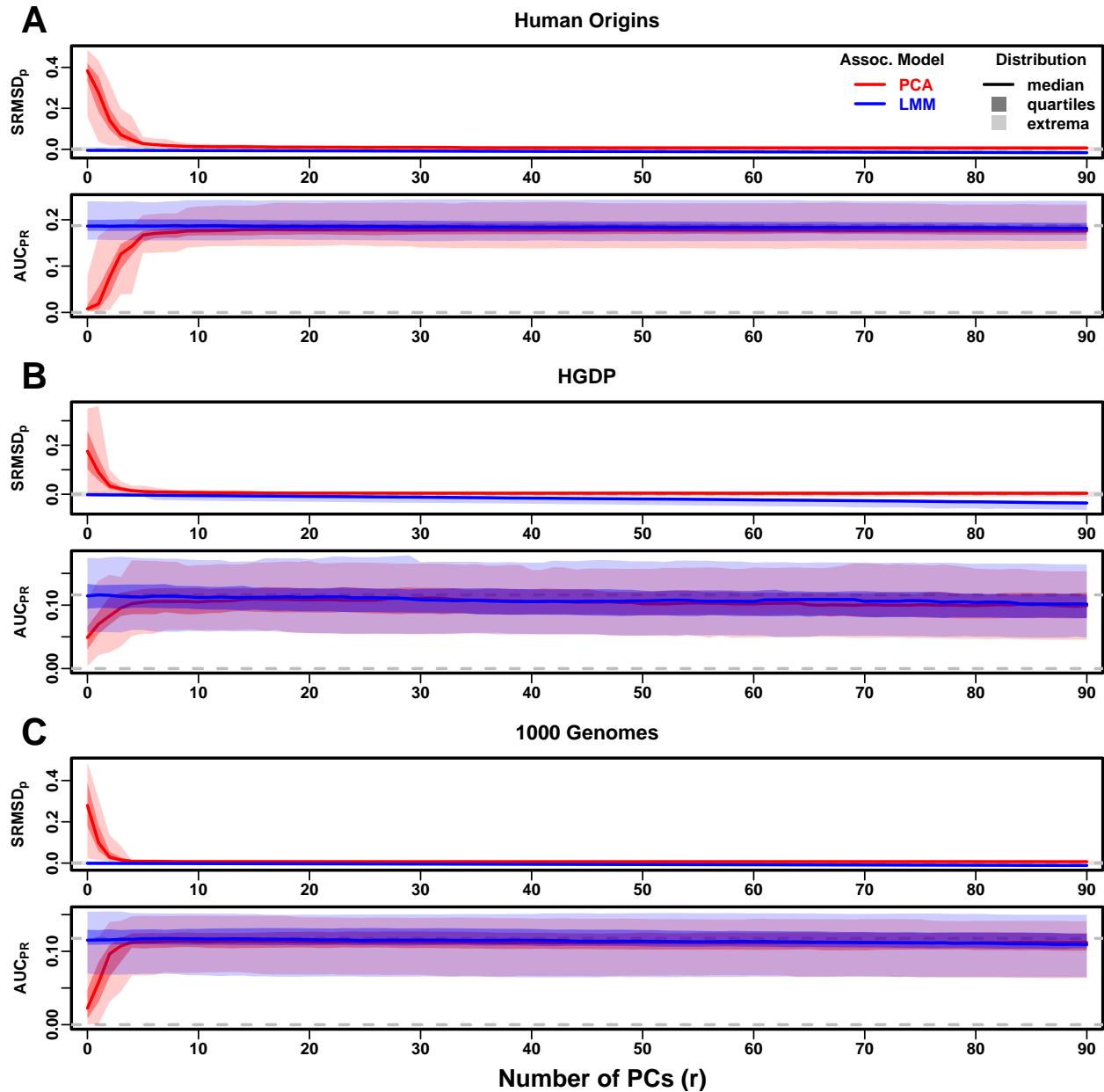


Figure S5: Evaluations in real human genotype datasets with RC traits. Traits simulated from RC model, otherwise the same as Fig. 4.

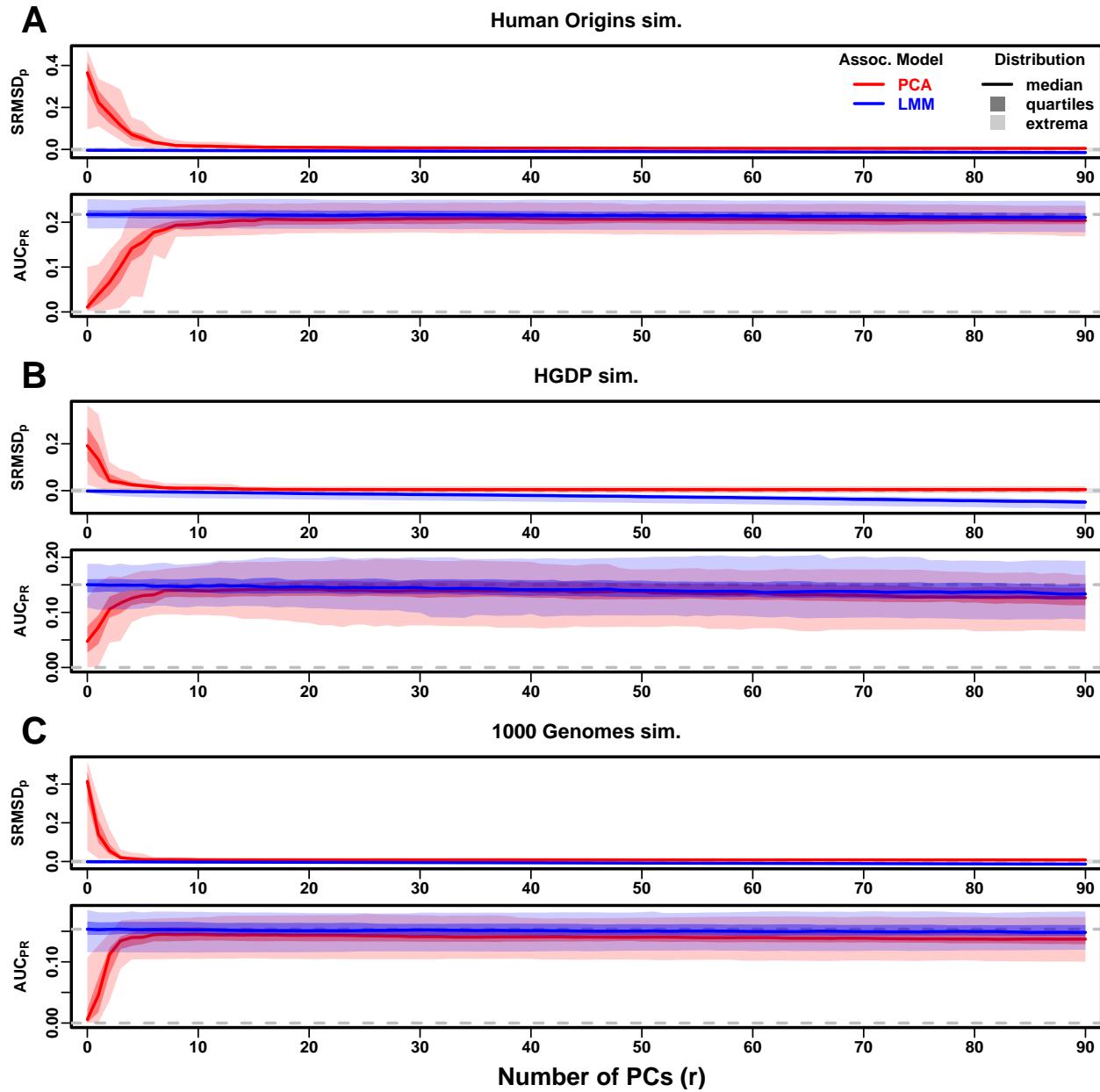


Figure S6: Evaluations in tree simulations fit to human data with RC traits. Traits simulated from RC model, otherwise the same as Fig. 5.

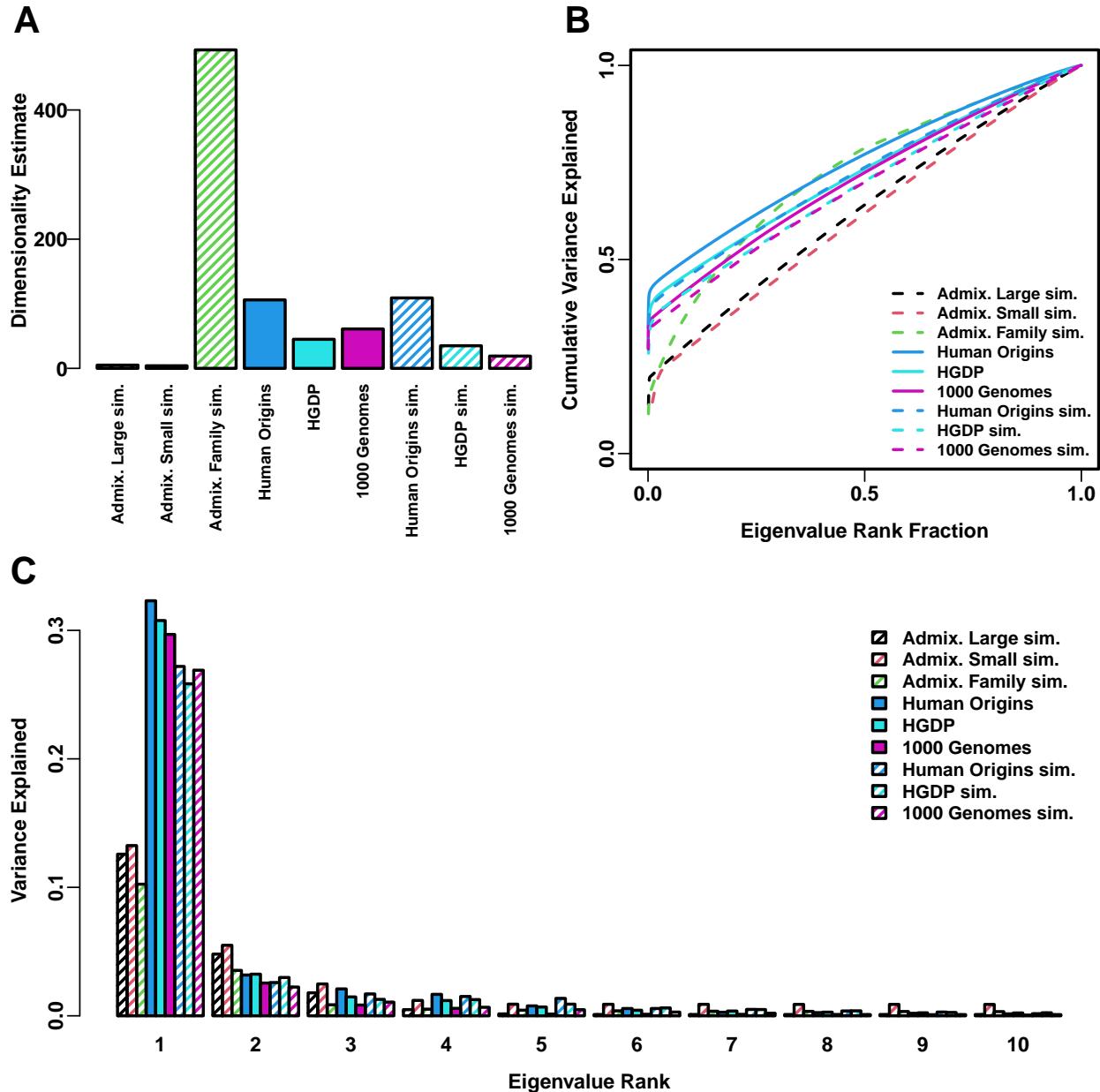


Figure S7: **Estimated dimensionality of datasets.** **A.** Kinship dimensionalities estimated with the Tracy-Widom test with $p < 0.01$. **B.** Cumulative variance explained versus eigenvalue rank fraction. **C.** Variance explained by first 10 eigenvalues.

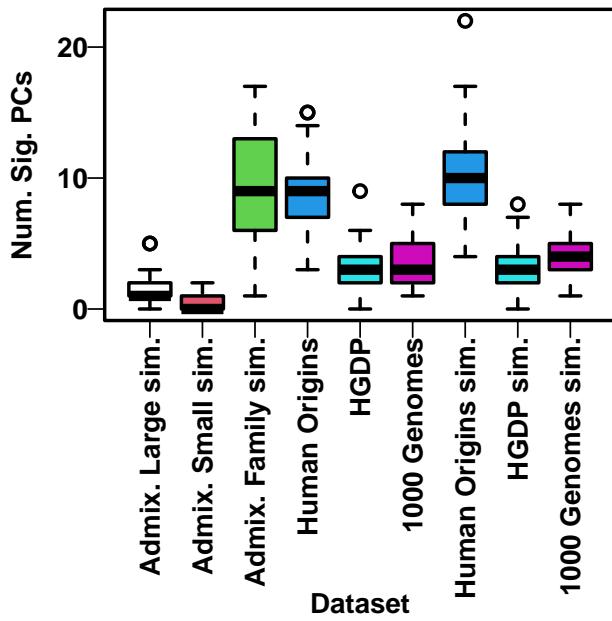


Figure S8: **Number of PCs significantly associated with traits.** PCs are tested using an ordinary linear regression sequentially, with the k th PC tested conditionally on the previous $k - 1$ PCs and the intercept. Q-values are estimated from the 90 p-values (one for each PC in a given dataset and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

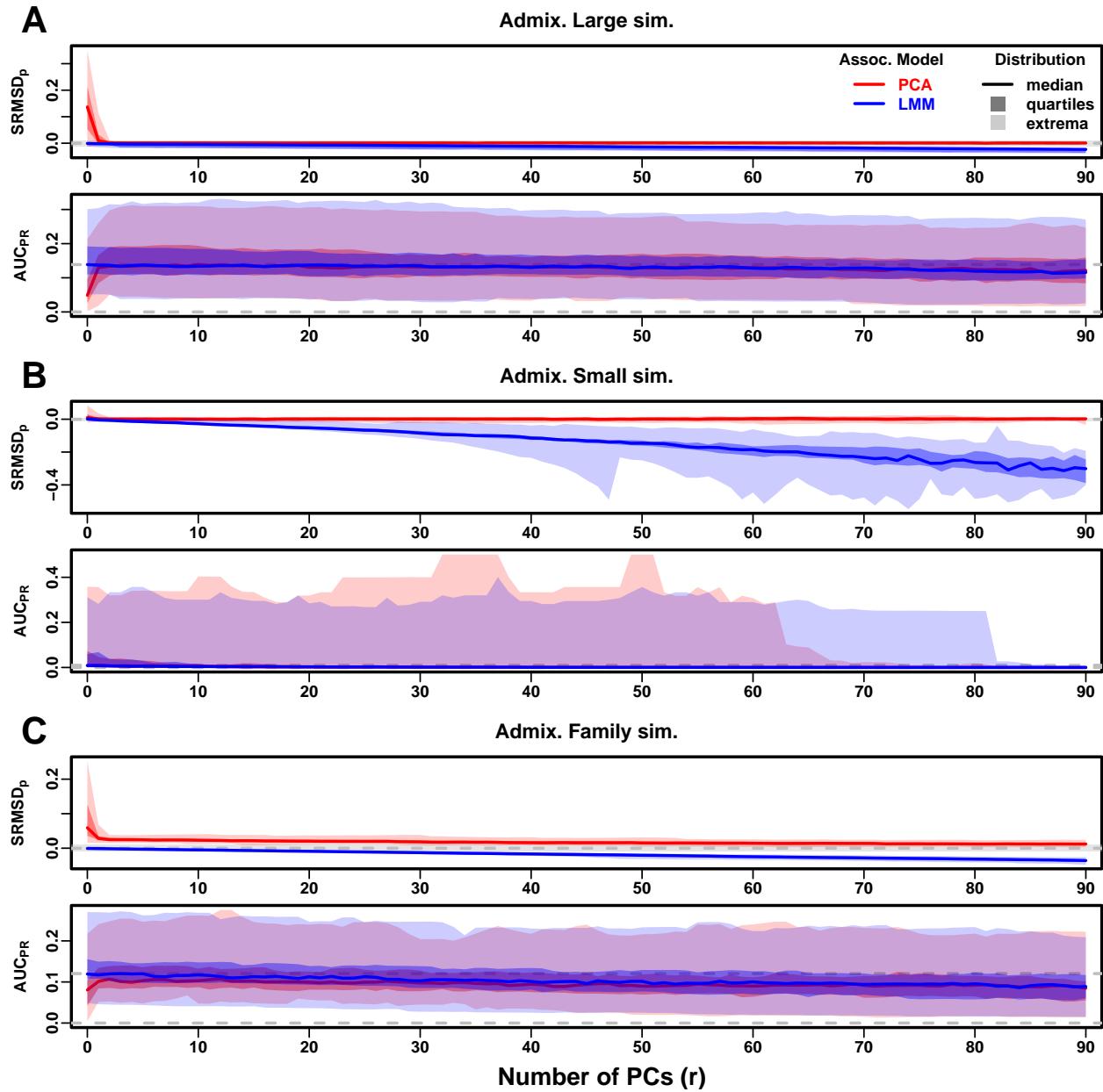


Figure S9: Evaluations in admixture simulations with FES traits, low heritability. Traits simulated using $h^2 = 0.3$, otherwise the same as Fig. 3.

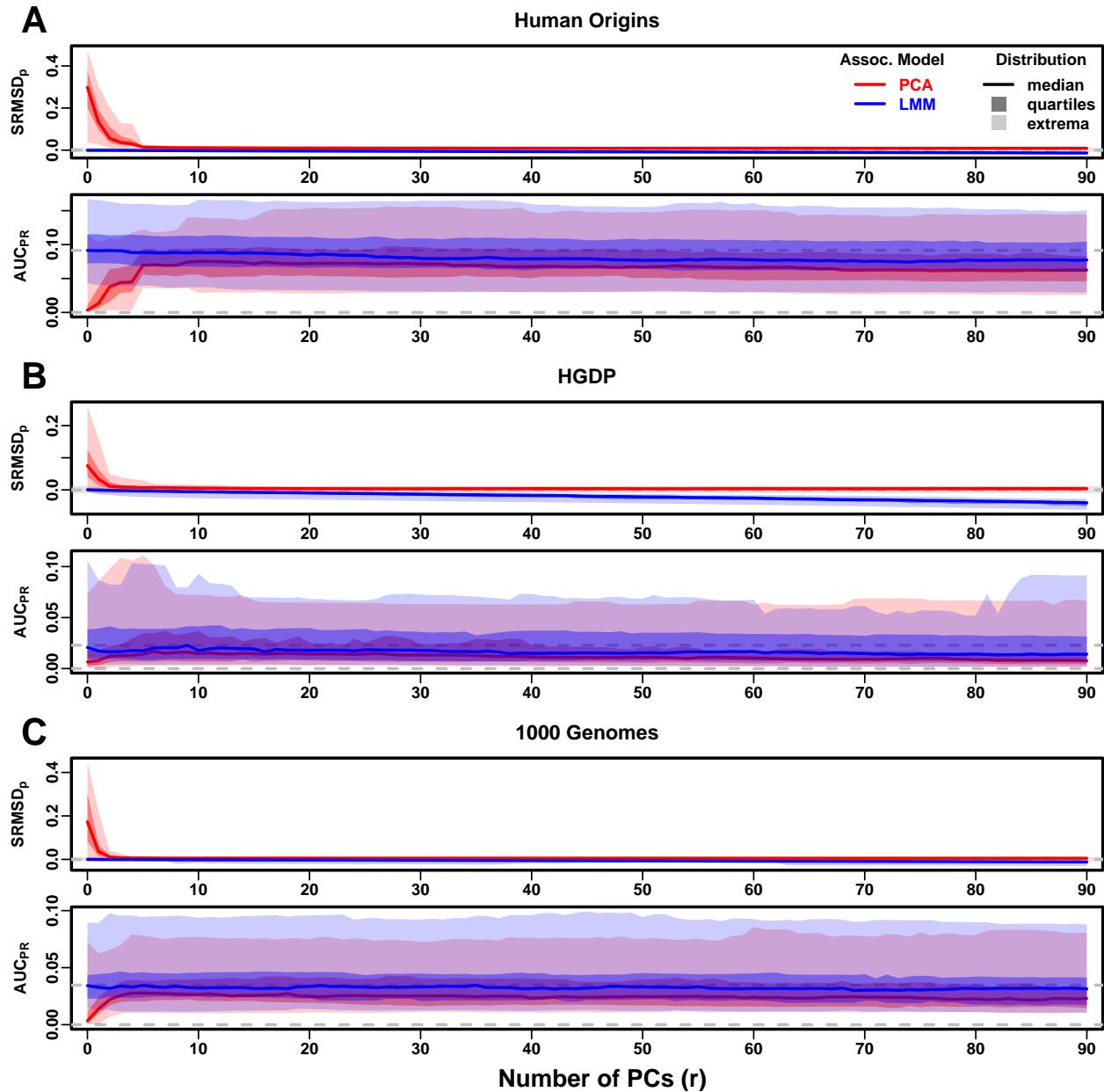


Figure S10: Evaluations in real human genotype datasets with FES traits, low heritability. Traits simulated using $h^2 = 0.3$, otherwise the same as Fig. 4.

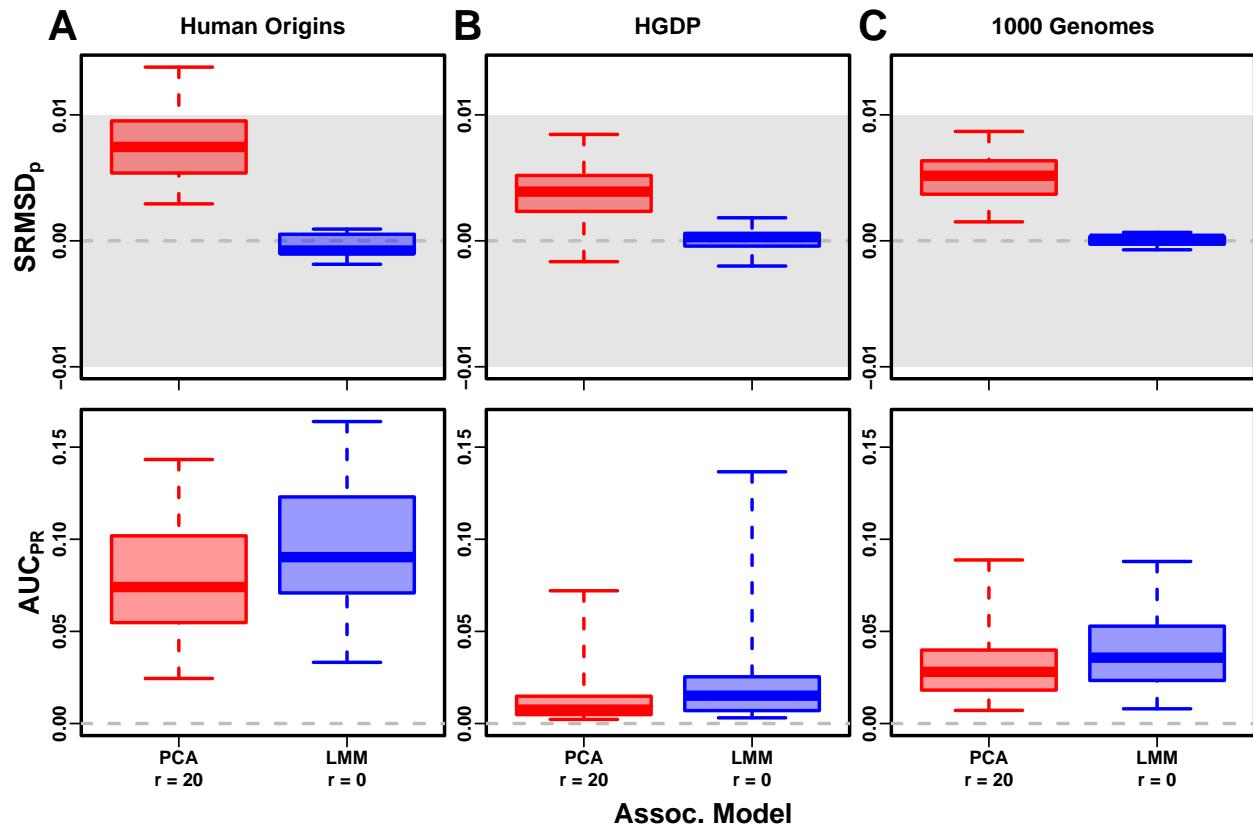


Figure S11: Evaluation in real datasets excluding 4th degree relatives, FES traits, low heritability. Traits simulated using $h^2 = 0.3$, otherwise the same as Fig. 7.

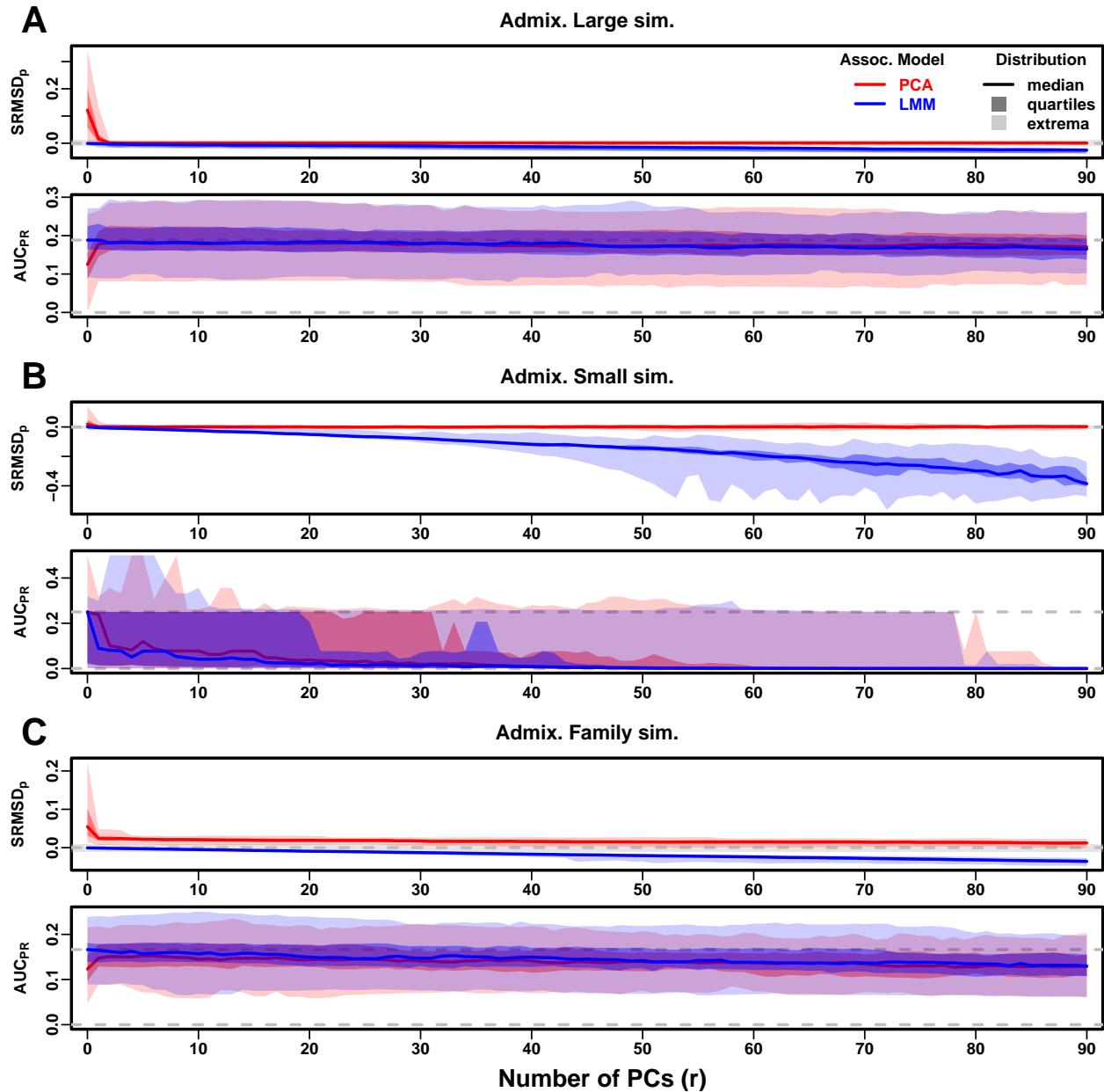


Figure S12: **Evaluations in admixture simulations with RC traits, low heritability.** Traits simulated using $h^2 = 0.3$, otherwise the same as Fig. S4.

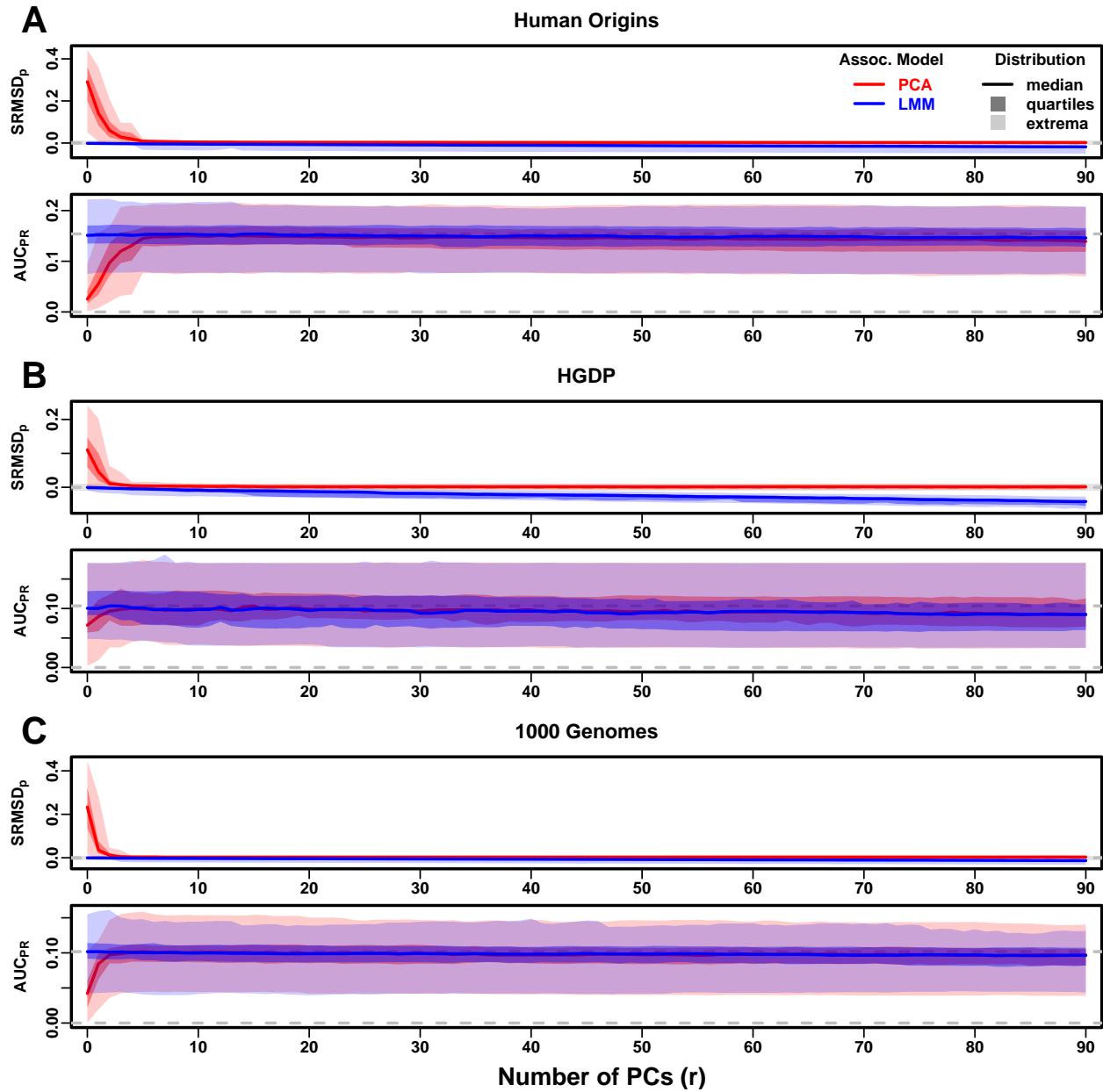


Figure S13: Evaluations in real human genotype datasets with RC traits, low heritability. Traits simulated using $h^2 = 0.3$, otherwise the same as Fig. S5.

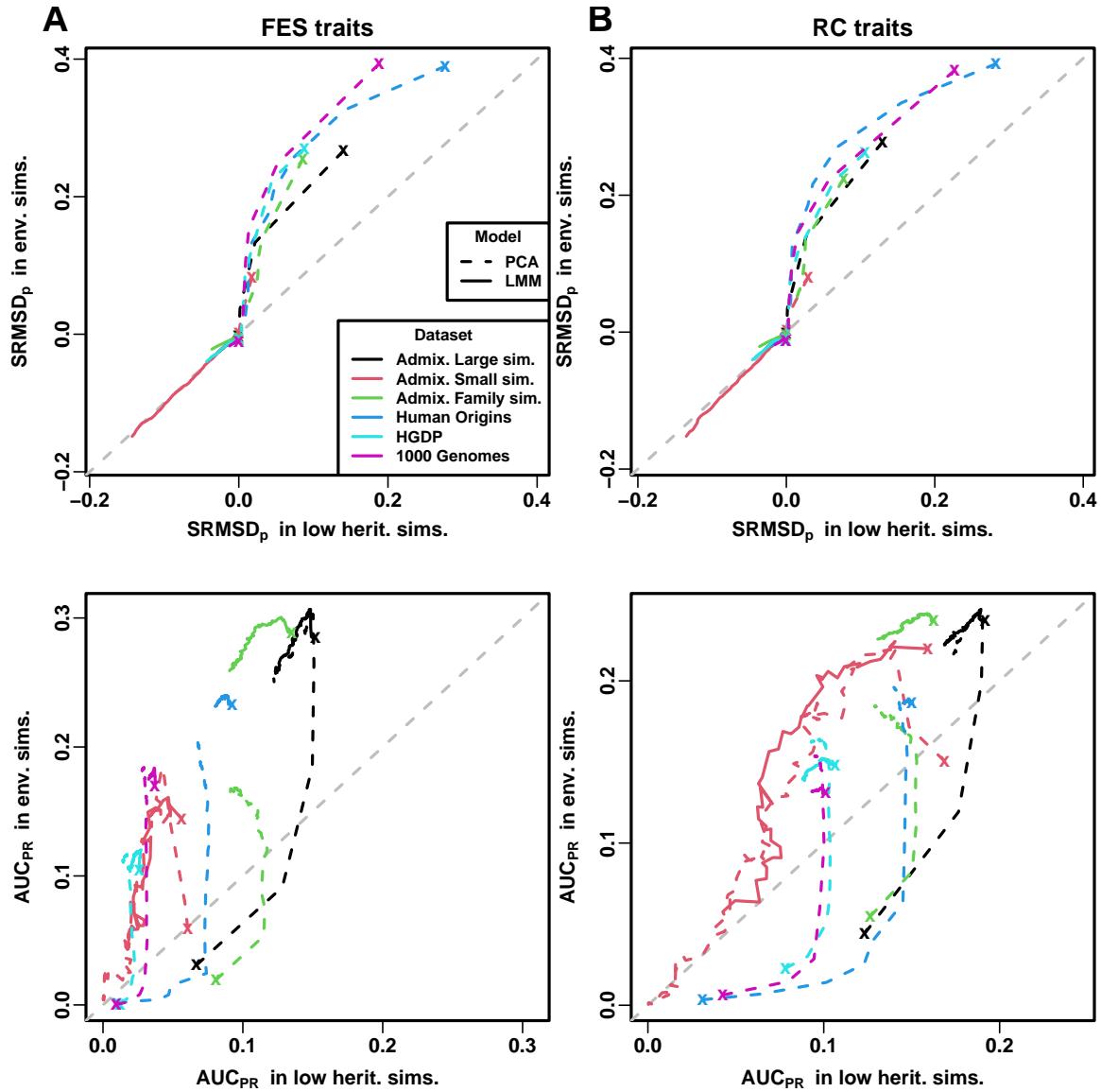


Figure S14: Comparison of performance in low heritability vs environment simulations. Each curve traces as the number of PCs r is increased from $r = 0$ (marked with an “ x ”) until $r = 90$ (unmarked end), on one axis is the mean value over replicates of either SRMSD_p or AUC_{PR} , for low heritability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

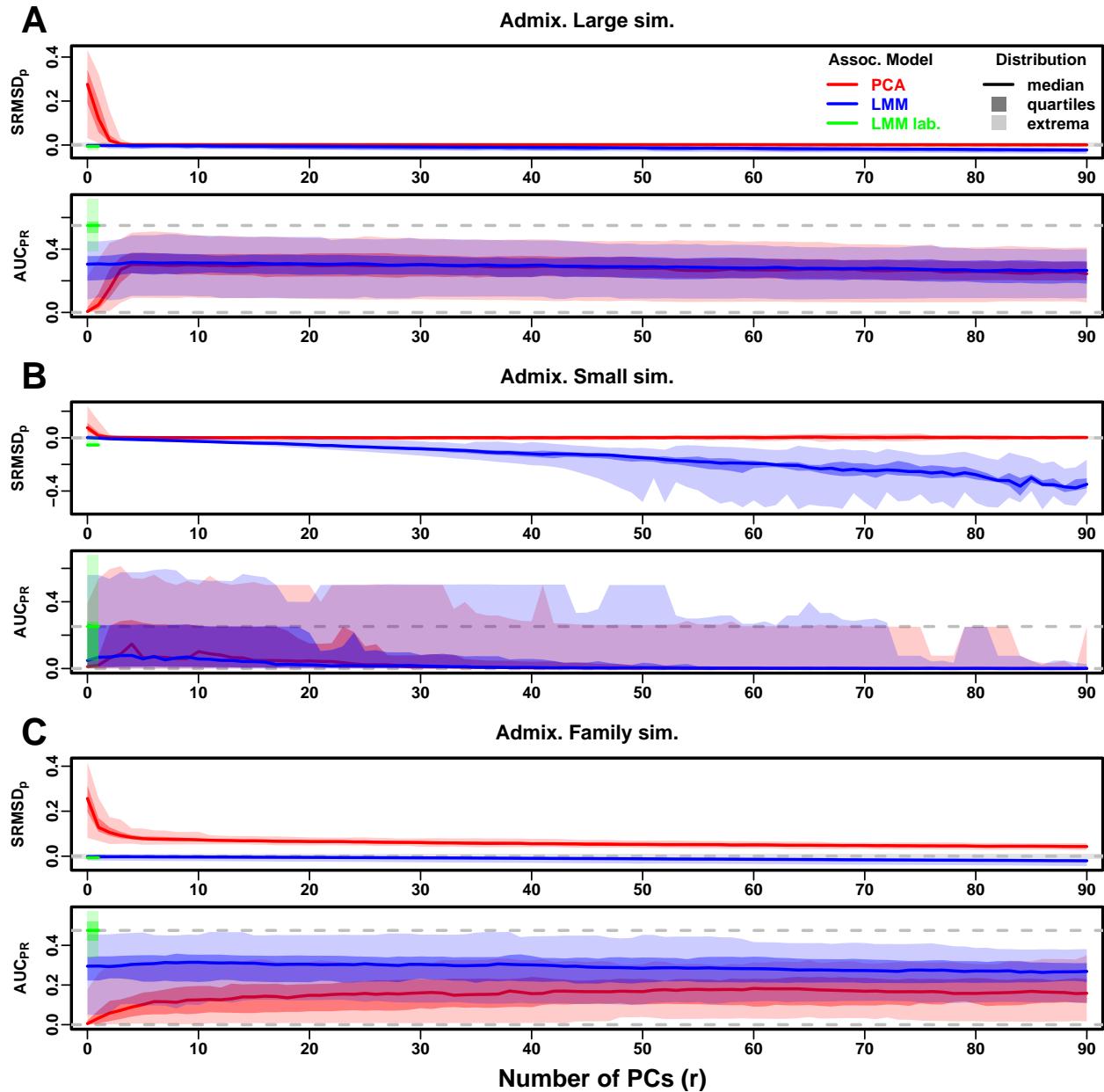


Figure S15: Evaluations in admixture simulations with FES traits, environment. Traits simulated with environment effects, otherwise the same as Fig. S9.

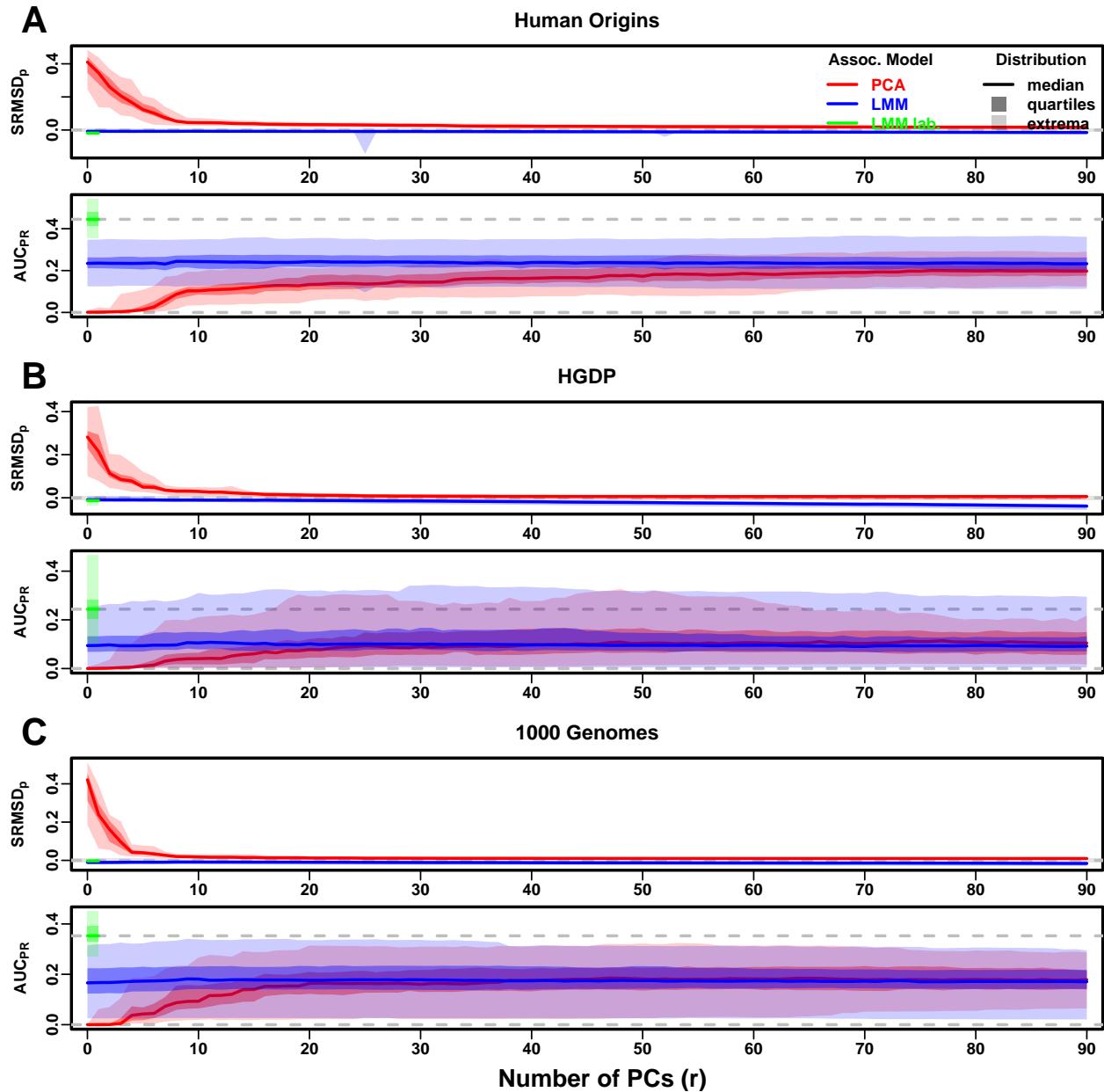


Figure S16: Evaluations in real human genotype datasets with FES traits, environment. Traits simulated with environment effects, otherwise the same as Fig. S10.

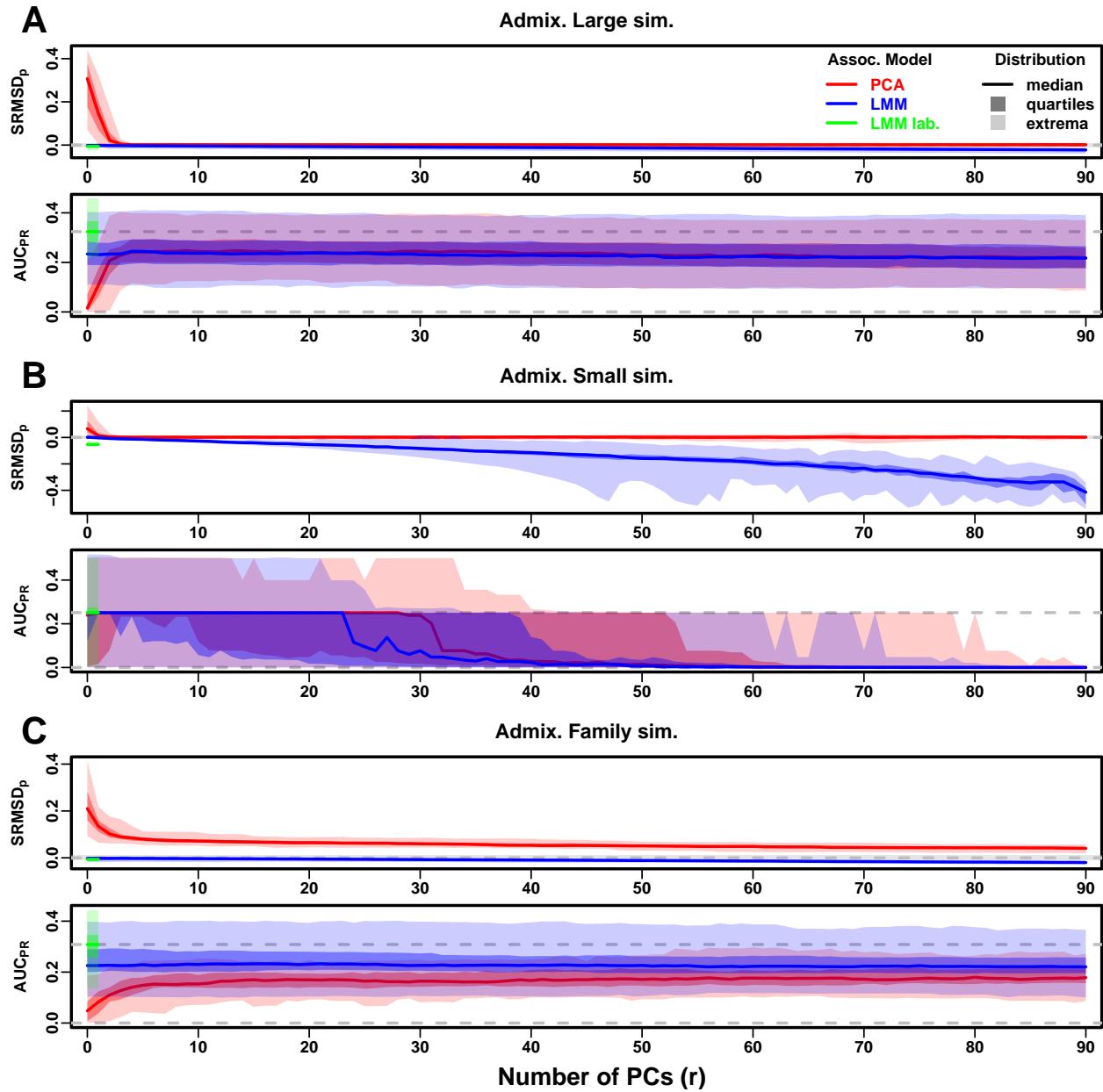


Figure S17: Evaluations in admixture simulations with RC traits, environment. Traits simulated with environment effects, otherwise the same as Fig. S12.

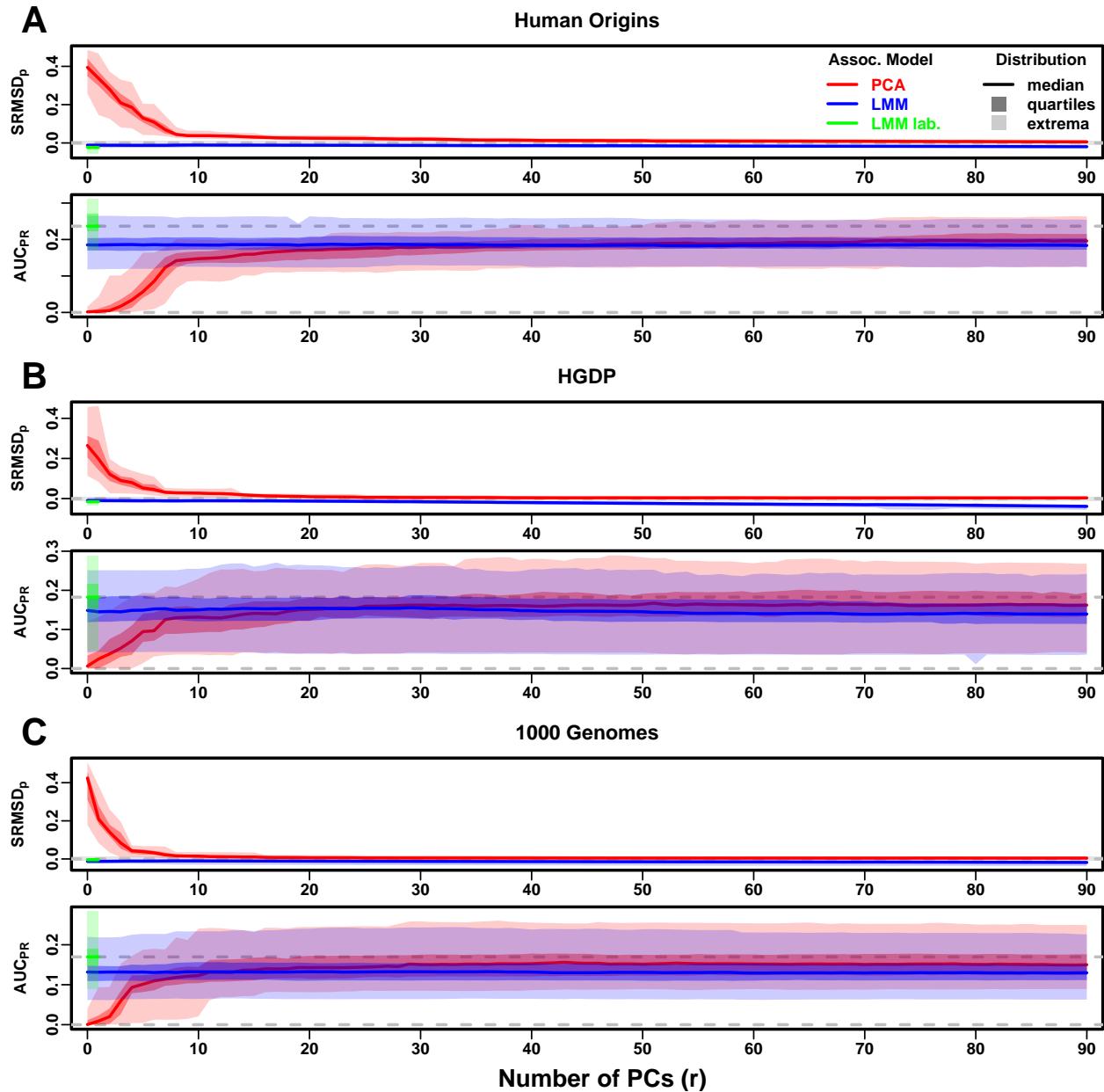


Figure S18: Evaluations in real human genotype datasets with RC traits, environment. Traits simulated with environment effects, otherwise the same as Fig. S13.

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. 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.