

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

3           Yiqi Yao,<sup>1,3</sup> Alejandro Ochoa<sup>1,2,\*</sup>

4           <sup>1</sup> Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27705, USA

5           <sup>2</sup> Duke Center for Statistical Genetics and Genomics, Duke University, Durham, NC 27705, USA

6           <sup>3</sup> 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 new realistic genotype and complex trait simulations including ad-  
15          mixed families, trees, and real multiethnic human genotype datasets (1000 Genomes, Human  
16          Genome Diversity Panel, Human Origins) with simulated traits. We find that LMM without  
17          PCs always performs best, with the largest effects in the family simulation and real human  
18          datasets. The large gaps in PCA-LMM performance on human datasets is driven by the high-  
19          dimensional effect of large numbers of distant relatives more than the smaller number of highly  
20          related pairs. While PCA was known to fail on family data, we report strong effects of family  
21          relatedness in several genetically diverse human datasets, not avoided by pruning highly related  
22          individual pairs. This work better characterizes the severe limitations of PCA compared to  
23          LMM in modeling the complex relatedness structures of multiethnic human data for association  
24          studies.

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

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

## 28 1 Introduction

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

37 Genetic association with PCA consists of including the top eigenvectors of the population kin-  
38 ship matrix as covariates in a generalized linear model [4–6]. These top eigenvectors are commonly  
39 referred to as PCs in genetics [7], the convention adopted here, but in other fields PCs denote  
40 the projections of loci onto eigenvectors [8]. The direct ancestor of PCA association is structured  
41 association, in which inferred ancestry or admixture proportions are used as regression covariates  
42 [9]. These models are deeply connected because PCs map to ancestry empirically [10, 11] and the-  
43oretically [12–15], and they work as well as global ancestry in association studies but are estimated  
44 more easily [6, 7, 10, 16]. The strength of PCA is its simplicity, which as covariates can be readily  
45 included in more complex models, such as haplotype association [17] and polygenic models [18].  
46 However, PCA assumes that relatedness is low-dimensional, which may limit its applicability. PCA  
47 is known to be inadequate for family data [7, 19, 20], which is called “cryptic relatedness” when  
48 it is unknown to the researchers, but no other troublesome cases have been confidently identified.  
49 Recent work has focused on developing more scalable versions of the PCA algorithm [21–25]. PCA  
50 remains a popular and powerful approach for association studies.

51 The other dominant association model under relatedness is the LMM, which includes a random  
52 effect parametrized by the kinship matrix. Unlike PCA, LMM does not assume that relatedness

53 is low-dimensional, and explicitly models families via the kinship matrix. Early LMMs required  
54 kinship matrices estimated from known pedigrees or which otherwise captured recent relatedness  
55 only [16, 26]. Modern LMMs estimate kinship from genotypes using a non-parametric estimator,  
56 often referred to as a genetic relationship matrix, that captures the combined covariance due to  
57 recent family relatedness and ancestral population structure [1, 27, 28]. The classic LMM assumes  
58 a quantitative (continuous) complex trait, the focus of our work. Although case-control (binary)  
59 traits and their underlying ascertainment are theoretically a challenge [29], LMMs have been applied  
60 successfully to balanced case-control studies [1, 30] and simulations [20, 31, 32], and have been  
61 adapted for unbalanced case-control studies [33]. However, LMMs tend to be considerably slower  
62 than PCA and other models, so much effort has focused on improving their runtime and scalability  
63 [27, 30, 33–41].

64 An LMM variant that incorporates PCs as fixed covariates is tested thoroughly in our work.  
65 Since PCs are the top eigenvectors of the same kinship matrix estimate used in modern LMMs [1,  
66 42], then population structure is modeled twice in an LMM with PCs. However, some previous  
67 work has found the apparent redundancy of an LMM with PCs beneficial [20, 43], while others  
68 did not [44], and the approach continues to be used [45]. Recall that early LMMs used kinship to  
69 model family relatedness only, so population structure had to be modeled separately, in practice as  
70 admixture fractions instead of PCs [16, 26].

71 LMM and PCA are closely related models [1, 42], so similar performance is expected particu-  
72 larly under low-dimensional relatedness. Direct comparisons have yielded mixed results, with several  
73 studies finding superior performance for LMM (notably from papers promoting advances in LMMs)  
74 while many others report comparable performance (Table 1). No papers find that PCA outper-  
75 forms LMM decisively, although PCA occasionally performs better in isolated and artificial cases  
76 or individual measures (often with unknown significance). Previous studies were generally divided  
77 those that employed simulated versus real genotypes (only one study used both). The simulated  
78 genotype studies, which tended to have low dimensionalities and differentiation ( $F_{ST}$ ), were more  
79 likely to report ties or mixed results (6/7), whereas real genotypes tended to clearly favor LMMs  
80 (5/7). Similarly, 6/8 papers with quantitative traits favor LMMs, whereas 5/7 papers with case-

81 control traits gave ties or mixed results (the only factor we do not explore). Additionally, although  
 82 all previous evaluations measured type I error (or proxies such as inflation factors or QQ plots), a  
 83 large fraction (5/13) did not measure power (including proxies such as ROC curves), and only two  
 84 used more than one number of PCs for PCA. Lastly, no consensus has emerged as to why LMM  
 85 might outperform PCA or vice versa [20, 32, 42, 49], or which features of the real datasets are  
 86 critical for the LMM advantage other than cryptic relatedness, resulting in unclear guidance for  
 87 using PCA. Hence, our work includes real and simulated genotypes with higher dimensionalities  
 88 and differentiation matching that of multiethnic human cohorts, we vary the number of PCs, and  
 89 measure robust proxies for type I error control and power.

90 In this work, we evaluate the PCA and LMM association models under various numbers of  
 91 PCs (included in LMM too). We use genotype simulations (admixture, family, and tree models)  
 92 and three real datasets: the 1000 Genomes Project [50, 51], the Human Genome Diversity Panel  
 93 (HGDP) [52–54], and Human Origins [55–58]. We simulate quantitative traits from two models:

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

Publication	Sim. Genotypes			Real <sup>d</sup>	Trait <sup>e</sup>	Power	PCs ( $r$ )	Best
	Type <sup>a</sup>	$K$ <sup>b</sup>	$F_{ST}$ <sup>c</sup>					
Zhao et al. [16]				✓	Q	✓	8	LMM
Astle and Balding [1]	I	3	0.10		CC	✓	10	Tie
Kang et al. [30]				✓	Both		2-100	LMM
Price et al. [20]	I, F	2	0.01		CC		1	Mixed
Wu et al. [31]	I, A	2-4	0.01		CC	✓	10	Mixed
Liu et al. [44]	S, A	2-3	R		Q	✓	10	Tie
Sul and Eskin [32]	I	2	0.01		CC		1	Tie
Tucker, Price, and Berger [43]	I	2	0.05	✓	Both	✓	5	Tie
Yang et al. [29]				✓	CC	✓	5	Tie
Song, Hao, and Storey [46]	S, A	2-3	R		Q		3	LMM
Loh et al. [41]				✓	Q	✓	10	LMM
Liu et al. [47]				✓	Q	✓	3-6	LMM
Sul, Martin, and Eskin [48]				✓	Q		100	LMM
This work	A, T, F	10-243	$\leq 0.25$	✓	Q	✓	0-90	LMM

<sup>a</sup>Genotype simulation types. I: Independent subpopulations; S: subpopulations (with parameters drawn from real data); A: Admixture; T: Tree; F: Family.

<sup>b</sup>Model dimensionality (number of subpopulations or ancestries)

<sup>c</sup>R: simulated parameters based on real data,  $F_{ST}$  not reported.

<sup>d</sup>Evaluations using unmodified real genotypes.

<sup>e</sup>Q: quantitative; CC: case-control.

94 fixed effect sizes (FES; coefficients inverse to allele frequency) that matches real data [45, 59, 60]  
 95 and corresponds to high pleiotropy and strong balancing selection [61] and strong negative selection  
 96 [45, 60], which are appropriate assumptions for diseases; and random coefficients (RC; independent  
 97 of allele frequency) that corresponds to neutral traits [45, 61]. LMM without PCs consistently  
 98 performs best, and greatly outperforms PCA in the family simulation and in all real datasets. The  
 99 tree simulations do not recapitulate the real data results, suggesting that family relatedness in real  
 100 data is the reason for poor PCA performance. Lastly, removing up to 4th degree relatives in the real  
 101 datasets recapitulates poor PCA performance, showing that the more numerous distant relatives  
 102 explain the result, and suggesting that PCA is generally not an appropriate model for real data. All  
 103 together, we find that LMMs without PCs are generally a preferable association model, and present  
 104 novel simulation and evaluation approaches to measure the performance of these and other genetic  
 105 association approaches.

## 106 2 Materials and Methods

### 107 2.1 The complex trait model and PCA and LMM approximations

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

$$112 \quad \mathbf{y} = \mathbf{1}\alpha + \mathbf{X}^\top \boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

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

118 The full model of Eq. (1), which has a coefficient for each of the  $m$  loci, is underdetermined

119 in current datasets where  $m \gg n$ . The PCA and LMM models, respectively, approximate the full  
 120 model fit at a single locus  $i$ :

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

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

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

$$128 \quad \mathbb{E}[\mathbf{x}_i|T] = 2p_i^T \mathbf{1}, \quad \text{Cov}(\mathbf{x}_i|T) = 4p_i^T(1 - p_i^T)\boldsymbol{\Phi}^T, \quad (4)$$

129 where  $p_i^T$  is the ancestral allele frequency of locus  $i$  [1, 62–64]. Assuming independent loci, the  
 130 covariance of the polygenic effect is

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

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

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

135 which follows from the affine transformation property of multivariate normal distributions. There-  
 136 fore, the PCA term  $\mathbf{U}_r\gamma_r$  can be derived from the above equation under the additional assumption

137 that the kinship matrix has dimensionality  $r$  and the coefficients  $\gamma_r$  are fit without constraints.  
 138 In contrast, the LMM uses all eigenvectors, while effectively shrinking their coefficients  $\gamma_{\text{LMM}}$  as  
 139 all random effects models do, although these parameters are marginalized [1, 42]. PCA has more  
 140 parameters than LMM, so it may overfit more: ignoring the shared terms in Eqs. (2) and (3), PCA  
 141 fits  $r$  parameters (length of  $\gamma$ ), whereas LMMs fit only one ( $\sigma_s^2$ ).

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

$$144 \quad \mathbf{X}_S = \left( \frac{x_{ij} - 2\hat{p}_i^T}{\sqrt{4\hat{p}_i^T(1-\hat{p}_i^T)}} \right), \quad \hat{\Phi}^T = \frac{1}{m} \mathbf{X}_S^\top \mathbf{X}_S, \quad (5)$$

145 where the unknown  $p_i^T$  is estimated by  $\hat{p}_i^T = \frac{1}{2n} \sum_{j=1}^n x_{ij}$  [5, 27, 29, 30, 33, 37, 39, 41, 48]. However,  
 146 this kinship estimator has a complex bias that differs for every individual pair, which arises due  
 147 to the use of this estimated  $\hat{p}_i^T$  [28, 65]. Nevertheless, in PCA and LMM these biased estimates  
 148 perform as well as unbiased ones, an observation that will be explored in future work (data not  
 149 shown).

150 We selected fast and robust software implementing the basic PCA and LMM models. PCA  
 151 association was performed with `plink2` [66]. The quantitative trait association model is a linear  
 152 regression with covariates, evaluated using the t-test. PCs were calculated with `plink2`, which equal  
 153 the top eigenvectors of Eq. (5) after removing loci with minor allele frequency MAF < 0.1.

154 LMM association was performed using GCTA [29, 37]. Its kinship estimator equals Eq. (5).  
 155 PCs were calculated using GCTA from its kinship estimate. Association significance is evaluated  
 156 with a score test. GCTA with large numbers of PCs (small simulation only) had convergence and  
 157 singularity errors in some replicates, which were treated as missing data.

## 158 2.2 Simulations

159 Every simulation was replicated 50 times, drawing anew all genotypes (except for real datasets)  
 160 and traits. Below we use the notation  $f_A^B$  for the inbreeding coefficient of a subpopulation  $A$  from  
 161 another subpopulation  $B$  ancestral to  $A$ . In the special case of the *total* inbreeding of  $A$ ,  $f_A^T$ ,  $T$  is

162 an overall ancestral population (ancestral to every individual under consideration, such as the most  
163 recent common ancestor (MRCA) population).

164 **2.2.1 Genotype simulation from the admixture model**

165 The basic admixture model is as described previously [28] and is implemented in the R package  
166 **bnpstd**. Large and Family have  $n = 1,000$  individuals, while Small has  $n = 100$ . The number of loci  
167 is  $m = 100,000$ . Individuals are admixed from  $K = 10$  intermediate subpopulations, or ancestries.  
168 Each subpopulation  $S_u$  ( $u \in \{1, \dots, K\}$ ) is at coordinate  $u$  and has an inbreeding coefficient  $f_{S_u}^T = u\tau$   
169 for some  $\tau$ . Ancestry proportions  $q_{ju}$  for individual  $j$  and  $S_u$  arise from a random walk with spread  
170  $\sigma$  on the 1D geography, and  $\tau$  and  $\sigma$  are fit to give  $F_{ST} = 0.1$  and mean kinship  $\bar{\theta}^T = 0.5F_{ST}$  for the  
171 admixed individuals [28]. Random ancestral allele frequencies  $p_i^T$ , subpopulation allele frequencies  
172  $p_i^{S_u}$ , individual-specific allele frequencies  $\pi_{ij}$ , and genotypes  $x_{ij}$  are drawn from this hierarchical  
173 model:

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

174 where this Beta is the Balding-Nichols distribution [67] with mean  $p_i^T$  and variance  $p_i^T (1 - p_i^T) f_{S_u}^T$ .  
175 Fixed loci ( $i$  where  $x_{ij} = 0$  for all  $j$ , or  $x_{ij} = 2$  for all  $j$ ) are drawn again from the model, starting  
176 from  $p_i^T$ , iterating until no loci are fixed. Each replicate draws a genotypes starting from  $p_i^T$ .

177 As a brief aside, we prove that global ancestry proportions as covariates is equivalent in expecta-  
178 tion to using PCs under the admixture model. Note that the latent space of  $\mathbf{X}$ , given by  $(\pi_{ij})$ ,  
179 has  $K$  dimensions (number of columns of  $\mathbf{Q} = (q_{ju})$ ), so the top  $K$  PCs span this space. Since  
180 associations include an intercept term ( $\mathbf{1}\alpha$  in Eq. (2)), estimated PCs are orthogonal to  $\mathbf{1}$  (note  
181  $\hat{\Phi}^T \mathbf{1} = \mathbf{0}$  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

182 intercept) are needed to span the latent space of this admixture model.

183 **2.2.2 Genotype simulation from random admixed families**

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

202 **2.2.3 Genotype simulation from a tree model**

203 This model draws subpopulations allele frequencies from a hierarchical model parametrized by a  
204 tree, which is also implemented in `bnpst` and relies on `ape` for general tree data structures and  
205 methods [68]. The ancestral population  $T$  is the root, and each node is a subpopulation  $S_w$  indexed  
206 arbitrarily. Each edge between  $S_w$  and its parent population  $P_w$  has an inbreeding coefficient  $f_{S_w}^{P_w}$ .  
207  $p_i^T$  are drawn from a given distribution (constructed to mimic each real dataset in Appendix A).

208 Given the allele frequencies  $p_i^{P_w}$  of the parent population,  $S_w$ 's allele frequencies are drawn from:

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

209 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  
210 with MAF < 0.01 are drawn again starting from the  $p_i^T$  distribution, iterating until no such loci  
211 remain.

#### 212 2.2.4 Fitting tree to real data

213 We developed new methods to fit trees to real data based on unbiased kinship estimates from  
214 `popkin`, implemented in `bnpsd`. A tree with given inbreeding edges  $f_{S_w}^{P_w}$  gives rise to a coancestry  
215 matrix  $\vartheta_{uv}^T$  for a subpopulation pair  $(S_u, S_v)$ , and the goal is to recover the inbreeding edges from  
216 coancestry estimates. Coancestry values are total inbreeding coefficients of the MRCA population  
217 of each subpopulation pair. Therefore, we calculate  $f_{S_w}^T$  for every  $S_w$  recursively from the root as  
218 follows. Nodes with parent  $P_w = T$  are already as desired. Given  $f_{P_w}^T$ , the desired  $f_{S_w}^T$  is calculated  
219 via the additive edge  $\delta_w$  [28]:

$$\small{220} \quad f_{S_w}^T = f_{P_w}^T + \delta_w, \quad \delta_w = f_{S_w}^{P_w} (1 - f_{P_w}^T). \quad (6)$$

221 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  
222 edges:  $f_{S_w}^{P_w} = \delta_w / (1 - f_{P_w}^T)$ . Overall, coancestry values are sums of  $\delta_w$  over common ancestor nodes,

$$\small{223} \quad \vartheta_{uv}^T = \sum_w \delta_w I_w(u, v), \quad (7)$$

224 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.

225 Note that  $I_w(u, v)$  reflects tree topology and  $\delta_w$  edge values.

226 To estimate population-level coancestry, first kinship ( $\hat{\varphi}_{jk}^T$ ) is estimated using `popkin` [28]. In-

227 individual coancestry ( $\hat{\theta}_{jk}^T$ ) is estimated from kinship using

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

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

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

230 Topology is estimated with hierarchical clustering using the weighted pair group method with  
231 arithmetic mean [69], with distance function  $d(S_u, S_v) = \max \left\{ \hat{\vartheta}_{uv}^T \right\} - \hat{\vartheta}_{uv}^T$ , which succeeds due to  
232 the monotonic relationship between node depth and coancestry (Eq. (7)). This algorithm recovers  
233 the true topology from the true coancestry values, and performs well for estimates from genotypes.

234 To estimate tree edge lengths, first  $\delta_w$  are estimated from  $\hat{\vartheta}_{uv}^T$  and the topology using Eq. (7) and  
235 non-negative least squares linear regression [70] (implemented in `nnls` [71]) to yield non-negative  
236  $\delta_w$ , and  $f_{S_w}^{P_w}$  are calculated from  $\delta_w$  by reversing Eq. (6). To account for small biases in coancestry  
237 estimation, an intercept term  $\delta_0$  is included ( $I_0(u, v) = 1$  for all  $u, v$ ), and when converting  $\delta_w$  to  
238  $f_{S_w}^{P_w}$ ,  $\delta_0$  is treated as an additional edge to the root, but is ignored when drawing allele frequencies  
239 from the tree.

## 240 2.2.5 Trait Simulation

241 Traits are simulated from the quantitative trait model of Eq. (1), with novel bias corrections for  
242 simulating the desired heritability from real data relying on the unbiased kinship estimator `popkin`  
243 [28]. This simulation is implemented in the R package `simtrait`. All simulations have a narrow-  
244 sense heritability of  $h^2 = 0.8$  and  $\epsilon_j \sim \text{Normal}(0, 1 - h^2)$ . To balance power while varying  $n$ , the  
245 number of causal loci is  $m_1 = n/10$ . The set of causal loci  $C$  is drawn anew for each replicate, from  
246 loci with MAF  $\geq 0.01$  to avoid rare causal variants (inappropriate for PCA and LMM). Letting  
247  $v_i^T = p_i^T (1 - p_i^T)$ , the effect size of locus  $i$  equals  $2v_i^T \beta_i^2$ , its contribution of the trait variance [72].

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

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

249 for known  $p_i^T$ ; otherwise  $v_i^T$  is replaced by the unbiased estimator [28]  $\hat{v}_i^T = \hat{p}_i^T (1 - \hat{p}_i^T) / (1 - \bar{\varphi}^T)$ ,  
250 where  $\bar{\varphi}^T$  is the mean kinship estimated with `popkin`. Each causal locus is multiplied by -1 with  
251 probability 0.5. Alternatively, under the *random coefficients* (RC) model, initial causal coefficients  
252 are drawn independently from  $\beta_i \sim \text{Normal}(0, 1)$ . For both models, the initial genetic variance is  
253  $\sigma_0^2 = \sum_{i \in C} 2v_i^T \beta_i^2$ , replacing  $v_i^T$  with  $\hat{v}_i^T$  for unknown  $p_i^T$  (so  $\sigma_0^2$  is an unbiased estimate), so we  
254 multiply every initial  $\beta_i$  by  $\frac{h}{\sigma_0}$  to have the desired heritability. Lastly, for known  $p_i^T$ , the intercept  
255 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  
256 the trait covariance (for the same reason the standard kinship estimator in Eq. (5) is biased), which  
257 is avoided with

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

### 258 2.3 Real human genotype datasets

259 The three datasets were processed as before [65] (summarized below), except with an additional filter  
260 so loci are in approximate linkage equilibrium and rare variants are removed. All processing was

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

Dataset	Type	Loci (m)	Ind. (n)	Subpops. <sup>a</sup> (K)	Causal loci <sup>b</sup> (m <sub>1</sub> )	FST <sup>c</sup>
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

<sup>a</sup>For admixed family, ignores dimensionality of 20 generation pedigree structure. For real datasets, lower range is continental subpopulations, upper range is number of fine-grained subpopulations.

<sup>b</sup> $m_1 = n/10$  to balance power across datasets.

<sup>c</sup>Model parameter for simulations, estimated value on real datasets.

261 performed with `plink2` [66], and analysis was uniquely enabled by the R packages `BEDMatrix` [73]  
262 and `genio`. Each dataset groups individuals in a two-level hierarchy: continental and fine-grained  
263 subpopulations. Final dataset sizes are in Table 2.

264 We obtained the full (including non-public) Human Origins by contacting the authors and  
265 agreeing to their usage restrictions. The Pacific data [58] was obtained separately from the rest [56,  
266 57], and datasets were merged using the intersection of loci. We removed ancient individuals, and  
267 individuals from singleton and non-native subpopulations. Non-autosomal loci were removed. Our  
268 analysis of the whole-genome sequencing (WGS) version of HGDP [54] was restricted to autosomal  
269 biallelic SNP loci with filter “PASS”. Our analysis of the high-coverage NYGC version of 1000  
270 Genomes [74] was restricted to autosomal biallelic SNP loci with filter “PASS”.

271 Since our evaluations assume uncorrelated loci, we filtered each dataset with `plink2` using  
272 parameters “`--indep-pairwise 1000kb 0.3`”, which iteratively removes loci that have a greater  
273 than 0.3 squared correlation coefficient with another locus that is within 1000kb, stopping until no  
274 such loci remain. Since all real datasets have numerous rare variants, while PCA and LMM are not  
275 able to detect associations involving rare variants, we removed all loci with  $\text{MAF} < 0.01$ . Lastly,  
276 only HGDP had loci with over 10% missingness removed, as they were otherwise 17% of remaining  
277 loci (for Human Origins and 1000 Genomes they were under 1% of loci so they were not removed).  
278 Kinship dimensionality and eigenvalues were calculated from `popkin` kinship estimates. Eigenvalues  
279 were assigned p-values with `twstats` of the Eigensoft package [7], and dimensionality was estimated  
280 as the largest number of consecutive eigenvalue from the start that all satisfy  $p < 0.01$  (p-values  
281 did not increase monotonically). For the evaluation with close relatives removed, each dataset was  
282 filtered with `plink2` with option “`--king-cutoff`” with cutoff  $0.02209709 (= 2^{-11/2})$  for removing  
283 up to 4th degree relatives using KING-robust [75], and  $\text{MAF} < 0.01$  filter is reapplied (Table S1).

284 **2.4 Evaluation of performance**

285 All approaches are evaluated in two orthogonal dimensions:  $\text{SRMSD}_p$  quantifies p-value uniformity,  
286 and  $\text{AUC}_{\text{PR}}$  measures causal locus classification performance and reflects power while ranking mis-  
287 calibrated models fairly. These measures are more robust alternatives to previous measures from

288 the literature (see Appendix B), and are implemented in `simtrait`.

289 P-values for continuous test statistics have a uniform distribution when the null hypothesis  
290 holds, a crucial assumption for type I error and FDR control [76, 77]. We use the Signed Root  
291 Mean Square Deviation ( $\text{SRMSD}_p$ ) to measure the difference between the observed null p-value  
292 quantiles and the expected uniform quantiles:

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

293 where  $m_0 = m - m_1$  is the number of null (non-causal) loci, here  $i$  indexes null loci only,  $p_{(i)}$  is  
294 the  $i$ th ordered null p-value,  $u_i = (i - 0.5)/m_0$  is its expectation,  $p_{\text{median}}$  is the median observed  
295 null p-value,  $u_{\text{median}} = \frac{1}{2}$  is its expectation, and  $\text{sgn}$  is the sign function (1 if  $u_{\text{median}} \geq p_{\text{median}}$ ,  
296 -1 otherwise). Thus,  $\text{SRMSD}_p = 0$  corresponds to calibrated p-values,  $\text{SRMSD}_p > 0$  indicate anti-  
297 conservative p-values, and  $\text{SRMSD}_p < 0$  are conservative p-values. The maximum  $\text{SRMSD}_p$  is  
298 achieved when all p-values are zero (the limit of anti-conservative p-values), which for infinite loci  
299 approaches

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

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

301 Precision and recall are standard performance measures for binary classifiers that do not require  
302 calibrated p-values [78]. Given the total numbers of true positives (TP), false positives (FP) and  
303 false negatives (FN) at some threshold or parameter  $t$ , precision and recall are

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

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

308 to the overall proportion of causal loci:  $\frac{m_1}{m}$ .

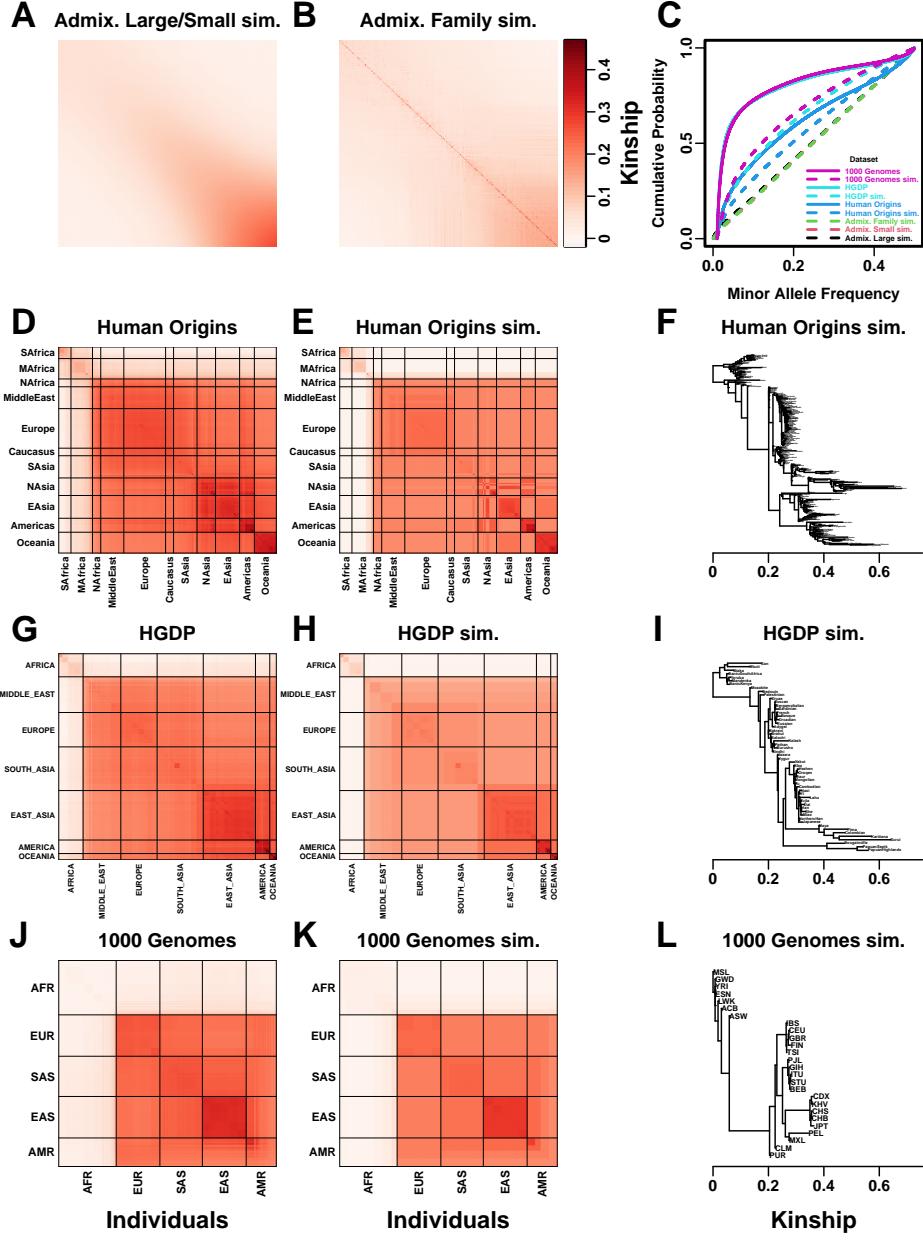
## 309 3 Results

### 310 3.1 Overview of evaluations

311 We use three real genotype datasets and simulated genotypes from six population structure scenarios  
312 to cover various features of interest (Table 2). We introduce them in sets of three, as they appear  
313 in the rest of our results. Population kinship matrices, which combine population and family  
314 relatedness, are estimated without bias using `popkin` [28] (Fig. 1). The first set of three simulated  
315 genotypes are based on an admixture model with 10 ancestries (Fig. 1A) [14, 28, 79]. The “large”  
316 version (1000 individuals) illustrates asymptotic performance, while the “small” simulation (100  
317 individuals) illustrates model overfitting. The “family” simulation has admixed founders and draws  
318 a 20-generation random pedigree with assortative mating, resulting in a complex joint family and  
319 ancestry structure in the last generation (Fig. 1B). The second set of three are the real human  
320 datasets representing global human diversity: Human Origins (Fig. 1D), HGDP (Fig. 1G), and  
321 1000 Genomes (Fig. 1J), which are enriched for small minor allele frequencies even after  $MAF < 1\%$   
322 filter (Fig. 1C). Last are tree simulations (Fig. 1F,I,L) fit to the kinship (Fig. 1E,H,K) and MAF  
323 (Fig. 1C) of each real human dataset, which by design do not have family structure.

324 All traits in this work are simulated. We repeated all evaluations on two additive quantitative  
325 trait models, *fixed effect sizes* (FES) and *random coefficients* (RC), which differ in how causal coef-  
326 ficients are constructed. The FES model captures the rough inverse relationship between coefficient  
327 and minor allele frequency that arises under strong negative and balancing selection and has been  
328 observed in numerous diseases and other traits [45, 59–61], so it is the focus of our results. The  
329 RC model draws coefficients independent of allele frequency, corresponding to neutral traits [45,  
330 61], which results in a wider effect size distribution that reduces association power and effective  
331 polygenicity compared to FES.

332 We evaluate using two complementary measures: (1)  $SRMSD_p$  (p-value signed root mean square  
333 deviation) measures p-value calibration (closer to zero is better), and (2)  $AUC_{PR}$  (precision-recall



**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.

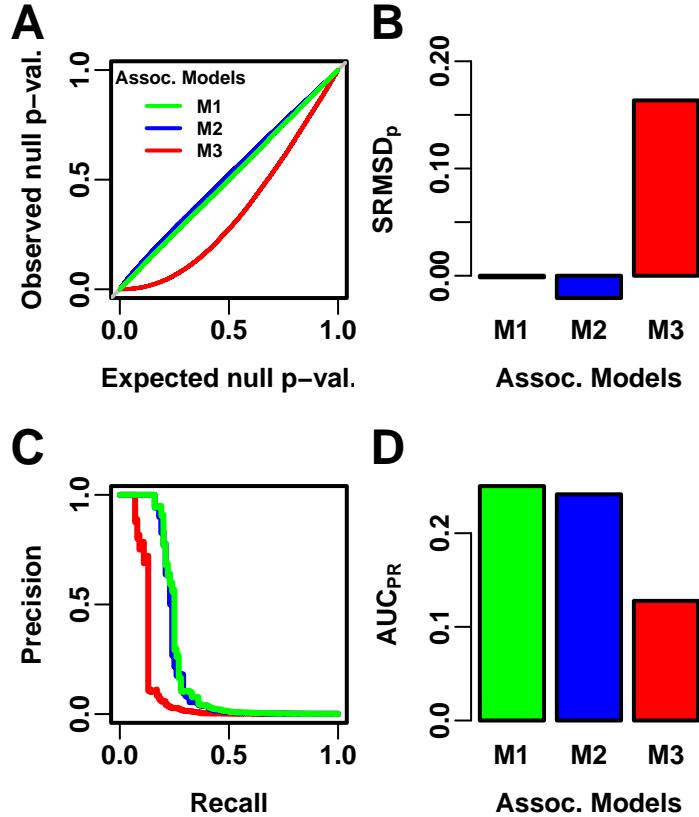


Figure 2: **Illustration of evaluation measures.** Three archetypal models illustrate our complementary measures: M1 is ideal, M2 overfits slightly, M3 is naive. **A.** QQ plot of p-values of “null” (non-causal) loci. M1 has desired uniform p-values, M2/M3 are miscalibrated. **B.** SRMSD<sub>p</sub>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

334 area under the curve) measures causal locus classification performance (higher is better; Fig. 2).  
 335 SRMSD<sub>p</sub> is a more robust alternative to the common inflation factor  $\lambda$  and type I error control  
 336 measures; there is a correspondence between  $\lambda$  and SRMSD<sub>p</sub>, with SRMSD<sub>p</sub> > 0.01 giving  $\lambda > 1.06$   
 337 (Fig. S1) and thus evidence of miscalibration close to the rule of thumb of  $\lambda > 1.05$  [20]. AUC<sub>PR</sub>  
 338 has been used to evaluate association models [80], and reflects statistical power while being robust  
 339 to miscalibrated models (Appendix B).

340 Both PCA and LMM were evaluated in each replicate dataset including a number of PCs  $r$   
 341 between 0 and 90 as fixed covariates. In terms of p-value calibration, for PCA the best number of

Table 3: Overview of PCA and LMM evaluation results

Dataset	Trait model <sup>b</sup>	SRMSD <sub>p</sub>   <sup>a</sup>			AUC <sub>PR</sub>		
		Best (min <sup>c</sup> ) PCs PCA	Best (min <sup>c</sup> ) PCs LMM	Best <sup>d</sup>	Best (min <sup>c</sup> ) PCs PCA	Best (min <sup>c</sup> ) PCs LMM	Best <sup>d</sup>
Admix. Large sim.	FES	12* (4*)	0*	tie	3	0	LMM
Admix. Small sim.	FES	4* (3*)	0*	tie	2 (0)	0	tie (LMM)
Admix. Family sim.	FES	90 (88)	0*	LMM	22 (14)	1 (0)	LMM
Human Origins	FES	89 (73)	0*	LMM	34 (16)	0	LMM
HGDP	FES	87* (30*)	0*	LMM	16 (13)	1 (0)	LMM
1000 Genomes	FES	90 (54)	0*	LMM	8 (7)	1 (0)	LMM
Human Origins sim.	FES	88* (80*)	0*	tie	36 (21)	0	LMM
HGDP sim.	FES	47* (21*)	0*	tie (LMM)	17 (8)	0	LMM
1000 Genomes sim.	FES	78* (12*)	0*	LMM	10 (6)	0	LMM
Admix. Large sim.	RC	26* (3*)	0*	tie	3	0	LMM
Admix. Small sim.	RC	4* (2*)	0*	LMM	1 (0)	0	tie (LMM)
Admix. Family sim.	RC	90 (89)	0*	LMM	34 (18)	0	LMM
Human Origins	RC	90* (71*)	0*	LMM	36 (14)	3 (0)	LMM
HGDP	RC	37* (26*)	0*	LMM	16 (5)	4 (0)	LMM
1000 Genomes	RC	76* (23*)	0*	LMM	9 (4)	5 (2)	LMM
Human Origins sim.	RC	85* (75*)	0*	tie	37 (29)	0	LMM
HGDP sim.	RC	44* (17*)	0*	LMM	17 (15)	3 (0)	LMM
1000 Genomes sim.	RC	90* (44*)	0*	LMM	10 (6)	3 (0)	LMM

<sup>a</sup> $r$  for which mean  $|\text{SRMSD}_p| < 0.01$  marked with asterisk.

<sup>b</sup>FES: Fixed Effect Sizes, RC: Random Coefficients.

<sup>c</sup>Smallest  $r$  (number of PCs) whose distribution ( $|\text{SRMSD}_p|$  or AUC<sub>PR</sub>) was not significantly different (Wilcoxon paired 1-tailed  $p > 0.01$ ) from best  $r$  (if any).

<sup>d</sup>Tie if distributions of best PCA and LMM version did not differ significantly (Wilcoxon paired 1-tailed  $p > 0.01$ ). Same result for “min” except cases in parentheses.

342 PCs  $r$  (minimizing mean  $|\text{SRMSD}_p|$  over replicates) is typically large across all datasets, but much  
343 smaller “min”  $r$  values often performed as well (numbers in parentheses in Table 3). Most cases  
344 had a mean  $|\text{SRMSD}_p| < 0.01$  (marked with asterisks in Table 3), whose p-values are effectively  
345 calibrated. However, PCA best and min  $r$  values tended to be large on the family simulation and  
346 real datasets, and those cases were often miscalibrated. In contrast, for LMM,  $r = 0$  (no PCs)  
347 was always best, and was always calibrated. Comparing LMM with  $r = 0$  to PCA with its best  $r$ ,  
348 LMM always had significantly smaller  $|\text{SRMSD}_p|$  than PCA or was statistically tied. For  $\text{AUC}_{\text{PR}}$   
349 and PCA, the best  $r$  was always smaller than the best  $r$  for  $|\text{SRMSD}_p|$ , so there is often a tradeoff  
350 between calibrated p-values versus classification performance. For LMM there is no tradeoff, as  
351  $r = 0$  had  $\text{AUC}_{\text{PR}}$  not significantly different from the best  $r$  in all cases except two (the min  $r$  was  
352 2 for both 1000 Genomes simulation with FES trait and 1000 Genomes real dataset with RC trait).  
353 Lastly, LMM with its best  $r$  always had significantly greater  $\text{AUC}_{\text{PR}}$  than PCA with its best  $r$ .

### 354 3.2 Evaluations in admixture simulations

355 Now we look more closely at results per dataset. The complete  $\text{SRMSD}_p$  and  $\text{AUC}_{\text{PR}}$  distributions  
356 for the admixture simulations and FES traits are in Fig. 3. RC traits gave qualitatively similar  
357 results (Fig. S2).

358 In the large admixture simulation, the  $\text{SRMSD}_p$  of PCA is largest when  $r = 0$  (no PCs) and  
359 decreases rapidly to near zero at  $r = 3$ , where it stays for up to  $r = 90$  (Fig. 3A). Thus, PCA  
360 has calibrated p-values for  $r \geq 3$ , smaller than the theoretical optimum for this simulation of  
361  $r = K - 1 = 9$ . In contrast, the  $\text{SRMSD}_p$  for LMM starts near zero for  $r = 0$ , but becomes negative  
362 as  $r$  increases (p-values are conservative). The  $\text{AUC}_{\text{PR}}$  distribution of PCA is similarly worst at  
363  $r = 0$ , increases rapidly and peaks at  $r = 3$ , then decreases slowly for  $r > 3$ , while the  $\text{AUC}_{\text{PR}}$   
364 distribution for LMM starts near its maximum at  $r = 0$  and decreases with  $r$ . Although the  $\text{AUC}_{\text{PR}}$   
365 distributions for LMM and PCA overlap considerably at each  $r$ , LMM with  $r = 0$  has significantly  
366 greater  $\text{AUC}_{\text{PR}}$  values than PCA with  $r = 3$  (Table 3). However, qualitatively PCA performs nearly  
367 as well as LMM in this simulation.

368 The observed robustness to large  $r$  led us to consider smaller sample sizes. A model with large

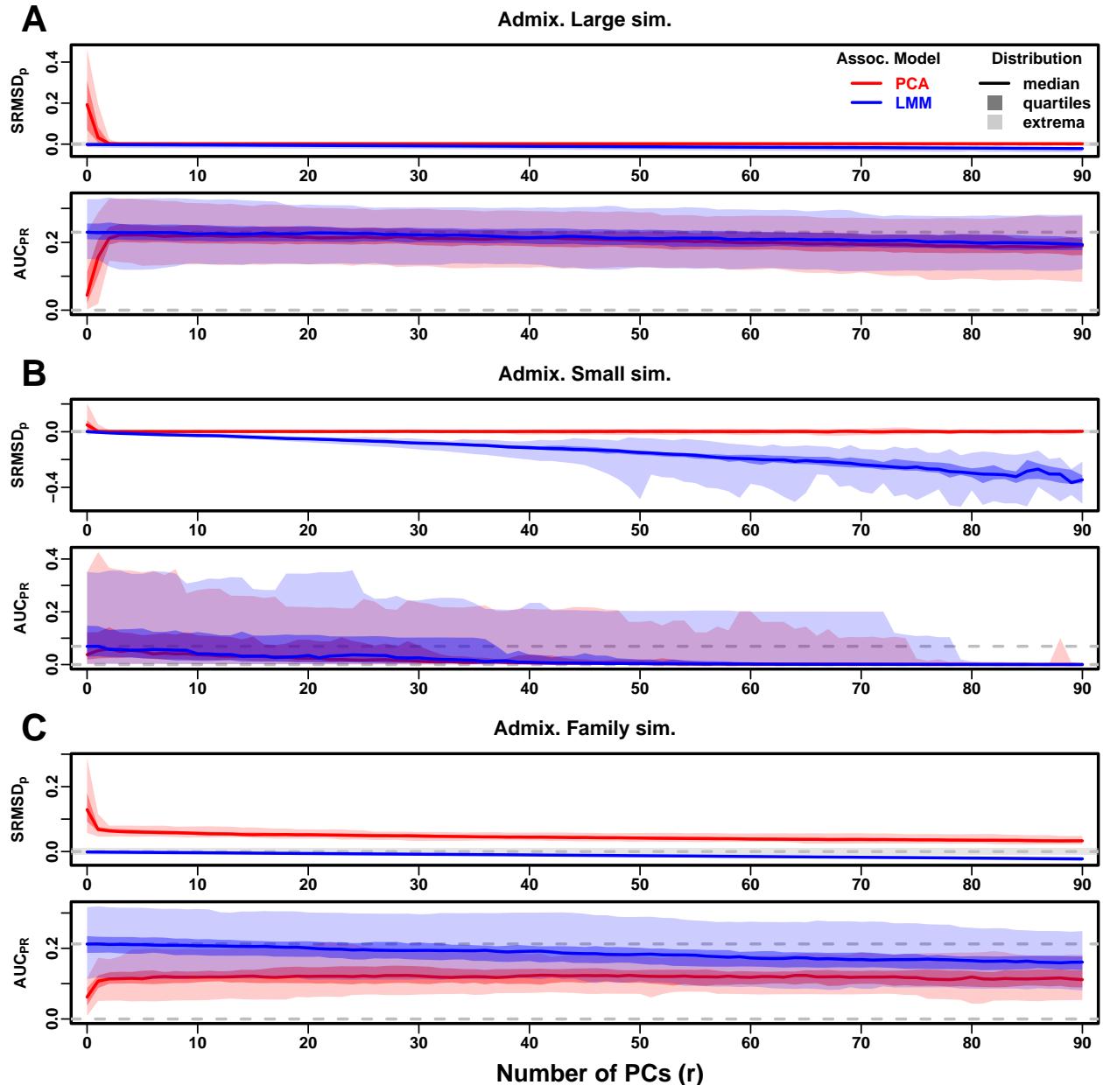


Figure 3: **Evaluations in admixture simulations.** Traits simulated from FES model. PCA and LMM models have varying number of PCs ( $r \in \{0, \dots, 90\}$  on x-axis), with the distributions (y-axis) of SRMSD<sub>p</sub> (top subpanel) and AUC<sub>PR</sub> (bottom subpanel) for 50 replicates. Best performance is zero SRMSD<sub>p</sub> and large AUC<sub>PR</sub>. Zero and maximum median AUC<sub>PR</sub> 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: SRMSD<sub>p</sub> > 0 for all  $r$  and large AUC<sub>PR</sub> gap.

369 numbers of parameters  $r$  should overfit more as  $r$  approaches the sample size  $n$ . Rather than increase  
370  $r$  beyond 90, we reduce individuals to  $n = 100$ , which is small for typical association studies but  
371 may occur in studies of rare diseases, pilot studies, or other constraints. To compensate for the  
372 loss of power due to reducing  $n$ , we also reduce the number of causal loci (fixed ratio  $n/m_1 = 10$ ),  
373 which increases per-locus effect sizes. We found a large decrease in performance for both models as  
374  $r$  increases, and best performance for  $r = 1$  for PCA and  $r = 0$  for LMM (Fig. 3B). Remarkably,  
375 LMM attains much larger negative SRMSD <sub>$p$</sub>  values than in our other evaluations. LMM with  $r = 0$   
376 is significantly better than PCA ( $r = 1$  to 4) in both measures (Table 3), but qualitatively the  
377 difference is negligible.

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

### 386 3.3 Evaluations in real human genotype datasets

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

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

395 HGDP has the fewest individuals among real datasets, but compared to Human Origins contains

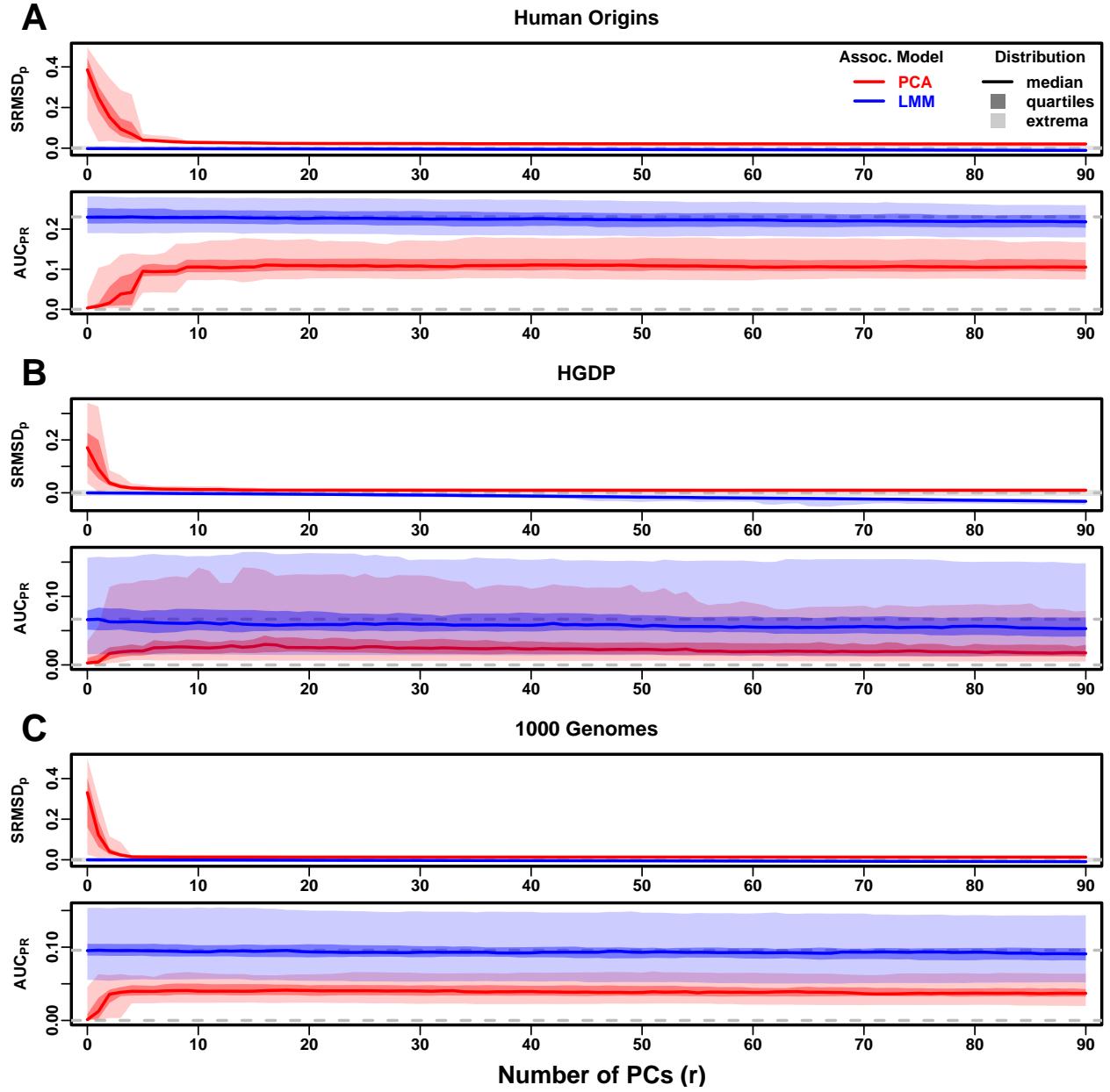


Figure 4: **Evaluations in real human genotype datasets.** Traits simulated from FES model. Same setup as Fig. 3, see that for details. These datasets strongly favor LMM with no PCs over PCA, with distributions that most resemble the family simulation. **A.** Human Origins. **B.** Human Genome Diversity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

396 more loci and low-frequency variants. Performance (Fig. 4B) again most resembled the family sim-  
397 ulations. In particular, LMM with  $r = 0$  achieves mean  $|\text{SRMSD}_p| < 0.01$  (p-values are calibrated),  
398 while PCA does not, and there is a sizable  $\text{AUC}_{\text{PR}}$  gap between LMM and PCA. Maximum  $\text{AUC}_{\text{PR}}$   
399 values were lowest in HGDP compared to the two other real datasets.

400 1000 Genomes has the fewest subpopulations but largest number of individuals per subpopula-  
401 tion. Thus, although this dataset has the simplest subpopulation structure among the real datasets,  
402 we find  $\text{SRMSD}_p$  and  $\text{AUC}_{\text{PR}}$  distributions (Fig. 4C) that again most resemble our earlier family  
403 simulation, with mean  $|\text{SRMSD}_p| < 0.01$  for LMM only and large  $\text{AUC}_{\text{PR}}$  gaps between LMM and  
404 PCA.

405 Our results are qualitatively different for RC traits, which had smaller  $\text{AUC}_{\text{PR}}$  gaps between  
406 LMM and PCA (Fig. S3). Maximum  $\text{AUC}_{\text{PR}}$  were smaller in RC compared to FES in Human Origins  
407 and 1000 Genomes, suggesting lower power for RC traits across association models. Nevertheless,  
408 LMM with  $r = 0$  was significantly better than PCA for all measures in the real datasets and RC  
409 traits (Table 3).

#### 410 3.4 Evaluations in tree simulations fit to human data

411 To better understand which features of the real datasets lead to the large differences in performance  
412 between LMM and PCA, we carried out tree simulations. Human subpopulations are related roughly  
413 by trees, which induce the strongest correlations and have numerous tips, so we fit trees to each  
414 real dataset and tested if data simulated from these complex tree structures could recapitulate our  
415 previous results (Fig. 1). These tree simulations also feature non-uniform ancestral allele frequency  
416 distributions, which recapitulated some of the skew for smaller minor allele frequencies of the real  
417 datasets (Fig. 1C). The  $\text{SRMSD}_p$  and  $\text{AUC}_{\text{PR}}$  distributions for these tree simulations (Fig. 5)  
418 resembled our admixture simulation more than either the family simulation (Fig. 3) or real data  
419 results (Fig. 4). Both LMM with  $r = 0$  and PCA (various  $r$ ) achieve mean  $|\text{SRMSD}_p| < 0.01$   
420 (Table 3). The  $\text{AUC}_{\text{PR}}$  distributions of both LMM and PCA track closely as  $r$  is varied, although  
421 there is a small gap resulting in LMM ( $r = 0$ ) besting PCA in all three simulations. The results  
422 are qualitatively similar for RC traits (Fig. S4 and Table 3). Overall, these tree simulations do not

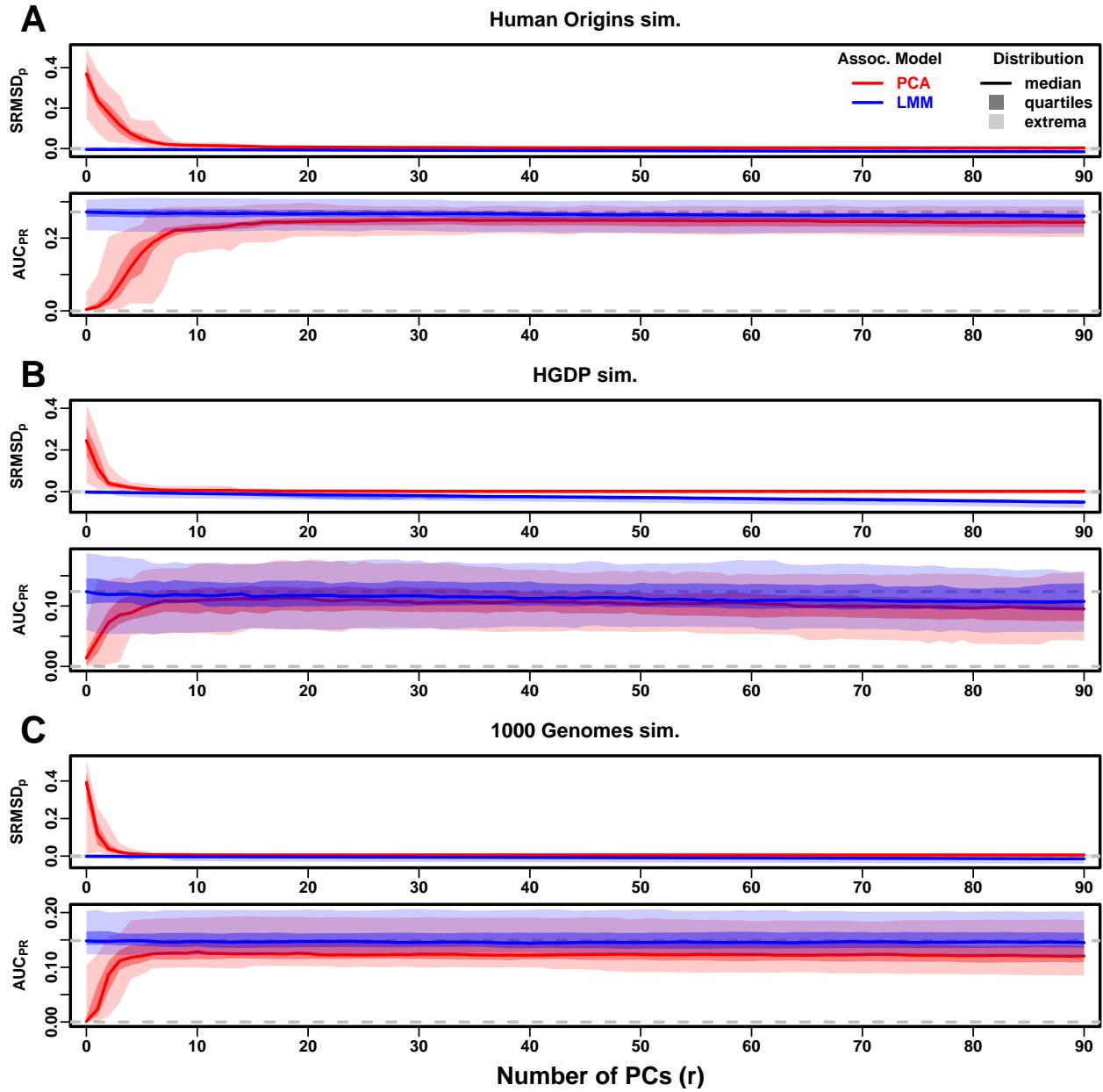


Figure 5: **Evaluations in tree simulations fit to human data.** Traits simulated from FES model. Same setup as Fig. 3, see that for details. These tree simulations, which exclude family structure by design, do not explain the large gaps in LMM-PCA performance observed in the real data. **A.** Human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

423 recapitulate the large LMM advantage over PCA observed on the real data.

### 424 3.5 Numerous distant relatives explain poor PCA performance in real data

425 In principle, PCA performance should be determined by the dimensionality of relatedness, since  
426 PCA is a low-dimensional model whereas LMM can model high-dimensional relatedness without  
427 overfitting. We used the Tracy-Widom test [7] with  $p < 0.01$  to estimate dimensionality as the  
428 number of significant PCs (Fig. S5A). The true dimensionality of our simulations is slightly un-  
429 derestimated (Table 2), but we confirm that the family simulation has the greatest dimensionality,  
430 and real datasets have greater estimates than their respective tree simulations, which confirms our  
431 hypothesis to some extent. However, estimated dimensionalities do not separate real datasets from  
432 tree simulations, as required to predict the observed PCA performance. Moreover, the HGDP and  
433 1000 Genomes dimensionality estimates are 45 and 61, respectively, yet PCA performed poorly  
434 for all  $r \leq 90$  numbers of PCs (Fig. 4). The top eigenvalue explained a proportion of variance  
435 proportional to  $F_{ST}$  (Table 2), but the rest of the top 10 eigenvalues show no clear differences  
436 between datasets, except the small simulation had larger variances explained per eigenvalue (ex-  
437 pected since it has fewer eigenvalues; Fig. S5C). Comparing cumulative variance explained versus  
438 rank fraction across all eigenvalues, all datasets increase from their starting point almost linearly  
439 until they reach 1, except the family simulation has much greater variance explained by mid-rank  
440 eigenvalues (Fig. S5B). Overall, there is no separation between real datasets (where PCA performed  
441 poorly) and tree simulations (where PCA performed relatively well) in terms of their eigenvalues or  
442 dimensionality estimates.

443 Local kinship, which is recent relatedness due to family structure excluding population structure,  
444 is the presumed cause of the LMM to PCA performance gap observed in real datasets but not their  
445 tree simulation counterparts. Instead of inferring local kinship through increased dimensionality, as  
446 attempted in the last paragraph, now we measure it directly using the KING-robust estimator [75].  
447 We observe more large local kinship in the real datasets and the family simulation compared to the  
448 other simulations (Fig. 6). However, for real data this distribution depends on the subpopulation  
449 structure, since locally related pairs are most likely in the same subpopulation. Therefore, the

450 only comparable curve to each real dataset is their corresponding tree simulation, which matches  
 451 subpopulation structure. In all real datasets we identified highly related individual pairs with  
 452 kinship above the 4th degree relative threshold of 0.022 [75, 81]. However, these highly related pairs  
 453 are vastly outnumbered by more distant pairs with evident non-zero local kinship as compared to  
 454 the extreme tree simulation values.

455 To try to improve PCA performance, we followed the standard practice of removing 4th degree  
 456 relatives, which reduced sample sizes between 5% and 10% (Table S1). Only  $r = 0$  for LMM  
 457 and  $r = 20$  for PCA were tested, as these performed well in our earlier evaluation, and only  
 458 FES traits were tested because they previously displayed the large PCA-LMM performance gap.  
 459 LMM significantly outperforms PCA in all these cases (Wilcoxon paired 1-tailed  $p < 0.01$ ; Fig. 7).  
 460 Notably, PCA still had miscalibrated p-values in all real datasets ( $|SRMSD_p| > 0.01$ ). Otherwise,  
 461 AUC<sub>PR</sub> and SRMSD<sub>p</sub> ranges were similar here as in our earlier evaluation. Therefore, the removal  
 462 of the small number of highly related individual pairs had a negligible effect in PCA performance,

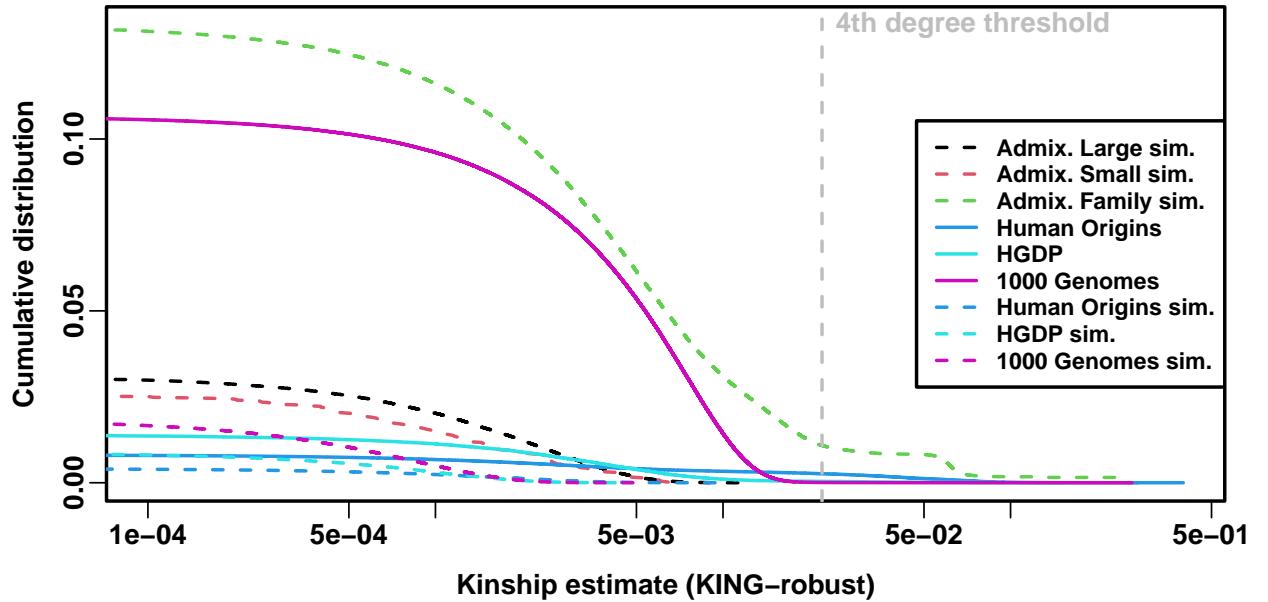


Figure 6: **Local kinship distributions.** Curves are complementary cumulative distribution of lower triangular kinship matrix (self kinship excluded) from KING-robust estimator. Note log x-axis; negative estimates are counted but not shown. Most values are below 4th degree relative threshold. Each real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

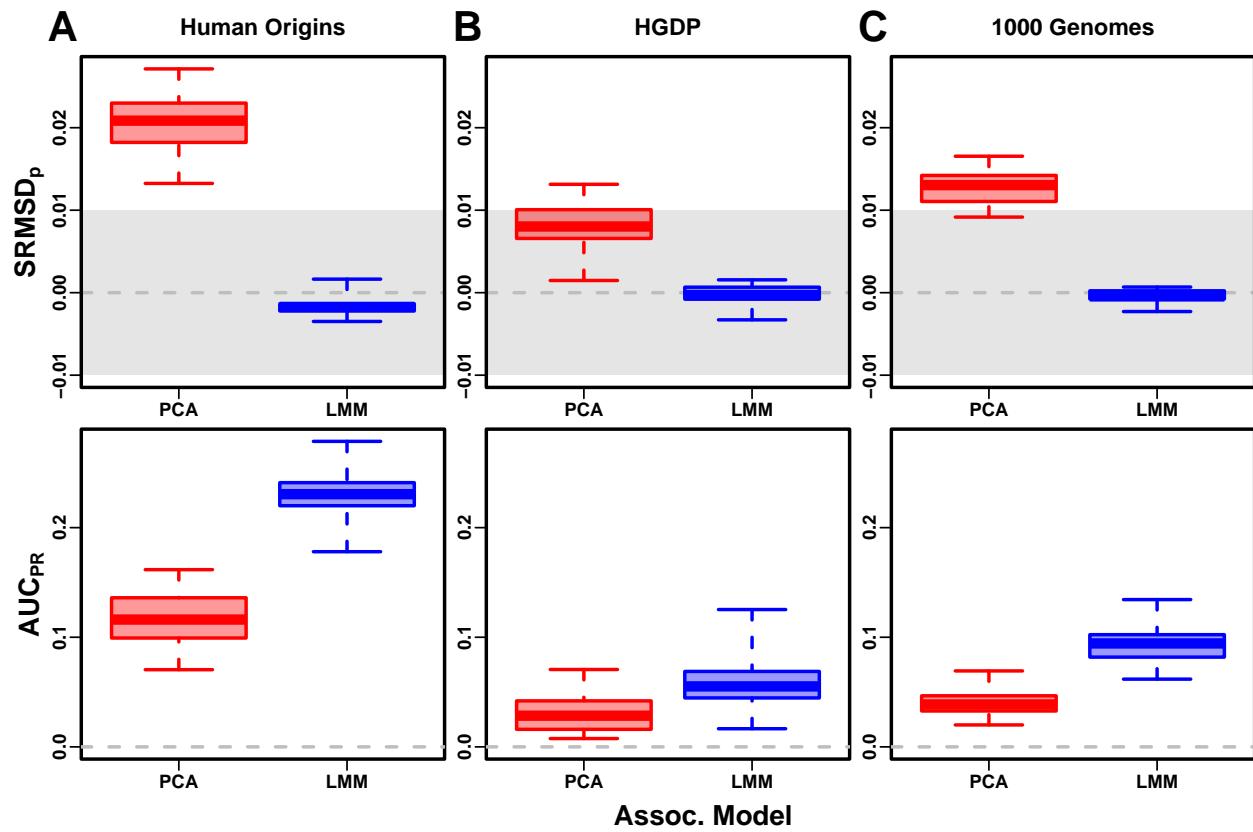


Figure 7: **Evaluation in real datasets excluding 4th degree relatives.** LMM had  $r = 0$  PCs and PCA had  $r = 20$ . FES traits only. Each dataset is a column, rows are measures. First row has  $|\text{SRMSD}_p| < 0.01$  band marked as gray area.

463 so the larger number of more distantly related pairs explain the poor PCA performance in the real  
464 datasets.

## 465 4 Discussion

466 Our evaluations conclusively determined that LMM without PCs performs better than PCA (for  
467 any number of PCs) across all scenarios, including all real and simulated genotypes and two trait  
468 simulation models. Although the addition of a few PCs to LMM does not greatly hurt its perfor-  
469 mance (except for small sample sizes), they generally did not improve it either (Table 3), which  
470 agrees with previous observations [44] but contradicts others [16, 20]. Our findings make sense since  
471 PCs are the eigenvectors of the same kinship matrix that parametrizes random effects, so including  
472 both is redundant.

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

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

487 PCA assumes that genetic relatedness is low-dimensional, whereas LMM can handle high-  
488 dimensional relatedness. Thus, PCA performs well in the admixture simulation, which is explicitly  
489 low-dimensional (see Materials and Methods), and our tree simulations, which, although complex in

principle due to the large number of nodes, had few long branches so a low-dimensional approximation suffices. Conversely, PCA performs poorly under family structure because its kinship matrix is high-dimensional (Fig. S5). However, estimating the dimensionality of real datasets is challenging because estimated eigenvalues have biased distributions. Dimensionality estimated using the Tracy-Widom test [7] did not fully predict the datasets that PCA performs well on. In contrast, estimated local kinship finds considerable cryptic relatedness in all real human datasets and better explains why PCA performs poorly there. The trait model also influences the relative performance of PCA, so genotype-only parameters (eigenvalues or local kinship) alone do not tell the full story.

PCA is at best underpowered relative to LMMs, and at worst miscalibrated regardless of the numbers of PCs included, in real human genotype tests. Among our simulations, such poor performance occurred only in the admixed family. Local kinship estimates reveal considerable family relatedness in the real datasets absent in the corresponding tree simulations. Admixture is also absent in our tree simulations, but our simulations and theory show that admixture is handled well by PCA. Hundreds of close relative pairs have been identified in 1000 Genomes [82–85], but their removal does not improve PCA performance sufficiently in our tests, so the larger number of more distantly related pairs are PCA’s most serious obstacle in practice. Distant relatives are expected to be numerous in any large human dataset [86, 87]. Our FES trait tests show that cryptic relatedness is more challenging when rarer variants have larger coefficients. Overall, the high dimensionality induced by cryptic relatedness is the key challenge for PCA association in modern datasets that is readily overcome by LMM.

Our tests also found PCA robust to large numbers of PCs, far beyond the optimal choice, agreeing with previous anecdotal observations [5, 30], in contrast to using too few PCs for which there is a large performance penalty. The exception was the small sample size simulation, where only small numbers of PCs performed well. In contrast, LMM is simpler since there is no need to choose the number of PCs. However, an LMM with a large number of covariates may have conservative p-values (as observed for LMM with large numbers of PCs), a weakness of the score test used by the LMM we evaluated that may be overcome with other statistical tests. Simulations or post hoc evaluations remain crucial for ensuring that statistics are calibrated.

518        The largest limitation of our work is that we only considered quantitative traits. We noted that  
519    previous evaluations involving case-control traits tended to report PCA-LMM ties or mixed results,  
520    an observation potentially confounded by the use of low-dimensional simulations without family  
521    relatedness (Table 1). An additional concern is case-control ascertainment bias, which appears to  
522    affect LMMs more severely, although recent work appears to solve this problem [29, 33]. Future  
523    evaluations should aim to include our simulations and real datasets, to ensure that previous results  
524    were not biased in favor of PCA by employing unrealistic low-dimensional genotype simulations,  
525    or by not simulating large coefficients for rare variants expected for diseases by various selection  
526    models.

527        Overall, our results lead us to recommend LMM over PCA for association studies in general. Al-  
528    though PCA offer flexibility and speed compared to LMM, additional work is required to ensure that  
529    PCA is adequate, including removal of close relatives (lowering sample size and wasting resources)  
530    followed by simulations or other evaluations of statistics, and even then PCA may perform poorly  
531    in terms of both type I error control and power. The large numbers of distant relatives expected of  
532    any real dataset all but ensures that PCA will perform poorly compared to LMM. Our findings also  
533    suggest that related applications such as polygenic models may enjoy gains in power and accuracy  
534    by employing an LMM instead of PCA to model relatedness [18, 80]. PCA remains indispensable  
535    across population genetics, from visualizing population structure and performing quality control  
536    to its deep connection to admixture models, but the time has come to limit its use in association  
537    testing in favor of LMM or other, richer models capable of modeling all forms of relatedness.

## 538        5 Appendices

### 539        5.1 Appendix A: Fitting ancestral allele frequency distribution to real data

540        We calculated  $\hat{p}_i^T$  distributions of each real dataset. However, differentiation increases the variance  
541    of these sample  $\hat{p}_i^T$  relative to the true  $p_i^T$  [28]. We present a new algorithm for constructing an  
542    “undifferentiated” distribution based on the input data but with the lower variance of the true  
543    ancestral distribution. Suppose the  $p_i^T$  distribution over loci  $i$  satisfies  $E[p_i^T] = \frac{1}{2}$  and  $\text{Var}(p_i^T) =$

544  $V^T$ . The sample allele frequency  $\hat{p}_i^T$ , conditioned on  $p_i^T$ , satisfies

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

545 where  $\bar{\varphi}^T = \frac{1}{n^2} \sum_{j=1}^n \sum_{k=1}^n \varphi_{jk}^T$  is the mean kinship over all individual [28]. The unconditional  
546 moments of  $\hat{p}_i^T$  follow from the laws of total expectation and variance:  $\mathrm{E} [\hat{p}_i^T] = \frac{1}{2}$  and

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

547 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  
548 the original, lower variance of

$$549 \quad V^T = \frac{W^T - \frac{1}{4} \bar{\varphi}^T}{1 - \bar{\varphi}^T}. \quad (9)$$

550 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  
551 parameters: the subpopulation coancestry matrix (Eq. (7)), expanded from subpopulations to indi-  
552 viduals, the diagonal converted to kinship (reversing Eq. (8)), and the matrix averaged. However,  
553 since our model ignores the MAF filters imposed in our simulations,  $\bar{\varphi}^T$  was adjusted. For Human  
554 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  
555 and 0.124, respectively, but 0.4 for both produced a better fit.

556 Lastly, we construct new allele frequencies,

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

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

558  $\mathrm{E}[q] = \frac{1}{2}$  is required to have  $\mathrm{E}[p'] = \frac{1}{2}$ . The resulting variance is

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

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

560 which is achieved with:

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

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

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

562 **5.2 Appendix B: comparisons between SRMSD<sub>p</sub>, AUC<sub>PR</sub>, and evaluation mea-  
563 sures from the literature**

564 **5.2.1 The inflation factor  $\lambda$**

565 Test statistic inflation has been used to measure model calibration [1, 20]. The inflation factor  
566  $\lambda$  is defined as the median  $\chi^2$  association statistic divided by theoretical median under the null  
567 hypothesis [2]. To compare p-values from non- $\chi^2$  tests (such as t-statistics),  $\lambda$  can be calculated  
568 from p-values using

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

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

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

580 **5.2.2 Empirical comparison of SRMSD<sub>p</sub> and λ**

581 There is a near one-to-one correspondence between λ and SRMSD<sub>p</sub> in our data (Fig. S1). PCA  
582 tended to be inflated ( $\lambda > 1$  and SRMSD<sub>p</sub> > 0) whereas LMM tended to be deflated ( $\lambda < 1$  and  
583 SRMSD<sub>p</sub> < 0), otherwise the data for both models fall on the same contiguous curve. We fit a  
584 sigmoidal function to this data,

585 
$$\text{SRMSD}_p(\lambda) = a \frac{\lambda^b - 1}{\lambda^b + 1}, \quad (10)$$

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

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

587 We fit this model to  $\lambda > 1$  only since it was less noisy and of greater interest, and obtained the  
588 curve shown in Fig. S1 with  $a = 0.563$  and  $b = 0.622$ . The value  $\lambda = 1.05$ , a common threshold  
589 for benign inflation [20], corresponds to  $\text{SRMSD}_p = 0.0085$  according to Eq. (10). Conversely,  
590  $\text{SRMSD}_p = 0.01$ , serving as a simpler rule of thumb, corresponds to  $\lambda = 1.06$ .

591 **5.2.3 Type I error rate**

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

597 **5.2.4 Statistical power and comparison to AUC<sub>PR</sub>**

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

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

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

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

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

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

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

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

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

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

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

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

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

608 Lastly, consider a simple yet common case in which model  $A$  is uniformly more powerful than  
609 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.

610 This ensures that the precision of  $A$  is greater than that of  $B$  at every recall value, so  $\text{AUC}_{\text{PR}}$  is  
611 greater for  $A$  than  $B$ . Thus,  $\text{AUC}_{\text{PR}}$  ranks calibrated models according to power.

## 612 Competing interests

613 The authors declare no competing interests.

614 **Acknowledgments**

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

618 **Web resources**

619 plink2, <https://www.cog-genomics.org/plink/2.0/>  
620 GCTA, <https://yanglab.westlake.edu.cn/software/gcta/>  
621 Eigensoft, <https://github.com/DReichLab/EIG>  
622 bnpsd, <https://cran.r-project.org/package=bnpsd>  
623 simfam, <https://cran.r-project.org/package=simfam>  
624 simtrait, <https://cran.r-project.org/package=simtrait>  
625 genio, <https://cran.r-project.org/package=genio>  
626 popkin, <https://cran.r-project.org/package=popkin>  
627 ape, <https://cran.r-project.org/package=ape>  
628 nnls, <https://cran.r-project.org/package=nnls>  
629 PRROC, <https://cran.r-project.org/package=PRROC>  
630 BEDMatrix, <https://cran.r-project.org/package=BEDMatrix>

631 **Data and code availability**

632 The data and code generated during this study are available on GitHub at <https://github.com/>  
633 OchoaLab/pca-assoc-paper. The public subset of Human Origins is available on the Reich Lab  
634 website at <https://reich.hms.harvard.edu/datasets>; non-public samples have to be requested  
635 from David Reich. The WGS version of HGDP was downloaded from the Wellcome Sanger In-  
636 stitute FTP site at [ftp://ngs.sanger.ac.uk/production/hgdp/hgdp\\_wgs.20190516/](ftp://ngs.sanger.ac.uk/production/hgdp/hgdp_wgs.20190516/). The high-  
637 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/](ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20190425_NYGC_GATK/).

639 **References**

- 640 [1] W. Astle and D. J. Balding. “Population Structure and Cryptic Relatedness in Genetic Asso-  
641 ciation Studies”. *Statist. Sci.* 24(4) (2009), pp. 451–471. DOI: 10.1214/09-STS307.
- 642 [2] B. Devlin and K. Roeder. “Genomic Control for Association Studies”. *Biometrics* 55(4) (1999),  
643 pp. 997–1004. DOI: 10.1111/j.0006-341X.1999.00997.x.
- 644 [3] B. F. Voight and J. K. Pritchard. “Confounding from Cryptic Relatedness in Case-Control  
645 Association Studies”. *PLOS Genetics* 1(3) (2005), e32. DOI: 10.1371/journal.pgen.0010032.
- 646 [4] S. Zhang, X. Zhu, and H. Zhao. “On a semiparametric test to detect associations between  
647 quantitative traits and candidate genes using unrelated individuals”. *Genetic Epidemiology*  
648 24(1) (2003), pp. 44–56. DOI: 10.1002/gepi.10196.
- 649 [5] A. L. Price et al. “Principal components analysis corrects for stratification in genome-wide  
650 association studies”. *Nat. Genet.* 38(8) (2006), pp. 904–909. DOI: 10.1038/ng1847.
- 651 [6] M. Bouaziz, C. Ambroise, and M. Guedj. “Accounting for Population Stratification in Practice:  
652 A Comparison of the Main Strategies Dedicated to Genome-Wide Association Studies”. *PLOS  
653 ONE* 6(12) (2011), e28845. DOI: 10.1371/journal.pone.0028845.
- 654 [7] N. Patterson, A. L. Price, and D. Reich. “Population Structure and Eigenanalysis”. *PLoS  
655 Genet* 2(12) (2006), e190. DOI: 10.1371/journal.pgen.0020190.
- 656 [8] I. T. Jolliffe. *Principal Component Analysis*. 2nd ed. New York: Springer-Verlag, 2002.
- 657 [9] J. K. Pritchard et al. “Association Mapping in Structured Populations”. *The American Journal  
658 of Human Genetics* 67(1) (2000), pp. 170–181. DOI: 10.1086/302959.
- 659 [10] D. H. Alexander, J. Novembre, and K. Lange. “Fast model-based estimation of ancestry in  
660 unrelated individuals”. *Genome Res.* 19(9) (2009), pp. 1655–1664. DOI: 10.1101/gr.094052.  
661 109.
- 662 [11] Q. Zhou, L. Zhao, and Y. Guan. “Strong Selection at MHC in Mexicans since Admixture”.  
663 *PLoS Genet.* 12(2) (2016), e1005847. DOI: 10.1371/journal.pgen.1005847.

- 664 [12] G. McVean. “A genealogical interpretation of principal components analysis”. *PLoS Genet*  
665 5(10) (2009), e1000686. DOI: [10.1371/journal.pgen.1000686](https://doi.org/10.1371/journal.pgen.1000686).
- 666 [13] X. Zheng and B. S. Weir. “Eigenanalysis of SNP data with an identity by descent interpreta-  
667 tion”. *Theor Popul Biol* 107 (2016), pp. 65–76. DOI: [10.1016/j.tpb.2015.09.004](https://doi.org/10.1016/j.tpb.2015.09.004).
- 668 [14] I. Cabreros and J. D. Storey. “A Likelihood-Free Estimator of Population Structure Bridging  
669 Admixture Models and Principal Components Analysis”. *Genetics* 212(4) (2019), pp. 1009–  
670 1029. DOI: [10.1534/genetics.119.302159](https://doi.org/10.1534/genetics.119.302159).
- 671 [15] A. M. Chiu et al. “Inferring population structure in biobank-scale genomic data”. *The Amer-  
672 ican Journal of Human Genetics* 0(0) (2022). DOI: [10.1016/j.ajhg.2022.02.015](https://doi.org/10.1016/j.ajhg.2022.02.015).
- 673 [16] K. Zhao et al. “An Arabidopsis Example of Association Mapping in Structured Samples”.  
674 *PLOS Genetics* 3(1) (2007), e4. DOI: [10.1371/journal.pgen.0030004](https://doi.org/10.1371/journal.pgen.0030004).
- 675 [17] H. Xu and Y. Guan. “Detecting Local Haplotype Sharing and Haplotype Association”. *Ge-  
676 netics* 197(3) (2014), pp. 823–838. DOI: [10.1534/genetics.114.164814](https://doi.org/10.1534/genetics.114.164814).
- 677 [18] J. Qian et al. “A fast and scalable framework for large-scale and ultrahigh-dimensional sparse  
678 regression with application to the UK Biobank”. *PLOS Genetics* 16(10) (2020), e1009141.  
679 DOI: [10.1371/journal.pgen.1009141](https://doi.org/10.1371/journal.pgen.1009141).
- 680 [19] T. Thornton and M. S. McPeek. “ROADTRIPS: case-control association testing with partially  
681 or completely unknown population and pedigree structure”. *Am. J. Hum. Genet.* 86(2) (2010),  
682 pp. 172–184. DOI: [10.1016/j.ajhg.2010.01.001](https://doi.org/10.1016/j.ajhg.2010.01.001).
- 683 [20] A. L. Price et al. “New approaches to population stratification in genome-wide association  
684 studies”. *Nature Reviews Genetics* 11(7) (2010), pp. 459–463. DOI: [10.1038/nrg2813](https://doi.org/10.1038/nrg2813).
- 685 [21] S. Lee et al. “Sparse Principal Component Analysis for Identifying Ancestry-Informative Mark-  
686 ers in Genome-Wide Association Studies”. *Genetic Epidemiology* 36(4) (2012), pp. 293–302.  
687 DOI: [10.1002/gepi.21621](https://doi.org/10.1002/gepi.21621).
- 688 [22] G. Abraham and M. Inouye. “Fast Principal Component Analysis of Large-Scale Genome-  
689 Wide Data”. *PLOS ONE* 9(4) (2014), e93766. DOI: [10.1371/journal.pone.0093766](https://doi.org/10.1371/journal.pone.0093766).

- 690 [23] K. Galinsky et al. “Fast Principal-Component Analysis Reveals Convergent Evolution of  
691 ADH1B in Europe and East Asia”. *The American Journal of Human Genetics* 98(3) (2016),  
692 pp. 456–472. DOI: 10.1016/j.ajhg.2015.12.022.
- 693 [24] G. Abraham, Y. Qiu, and M. Inouye. “FlashPCA2: principal component analysis of Biobank-  
694 scale genotype datasets”. *Bioinformatics* 33(17) (2017), pp. 2776–2778. DOI: 10.1093/bioinformatics/  
695 btx299.
- 696 [25] A. Agrawal et al. “Scalable probabilistic PCA for large-scale genetic variation data”. *PLOS  
697 Genetics* 16(5) (2020), e1008773. DOI: 10.1371/journal.pgen.1008773.
- 698 [26] J. Yu et al. “A unified mixed-model method for association mapping that accounts for multiple  
699 levels of relatedness”. *Nat. Genet.* 38(2) (2006), pp. 203–208. DOI: 10.1038/ng1702.
- 700 [27] H. M. Kang et al. “Efficient control of population structure in model organism association  
701 mapping”. *Genetics* 178(3) (2008), pp. 1709–1723. DOI: 10.1534/genetics.107.080101.
- 702 [28] A. Ochoa and J. D. Storey. “Estimating FST and kinship for arbitrary population structures”.  
703 *PLoS Genet* 17(1) (2021), e1009241. DOI: 10.1371/journal.pgen.1009241.
- 704 [29] J. Yang et al. “Advantages and pitfalls in the application of mixed-model association methods”.  
705 *Nat Genet* 46(2) (2014), pp. 100–106. DOI: 10.1038/ng.2876.
- 706 [30] H. M. Kang et al. “Variance component model to account for sample structure in genome-wide  
707 association studies”. *Nat. Genet.* 42(4) (2010), pp. 348–354. DOI: 10.1038/ng.548.
- 708 [31] C. Wu et al. “A Comparison of Association Methods Correcting for Population Stratification  
709 in Case–Control Studies”. *Annals of Human Genetics* 75(3) (2011), pp. 418–427. DOI: 10.  
710 1111/j.1469-1809.2010.00639.x.
- 711 [32] J. H. Sul and E. Eskin. “Mixed models can correct for population structure for genomic regions  
712 under selection”. *Nature Reviews Genetics* 14(4) (2013), p. 300. DOI: 10.1038/nrg2813-c1.
- 713 [33] W. Zhou et al. “Efficiently controlling for case-control imbalance and sample relatedness in  
714 large-scale genetic association studies”. *Nat Genet* 50(9) (2018), pp. 1335–1341. DOI: 10.1038/  
715 s41588-018-0184-y.

- 716 [34] Y. S. Aulchenko, D.-J. de Koning, and C. Haley. “Genomewide rapid association using mixed  
717 model and regression: a fast and simple method for genomewide pedigree-based quantitative  
718 trait loci association analysis”. *Genetics* 177(1) (2007), pp. 577–585. DOI: 10.1534/genetics.  
719 107.075614.
- 720 [35] Z. Zhang et al. “Mixed linear model approach adapted for genome-wide association studies”.  
721 *Nat Genet* 42(4) (2010), pp. 355–360. DOI: 10.1038/ng.546.
- 722 [36] C. Lippert et al. “FaST linear mixed models for genome-wide association studies”. *Nat. Meth-*  
723 *ods* 8(10) (2011), pp. 833–835. DOI: 10.1038/nmeth.1681.
- 724 [37] J. Yang et al. “GCTA: a tool for genome-wide complex trait analysis”. *Am. J. Hum. Genet.*  
725 88(1) (2011), pp. 76–82. DOI: 10.1016/j.ajhg.2010.11.011.
- 726 [38] J. Listgarten et al. “Improved linear mixed models for genome-wide association studies”. *Nat*  
727 *Methods* 9(6) (2012), pp. 525–526. DOI: 10.1038/nmeth.2037.
- 728 [39] X. Zhou and M. Stephens. “Genome-wide efficient mixed-model analysis for association stud-  
729 ies”. *Nat. Genet.* 44(7) (2012), pp. 821–824. DOI: 10.1038/ng.2310.
- 730 [40] G. R. Svishcheva et al. “Rapid variance components-based method for whole-genome associ-  
731 ation analysis”. *Nat Genet* 44(10) (2012), pp. 1166–1170. DOI: 10.1038/ng.2410.
- 732 [41] P.-R. Loh et al. “Efficient Bayesian mixed-model analysis increases association power in large  
733 cohorts”. *Nat. Genet.* 47(3) (2015), pp. 284–290. DOI: 10.1038/ng.3190.
- 734 [42] G. E. Hoffman. “Correcting for population structure and kinship using the linear mixed model:  
735 theory and extensions”. *PLoS ONE* 8(10) (2013), e75707. DOI: 10.1371/journal.pone.  
736 0075707.
- 737 [43] G. Tucker, A. L. Price, and B. Berger. “Improving the Power of GWAS and Avoiding Con-  
738 founding from Population Stratification with PC-Select”. *Genetics* 197(3) (2014), pp. 1045–  
739 1049. DOI: 10.1534/genetics.114.164285.
- 740 [44] N. Liu et al. “Controlling Population Structure in Human Genetic Association Studies with  
741 Samples of Unrelated Individuals”. *Stat Interface* 4(3) (2011), pp. 317–326. DOI: 10.4310/  
742 sii.2011.v4.n3.a6.

- 743 [45] J. Zeng et al. “Signatures of negative selection in the genetic architecture of human complex  
744 traits”. *Nature Genetics* 50(5) (2018), pp. 746–753. DOI: 10.1038/s41588-018-0101-4.
- 745 [46] M. Song, W. Hao, and J. D. Storey. “Testing for genetic associations in arbitrarily structured  
746 populations”. *Nat. Genet.* 47(5) (2015), pp. 550–554. DOI: 10.1038/ng.3244.
- 747 [47] X. Liu et al. “Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient  
748 Genome-Wide Association Studies”. *PLOS Genet.* 12(2) (2016), e1005767. DOI: 10.1371/journal.  
749 pgen.1005767.
- 750 [48] J. H. Sul, L. S. Martin, and E. Eskin. “Population structure in genetic studies: Confounding  
751 factors and mixed models”. *PLoS Genet.* 14(12) (2018), e1007309. DOI: 10.1371/journal.  
752 pgen.1007309.
- 753 [49] A. L. Price et al. “Response to Sul and Eskin”. *Nature Reviews Genetics* 14(4) (2013), p. 300.  
754 DOI: 10.1038/nrg2813-c2.
- 755 [50] T. G. P. Consortium. “A map of human genome variation from population-scale sequencing”.  
756 *Nature* 467(7319) (2010), pp. 1061–1073. DOI: 10.1038/nature09534.
- 757 [51] 1000 Genomes Project Consortium et al. “An integrated map of genetic variation from 1,092  
758 human genomes”. *Nature* 491(7422) (2012), pp. 56–65. DOI: 10.1038/nature11632.
- 759 [52] H. M. Cann et al. “A human genome diversity cell line panel”. *Science* 296(5566) (2002),  
760 pp. 261–262. DOI: 10.1126/science.296.5566.261b.
- 761 [53] N. A. Rosenberg et al. “Genetic Structure of Human Populations”. *Science* 298(5602) (2002),  
762 pp. 2381–2385. DOI: 10.1126/science.1078311.
- 763 [54] A. Bergström et al. “Insights into human genetic variation and population history from 929  
764 diverse genomes”. *Science* 367(6484) (2020). DOI: 10.1126/science.aay5012.
- 765 [55] N. Patterson et al. “Ancient admixture in human history”. *Genetics* 192(3) (2012), pp. 1065–  
766 1093. DOI: 10.1534/genetics.112.145037.
- 767 [56] I. Lazaridis et al. “Ancient human genomes suggest three ancestral populations for present-day  
768 Europeans”. *Nature* 513(7518) (2014), pp. 409–413. DOI: 10.1038/nature13673.

- 769 [57] I. Lazaridis et al. “Genomic insights into the origin of farming in the ancient Near East”.  
770        *Nature* 536(7617) (2016), pp. 419–424. DOI: 10.1038/nature19310.
- 771 [58] P. Skoglund et al. “Genomic insights into the peopling of the Southwest Pacific”. *Nature*  
772        538(7626) (2016), pp. 510–513. DOI: 10.1038/nature19844.
- 773 [59] J.-H. Park et al. “Distribution of allele frequencies and effect sizes and their interrelationships  
774        for common genetic susceptibility variants”. *PNAS* 108(44) (2011), pp. 18026–18031. DOI:  
775        10.1073/pnas.1114759108.
- 776 [60] L. J. O’Connor et al. “Extreme Polygenicity of Complex Traits Is Explained by Negative  
777        Selection”. *The American Journal of Human Genetics* 0(0) (2019). DOI: 10.1016/j.ajhg.  
778        2019.07.003.
- 779 [61] Y. B. Simons et al. “A population genetic interpretation of GWAS findings for human quanti-  
780        tative traits”. *PLOS Biology* 16(3) (2018), e2002985. DOI: 10.1371/journal.pbio.2002985.
- 781 [62] G. Malécot. *Mathématiques de l'hérédité*. Masson et Cie, 1948.
- 782 [63] S. Wright. “The Genetical Structure of Populations”. *Annals of Eugenics* 15(1) (1949), pp. 323–  
783        354. DOI: 10.1111/j.1469-1809.1949.tb02451.x.
- 784 [64] A. Jacquard. *Structures génétiques des populations*. Paris: Masson et Cie, 1970.
- 785 [65] A. Ochoa and J. D. Storey. *New kinship and FST estimates reveal higher levels of differenti-  
786        ation in the global human population*. 2019. DOI: 10.1101/653279.
- 787 [66] C. C. Chang et al. “Second-generation PLINK: rising to the challenge of larger and richer  
788        datasets”. *GigaScience* 4(1) (2015), p. 7. DOI: 10.1186/s13742-015-0047-8.
- 789 [67] D. J. Balding and R. A. Nichols. “A method for quantifying differentiation between populations  
790        at multi-allelic loci and its implications for investigating identity and paternity”. *Genetica*  
791        96(1-2) (1995), pp. 3–12. DOI: <https://doi.org/10.1007/BF01441146>.
- 792 [68] E. Paradis and K. Schliep. “ape 5.0: an environment for modern phylogenetics and evolutionary  
793        analyses in R”. *Bioinformatics* 35(3) (2019), pp. 526–528. DOI: 10.1093/bioinformatics/  
794        bty633.

- 795 [69] R. R. Sokal and C. D. Michener. “A statistical method for evaluating systematic relationships.”
- 796       *Univ. Kansas, Sci. Bull.* 38 (1958), pp. 1409–1438.
- 797 [70] C. L. Lawson and R. J. Hanson. *Solving least squares problems*. Englewood Cliffs: Prentice
- 798       Hall, 1974.
- 799 [71] K. M. Mullen and I. H. M. v. Stokkum. *nnls: The Lawson-Hanson algorithm for non-negative*
- 800       *least squares (NNLS)*. 2012.
- 801 [72] J.-H. Park et al. “Estimation of effect size distribution from genome-wide association studies
- 802       and implications for future discoveries”. *Nature Genetics* 42(7) (2010), pp. 570–575. DOI:
- 803       10.1038/ng.610.
- 804 [73] A. Grueneberg and G. d. l. Campos. “BGData - A Suite of R Packages for Genomic Analysis
- 805       with Big Data”. *G3: Genes, Genomes, Genetics* 9(5) (2019), pp. 1377–1383. DOI: 10.1534/
- 806       g3.119.400018.
- 807 [74] S. Fairley et al. “The International Genome Sample Resource (IGSR) collection of open human
- 808       genomic variation resources”. *Nucleic Acids Research* 48(D1) (2020), pp. D941–D947. DOI:
- 809       10.1093/nar/gkz836.
- 810 [75] A. Manichaikul et al. “Robust relationship inference in genome-wide association studies”.
- 811       *Bioinformatics* 26(22) (2010), pp. 2867–2873. DOI: 10.1093/bioinformatics/btq559.
- 812 [76] J. D. Storey. “The positive false discovery rate: a Bayesian interpretation and the q-value”.
- 813       *Ann. Statist.* 31(6) (2003), pp. 2013–2035. DOI: 10.1214/aos/1074290335.
- 814 [77] J. D. Storey and R. Tibshirani. “Statistical significance for genomewide studies”. *Proceedings*
- 815       *of the National Academy of Sciences of the United States of America* 100(16) (2003), pp. 9440–
- 816       9445. DOI: 10.1073/pnas.1530509100.
- 817 [78] J. Grau, I. Grosse, and J. Keilwagen. “PRROC: computing and visualizing precision-recall and
- 818       receiver operating characteristic curves in R”. *Bioinformatics* 31(15) (2015), pp. 2595–2597.
- 819       DOI: 10.1093/bioinformatics/btv153.
- 820 [79] P. Gopalan et al. “Scaling probabilistic models of genetic variation to millions of humans”.
- 821       *Nat. Genet.* 48(12) (2016), pp. 1587–1590. DOI: 10.1038/ng.3710.

- 822 [80] B. Rakitsch et al. “A Lasso multi-marker mixed model for association mapping with population  
823 structure correction”. *Bioinformatics* 29(2) (2013), pp. 206–214. DOI: 10.1093/bioinformatics/  
824 bts669.
- 825 [81] M. Conomos et al. “Model-free Estimation of Recent Genetic Relatedness”. *The American  
826 Journal of Human Genetics* 98(1) (2016), pp. 127–148. DOI: 10.1016/j.ajhg.2015.11.022.
- 827 [82] S. Gazal et al. “High level of inbreeding in final phase of 1000 Genomes Project”. *Sci Rep* 5(1)  
828 (2015), p. 17453. DOI: 10.1038/srep17453.
- 829 [83] A. Al-Khudhair et al. “Inference of Distant Genetic Relations in Humans Using “1000 Genomes””.  
830 *Genome Biology and Evolution* 7(2) (2015), pp. 481–492. DOI: 10.1093/gbe/evv003.
- 831 [84] L. Fedorova et al. “Atlas of Cryptic Genetic Relatedness Among 1000 Human Genomes”.  
832 *Genome Biology and Evolution* 8(3) (2016), pp. 777–790. DOI: 10.1093/gbe/evw034.
- 833 [85] D. Schlauch, H. Fier, and C. Lange. “Identification of genetic outliers due to sub-structure  
834 and cryptic relationships”. *Bioinformatics* 33(13) (2017), pp. 1972–1979. DOI: 10.1093/  
835 bioinformatics/btx109.
- 836 [86] B. M. Henn et al. “Cryptic Distant Relatives Are Common in Both Isolated and Cosmopolitan  
837 Genetic Samples”. *PLOS ONE* 7(4) (2012), e34267. DOI: 10.1371/journal.pone.0034267.
- 838 [87] V. Shchur and R. Nielsen. “On the number of siblings and p-th cousins in a large population  
839 sample”. *J Math Biol* 77(5) (2018), pp. 1279–1298. DOI: 10.1007/s00285-018-1252-8.

## Supplemental figures

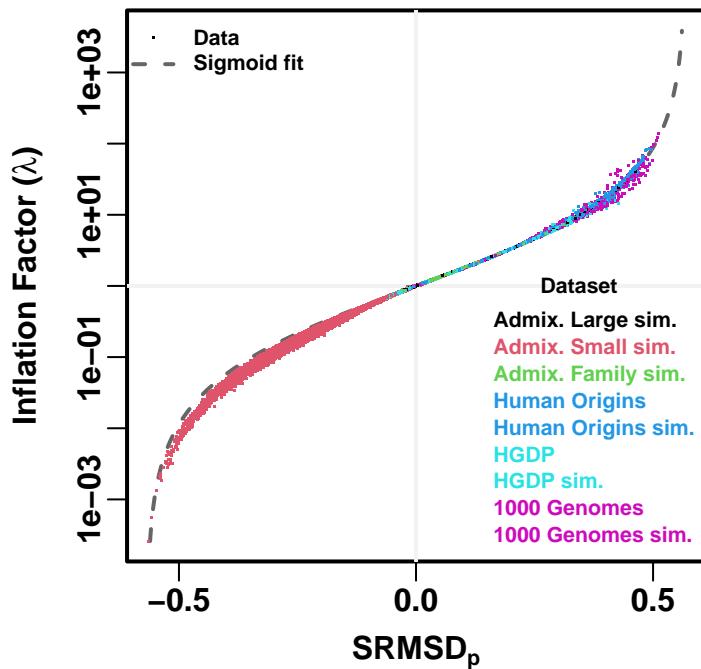


Figure S1: **Comparison between  $\text{SRMSD}_p$  and inflation factor.** Each point is a pair of statistics for one replicate, one association model (PCA or LMM with some number of PCs  $r$ ), one trait model (FES vs RC), and one dataset (color coded by dataset). Note log y-axis ( $\lambda$ ). The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

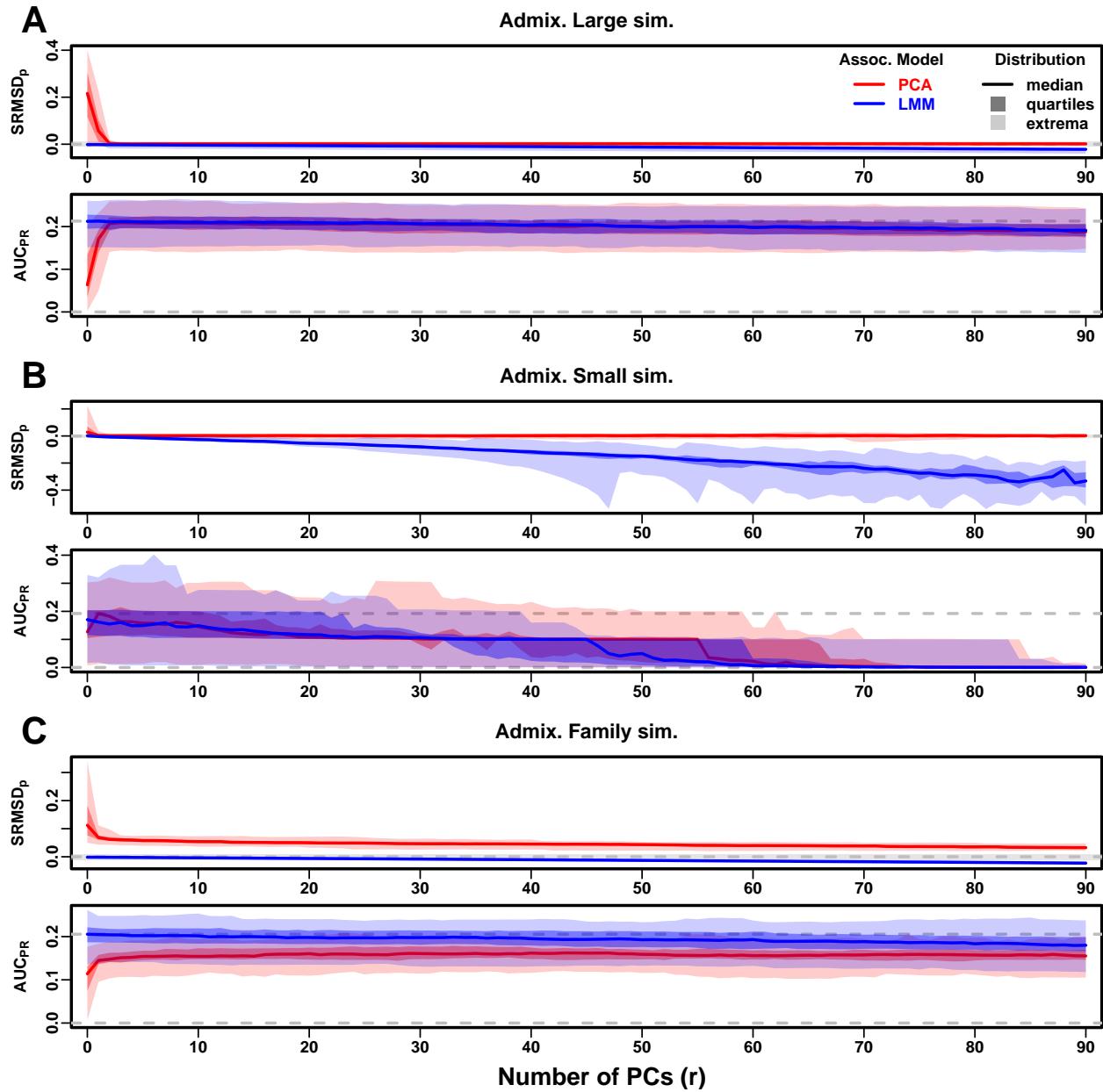


Figure S2: **Evaluations in admixture simulations with RC traits.** Traits simulated from RC model, otherwise the same as Fig. 3.

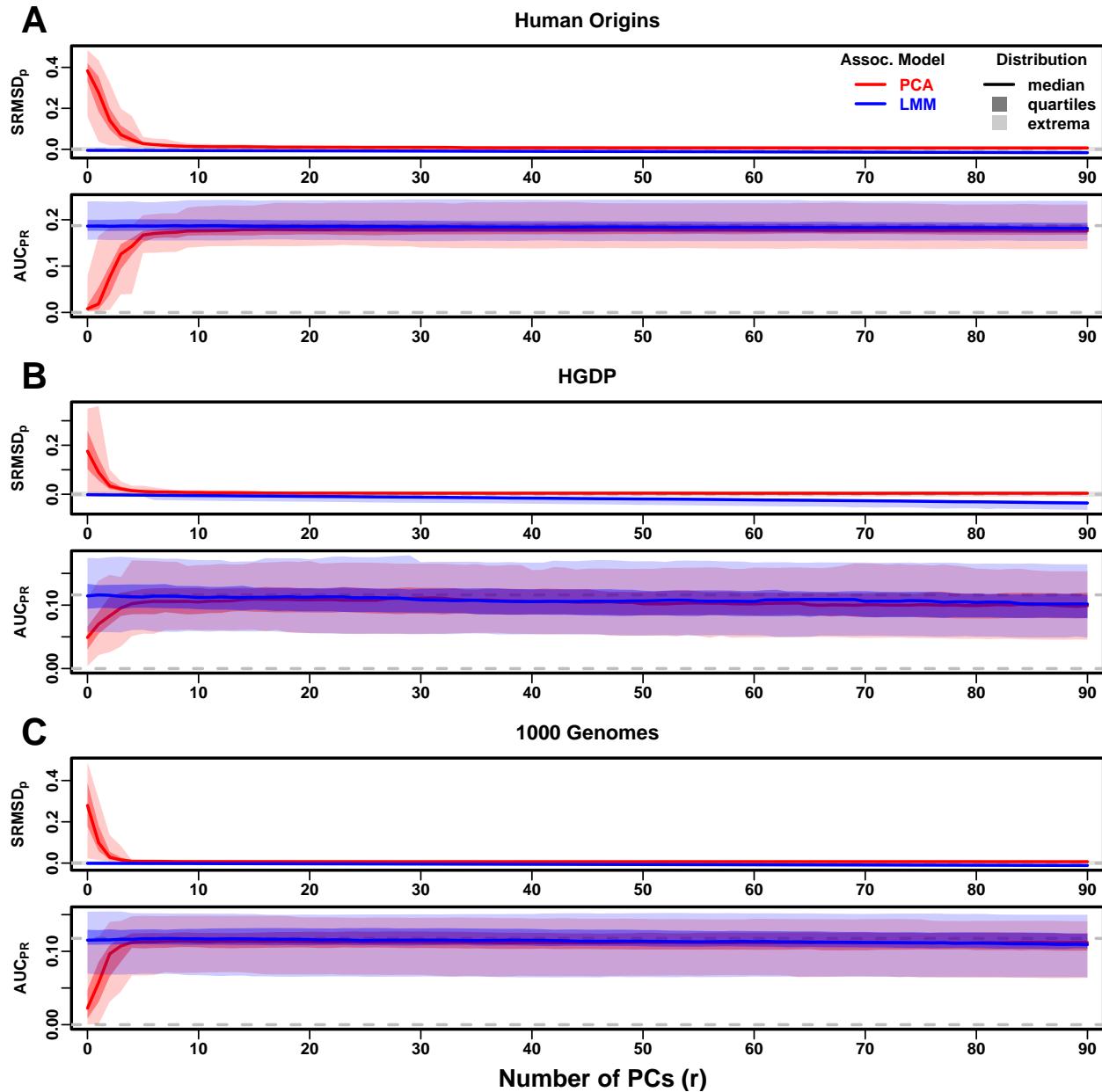


Figure S3: Evaluations in real human genotype datasets with RC traits. Traits simulated from RC model, otherwise the same as Fig. 4.

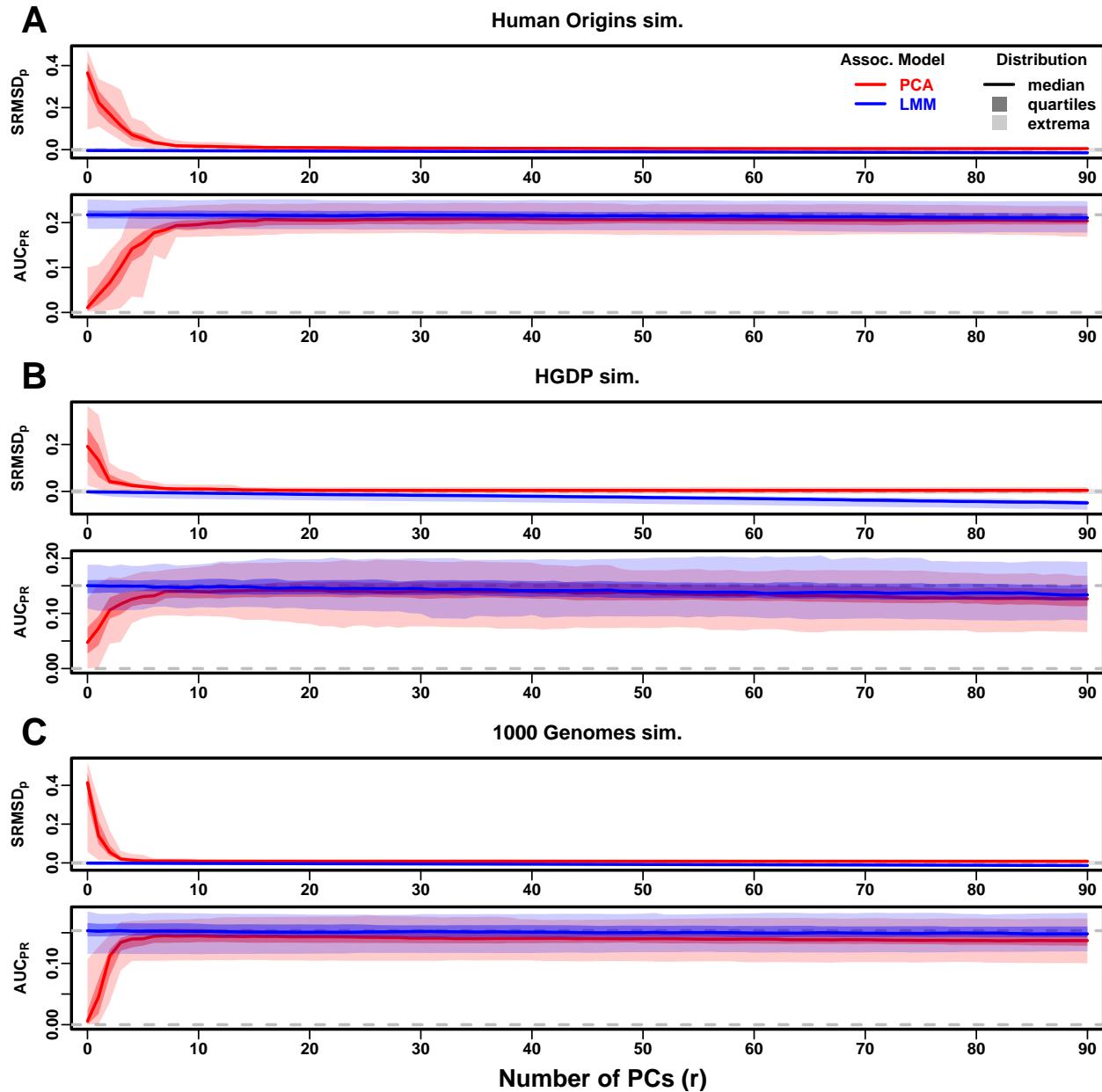


Figure S4: Evaluations in tree simulations fit to human data with RC traits. Traits simulated from RC model, otherwise the same as Fig. 5.

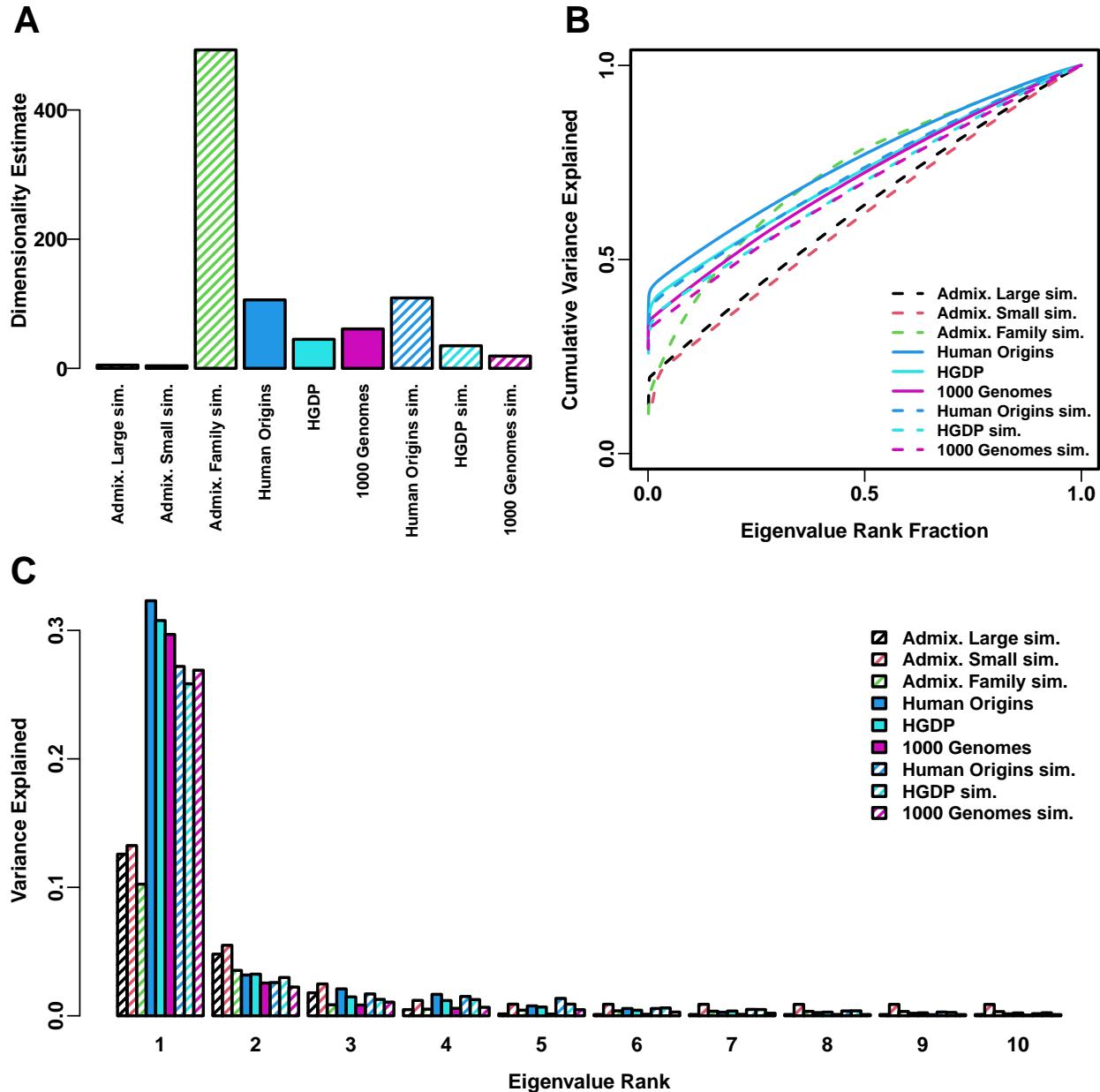


Figure S5: **Estimated dimensionality of datasets.** **A.** Kinship dimensionalities estimated with the Tracy-Widom test with  $p < 0.01$ . **B.** Cumulative variance explained versus eigenvalue rank fraction. **C.** Variance explained by first 10 eigenvalues.

## 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