

# **1 Limitations of principal components 2 in quantitative genetic association 3 models for human studies**

**4 Yiqi Yao<sup>1,5</sup> and Alejandro Ochoa<sup>1,2,\*</sup>**

**\*For correspondence:**

[alejandro.ochoa@duke.edu](mailto:alejandro.ochoa@duke.edu) (AO)

**Present address:** <sup>5</sup>BenHealth Consulting, Shanghai, Shanghai, 200023, China

**5** <sup>1</sup>Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27705, USA; **6** <sup>2</sup>Duke Center for Statistical Genetics and Genomics, Duke University, Durham, NC **7** 27705, USA

**8** 

---

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

**25** 

---

## **26 Introduction**

**27** The goal of a genetic association study is to identify loci whose genotype variation is significantly  
**28** correlated to given trait. Naive association tests assume that genotypes are drawn independently  
**29** from a common allele frequency. This assumption does not hold for structured populations, which  
**30** includes multiethnic cohorts and admixed individuals (ancient relatedness), and for family data  
**31** (recent relatedness) (*Astle and Balding, 2009*). Association studies of admixed and multiethnic co-  
**32** horts, the focus of this work, are becoming more common, are believed to be more powerful, and  
**33** are necessary to bring more equity to genetic medicine (*Rosenberg et al., 2010; Hoffman, 2013;*  
**34** *Coram et al., 2013; Medina-Gomez et al., 2015; Conomos et al., 2016a; Hodonsky et al., 2017; Mar-*  
**35** *tin et al., 2017a,b; Hindorff et al., 2018; Hoffmann et al., 2018; Mogil et al., 2018; Roselli et al.,*  
**36** *2018; Wojcik et al., 2019; Peterson et al., 2019; Zhong et al., 2019; Hu et al., 2020; Simonin-Wilmer*  
**37** *et al., 2021; Kamariza et al., 2021; Lin et al., 2021; Mahajan et al., 2022; Hou et al., 2023*). When  
**38** insufficient approaches are applied to data with relatedness, their association statistics are mis-  
**39** calibrated, resulting in excess false positives and loss of power (*Devlin and Roeder, 1999; Voight*  
**40** *and Pritchard, 2005; Astle and Balding, 2009*). Therefore, many specialized approaches have been

41 developed for genetic association under relatedness, of which PCA and LMM are the most popular.  
42 Genetic association with PCA consists of including the top eigenvectors of the population kin-  
43 ship matrix as covariates in a generalized linear model (*Zhang et al., 2003; Price et al., 2006; Bouaziz*  
44 *et al., 2011*). These top eigenvectors are a new set of coordinates for individuals that are commonly  
45 referred to as PCs in genetics (*Patterson et al., 2006*), the convention adopted here, but in other  
46 fields PCs instead denote what in genetics would be the projections of loci onto eigenvectors, which  
47 are new independent coordinates for loci (*Jolliffe, 2002*). The direct ancestor of PCA association is  
48 structured association, in which inferred ancestry (genetic cluster membership, often correspond-  
49 ing with labels such as “European”, “African”, “Asian”, etc.) or admixture proportions of these ances-  
50 tries are used as regression covariates (*Pritchard et al., 2000*). These models are deeply connected  
51 because PCs map to ancestry empirically (*Alexander et al., 2009; Zhou et al., 2016*) and theoretically  
52 (*McVean, 2009; Zheng and Weir, 2016; Cabreros and Storey, 2019; Chiu et al., 2022*), and they work  
53 as well as global ancestry in association studies but are estimated more easily (*Patterson et al.,*  
54 *2006; Zhao et al., 2007; Alexander et al., 2009; Bouaziz et al., 2011*). Another approach closely  
55 related to PCA is nonmetric multidimensional scaling (*Zhu and Yu, 2009*). PCs are also proposed  
56 for modeling environment effects that are correlated to ancestry, for example, through geography  
57 (*Novembre et al., 2008; Zhang and Pan, 2015; Lin et al., 2021*). The strength of PCA is its simplicity,  
58 which as covariates can be readily included in more complex models, such as haplotype association  
59 (*Xu and Guan, 2014*) and polygenic models (*Qian et al., 2020*). However, PCA assumes that the un-  
60 derlying relatedness space is low dimensional (or low rank), so it can be well modeled with a small  
61 number of PCs, which may limit its applicability. PCA is known to be inadequate for family data  
62 (*Patterson et al., 2006; Zhu and Yu, 2009; Thornton and McPeek, 2010; Price et al., 2010*), which is  
63 called “cryptic relatedness” when it is unknown to the researchers, but no other troublesome cases  
64 have been confidently identified. Recent work has focused on developing more scalable versions  
65 of the PCA algorithm (*Lee et al., 2012; Abraham and Inouye, 2014; Galinsky et al., 2016; Abraham*  
66 *et al., 2017; Agrawal et al., 2020*). PCA remains a popular and powerful approach for association  
67 studies.

68 The other dominant association model under relatedness is the LMM, which includes a random  
69 effect parameterized by the kinship matrix. Unlike PCA, LMM does not assume that relatedness  
70 is low-dimensional, and explicitly models families via the kinship matrix. Early LMMs used kinship  
71 matrices estimated from known pedigrees or using methods that captured recent relatedness only,  
72 and modeled population structure (ancestry) as fixed effects (*Yu et al., 2006; Zhao et al., 2007; Zhu*  
73 *and Yu, 2009*). Modern LMMs estimate kinship from genotypes using a non-parametric estima-  
74 tor, often referred to as a genetic relationship matrix, that captures the combined covariance due  
75 to family relatedness and ancestry (*Kang et al., 2008; Astle and Balding, 2009; Ochoa and Storey,*  
76 *2021*). Like PCA, LMM has also been proposed for modeling environment correlated to genetics  
77 (*Vilhjálmsson and Nordborg, 2013; Wang et al., 2022*). The classic LMM assumes a quantitative  
78 (continuous) complex trait, the focus of our work. Although case-control (binary) traits and their  
79 underlying ascertainment are theoretically a challenge (*Yang et al., 2014*), LMMs have been ap-  
80 plied successfully to balanced case-control studies (*Astle and Balding, 2009; Kang et al., 2010*) and  
81 simulations (*Price et al., 2010; Wu et al., 2011; Sul and Eskin, 2013*), and have been adapted for un-  
82 balanced case-control studies (*Zhou et al., 2018*). However, LMMs tend to be considerably slower  
83 than PCA and other models, so much effort has focused on improving their runtime and scalability  
84 (*Aulchenko et al., 2007; Kang et al., 2008, 2010; Zhang et al., 2010; Lippert et al., 2011; Yang et al.,*  
85 *2011; Listgarten et al., 2012; Zhou and Stephens, 2012; Svishcheva et al., 2012; Loh et al., 2015;*  
86 *Zhou et al., 2018*).

87 An LMM variant that incorporates PCs as fixed covariates is tested thoroughly in our work. Since  
88 PCs are the top eigenvectors of the same kinship matrix estimate used in modern LMMs (*Astle and*  
89 *Balding, 2009; Janss et al., 2012; Hoffman, 2013; Zhang and Pan, 2015*), then population structure  
90 is modeled twice in an LMM with PCs. However, some previous work has found the apparent  
91 redundancy of an LMM with PCs beneficial (*Price et al., 2010; Tucker et al., 2014; Zhang and Pan,*

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

Publication	Sim. Genotypes			Real <sup>d</sup>	Trait <sup>e</sup>	Power	PCs ( <i>r</i> )	Best
	Type <sup>a</sup>	K <sup>b</sup>	F <sub>ST</sub> <sup>c</sup>					
Zhao et al. (2007)				✓	Q	✓	8	LMM
Zhu and Yu (2009)	I, A, F	3, 8	≤0.15	✓	Q	✓	1-22	LMM
Astle and Balding (2009)	I	3	0.10		CC	✓	10	Tie
Kang et al. (2010)				✓	Both		2-100	LMM
Price et al. (2010)	I, F	2	0.01		CC		1	Mixed
Wu et al. (2011)	I, A	2-4	0.01		CC	✓	10	Mixed
Liu et al. (2011)	S, A	2-3	R		Q	✓	10	Tie
Sul and Eskin (2013)	I	2	0.01		CC		1	Tie
Tucker et al. (2014)	I	2	0.05	✓	Both	✓	5	Tie
Yang et al. (2014)				✓	CC	✓	5	Tie
Song et al. (2015)	S, A	2-3	R		Q		3	LMM
Loh et al. (2015)				✓	Q	✓	10	LMM
Zhang and Pan (2015)				✓	Q	✓	20-100	LMM
Liu et al. (2016)				✓	Q	✓	3-6	LMM
Sul et al. (2018)				✓	Q		100	LMM
Loh et al. (2018)				✓	Both	✓	20	LMM
Mbatchou et al. (2021)				✓	Both		1	LMM
This work	A, T, F	10-243	≤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: Subpopulation Tree; F: Family.

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

<sup>c</sup>R: simulated parameters based on real data, F<sub>ST</sub> not reported.

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

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

2015), while others did not (Liu et al., 2011; Janss et al., 2012), and the approach continues to be used (Zeng et al., 2018; Mbatchou et al., 2021) though not always (Matoba et al., 2020). (Recall that early LMMs used kinship to model family relatedness only, so population structure had to be modeled separately in those models, in practice as admixture fractions instead of PCs (Yu et al., 2006; Zhao et al., 2007; Zhu and Yu, 2009).) The LMM with PCs (vs no PCs) is also believed to help better model loci that have experienced selection (Price et al., 2010; Vilhjálmsson and Nordborg, 2013) and environment effects correlated with genetics (Zhang and Pan, 2015).

LMM and PCA are closely related models (Astle and Balding, 2009; Janss et al., 2012; Hoffman, 2013; Zhang and Pan, 2015), so similar performance is expected particularly under low-dimensional relatedness. Direct comparisons have yielded mixed results, with several studies finding superior performance for LMM, notably from papers promoting advances in LMMs, while many others report comparable performance (Table 1). No papers find that PCA outperforms LMM decisively, although PCA occasionally performs better in isolated and artificial cases or individual measures, often with unknown significance. Previous studies generally used either only simulated or only real genotypes, with only two studies using both. The simulated genotype studies, which tended to have low model dimensions and F<sub>ST</sub>, were more likely to report ties or mixed results (6/8), whereas real genotypes tended to clearly favor LMMs (9/11). Similarly, 10/12 papers with quantitative traits favor LMMs, whereas 6/9 papers with case-control traits gave ties or mixed results—the only factor we do not explore in this work. Additionally, although all previous evaluations measured type I error (or proxies such as genomic inflation factors (Devlin and Roeder, 1999) or QQ plots), a large fraction (6/17) did not measure power (or proxies such as ROC curves), and only four used more

113 than one number of PCs for PCA. Lastly, no consensus has emerged as to why LMM might outper-  
 114 form PCA or vice versa (*Price et al., 2010; Sul and Eskin, 2013; Price et al., 2013; Hoffman, 2013*),  
 115 or which features of the real datasets are critical for the LMM advantage other than family relat-  
 116 edness, resulting in unclear guidance for using PCA. Hence, our work includes real and simulated  
 117 genotypes with higher model dimensions and  $F_{ST}$  matching that of multiethnic human cohorts  
 118 (*Ochoa and Storey, 2021, 2019*), we vary the number of PCs, and measure robust proxies for type  
 119 I error control and calibrated power.

120 In this work, we evaluate the PCA and LMM association models under various numbers of PCs,  
 121 which are included in LMMs too. We use genotype simulations (admixture, family, and subpop-  
 122 ulation tree models) and three real datasets: the 1000 Genomes Project (*Consortium, 2010; 1000*  
 123 *Genomes Project Consortium et al., 2012*), the Human Genome Diversity Panel (HGDP) (*Cann et al.,*  
 124 *2002; Rosenberg et al., 2002; Bergström et al., 2020*), and Human Origins (*Patterson et al., 2012;*  
 125 *Lazaridis et al., 2014, 2016; Skoglund et al., 2016*). We simulate quantitative traits from two models:  
 126 fixed effect sizes (FES) construct coefficients inverse to allele frequency, which matches real data  
 127 (*Park et al., 2011; Zeng et al., 2018; O'Connor et al., 2019*) and corresponds to high pleiotropy and  
 128 strong balancing selection (*Simons et al., 2018*) and strong negative selection (*Zeng et al., 2018;*  
 129 *O'Connor et al., 2019*), which are appropriate assumptions for diseases; and random coefficients  
 130 (RC), which are drawn independent of allele frequency, and corresponds to neutral traits (*Zeng*  
 131 *et al., 2018; Simons et al., 2018*). LMM without PCs consistently performs best in simulations with-  
 132 out environment, and greatly outperforms PCA in the family simulation and in all real datasets. The  
 133 tree simulations, which model subpopulations with the tree but exclude family structure, do not  
 134 recapitulate the real data results, suggesting that family relatedness in real data is the reason for  
 135 poor PCA performance. Lastly, removing up to 4th degree relatives in the real datasets recapitu-  
 136 lates poor PCA performance, showing that the more numerous distant relatives explain the result,  
 137 and suggesting that PCA is generally not an appropriate model for real data. We find that both  
 138 LMM and PCA are able to model environment effects correlated with genetics, and LMM with PCs  
 139 gains a small advantage in this setting only, but direct modeling of environment performs much  
 140 better. All together, we find that LMMs without PCs are generally a preferable association model,  
 141 and present novel simulation and evaluation approaches to measure the performance of these  
 142 and other genetic association approaches.

## 143 Materials and Methods

### 144 The complex trait model and PCA and LMM approximations

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

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

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

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

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

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

157 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  $n \times r$   
 158 matrix of PCs,  $\gamma_r$  is the length- $r$  vector of PC coefficients,  $s$  is a length- $n$  vector of random effects,  
 159  $\Phi^T = (\varphi_{jk}^T)$  is the  $n \times n$  kinship matrix conditioned on the ancestral population  $T$ , and  $\sigma_s^2$  is a variance  
 160 factor. Both models condition the regression of the focal locus  $i$  on an approximation of the total  
 161 polygenic effect  $\mathbf{X}'\beta$  with the same covariance structure, which is parameterized by the kinship  
 162 matrix. Under the kinship model, genotypes are random variables obeying

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

163 where  $p_i^T$  is the ancestral allele frequency of locus  $i$  (*Malécot, 1948; Wright, 1949; Jacquard, 1970;*  
 164 *Astle and Balding, 2009*). Assuming independent loci, the covariance of the polygenic effect is

$$165 \quad \text{Cov}(\mathbf{X}'\beta) = 2\sigma_s^2\Phi^T, \quad \sigma_s^2 = \sum_{i=1}^m 2p_i^T(1-p_i^T)\beta_i^2,$$

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

$$170 \quad s = \mathbf{U}\gamma_{LMM}, \quad \gamma_{LMM} \sim \text{Normal}(\mathbf{0}, 2\sigma_s^2\Lambda),$$

171 which follows from the affine transformation property of multivariate normal distributions. There-  
 172 fore, the PCA term  $\mathbf{U}_r\gamma_r$  can be derived from the above equation under the additional assumption  
 173 that the kinship matrix has approximate rank  $r$  and the coefficients  $\gamma_r$  are fit without constraints.  
 174 In contrast, the LMM uses all eigenvectors, while effectively shrinking their coefficients  $\gamma_{LMM}$  as all  
 175 random effects models do, although these parameters are marginalized (*Astle and Balding, 2009;*  
 176 *Janss et al., 2012; Hoffman, 2013; Zhang and Pan, 2015*). PCA has more parameters than LMM, so  
 177 it may overfit more: ignoring the shared terms in *Equation 2* and *Equation 3*, PCA fits  $r$  parameters  
 178 (length of  $\gamma$ ), whereas LMMs fit only one ( $\sigma_s^2$ ).

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

$$181 \quad \mathbf{X}_S = \left( \frac{x_{ij} - 2\hat{p}_i^T}{\sqrt{4\hat{p}_i^T(1-\hat{p}_i^T)}} \right), \quad \hat{\Phi}^T = \frac{1}{m}\mathbf{X}'_S\mathbf{X}_S, \quad (5)$$

182 where the unknown  $p_i^T$  is estimated by  $\hat{p}_i^T = \frac{1}{2n} \sum_{j=1}^n x_{ij}$  (*Price et al., 2006; Kang et al., 2008, 2010;*  
 183 *Yang et al., 2011; Zhou and Stephens, 2012; Yang et al., 2014; Loh et al., 2015; Sul et al., 2018; Zhou*  
 184 *et al., 2018*). However, this kinship estimator has a complex bias that differs for every individual  
 185 pair, which arises due to the use of this estimated  $\hat{p}_i^T$  (*Ochoa and Storey, 2021, 2019*). Nevertheless,  
 186 in PCA and LMM these biased estimates perform as well as unbiased ones (*Hou and Ochoa, 2023*).

187 We selected fast and robust software implementing the basic PCA and LMM models. PCA as-  
 188 sociation was performed with plink2 (*Chang et al., 2015*). The quantitative trait association model  
 189 is a linear regression with covariates, evaluated using the t-test. PCs were calculated with plink2,  
 190 which equal the top eigenvectors of *Equation 5* after removing loci with minor allele frequency  
 191 MAF < 0.1.

192 LMM association was performed using GCTA (*Yang et al., 2011, 2014*). Its kinship estimator  
 193 equals *Equation 5*. PCs were calculated using GCTA from its kinship estimate. Association signifi-  
 194 cance is evaluated with a score test. In the small simulation only, GCTA with large numbers of PCs  
 195 had convergence and singularity errors in some replicates, which were treated as missing data.

## 195 Simulations

196 Every simulation was replicated 50 times, drawing anew all genotypes (except for real datasets)  
 197 and traits. Below we use the notation  $f_A^B$  for the inbreeding coefficient of a subpopulation  $A$  from

198 another subpopulation  $B$  ancestral to  $A$ . In the special case of the *total* inbreeding of  $A$ ,  $f_A^T$ ,  $T$  is  
 199 an overall ancestral population, which is ancestral to every individual under consideration, such as  
 200 the most recent common ancestor (MRCA) population.

201 Genotype simulation from the admixture model

202 The basic admixture model is as described previously (*Ochoa and Storey, 2021*) and is implemented  
 203 in the R package `bnpd`. Both Large and Family simulations have  $n = 1,000$  individuals, while Small  
 204 has  $n = 100$ . The number of loci is  $m = 100,000$ . Individuals are admixed from  $K = 10$  intermediate  
 205 subpopulations, or ancestries. Each subpopulation  $S_u$  ( $u \in \{1, \dots, K\}$ ) is at coordinate  $u$  and has an  
 206 inbreeding coefficient  $f_{S_u}^T = u\tau$  for some  $\tau$ . Ancestry proportions  $q_{ju}$  for individual  $j$  and  $S_u$  arise  
 207 from a random walk with spread  $\sigma$  on the 1D geography, and  $\tau$  and  $\sigma$  are fit to give  $F_{ST} = 0.1$  and  
 208 mean kinship  $\bar{\theta}^T = 0.5F_{ST}$  for the admixed individuals (*Ochoa and Storey, 2021*). Random ancestral  
 209 allele frequencies  $p_i^T$ , subpopulation allele frequencies  $p_i^{S_u}$ , individual-specific allele frequencies  $\pi_{ij}$ ,  
 210 and genotypes  $x_{ij}$  are drawn from this hierarchical model:

$$211 \quad p_i^T \sim \text{Uniform}(0.01, 0.5), \\ 212 \quad p_i^{S_u} | p_i^T \sim \text{Beta}\left(p_i^T \left(\frac{1}{f_{S_u}^T} - 1\right), (1 - p_i^T) \left(\frac{1}{f_{S_u}^T} - 1\right)\right), \\ 213 \quad \pi_{ij} = \sum_{u=1}^K q_{ju} p_i^{S_u}, \\ 214 \quad x_{ij} | \pi_{ij} \sim \text{Binomial}(2, \pi_{ij}),$$

215 where this Beta is the Balding-Nichols distribution (*Balding and Nichols, 1995*) with mean  $p_i^T$  and  
 216 variance  $p_i^T (1 - p_i^T) f_{S_u}^T$ . Fixed loci ( $i$  where  $x_{ij} = 0$  for all  $j$ , or  $x_{ij} = 2$  for all  $j$ ) are drawn again  
 217 from the model, starting from  $p_i^T$ , iterating until no loci are fixed. Each replicate draws a genotypes  
 218 starting from  $p_i^T$ .

219 As a brief aside, we prove that global ancestry proportions as covariates is equivalent in ex-  
 220 pectation to using PCs under the admixture model. Note that the latent space of  $\mathbf{X}$ , which is the  
 221 subspace to which the data is constrained by the admixture model, is given by  $(\pi_{ij})$ , which has  $K$   
 222 dimensions (number of columns of  $\mathbf{Q} = (q_{ju})$ ), so the top  $K$  PCs span this space. Since associations  
 223 include an intercept term ( $\mathbf{1}\alpha$  in *Equation 2*), estimated PCs are orthogonal to  $\mathbf{1}$  (note  $\hat{\Phi}^T \mathbf{1} = \mathbf{0}$  be-  
 224 cause  $\mathbf{X}_S \mathbf{1} = \mathbf{0}$ ), and the sum of rows of  $\mathbf{Q}$  sums to one, then only  $K - 1$  PCs plus the intercept are  
 225 needed to span the latent space of this admixture model.

226 Genotype simulation from random admixed families

227 We simulated a pedigree with admixed founders, no close relative pairings, assortative mating  
 228 based on a 1D geography (to preserve admixture structure), random family sizes, and arbitrary  
 229 numbers of generations (20 here). This simulation is implemented in the R package `simfam`. Gen-  
 230 erations are drawn iteratively. Generation 1 has  $n = 1000$  individuals from the above admixture  
 231 simulation ordered by their 1D geography. Local kinship measures pedigree relatedness; in the  
 232 first generation, everybody is locally unrelated and outbred. Individuals are randomly assigned  
 233 sex. In the next generation, individuals are paired iteratively, removing random males from the  
 234 pool of available males and pairing them with the nearest available female with local kinship  $< 1/4^3$   
 235 (stay unpaired if there are no matches), until there are no more available males or females. Let  
 236  $n = 1000$  be the desired population size,  $n_m = 1$  the minimum number of children per family and  $n_f$   
 237 the number of families (paired parents) in the current generation, then the number of additional  
 238 children (beyond the minimum) is drawn from Poisson( $n/n_f - n_m$ ). Let  $\delta$  be the difference between  
 239 desired and current population sizes. If  $\delta > 0$ , then  $\delta$  random families are incremented by 1. If  
 240  $\delta < 0$ , then  $|\delta|$  random families with at least  $n_m + 1$  children are decremented by 1. If  $|\delta|$  exceeds  
 241 the number of families, all families are incremented or decremented as needed and the process  
 242 is iterated. Children are assigned sex randomly, and are reordered by the average coordinate of

their parents. Children draw alleles from their parents independently per locus. A new random pedigree is drawn for each replicate, as well as new founder genotypes from the admixture model.

Genotype simulation from a subpopulation tree model

This model draws subpopulations allele frequencies from a hierarchical model parameterized by a tree, which is also implemented in `bnpnsd` and relies on the R package `ape` for general tree data structures and methods (*Paradis and Schliep, 2019*). The ancestral population  $T$  is the root, and each node is a subpopulation  $S_w$  indexed arbitrarily. Each edge between  $S_w$  and its parent population  $P_w$  has an inbreeding coefficient  $f_{S_w}^{P_w}$ .  $p_i^T$  are drawn from a given distribution, which is constructed to mimic each real dataset in **Appendix 1**. Given the allele frequencies  $p_i^{P_w}$  of the parent population,  $S_w$ 's allele frequencies are drawn from:

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

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

Fitting subpopulation tree to real data

We developed new methods to fit trees to real data based on unbiased kinship estimates from `popkin`, implemented in `bnpnsd`. A tree with given inbreeding coefficients  $f_{S_w}^{P_w}$  for its edges (between subpopulation  $S_w$  and its parent  $P_w$ ) gives rise to a coancestry matrix  $\theta_{uv}^T$  for a subpopulation pair  $(S_u, S_v)$ , and the goal is to recover these edge inbreeding coefficients from coancestry estimates. Coancestry values are total inbreeding coefficients of the MRCA population of each subpopulation pair. Therefore, we calculate  $f_{S_w}^T$  for every  $S_w$  recursively from the root as follows. Nodes with parent  $P_w = T$  are already as desired. Given  $f_{P_w}^T$ , the desired  $f_{S_w}^T$  is calculated via the "additive edge"  $\delta_w$  (*Ochoa and Storey, 2021*):

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

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

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

where the sum includes all  $w$ , and  $I_w(u, v)$  equals 1 if  $S_w$  is a common ancestor of  $S_u, S_v$ , 0 otherwise. Note that  $I_w(u, v)$  reflects tree topology and  $\delta_w$  edge values.

To estimate population-level coancestry, first kinship ( $\hat{\phi}_{jk}^T$ ) is estimated using `popkin` (*Ochoa and Storey, 2021*). Individual coancestry ( $\hat{\theta}_{jk}^T$ ) is estimated from kinship using

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

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

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

Topology is estimated with hierarchical clustering using the weighted pair group method with arithmetic mean (*Sokal and Michener, 1958*), with distance function  $d(S_u, S_v) = \max\{\hat{\theta}_{uv}^T\} - \hat{\theta}_{uv}^T$ , which succeeds due to the monotonic relationship between node depth and coancestry (**Equation 7**). This algorithm recovers the true topology from the true coancestry values, and performs well for estimates from genotypes.

279 To estimate tree edge lengths, first  $\delta_w$  are estimated from  $\hat{\phi}_w^T$  and the topology using **Equation 7**  
 280 and non-negative least squares linear regression (**Lawson and Hanson, 1974**) (implemented in `nmls`  
 281 (**Mullen and Stokkum, 2012**)) to yield non-negative  $\delta_w$ , and  $f_{S_w}^{P_w}$  are calculated from  $\delta_w$  by reversing  
 282 **Equation 6**. To account for small biases in coancestry estimation, an intercept term  $\delta_0$  is included  
 283 ( $I_0(u, v) = 1$  for all  $u, v$ ), and when converting  $\delta_w$  to  $f_{S_w}^{P_w}$ ,  $\delta_0$  is treated as an additional edge to the  
 284 root, but is ignored when drawing allele frequencies from the tree.

#### 285 Trait Simulation

286 Traits are simulated from the quantitative trait model of **Equation 1**, with novel bias corrections  
 287 for simulating the desired heritability from real data relying on the unbiased kinship estimator  
 288 `popkin` (**Ochoa and Storey, 2021**). This simulation is implemented in the R package `simtrait`. All  
 289 simulations have a fixed narrow-sense heritability of  $h^2$ , a variance proportion due to environment  
 290 effects  $\sigma_\eta^2$ , and residuals are drawn from  $\epsilon_j \sim \text{Normal}(0, \sigma_\epsilon^2)$  with  $\sigma_\epsilon^2 = 1 - h^2 - \sigma_\eta^2$ . The number of  
 291 causal loci  $m_1$ , which determines the average coefficient size, is chosen with the heuristic formula  
 292  $m_1 = \text{round}(nh^2/8)$ , which empirically balances power well with varying  $n$  and  $h^2$ . The set of causal loci  
 293  $C$  is drawn anew for each replicate, from loci with MAF  $\geq 0.01$  to avoid rare causal variants, which  
 294 are not discoverable by PCA or LMM at the sample sizes we considered. Letting  $v_i^T = p_i^T(1 - p_i^T)$ ,  
 295 the effect size of locus  $i$  equals  $2v_i^T\beta_i^2$ , its contribution of the trait variance (**Park et al., 2010**). Under  
 296 the *fixed effect sizes* (FES) model, initial causal coefficients are

$$297 \quad \beta_i = \frac{1}{\sqrt{2v_i^T}}$$

298 for known  $p_i^T$ ; otherwise  $v_i^T$  is replaced by the unbiased estimator (**Ochoa and Storey, 2021**)  $\hat{v}_i^T =$   
 299  $\hat{p}_i^T(1 - \hat{p}_i^T)/(1 - \bar{\varphi}^T)$ , where  $\bar{\varphi}^T$  is the mean kinship estimated with `popkin`. Each causal locus is  
 300 multiplied by -1 with probability 0.5. Alternatively, under the *random coefficients* (RC) model, initial  
 301 causal coefficients are drawn independently from  $\beta_i \sim \text{Normal}(0, 1)$ . For both models, the initial  
 302 genetic variance is  $\sigma_0^2 = \sum_{i \in C} 2v_i^T\beta_i^2$ , replacing  $v_i^T$  with  $\hat{v}_i^T$  for unknown  $p_i^T$  (so  $\sigma_0^2$  is an unbiased  
 303 estimate), so we multiply every initial  $\beta_i$  by  $\frac{h}{\sigma_0}$  to have the desired heritability. Lastly, for known  $p_i^T$ ,  
 304 the intercept coefficient is  $\alpha = -\sum_{i \in C} 2p_i^T\beta_i$ . When  $p_i^T$  are unknown,  $\hat{p}_i^T$  should not replace  $p_i^T$  since  
 305 that distorts the trait covariance (for the same reason the standard kinship estimator in **Equation 5**  
 306 is biased), which is avoided with

$$307 \quad \alpha = -\frac{2}{m_1} \left( \sum_{i \in C} \hat{p}_i^T \right) \left( \sum_{i \in C} \beta_i \right).$$

308 Simulations optionally included multiple environment group effects, similarly to previous mod-  
 309 els (**Zhang and Pan, 2015; Wang et al., 2022**), as follows. Each independent environment  $i$  has  
 310 predefined groups, and each group  $g$  has random coefficients drawn independent from  $\eta_{gi} \sim$   
 311  $\text{Normal}(0, \sigma_{\eta i}^2)$  where  $\sigma_{\eta i}^2$  is a specified variance proportion for environment  $i$ .  $Z$  has individuals  
 312 along columns and environment-groups along rows, and it contains indicator variables: 1 if the  
 313 individual belongs to the environment-group, 0 otherwise.

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

#### 322 Real human genotype datasets

323 The three datasets were processed as before (**Ochoa and Storey, 2019**) (summarized below), ex-  
 324 cept with an additional filter so loci are in approximate linkage equilibrium and rare variants are

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

Trait variance type	$h^2$	$\sigma_\eta^2$	$\sigma_\epsilon^2$
High heritability	0.8	0.0	0.2
Low heritability	0.3	0.0	0.7
Environment	0.3	0.5	0.2

**Table 3.** Features of simulated and real human genotype datasets.

Dataset	Type	Loci ( $m$ )	Ind. ( $n$ )	Subpops. <sup>a</sup> ( $K$ )	Causal loci <sup>b</sup> ( $m_1$ )	$F_{ST}$ <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 additional model dimension of 20 generation pedigree structure. For real datasets, lower range is continental subpopulations, upper range is number of fine-grained subpopulations.

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

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

325 removed. All processing was performed with `plink2` (Chang *et al.*, 2015), and analysis was uniquely  
 326 enabled by the R packages `BEDMatrix` (Grueneberg and Campos, 2019) and `genio`. Each dataset  
 327 groups individuals in a two-level hierarchy: continental and fine-grained subpopulations. Final  
 328 dataset sizes are in **Table 3**.

329 We obtained the full (including non-public) Human Origins by contacting the authors and agree-  
 330 ing to their usage restrictions. The Pacific data (Skoglund *et al.*, 2016) was obtained separately from  
 331 the rest (Lazaridis *et al.*, 2014, 2016), and datasets were merged using the intersection of loci. We  
 332 removed ancient individuals, and individuals from singleton and non-native subpopulations. Non-  
 333 autosomal loci were removed. Our analysis of both the whole-genome sequencing (WGS) version  
 334 of HGDP (Bergström *et al.*, 2020) and the high-coverage NYGC version of 1000 Genomes (Fairley  
 335 *et al.*, 2020) was restricted to autosomal biallelic SNP loci with filter “PASS”.

336 Since our evaluations assume uncorrelated loci, we filtered each real dataset with `plink2` us-  
 337 ing parameters “`--indep-pairwise 1000kb 0.3`”, which iteratively removes loci that have a greater  
 338 than 0.3 squared correlation coefficient with another locus that is within 1000kb, stopping until  
 339 no such loci remain. Since all real datasets have numerous rare variants, while PCA and LMM are  
 340 not able to detect associations involving rare variants, we removed all loci with MAF < 0.01. Lastly,  
 341 only HGDP had loci with over 10% missingness removed, as they were otherwise 17% of remaining  
 342 loci (for Human Origins and 1000 Genomes they were under 1% of loci so they were not removed).  
 343 Kinship matrix rank and eigenvalues were calculated from `popkin` kinship estimates. Eigenvalues  
 344 were assigned p-values with `twstats` of the `Eigensoft` package (Patterson *et al.*, 2006), and kinship  
 345 matrix rank was estimated as the largest number of consecutive eigenvalue from the start that all  
 346 satisfy  $p < 0.01$  (p-values did not increase monotonically). For the evaluation with close relatives re-  
 347 moved, each dataset was filtered with `plink2` with option “`--king-cutoff`” with cutoff 0.02209709  
 348 ( $= 2^{-11/2}$ ) for removing up to 4th degree relatives using KING-robust (Manichaikul *et al.*, 2010), and  
 349 MAF < 0.01 filter is reapplied (**Table 4**).

**Table 4.** 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

### 350    **Evaluation of performance**

351    All approaches are evaluated using two complementary metrics: SRMSD<sub>p</sub> quantifies p-value uniformity, and AUC<sub>PR</sub> measures causal locus classification performance and reflects power while ranking 352    miscalibrated models fairly. These measures are more robust alternatives to previous measures 353    from the literature (*Appendix 2*), and are implemented in `simtrait`.

354    P-values for continuous test statistics have a uniform distribution when the null hypothesis 355    holds, a crucial assumption for type I error and FDR control (*Storey, 2003; Storey and Tibshirani, 356    2003*). We use the Signed Root Mean Square Deviation (SRMSD<sub>p</sub>) to measure the difference be- 357    tween the observed null p-value quantiles and the expected uniform quantiles:

$$359 \quad \text{SRMSD}_p = \text{sgn}(u_{\text{median}} - p_{\text{median}}) \sqrt{\frac{1}{m_0} \sum_{i=1}^{m_0} (u_i - p_{(i)})^2},$$

360    where  $m_0 = m - m_1$  is the number of null (non-causal) loci, here  $i$  indexes null loci only,  $p_{(i)}$  is the  $i$ th 361    ordered null p-value,  $u_i = (i - 0.5)/m_0$  is its expectation,  $p_{\text{median}}$  is the median observed null p-value, 362     $u_{\text{median}} = \frac{1}{2}$  is its expectation, and sgn is the sign function (1 if  $u_{\text{median}} \geq p_{\text{median}}$ , -1 otherwise). Thus, 363    SRMSD<sub>p</sub> = 0 corresponds to calibrated p-values, SRMSD<sub>p</sub> > 0 indicate anti-conservative p-values, 364    and SRMSD<sub>p</sub> < 0 are conservative p-values. The maximum SRMSD<sub>p</sub> is achieved when all p-values 365    are zero (the limit of anti-conservative p-values), which for infinite loci approaches

$$366 \quad \text{SRMSD}_p \rightarrow \sqrt{\int_0^1 u^2 du} = \frac{1}{\sqrt{3}} \approx 0.577.$$

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

368    Precision and recall are standard performance measures for binary classifiers that do not re- 369    quire calibrated p-values (*Grau et al., 2015*). Given the total numbers of true positives (TP), false 370    positives (FP) and false negatives (FN) at some threshold or parameter  $t$ , precision and recall are

$$371 \quad \text{Precision}(t) = \frac{\text{TP}(t)}{\text{TP}(t) + \text{FP}(t)},$$

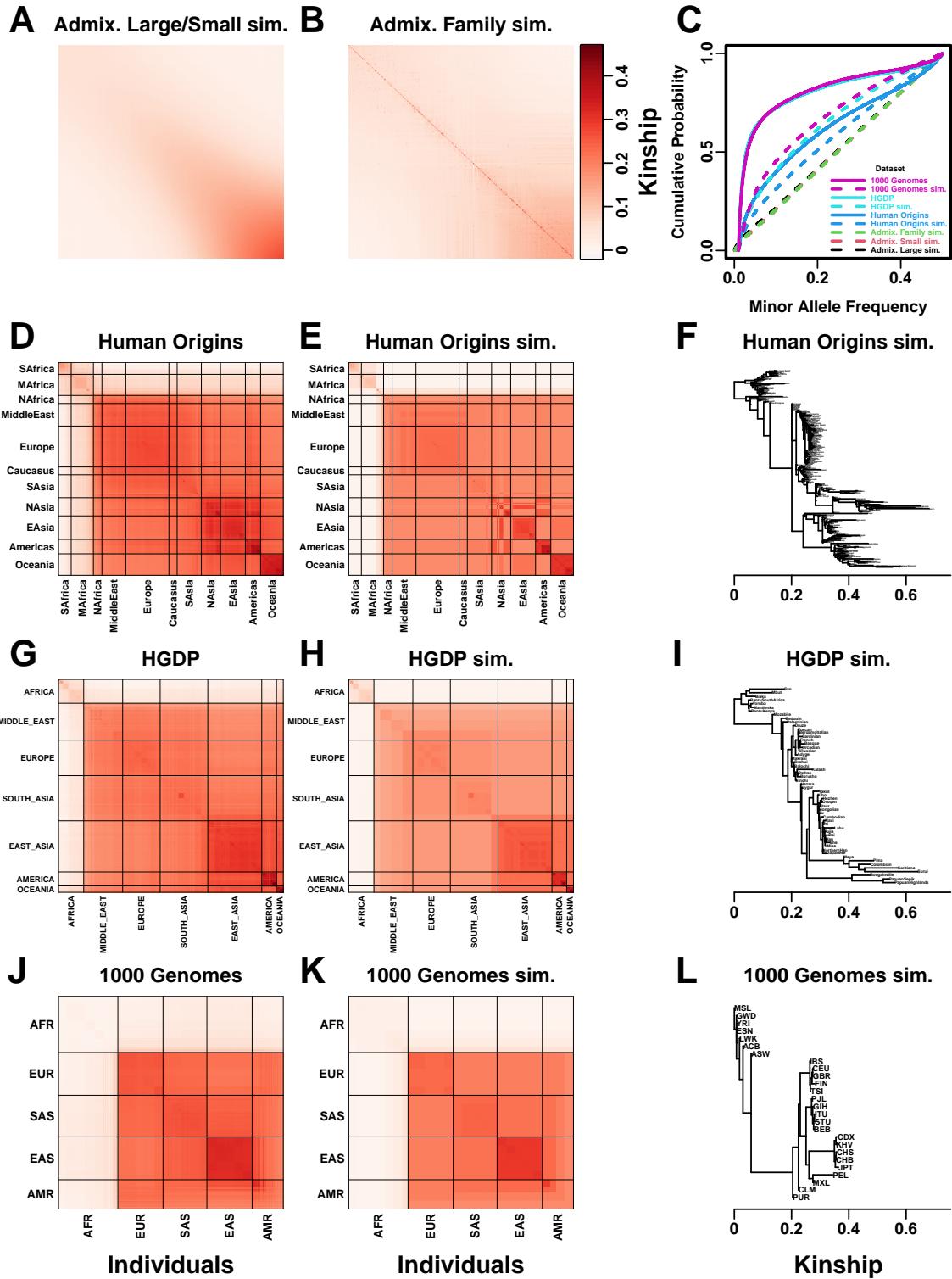
$$372 \quad \text{Recall}(t) = \frac{\text{TP}(t)}{\text{TP}(t) + \text{FN}(t)}.$$

373    Precision and Recall trace a curve as  $t$  is varied, and the area under this curve is AUC<sub>PR</sub>. We use the 374    R package PRROC to integrate the correct non-linear piecewise function when interpolating between 375    points. A model obtains the maximum AUC<sub>PR</sub> = 1 if there is a  $t$  that classifies all loci perfectly. In 376    contrast, the worst models, which classify at random, have an expected precision (= AUC<sub>PR</sub>) equal 377    to the overall proportion of causal loci:  $m_1/m$ .

## 378    **Results**

### 379    **Overview of evaluations**

380    We use three real genotype datasets and simulated genotypes from six population structure sce- 381    narios to cover various features of interest (*Table 3*). We introduce them in sets of three, as they 382    appear in the rest of our results. Population kinship matrices, which combine population and fam- 383    ily relatedness, are estimated without bias using `popkin` (*Ochoa and Storey, 2021*) (*Figure 1*). The



**Figure 1.** Population structures of simulated and real human genotype datasets. First two columns are population kinship matrices as heatmaps: individuals along x- and y-axis, kinship as color. Diagonal shows inbreeding values. **A.** Admixture scenario for both Large and Small simulations. **B.** Last generation of 20-generation admixed family, shows larger kinship values near diagonal corresponding to siblings, first cousins, etc. **C.** Minor allele frequency (MAF) distributions. Real datasets and subpopulation tree simulations had MAF  $\geq 0.01$  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.

384 first set of three simulated genotypes are based on an admixture model with 10 ancestries (*Fig-*  
385 *ure 1A*) (*Ochoa and Storey, 2021; Gopalan et al., 2016; Cabreros and Storey, 2019*). The “large”  
386 version (1000 individuals) illustrates asymptotic performance, while the “small” simulation (100 in-  
387 dividuals) illustrates model overfitting. The “family” simulation has admixed founders and draws  
388 a 20-generation random pedigree with assortative mating, resulting in a complex joint family and  
389 ancestry structure in the last generation (*Figure 1B*). The second set of three are the real human  
390 datasets representing global human diversity: Human Origins (*Figure 1D*), HGDP (*Figure 1G*), and  
391 1000 Genomes (*Figure 1J*), which are enriched for small minor allele frequencies even after MAF <  
392 1% filter (*Figure 1C*). Last are subpopulation tree simulations (*Figure 1F,I,L*) fit to the kinship (*Fig-*  
393 *ure 1E,H,K*) and MAF (*Figure 1C*) of each real human dataset, which by design do not have family  
394 structure.

395 All traits in this work are simulated. We repeated all evaluations on two additive quantitative  
396 trait models, *fixed effect sizes* (FES) and *random coefficients* (RC), which differ in how causal coeffi-  
397 cients are constructed. The FES model captures the rough inverse relationship between coefficient  
398 and minor allele frequency that arises under strong negative and balancing selection and has been  
399 observed in numerous diseases and other traits (*Park et al., 2011; Zeng et al., 2018; Simons et al.,*  
400 *2018; O’Connor et al., 2019*), so it is the focus of our results. The RC model draws coefficients  
401 independent of allele frequency, corresponding to neutral traits (*Zeng et al., 2018; Simons et al.,*  
402 *2018*), which results in a wider effect size distribution that reduces association power and effective  
403 polygenicity compared to FES.

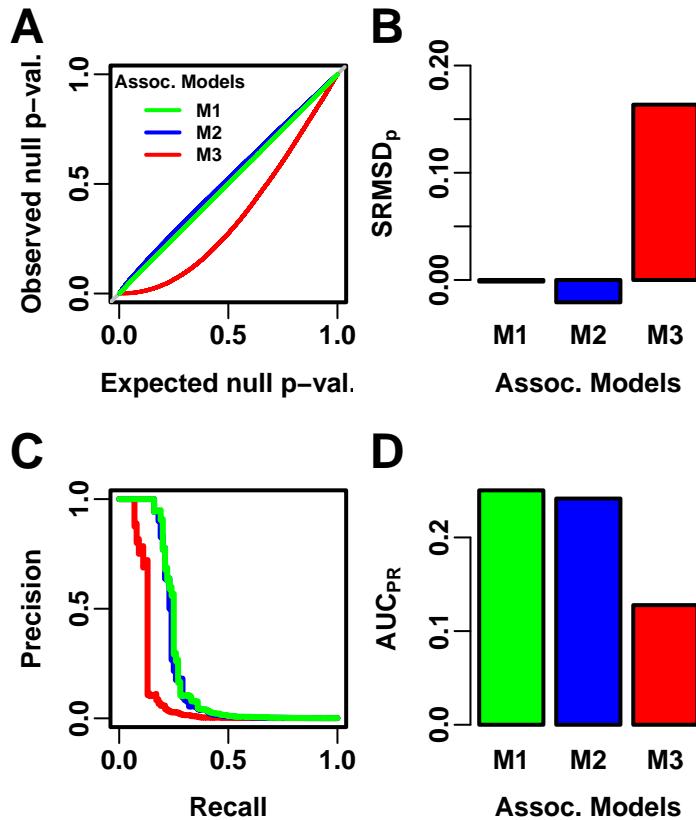
404 We evaluate using two complementary measures: (1) SRMSD<sub>p</sub> (p-value signed root mean square  
405 deviation) measures p-value calibration (closer to zero is better), and (2) AUC<sub>PR</sub> (precision-recall  
406 area under the curve) measures causal locus classification performance (higher is better; *Figure 2*).  
407 SRMSD<sub>p</sub> is a more robust alternative to the common inflation factor  $\lambda$  and type I error control  
408 measures; there is a correspondence between  $\lambda$  and SRMSD<sub>p</sub>, with SRMSD<sub>p</sub> > 0.01 giving  $\lambda$  > 1.06  
409 (*Figure 2—figure Supplement 1*) and thus evidence of miscalibration close to the rule of thumb of  
410  $\lambda$  > 1.05 (*Price et al., 2010*). There is also a monotonic correspondence between SRMSD<sub>p</sub> and type  
411 I error rate (*Figure 2—figure Supplement 2*). AUC<sub>PR</sub> has been used to evaluate association models  
412 (*Rakitsch et al., 2013*), and reflects calibrated statistical power (*Figure 2—figure Supplement 3*)  
413 while being robust to miscalibrated models (*Appendix 2*).

414 Both PCA and LMM are evaluated in each replicate dataset including a number of PCs  $r$  between  
415 0 and 90 as fixed covariates. In terms of p-value calibration, for PCA the best number of PCs  $r$  (mini-  
416 mizing mean |SRMSD<sub>p</sub>| over replicates) is typically large across all datasets (*Table 5*), although much  
417 smaller  $r$  values often performed as well (shown in following sections). Most cases have a mean  
418 |SRMSD<sub>p</sub>| < 0.01, whose p-values are effectively calibrated. However, PCA is often miscalibrated  
419 on the family simulation and real datasets (*Table 5*). In contrast, for LMM,  $r = 0$  (no PCs) is always  
420 best, and is always calibrated. Comparing LMM with  $r = 0$  to PCA with its best  $r$ , LMM always has  
421 significantly smaller |SRMSD<sub>p</sub>| than PCA or is statistically tied. For AUC<sub>PR</sub> and PCA, the best  $r$  is al-  
422 ways smaller than the best  $r$  for |SRMSD<sub>p</sub>|, so there is often a tradeoff between calibrated p-values  
423 versus classification performance. For LMM there is no tradeoff, as  $r = 0$  often has the best mean  
424 AUC<sub>PR</sub>, and otherwise is not significantly different from the best  $r$ . Lastly, LMM with  $r = 0$  always  
425 has significantly greater or statistically tied AUC<sub>PR</sub> than PCA with its best  $r$ .

#### 426 **Evaluations in admixture simulations**

427 Now we look more closely at results per dataset. The complete SRMSD<sub>p</sub> and AUC<sub>PR</sub> distributions for  
428 the admixture simulations and FES traits are in *Figure 3*. RC traits gave qualitatively similar results  
429 (*Figure 3—figure Supplement 1*).

430 In the large admixture simulation, the SRMSD<sub>p</sub> of PCA is largest when  $r = 0$  (no PCs) and de-  
431 creases rapidly to near zero at  $r = 3$ , where it stays for up to  $r = 90$  (*Figure 3A*). Thus, PCA has cali-  
432 brated p-values for  $r \geq 3$ , smaller than the theoretical optimum for this simulation of  $r = K - 1 = 9$ .  
433 In contrast, the SRMSD<sub>p</sub> for LMM starts near zero for  $r = 0$ , but becomes negative as  $r$  increases (p-



**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).

**Figure 2—figure supplement 1.** Comparison between SRMSD<sub>p</sub> and inflation factor.

**Figure 2—figure supplement 2.** Comparison between SRMSD<sub>p</sub> and type I error rate.

**Figure 2—figure supplement 3.** Comparison between AUC<sub>PR</sub> and calibrated power.

**Table 5.** Overview of PCA and LMM evaluations for high heritability simulations.

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

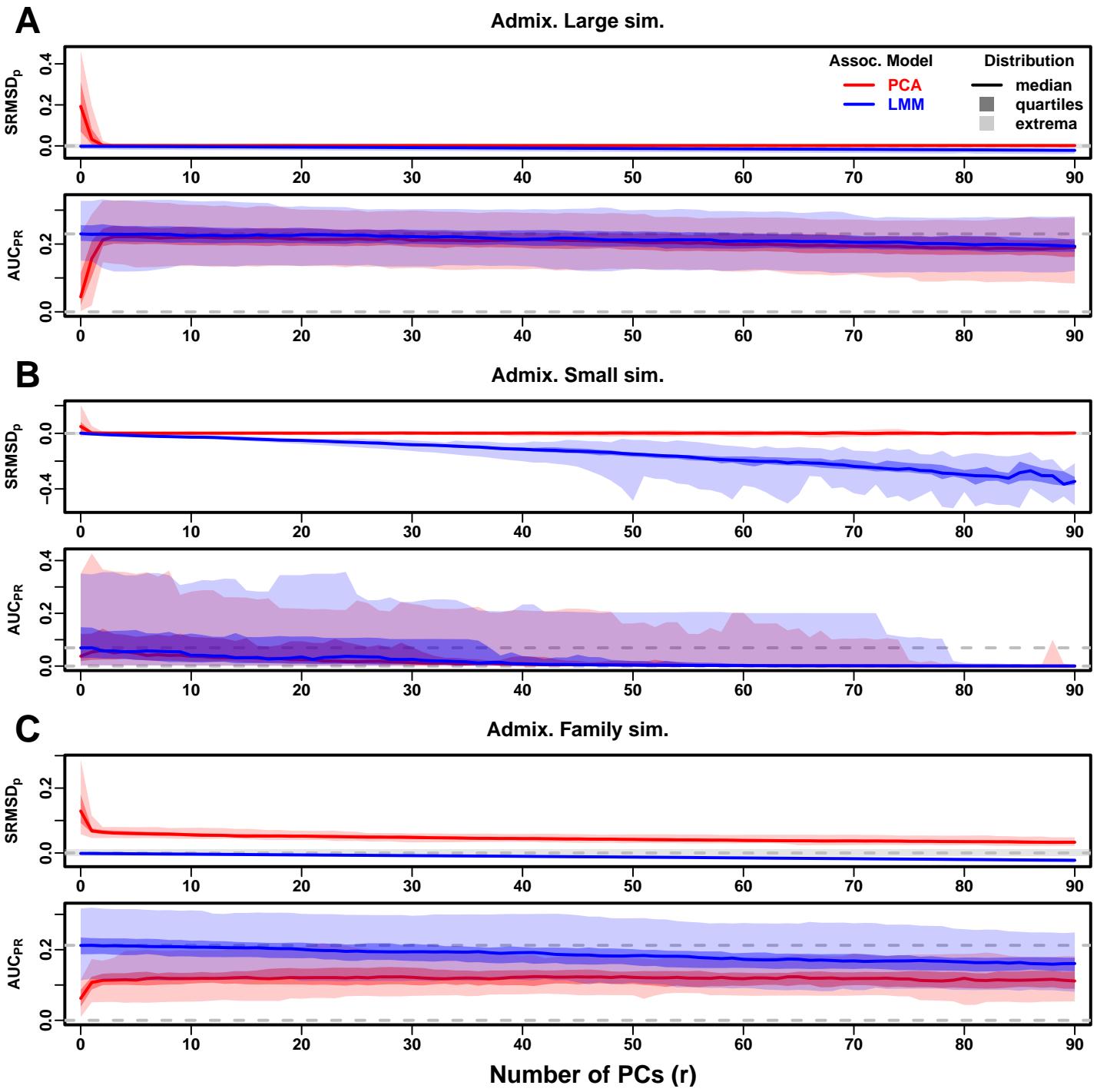
<sup>a</sup>FES: Fixed Effect Sizes, RC: Random Coefficients.

<sup>b</sup>Calibrated: whether mean  $|\text{SRMSD}_p| < 0.01$ .

<sup>c</sup>Value of  $r$  (number of PCs) with minimum mean  $|\text{SRMSD}_p|$  or maximum mean  $AUC_{\text{PR}}$ .

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

<sup>e</sup>Tie if no significant difference using Bonferroni threshold.



**Figure 3.** Evaluations in admixture simulations with FES traits, high heritability. PCA and LMM models have varying number of PCs ( $r \in \{0, \dots, 90\}$ ) on x-axis, with the distributions (y-axis) of SRMSD<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  $|SRMSD_p| < 0.01$  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.

**Figure 3—figure supplement 1.** Evaluations in admixture simulations with RC traits, high heritability.

**Figure 3—figure supplement 2.** Evaluations in admixture simulations with FES traits, low heritability.

**Figure 3—figure supplement 3.** Evaluations in admixture simulations with RC traits, low heritability.

**Figure 3—figure supplement 4.** Evaluations in admixture simulations with FES traits, environment.

**Figure 3—figure supplement 5.** Evaluations in admixture simulations with RC traits, environment.

434 values are conservative). The  $AUC_{PR}$  distribution of PCA is similarly worst at  $r = 0$ , increases rapidly  
435 and peaks at  $r = 3$ , then decreases slowly for  $r > 3$ , while the  $AUC_{PR}$  distribution for LMM starts  
436 near its maximum at  $r = 0$  and decreases with  $r$ . Although the  $AUC_{PR}$  distributions for LMM and  
437 PCA overlap considerably at each  $r$ , LMM with  $r = 0$  has significantly greater  $AUC_{PR}$  values than PCA  
438 with  $r = 3$  (*Table 5*). However, qualitatively PCA performs nearly as well as LMM in this simulation.

439 The observed robustness to large  $r$  led us to consider smaller sample sizes. A model with large  
440 numbers of parameters  $r$  should overfit more as  $r$  approaches the sample size  $n$ . Rather than  
441 increase  $r$  beyond 90, we reduce individuals to  $n = 100$ , which is small for typical association studies  
442 but may occur in studies of rare diseases, pilot studies, or other constraints. To compensate for  
443 the loss of power due to reducing  $n$ , we also reduce the number of causal loci (see Trait Simulation),  
444 which increases per-locus effect sizes. We found a large decrease in performance for both models  
445 as  $r$  increases, and best performance for  $r = 1$  for PCA and  $r = 0$  for LMM (*Figure 3B*). Remarkably,  
446 LMM attains much larger negative  $SRMSD_p$  values than in our other evaluations. LMM with  $r = 0$  is  
447 significantly better than PCA ( $r = 1$  to 4) in both measures (*Table 5*), but qualitatively the difference  
448 is negligible.

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

#### 457 Evaluations in real human genotype datasets

458 Next we repeat our evaluations with real human genotype data, which differs from our simula-  
459 tions in allele frequency distributions and more complex population structures with greater  $F_{ST}$ ,  
460 numerous correlated subpopulations, and potential cryptic family relatedness.

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

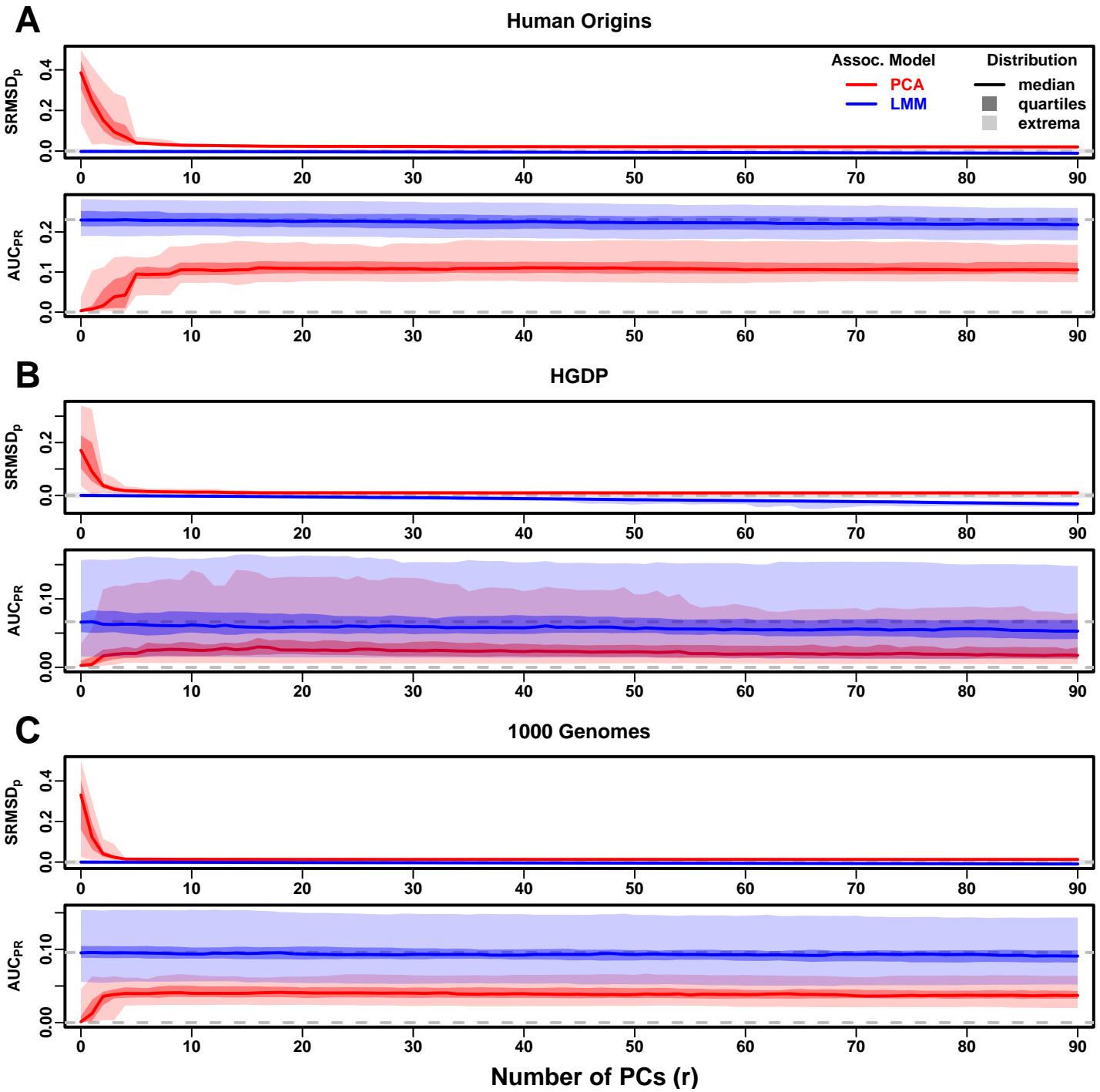
466 HGDP has the fewest individuals among real datasets, but compared to Human Origins contains  
467 more loci and low-frequency variants. Performance (*Figure 4B*) again most resembled the family  
468 simulations. In particular, LMM with  $r = 0$  achieves mean  $|SRMSD_p| < 0.01$  (p-values are calibrated),  
469 while PCA does not, and there is a sizable  $AUC_{PR}$  gap between LMM and PCA. Maximum  $AUC_{PR}$   
470 values were lowest in HGDP compared to the two other real datasets.

471 1000 Genomes has the fewest subpopulations but largest number of individuals per subpopula-  
472 tion. Thus, although this dataset has the simplest subpopulation structure among the real datasets,  
473 we find  $SRMSD_p$  and  $AUC_{PR}$  distributions (*Figure 4C*) that again most resemble our earlier family  
474 simulation, with mean  $|SRMSD_p| < 0.01$  for LMM only and large  $AUC_{PR}$  gaps between LMM and PCA.

475 Our results are qualitatively different for RC traits, which had smaller  $AUC_{PR}$  gaps between LMM  
476 and PCA (*Figure 4—figure Supplement 1*). Maximum  $AUC_{PR}$  were smaller in RC compared to FES  
477 in Human Origins and 1000 Genomes, suggesting lower power for RC traits across association  
478 models. Nevertheless, LMM with  $r = 0$  was significantly better than PCA for all measures in the real  
479 datasets and RC traits (*Table 5*).

#### 480 Evaluations in subpopulation tree simulations fit to human data

481 To better understand which features of the real datasets lead to the large differences in perfor-  
482 mance between LMM and PCA, we carried out subpopulation tree simulations. Human subpopula-



**Figure 4.** Evaluations in real human genotype datasets with FES traits, high heritability. Same setup as *Figure 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.

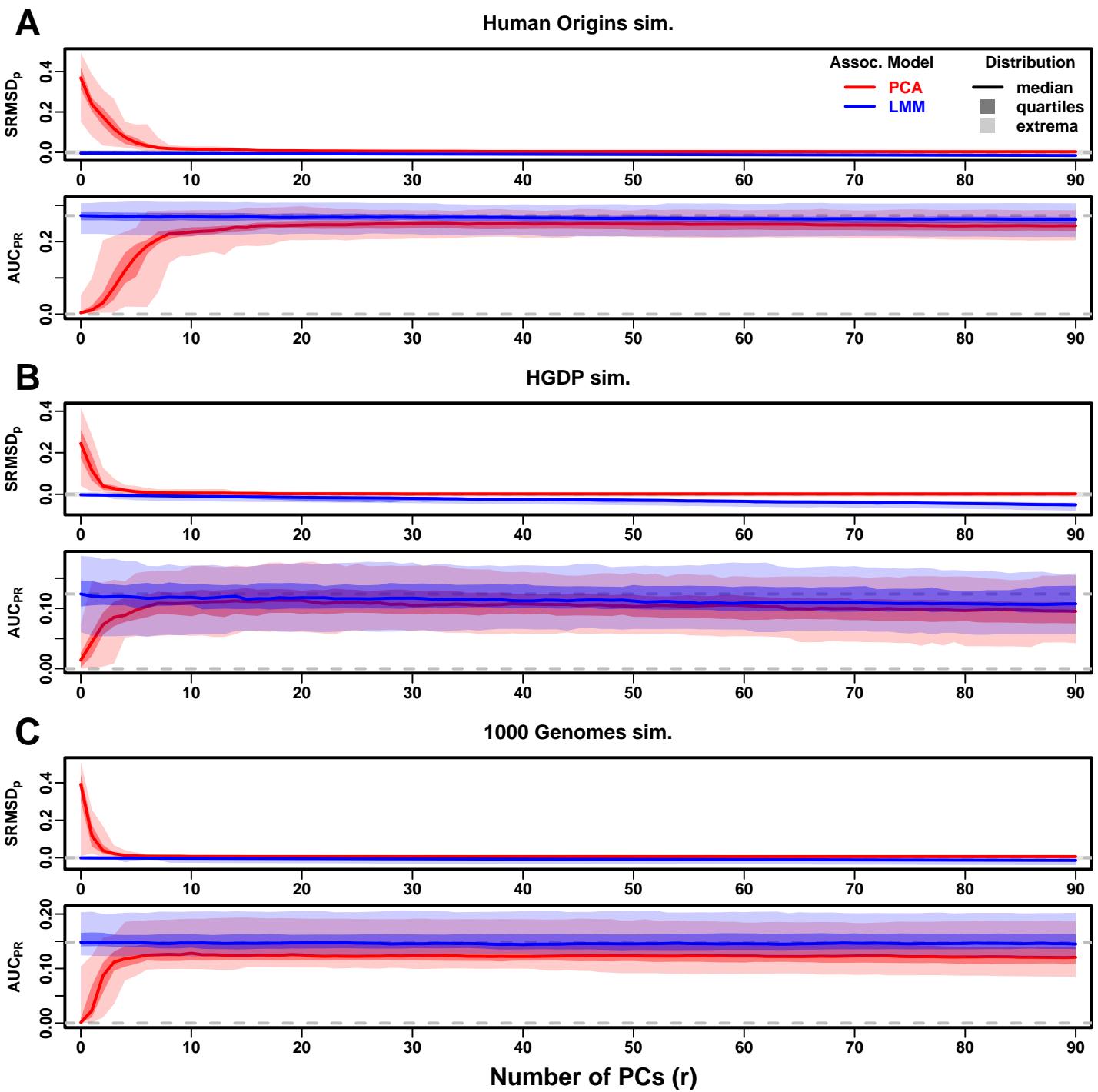
**Figure 4—figure supplement 1.** Evaluations in real human genotype datasets with RC traits, high heritability.

**Figure 4—figure supplement 2.** Evaluations in real human genotype datasets with FES traits, low heritability.

**Figure 4—figure supplement 3.** Evaluations in real human genotype datasets with RC traits, low heritability.

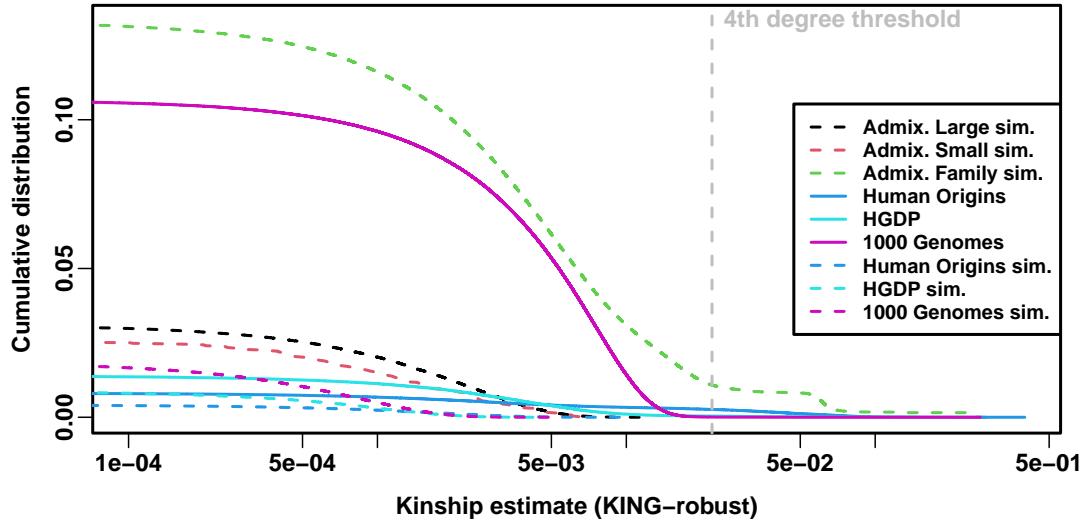
**Figure 4—figure supplement 4.** Evaluations in real human genotype datasets with FES traits, environment.

**Figure 4—figure supplement 5.** Evaluations in real human genotype datasets with RC traits, environment.



**Figure 5.** Evaluations in subpopulation tree simulations fit to human data with FES traits, high heritability. Same setup as *Figure 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.

**Figure 5—figure supplement 1.** Evaluations in subpopulation tree simulations fit to human data with RC traits, high heritability.



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

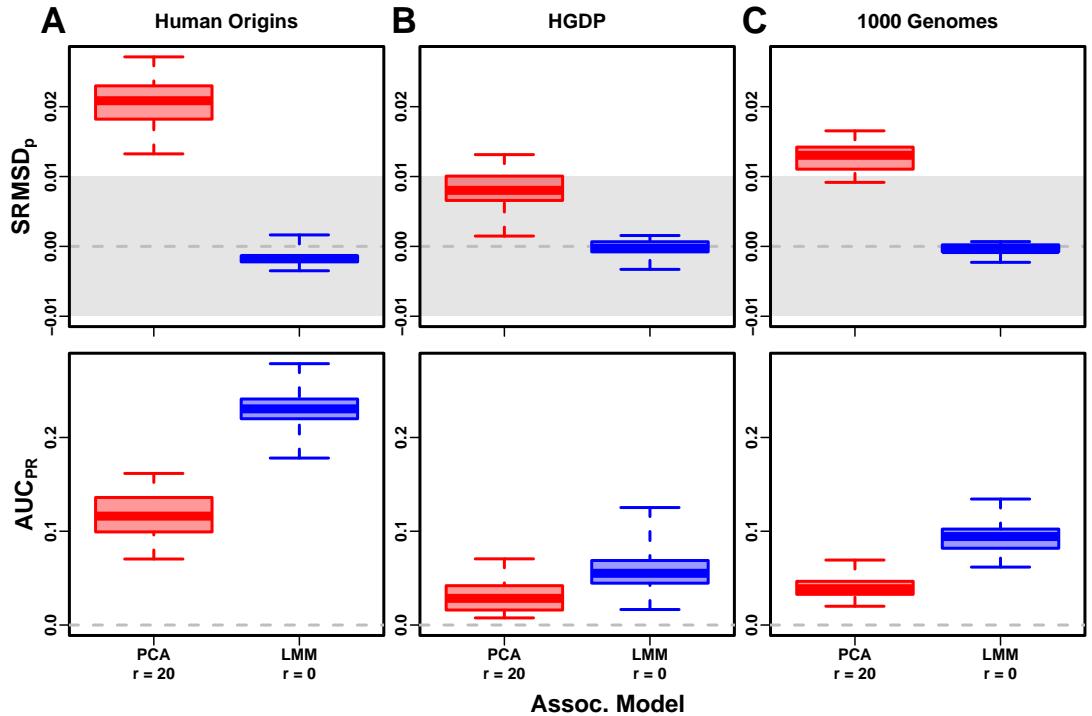
**Figure 6—figure supplement 1.** Estimated relatedness dimensions of datasets.

**Figure 6—figure supplement 2.** Number of PCs significantly associated with traits.

483 tions are related roughly by trees, which induce the strongest correlations, so we fit trees to each  
 484 real dataset and tested if data simulated from these complex tree structures could recapitulate  
 485 our previous results (*Figure 1*). These tree simulations also feature non-uniform ancestral allele  
 486 frequency distributions, which recapitulated some of the skew for smaller minor allele frequencies  
 487 of the real datasets (*Figure 1C*). The SRMSD<sub>p</sub> and AUC<sub>PR</sub> distributions for these tree simulations  
 488 (*Figure 5*) resembled our admixture simulation more than either the family simulation (*Figure 3*) or  
 489 real data results (*Figure 4*). Both LMM with  $r = 0$  and PCA (various  $r$ ) achieve mean  $|\text{SRMSD}_p| < 0.01$   
 490 (*Table 5*). The AUC<sub>PR</sub> distributions of both LMM and PCA track closely as  $r$  is varied, although there  
 491 is a small gap resulting in LMM ( $r = 0$ ) besting PCA in all three simulations. The results are qualita-  
 492 tively similar for RC traits (*Figure 5—figure Supplement 1, Table 5*). Overall, these subpopulation  
 493 tree simulations do not recapitulate the large LMM advantage over PCA observed on the real data.

#### 494 **Numerous distant relatives explain poor PCA performance in real data**

495 In principle, PCA performance should be determined by the dimension of relatedness, or kinship  
 496 matrix rank, since PCA is a low-dimensional model whereas LMM can model high-dimensional re-  
 497 latedness without overfitting. We used the Tracy-Widom test (*Patterson et al., 2006*) with  $p < 0.01$   
 498 to estimate kinship matrix rank as the number of significant PCs (*Figure 6—figure Supplement 1A*).  
 499 The true rank of our simulations is slightly underestimated (*Table 3*), but we confirm that the family  
 500 simulation has the greatest rank, and real datasets have greater estimates than their respective  
 501 subpopulation tree simulations, which confirms our hypothesis to some extent. However, esti-  
 502 mated ranks do not separate real datasets from tree simulations, as required to predict the ob-  
 503 served PCA performance. Moreover, the HGDP and 1000 Genomes rank estimates are 45 and 61,  
 504 respectively, yet PCA performed poorly for all  $r \leq 90$  numbers of PCs (*Figure 4*). The top eigen-  
 505 value explained a proportion of variance proportional to  $F_{ST}$  (*Table 3*), but the rest of the top 10  
 506 eigenvalues show no clear differences between datasets, except the small simulation had larger  
 507 variances explained per eigenvalue (expected since it has fewer eigenvalues; *Figure 6—figure Sup-*  
 508 *plement 1C*). Comparing cumulative variance explained versus rank fraction across all eigenvalues,  
 509 all datasets increase from their starting point almost linearly until they reach 1, except the family



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

**Figure 7—figure supplement 1.** Evaluation in real datasets excluding 4th degree relatives, FES traits, low heritability.

simulation has much greater variance explained by mid-rank eigenvalues (*Figure 6—figure Supplement 1B*). We also calculated the number of PCs that are significantly associated with the trait, and observed similar results, namely that while the family simulation has more significant PCs than the non-family admixture simulations, the real datasets and their tree simulated counterparts have similar numbers of significant PCs (*Figure 6—figure Supplement 2*). Overall, there is no separation between real datasets (where PCA performed poorly) and subpopulation tree simulations (where PCA performed relatively well) in terms of their eigenvalues or kinship matrix rank estimates.

Local kinship, which is recent relatedness due to family structure excluding population structure, is the presumed cause of the LMM to PCA performance gap observed in real datasets but not their subpopulation tree simulation counterparts. Instead of inferring local kinship through increased kinship matrix rank, as attempted in the last paragraph, now we measure it directly using the KING-robust estimator (*Manichaikul et al., 2010*). We observe more large local kinship in the real datasets and the family simulation compared to the other simulations (*Figure 6*). However, for real data this distribution depends on the subpopulation structure, since locally related pairs are most likely in the same subpopulation. Therefore, the only comparable curve to each real dataset is their corresponding subpopulation tree simulation, which matches subpopulation structure. In all real datasets we identified highly related individual pairs with kinship above the 4th degree relative threshold of 0.022 (*Manichaikul et al., 2010; Conomos et al., 2016b*). However, these highly related pairs are vastly outnumbered by more distant pairs with evident non-zero local kinship as compared to the extreme tree simulation values.

To try to improve PCA performance, we followed the standard practice of removing 4th degree relatives, which reduced sample sizes between 5% and 10% (*Table 4*). Only  $r = 0$  for LMM and  $r = 20$  for PCA were tested, as these performed well in our earlier evaluation, and only FES traits were tested because they previously displayed the large PCA-LMM performance gap. LMM significantly

**Table 6.** Overview of PCA and LMM evaluations for low heritability simulations

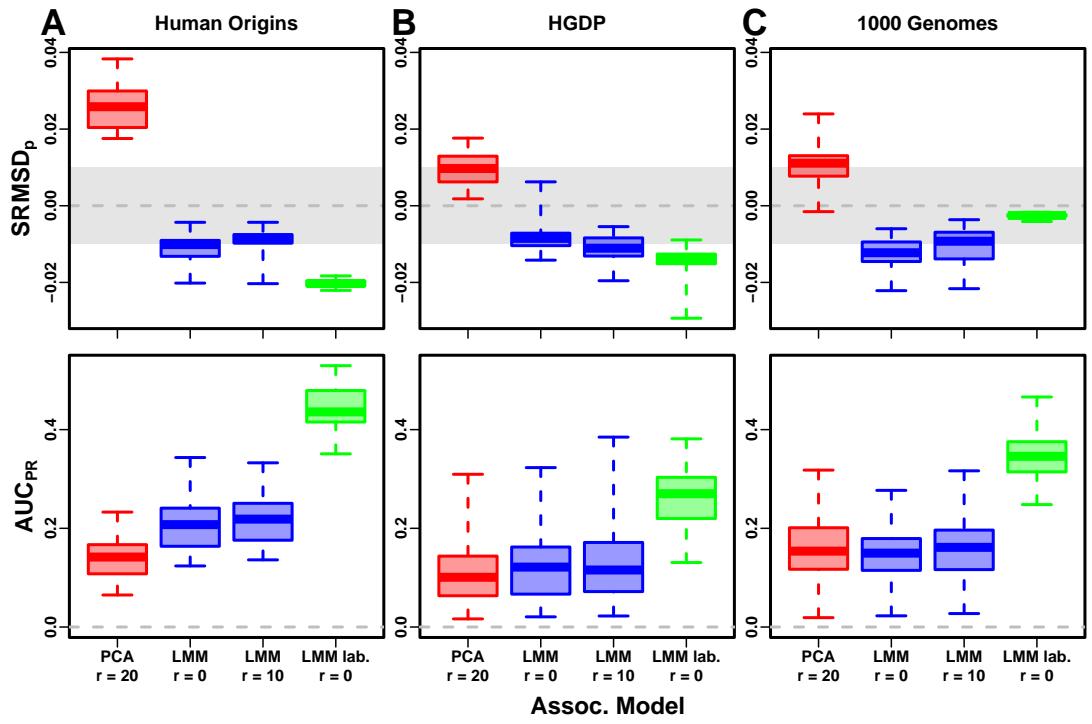
Dataset	Metric	Trait <sup>a</sup>	LMM $r = 0$ vs best $r$			PCA vs LMM $r = 0$			
			Cal. <sup>b</sup>	Best $r^c$	P-value <sup>d</sup>	Best $r^c$	Cal. <sup>b</sup>	P-value <sup>d</sup>	Best model <sup>e</sup>
Admix. Large sim.	$ \text{SRMSD}_p $	FES	True	0 1		62	True	0.00012*	LMM
Admix. Small sim.	$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

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

534 outperforms PCA in all these cases (Wilcoxon paired 1-tailed  $p < 0.01$ ; **Figure 7**). Notably, PCA still  
 535 had miscalibrated p-values two of the three real datasets ( $|\text{SRMSD}_p| > 0.01$ ), the only marginally  
 536 calibrated case being HGDP which is also the smallest of these datasets. Otherwise,  $AUC_{\text{PR}}$  and  
 537  $|\text{SRMSD}_p|$  ranges were similar here as in our earlier evaluation. Therefore, the removal of the small  
 538 number of highly related individual pairs had a negligible effect in PCA performance, so the larger  
 539 number of more distantly related pairs explain the poor PCA performance in the real datasets.

#### 540 Low heritability and environment simulations

541 Our main evaluations were repeated with traits simulated under a lower heritability value of  $h^2 =$   
 542 0.3. We reduced the number of causal loci in response to this change in heritability, to result in  
 543 equal average effect size per locus compared to the previous high heritability evaluations (see  
 544 Trait Simulation). Despite that, these low heritability evaluations measured lower  $AUC_{\text{PR}}$  values  
 545 than their high heritability counterparts (**Figure 3—figure Supplement 2**, **Figure 3—figure Supple-**  
 546 **ment 3**, **Figure 4—figure Supplement 2**, **Figure 4—figure Supplement 3**, **Figure 7—figure Supple-**  
 547 **ment 1**). The gap between LMM and PCA was reduced in these evaluations, but the main conclu-



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

**Figure 8—figure supplement 1.** Comparison of performance in low heritability vs environment simulations.

548 sion of the high heritability evaluation holds for low heritability as well, namely that LMM with  $r = 0$   
 549 significantly outperforms or ties LMM with  $r > 0$  and PCA in all cases (*Table 6*).

550 Lastly, we simulated traits with both low heritability and large environment effects determined  
 551 by geography and subpopulation labels, so they are strongly correlated to the low-dimensional  
 552 population structure (*Table 2*). For that reason, PCs may be expected to perform better in this set-  
 553 ting (in either PCA or LMM). However, we find that both PCA and LMM (even without PCs) increase  
 554 their  $AUC_{PR}$  values compared to the low-heritability evaluations (*Figure 8—figure Supplement 1*;  
 555 *Figure 8* also shows representative numbers of PCs, which performed optimally or nearly so in  
 556 individual simulations shown in *Figure 3—figure Supplement 4*, *Figure 3—figure Supplement 5*,  
 557 *Figure 4—figure Supplement 4*, *Figure 4—figure Supplement 5*). P-value calibration is comparable  
 558 with or without environment effects, for LMM for all  $r$  and for PCA once  $r$  is large enough (*Figure 8—*  
 559 *figure Supplement 1*). These simulations are the only where we occasionally observed for both  
 560 metrics a significant, though small, advantage of LMM with PCs versus LMM without PCs (*Table 7*).  
 561 Additionally, on RC traits only, PCA significantly outperforms LMM in the three real human datasets  
 562 (*Table 7*), the only cases in all of our evaluations where this is observed. For comparison, we also  
 563 evaluate an “oracle” LMM without PCs but with the finest group labels, the same used to simulate  
 564 environment, as fixed categorical covariates (“LMM lab.”), and see much larger  $AUC_{PR}$  values than  
 565 either LMM with PCs or PCA (*Figure 8*, *Figure 3—figure Supplement 4*, *Figure 3—figure Supple-*  
 566 *ment 5*, *Figure 4—figure Supplement 4*, *Figure 4—figure Supplement 5*, *Table 7*). However, LMM  
 567 with labels is often more poorly calibrated than LMM or PCA without labels, which may be since  
 568 these numerous labels are inappropriately modeled as fixed rather than random effects. Overall,  
 569 we find that association studies with correlated environment and genetic effects remain a chal-  
 570 lenge for PCA and LMM, that addition of PCs to an LMM improves performance only marginally,  
 571 and that if the environment effect is driven by geography or ethnicity then use of those labels

**Table 7.** Overview of PCA and LMM evaluations for environment simulations

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

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

572 greatly improves performance compared to using PCs.

## 573 Discussion

574 Our evaluations conclusively determined that LMM without PCs performs better than PCA (for any  
 575 number of PCs) across all scenarios without environment effects, including all real and simulated  
 576 genotypes and two trait simulation models. Although the addition of a few PCs to LMM does not  
 577 greatly hurt its performance (except for small sample sizes), they generally did not improve it either  
 578 (**Table 5**, **Table 6**), which agrees with previous observations (*Liu et al., 2011; Janss et al., 2012*) but  
 579 contradicts others (*Zhao et al., 2007; Price et al., 2010*). Our findings make sense since PCs are the  
 580 eigenvectors of the same kinship matrix that parameterized random effects, so including both is  
 581 redundant.

582 The presence of environment effects that are correlated to relatedness presents the only sce-  
 583 nario where occasionally PCA and LMM with PCs outperform LMM without PCs (**Table 7**). It is  
 584 commonly believed that PCs model such environment effects well (*Novembre et al., 2008; Zhang  
 585 and Pan, 2015; Lin et al., 2021*). However, we observe that LMM without PCs models environment  
 586 effects nearly as well as with PCs (**Figure 8**), consistent with previous findings (*Vilhjálmsson and  
 587 Nordborg, 2013; Wang et al., 2022*) and with environment inflating heritability estimates using LMM  
 588 (*Heckerman et al., 2016*). Moreover, modeling the true environment groups as fixed categorical  
 589 effects always substantially improved AUC<sub>PR</sub> compared to modeling them with PCs (**Figure 8, Table 7**).  
 590 Modeling numerous environment groups as fixed effects does result in deflated p-values (**Figure 8**,

591 **Table 7**), which we expect would be avoided by modeling them as random effects, a strategy we  
592 chose not to pursue here as it is both a circular evaluation (the true effects were drawn from that  
593 model) and out of scope. Overall, including PCs to model environment effects yields limited power  
594 gains if at all, even in an LMM, and is no replacement for more adequate modeling of environment  
595 whenever possible.

596 Previous studies found that PCA was better calibrated than LMM for unusually differentiated  
597 markers (*Price et al., 2010; Wu et al., 2011; Yang et al., 2014*), which as simulated were an artificial  
598 scenario not based on a population genetics model, and are otherwise believed to be unusual (*Sul  
599 and Eskin, 2013; Price et al., 2013*). Our evaluations on real human data, which contain such loci in  
600 relevant proportions if they exist, do not replicate that result. Family relatedness strongly favors  
601 LMM, an advantage that probably outweighs this potential PCA benefit in real data.

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

610 PCA assumes that genetic relatedness is restricted to a low-dimensional subspace, whereas  
611 LMM can handle high-dimensional relatedness. Thus, PCA performs well in the admixture sim-  
612 ulation, which is explicitly low-dimensional (see Genotype simulation from the admixture model), and  
613 our subpopulation tree simulations, which are likely well approximated by a few dimensions de-  
614 spite the large number of subpopulations because there are few long branches. Conversely, PCA  
615 performs poorly under family structure because its kinship matrix is high-dimensional (**Figure 6—**  
616 **figure Supplement 1**). However, estimating the latent space dimensions of real datasets is chal-  
617 lenging because estimated eigenvalues have biased distributions (*Hayashi et al., 2018*). Kinship matrix  
618 rank estimated using the Tracy-Widom test (*Patterson et al., 2006*) did not fully predict the datasets  
619 that PCA performs well on. In contrast, estimated local kinship finds considerable cryptic family re-  
620 latedness in all real human datasets and better explains why PCA performs poorly there. The trait  
621 model also influences the relative performance of PCA, so genotype-only parameters (eigenvalues  
622 or local kinship) alone cannot tell the full story. There are related tests for numbers of dimensions  
623 that consider the trait which we did not consider, including the Bayesian information criterion for  
624 the regression with PCs against the trait (*Zhu and Yu, 2009*). Additionally, PCA and LMM goodness  
625 of fit could be compared using the coefficient of determination generalized for LMMs (*Sun et al.,  
626 2010*).

627 PCA is at best underpowered relative to LMMs, and at worst miscalibrated regardless of the  
628 numbers of PCs included, in real human genotype tests. Among our simulations, such poor per-  
629 formance occurred only in the admixed family. Local kinship estimates reveal considerable family  
630 relatedness in the real datasets absent in the corresponding subpopulation tree simulations. Ad-  
631 mixture is also absent in our tree simulations, but our simulations and theory show that admixture  
632 is well handled by PCA. Hundreds of close relative pairs have been identified in 1000 Genomes  
633 (*Gazal et al., 2015; Al-Khudhair et al., 2015; Fedorova et al., 2016; Schlauch et al., 2017*), but their  
634 removal does not improve PCA performance sufficiently in our tests, so the larger number of more  
635 distantly related pairs are PCA's most serious obstacle in practice. Distant relatives are expected to  
636 be numerous in any large human dataset (*Henn et al., 2012; Shchur and Nielsen, 2018; Loh et al.,  
637 2018*). Our FES trait tests show that family relatedness is more challenging when rarer variants  
638 have larger coefficients. Overall, the high relatedness dimensions induced by family relatedness is  
639 the key challenge for PCA association in modern datasets that is readily overcome by LMM.

640 Our tests also found PCA robust to large numbers of PCs, far beyond the optimal choice, agree-  
641 ing with previous anecdotal observations (*Price et al., 2006; Kang et al., 2010*), in contrast to using

642 too few PCs for which there is a large performance penalty. The exception was the small sample  
643 size simulation, where only small numbers of PCs performed well. In contrast, LMM is simpler  
644 since there is no need to choose the number of PCs. However, an LMM with a large number of co-  
645 variates may have conservative p-values, as observed for LMM with large numbers of PCs, which  
646 is a weakness of the score test used by the LMM we evaluated that may be overcome with other  
647 statistical tests. Simulations or post hoc evaluations remain crucial for ensuring that statistics are  
648 calibrated.

649 There are several variants of the PCA and LMM analyses, most designed for better modeling  
650 linkage disequilibrium (LD), that we did not evaluate directly, in which PCs are no longer exactly the  
651 top eigenvectors of the kinship matrix (if estimated with different approaches), although this is not  
652 a crucial aspect of our arguments. We do not consider the case where samples are projected onto  
653 PCs estimated from an external sample (*Privé et al., 2020*), which is uncommon in association  
654 studies, and whose primary effect is shrinkage, so if all samples are projected then they are all  
655 equally affected and larger regression coefficients compensate for the shrinkage, although this  
656 will no longer be the case if only a portion of the sample is projected onto the PCs of the rest of the  
657 sample. Another approach tests PCs for association against every locus in the genome in order to  
658 identify and exclude PCs that capture LD structure (which is localized) instead of ancestry (which  
659 should be present across the genome) (*Privé et al., 2020*); a previous proposal removes LD using an  
660 autocorrelation model prior to estimating PCs (*Patterson et al., 2006*). These improved PCs remain  
661 inadequate models of family relatedness, so an LMM will continue to outperform them in that  
662 setting. Similarly, the leave-one-chromosome-out (LOCO) approach for estimating kinship matrices  
663 for LMMs prevents the test locus and loci in LD with it from being modeled by the random effect as  
664 well, which is called “proximal contamination” (*Lippert et al., 2011; Yang et al., 2014*). While LOCO  
665 kinship estimates vary for each chromosome, they continue to model family relatedness, thus  
666 maintaining their key advantage over PCA. The LDAK model estimates kinship instead by weighing  
667 loci taking LD into account (*Speed et al., 2012*). LD effects must be adjusted for, if present, so  
668 in unfiltered data we advise the previous methods be applied. However, in this work, simulated  
669 genotypes do not have LD, and the real datasets were filtered to remove LD, so here there is no  
670 proximal contamination and LD confounding is minimized if present at all, so these evaluations  
671 may be considered the ideal situation where LD effects have been adjusted successfully, and in  
672 this setting LMM outperforms PCA. Overall, these alternative PCs or kinship matrices differ from  
673 their basic counterparts by either the extent to which LD influences the estimates (which may be a  
674 confounder in a small portion of the genome, by definition) or by sampling noise, neither of which  
675 are expected to change our key conclusion.

676 One of the limitations of this work include relatively small sample sizes compared to modern  
677 association studies. However, our conclusions are not expected to change with larger sample sizes,  
678 as cryptic family relatedness will continue to be abundant in such data, if not increase in abundance,  
679 and thus give LMMs an advantage over PCA (*Henn et al., 2012; Shchur and Nielsen, 2018; Loh  
680 et al., 2018*). One reason PCA has been favored over classic LMMs is because PCA’s runtime scales  
681 much better with increasing sample size. However, recent approaches not tested in this work  
682 have made LMMs more scalable and applicable to biobank-scale data (*Loh et al., 2015; Zhou et al.,  
683 2018; Mbatchou et al., 2021*), so one clear next step is carefully evaluating these approaches in  
684 simulations with larger sample sizes. A different benefit for including PCs were recently reported  
685 for BOLT-LMM, which does not result in greater power but rather in reduced runtime, a property  
686 that may be specific to its use of scalable algorithms such as conjugate gradient and variational  
687 Bayes (*Loh et al., 2018*). Many of these newer LMMs also no longer follow the infinitesimal model  
688 of the basic LMM (*Loh et al., 2015; Mbatchou et al., 2021*), and employ novel approximations, which  
689 are features not evaluated in this work and worthy of future study.

690 Another limitation of this work is ignoring rare variants, a necessity given our smaller sample  
691 sizes, where rare variant association is miscalibrated and underpowered. Using simulations mim-  
692 icking the UK Biobank, recent work has found that rare variants can have a more pronounced

693 structure than common variants, and that modeling this rare variant structure (with either PCA  
694 and LMM) may better model environment confounding, reduce inflation in association studies,  
695 and ameliorate stratification in polygenic risk scores (*Zaidi and Mathieson, 2020*). Better modeling  
696 rare variants and their structure is a key next step in association studies.

697 The largest limitation of our work is that we only considered quantitative traits. Previous evalua-  
698 tions involving case-control traits tended to report PCA-LMM ties or mixed results, an observation  
699 potentially confounded by the use of low-dimensional simulations without family relatedness (*Ta-*  
700 *ble 1*). An additional concern is case-control ascertainment bias and imbalance, which appears to  
701 affect LMMs more severely, although recent work appears to solve this problem (*Yang et al., 2014*;  
702 *Zhou et al., 2018*). Future evaluations should aim to include our simulations and real datasets, to  
703 ensure that previous results were not biased in favor of PCA by not simulating family structure or  
704 larger coefficients for rare variants that are expected for diseases by various selection models.

705 Overall, our results lead us to recommend LMM over PCA for association studies in general. Al-  
706 though PCA offer flexibility and speed compared to LMM, additional work is required to ensure that  
707 PCA is adequate, including removal of close relatives (lowering sample size and wasting resources)  
708 followed by simulations or other evaluations of statistics, and even then PCA may perform poorly  
709 in terms of both type I error control and power. The large numbers of distant relatives expected of  
710 any real dataset all but ensures that PCA will perform poorly compared to LMM (*Henn et al., 2012*;  
711 *Shchur and Nielsen, 2018; Loh et al., 2018*). Our findings also suggest that related applications  
712 such as polygenic models may enjoy gains in power and accuracy by employing an LMM instead  
713 of PCA to model relatedness (*Rakitsch et al., 2013; Qian et al., 2020*). PCA remains indispensable  
714 across population genetics, from visualizing population structure and performing quality control  
715 to its deep connection to admixture models, but the time has come to limit its use in association  
716 testing in favor of LMM or other, richer models capable of modeling all forms of relatedness.

## 717 Competing interests

718 The authors declare no competing interests.

## 719 Acknowledgments

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

## 724 Web resources

725 plink2, <https://www.cog-genomics.org/plink/2.0/>  
726 GCTA, <https://yanglab.westlake.edu.cn/software/gcta/>  
727 Eigensoft, <https://github.com/DReichLab/EIG>  
728 bnpsd, <https://cran.r-project.org/package=bnpsd>  
729 simfam, <https://cran.r-project.org/package=simfam>  
730 simtrait, <https://cran.r-project.org/package=simtrait>  
731 genio, <https://cran.r-project.org/package=genio>  
732 popkin, <https://cran.r-project.org/package=popkin>  
733 ape, <https://cran.r-project.org/package=ape>  
734 nnls, <https://cran.r-project.org/package=nnls>  
735 PRROC, <https://cran.r-project.org/package=PRROC>  
736 BEDMatrix, <https://cran.r-project.org/package=BEDMatrix>

737 **Data and code availability**

738 The data and code generated during this study are available on GitHub at [https://github.com/  
739 OchoaLab/pca-assoc-paper](https://github.com/OchoaLab/pca-assoc-paper). The public subset of Human Origins is available on the Reich Lab web-  
740 site at <https://reich.hms.harvard.edu/datasets>; non-public samples have to be requested from David  
741 Reich. The WGS version of HGDP was downloaded from the Wellcome Sanger Institute FTP site at  
742 [ftp://ngs.sanger.ac.uk/production/hgdp/hgdp\\_wgs.20190516/](ftp://ngs.sanger.ac.uk/production/hgdp/hgdp_wgs.20190516/). The high-coverage version of the 1000  
743 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/).  
744

745 **References**

- 746 **1000 Genomes Project Consortium**, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker  
747 RE, Kang HM, Marth GT, McVean GA. An integrated map of genetic variation from 1,092 human genomes.  
748 Nature. 2012 Nov; 491(7422):56–65. doi: 10.1038/nature11632.
- 749 **Abraham G**, Inouye M. Fast Principal Component Analysis of Large-Scale Genome-Wide Data. PLOS ONE. 2014  
750 Apr; 9(4):e93766. <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0093766>, doi: 10.1371/jour-  
751 nal.pone.0093766.
- 752 **Abraham G**, Qiu Y, Inouye M. FlashPCA2: principal component analysis of Biobank-scale genotype datasets.  
753 Bioinformatics. 2017 Sep; 33(17):2776–2778. <https://academic.oup.com/bioinformatics/article/33/17/2776/3798630>, doi: 10.1093/bioinformatics/btx299.
- 755 **Agrawal A**, Chiu AM, Le M, Halperin E, Sankararaman S. Scalable probabilistic PCA for large-scale genetic  
756 variation data. PLOS Genetics. 2020 May; 16(5):e1008773. <https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1008773>, doi: 10.1371/journal.pgen.1008773.
- 758 **Al-Khudhair A**, Qiu S, Wyse M, Chowdhury S, Cheng X, Bekbolsynov D, Saha-Mandal A, Dutta R, Fedorova L,  
759 Fedorov A. Inference of Distant Genetic Relations in Humans Using “1000 Genomes”. Genome Biology and  
760 Evolution. 2015 Feb; 7(2):481–492. <https://doi.org/10.1093/gbe/evv003>, doi: 10.1093/gbe/evv003.
- 761 **Alexander DH**, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals.  
762 Genome Res. 2009 Sep; 19(9):1655–1664. doi: 10.1101/gr.094052.109.
- 763 **Astle W**, Balding DJ. Population Structure and Cryptic Relatedness in Genetic Association Studies. Statist Sci.  
764 2009 Nov; 24(4):451–471. <http://projecteuclid.org/euclid.ss/1271770342>, doi: 10.1214/09-STS307.
- 765 **Aulchenko YS**, de Koning DJ, Haley C. Genomewide rapid association using mixed model and regression: a  
766 fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. Genetics.  
767 2007 Sep; 177(1):577–585. doi: 10.1534/genetics.107.075614.
- 768 **Balding DJ**, Nichols RA. A method for quantifying differentiation between populations at multi-allelic  
769 loci and its implications for investigating identity and paternity. Genetica. 1995; 96(1-2):3–12. doi:  
770 <https://doi.org/10.1007/BF01441146>.
- 771 **Bergström A**, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, Chen Y, Felkel S, Hallast P, Kamm J, Blanché  
772 H, Deleuze JF, Cann H, Mallick S, Reich D, Sandhu MS, Skoglund P, Scally A, Xue Y, Durbin R, et al. Insights into  
773 human genetic variation and population history from 929 diverse genomes. Science. 2020; 367(6484). doi:  
774 <https://doi.org/10.1126/science.aay5012>.
- 775 **Bouaziz M**, Ambroise C, Guedj M. Accounting for Population Stratification in Practice: A Comparison of the  
776 Main Strategies Dedicated to Genome-Wide Association Studies. PLOS ONE. 2011 Dec; 6(12):e28845. <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0028845>, doi: 10.1371/journal.pone.0028845.
- 778 **Cabrero I**, Storey JD. A Likelihood-Free Estimator of Population Structure Bridging Admixture Models and  
779 Principal Components Analysis. Genetics. 2019; 212(4):1009–1029. doi: 10.1534/genetics.119.302159.
- 780 **Cann HM**, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-  
781 Thomsen A, Chen Z, Chu J, Carcassi C, Contu L, Du R, Excoffier L, Ferrara GB, Friedlaender JS, Groot H, Gurwitz  
782 D, et al. A human genome diversity cell line panel. Science. 2002 Apr; 296(5566):261–262. doi: 10.1126/sci-  
783 ence.296.5566.261b.
- 784 **Chang CC**, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge  
785 of larger and richer datasets. Gigascience. 2015 Feb; 4(1):7. <http://www.gigasciencejournal.com/content/4/1/7/abstract>, doi: 10.1186/s13742-015-0047-8.

- 787 **Chiu AM**, Molloy EK, Tan Z, Talwalkar A, Sankararaman S. Inferring population structure in biobank-scale ge-  
788 nomic data. *Am J Hum Genet*. 2022 Apr; 109(4):727–737. doi: [10.1016/j.ajhg.2022.02.015](https://doi.org/10.1016/j.ajhg.2022.02.015).
- 789 **Conomos M**, Laurie C, Stilp A, Gogarten S, McHugh C, Nelson S, Sofer T, Fernández-Rhodes L, Justice A, Graff  
790 M, Young K, Seyerle A, Avery C, Taylor K, Rotter J, Talavera G, Daviglus M, Wassertheil-Smoller S, Schnei-  
791 derman N, Heiss G, et al. Genetic Diversity and Association Studies in US Hispanic/Latino Populations: Ap-  
792 plications in the Hispanic Community Health Study/Study of Latinos. *The American Journal of Human Ge-  
793 netics*. 2016 Jan; 98(1):165–184. <http://www.sciencedirect.com/science/article/pii/S0002929715004966>, doi:  
794 [10.1016/j.ajhg.2015.12.001](https://doi.org/10.1016/j.ajhg.2015.12.001).
- 795 **Conomos M**, Reiner A, Weir B, Thornton T. Model-free Estimation of Recent Genetic Relatedness. *The Ameri-  
796 can Journal of Human Genetics*. 2016 Jan; 98(1):127–148. <http://www.sciencedirect.com/science/article/pii/S0002929715004930>, doi:  
797 [10.1016/j.ajhg.2015.11.022](https://doi.org/10.1016/j.ajhg.2015.11.022).
- 798 **Consortium TGP**. A map of human genome variation from population-scale sequencing. *Nature*. 2010  
799 Oct; 467(7319):1061–1073. <http://www.nature.com/nature/journal/v467/n7319/abs/nature09534.html>, doi:  
800 [10.1038/nature09534](https://doi.org/10.1038/nature09534).
- 801 **Coram MA**, Duan Q, Hoffmann TJ, Thornton T, Knowles JW, Johnson NA, Ochs-Balcom HM, Donlon TA, Mar-  
802 tin LW, Eaton CB, Robinson JG, Risch NJ, Zhu X, Kooperberg C, Li Y, Reiner AP, Tang H. Genome-wide  
803 Characterization of Shared and Distinct Genetic Components that Influence Blood Lipid Levels in Ethni-  
804 cally Diverse Human Populations. *The American Journal of Human Genetics*. 2013 Jun; 92(6):904–916.  
805 [https://www.cell.com/ajhg/abstract/S0002-9297\(13\)00212-7](https://www.cell.com/ajhg/abstract/S0002-9297(13)00212-7), doi: [10.1016/j.ajhg.2013.04.025](https://doi.org/10.1016/j.ajhg.2013.04.025).
- 806 **Devlin B**, Roeder K. Genomic Control for Association Studies. *Biometrics*. 1999 Dec; 55(4):997–  
807 1004. <http://onlinelibrary.wiley.com/doi/10.1111/j.0006-341X.1999.00997.x/abstract>, doi: [10.1111/j.0006-341X.1999.00997.x](https://doi.org/10.1111/j.0006-341X.1999.00997.x).
- 809 **Fairley S**, Lowy-Gallego E, Perry E, Flück P. The International Genome Sample Resource (IGSR) collec-  
810 tion of open human genomic variation resources. *Nucleic Acids Res*. 2020 Jan; 48(D1):D941–D947. doi:  
811 [10.1093/nar/gkz836](https://doi.org/10.1093/nar/gkz836).
- 812 **Fedorova L**, Qiu S, Dutta R, Fedorov A. Atlas of Cryptic Genetic Relatedness Among 1000 Human  
813 Genomes. *Genome Biology and Evolution*. 2016 Mar; 8(3):777–790. <https://doi.org/10.1093/gbe/evw034>,  
814 doi: [10.1093/gbe/evw034](https://doi.org/10.1093/gbe/evw034).
- 815 **Galinsky K**, Bhatia G, Loh PR, Georgiev S, Mukherjee S, Patterson N, Price A. Fast Principal-Component Anal-  
816 ysis Reveals Convergent Evolution of ADH1B in Europe and East Asia. *The American Journal of Human Ge-  
817 netics*. 2016 Mar; 98(3):456–472. <http://www.sciencedirect.com/science/article/pii/S0002929716000033>, doi:  
818 [10.1016/j.ajhg.2015.12.022](https://doi.org/10.1016/j.ajhg.2015.12.022).
- 819 **Gazal S**, Sahbatou M, Babron MC, Génin E, Leutenegger AL. High level of inbreeding in final phase of  
820 1000 Genomes Project. *Sci Rep*. 2015 Dec; 5(1):17453. <https://doi.org/10.1038/srep17453>, doi:  
821 [10.1038/srep17453](https://doi.org/10.1038/srep17453).
- 822 **Gopalan P**, Hao W, Blei DM, Storey JD. Scaling probabilistic models of genetic variation to millions of humans.  
823 *Nat Genet*. 2016; 48(12):1587–1590. doi: [10.1038/ng.3710](https://doi.org/10.1038/ng.3710).
- 824 **Grau J**, Grosse I, Keilwagen J. PRRoc: computing and visualizing precision-recall and receiver operating char-  
825 acteristic curves in R. *Bioinformatics*. 2015 Aug; 31(15):2595–2597. <https://academic.oup.com/bioinformatics/article/31/15/2595/187781>, doi: [10.1093/bioinformatics/btv153](https://doi.org/10.1093/bioinformatics/btv153).
- 827 **Grueneberg A**, Campos Gdl. BGData - A Suite of R Packages for Genomic Analysis with Big Data. *G3:*  
828 *Genes, Genomes, Genetics*. 2019 May; 9(5):1377–1383. <https://doi.org/10.1534/g3.119.400018>, doi: [10.1534/g3.119.400018](https://doi.org/10.1534/g3.119.400018).
- 830 **Hayashi K**, Yuan KH, Liang L. On the Bias in Eigenvalues of Sample Covariance Matrix. In: Wiberg M, Culpepper  
831 S, Janssen R, González J, Molenaar D, editors. *Quantitative Psychology Springer Proceedings in Mathematics  
832 & Statistics*, Cham: Springer International Publishing; 2018. p. 221–233. doi: [10.1007/978-3-319-77249-3\\_19](https://doi.org/10.1007/978-3-319-77249-3_19).
- 833 **Heckerman D**, Gurdasani D, Kadie C, Pomilla C, Carstensen T, Martin H, Ekoru K, Nsubuga RN, Ssenyomo  
834 G, Kamali A, Kaleebu P, Widmer C, Sandhu MS. Linear mixed model for heritability estimation that ex-  
835 plicitly addresses environmental variation. *Proc Natl Acad Sci USA*. 2016 Jul; 113(27):7377–7382. doi:  
836 [10.1073/pnas.1510497113](https://doi.org/10.1073/pnas.1510497113).

- 837 Henn BM, Hon L, Macpherson JM, Eriksson N, Saxonov S, Pe'er I, Mountain JL. Cryptic Distant Relatives Are  
838 Common in Both Isolated and Cosmopolitan Genetic Samples. PLOS ONE. 2012 Apr; 7(4):e34267. <https://doi.org/10.1371/journal.pone.0034267>.
- 840 Hindorff LA, Bonham VL, Brody LC, Ginoza MEC, Hutter CM, Manolio TA, Green ED. Prioritizing diversity in  
841 human genomics research. Nature Reviews Genetics. 2018 Mar; 19(3):175–185. <https://doi.org/10.1038/nrg.2017.89>.
- 843 Hodonsky CJ, Jain D, Schick UM, Morrison JV, Brown L, McHugh CP, Schurmann C, Chen DD, Liu YM, Auer  
844 PL, Laurie CA, Taylor KD, Browning BL, Li Y, Papanicolaou G, Rotter JI, Kurita R, Nakamura Y, Browning SR,  
845 Loos RJF, et al. Genome-wide association study of red blood cell traits in Hispanics/Latinos: The Hispanic  
846 Community Health Study/Study of Latinos. PLOS Genetics. 2017 Apr; 13(4):e1006760. [http://doi.org/10.1371/journal.pgen.1006760](https://doi.org/10.1371/journal.pgen.1006760).
- 848 Hoffman GE. Correcting for population structure and kinship using the linear mixed model: theory and exten-  
849 sions. PLoS ONE. 2013; 8(10):e75707. doi: [10.1371/journal.pone.0075707](https://doi.org/10.1371/journal.pone.0075707).
- 850 Hoffmann TJ, Choquet H, Yin J, Banda Y, Kvale MN, Glymour M, Schaefer C, Risch N, Jorgenson E. A Large  
851 Multiethnic Genome-Wide Association Study of Adult Body Mass Index Identifies Novel Loci. Genetics. 2018  
852 Oct; 210(2):499–515. [http://doi.org/10.1534/genetics.118.301479](https://doi.org/10.1534/genetics.118.301479).
- 853 Hou K, Ding Y, Xu Z, Wu Y, Bhattacharya A, Mester R, Belbin GM, Buyske S, Conti DV, Darst BF, Fornage M,  
854 Gignoux C, Guo X, Haiman C, Kenny EE, Kim M, Kooperberg C, Lange L, Manichaikul A, North KE, et al.  
855 Causal effects on complex traits are similar for common variants across segments of different continental  
856 ancestries within admixed individuals. Nat Genet. 2023 Mar; p. 1–10. <https://doi.org/10.1038/s41588-023-01338-6>, doi: 10.1038/s41588-023-01338-6.
- 858 Hou Z, Ochoa A. Genetic association models are robust to common population kinship estimation biases.  
859 Genetics. 2023 Feb; p. iyad030. doi: [10.1093/genetics/iyad030](https://doi.org/10.1093/genetics/iyad030).
- 860 Hu Y, Graff M, Haessler J, Buyske S, Bien SA, Tao R, Highland HM, Nishimura KK, Zubair N, Lu Y, Verbanck M,  
861 Hilliard AT, Klarin D, Damrauer SM, Ho YL, Program tVMV, Wilson PWF, Chang KM, Tsao PS, Cho K, et al.  
862 Minority-centric meta-analyses of blood lipid levels identify novel loci in the Population Architecture using  
863 Genomics and Epidemiology (PAGE) study. PLOS Genetics. 2020 Mar; 16(3):e1008684. <https://doi.org/10.1371/journal.pgen.1008684>.
- 865 Jacquard A. Structures génétiques des populations. Paris: Masson et Cie; 1970.
- 866 Janss L, Campos Gdl, Sheehan N, Sorensen D. Inferences from Genomic Models in Stratified Populations.  
867 Genetics. 2012 Oct; 192(2):693–704. <https://doi.org/10.1534/genetics.112.141143>.
- 869 Jolliffe IT. Principal Component Analysis. 2 ed. New York: Springer-Verlag; 2002. <http://www.springer.com/gp/book/9780387954424>.
- 871 Kamariza M, Crawford L, Jones D, Finucane H. Misuse of the term ‘trans-ethnic’ in genomics re-  
872 search. Nat Genet. 2021 Nov; 53(11):1520–1521. <https://doi.org/10.1038/s41588-021-00952-6>, doi:  
873 10.1038/s41588-021-00952-6.
- 874 Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E. Variance component model to  
875 account for sample structure in genome-wide association studies. Nat Genet. 2010 Apr; 42(4):348–354. doi:  
876 [10.1038/ng.548](https://doi.org/10.1038/ng.548).
- 877 Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E. Efficient control of population struc-  
878 ture in model organism association mapping. Genetics. 2008 Mar; 178(3):1709–1723. doi: [10.1534/genetics.107.080101](https://doi.org/10.1534/genetics.107.080101).
- 880 Lawson CL, Hanson RJ. Solving least squares problems. Englewood Cliffs: Prentice Hall; 1974.
- 881 Lazaridis I, Nadel D, Rollefson G, Merrett DC, Rohland N, Mallick S, Fernandes D, Novak M, Gamarra B, Sirak K,  
882 Connell S, Stewardson K, Harney E, Fu Q, Gonzalez-Fortes G, Jones ER, Roodenberg SA, Lengyel G, Bocquentin  
883 F, Gasparian B, et al. Genomic insights into the origin of farming in the ancient Near East. Nature. 2016;  
884 536(7617):419–424. doi: [10.1038/nature19310](https://doi.org/10.1038/nature19310).

- 885 Lazaridis I, Patterson N, Mitnik A, Renaud G, Mallick S, Kirsanow K, Sudmant PH, Schraiber JG, Castellano S,  
886 Lipson M, Berger B, Economou C, Bollongino R, Fu Q, Bos KL, Nordenfelt S, Li H, de Filippo C, Prüfer K, Sawyer  
887 S, et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*.  
888 2014 Sep; 513(7518):409–413. doi: 10.1038/nature13673.
- 889 Lee S, Epstein MP, Duncan R, Lin X. Sparse Principal Component Analysis for Identifying Ancestry-Informative  
890 Markers in Genome-Wide Association Studies. *Genetic Epidemiology*. 2012; 36(4):293–302. <https://onlinelibrary.wiley.com/doi/abs/10.1002/gepi.21621>, doi: 10.1002/gepi.21621.
- 892 Lin M, Park DS, Zaitlen NA, Henn BM, Gignoux CR. Admixed Populations Improve Power for Variant Discovery  
893 and Portability in Genome-Wide Association Studies. *Frontiers in Genetics*. 2021; 12. <https://www.frontiersin.org/articles/10.3389/fgene.2021.673167>.
- 895 Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. FaST linear mixed models for genome-wide  
896 association studies. *Nat Methods*. 2011 Sep; 8(10):833–835. doi: 10.1038/nmeth.1681.
- 897 Listgarten J, Lippert C, Kadie CM, Davidson RI, Eskin E, Heckerman D. Improved linear mixed models for  
898 genome-wide association studies. *Nat Methods*. 2012 Jun; 9(6):525–526. <https://www.nature.com/articles/nmeth.2037>, doi: 10.1038/nmeth.2037.
- 900 Liu N, Zhao H, Patki A, Limdi NA, Allison DB. Controlling Population Structure in Human Genetic As-  
901 sociation Studies with Samples of Unrelated Individuals. *Stat Interface*. 2011; 4(3):317–326. doi:  
902 10.4310/sii.2011.v4.n3.a6.
- 903 Liu X, Huang M, Fan B, Buckler ES, Zhang Z. Iterative Usage of Fixed and Random Effect Models for Powerful  
904 and Efficient Genome-Wide Association Studies. *PLOS Genet*. 2016 Feb; 12(2):e1005767. <http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1005767>, doi: 10.1371/journal.pgen.1005767.
- 906 Loh PR, Kichaev G, Gazal S, Schoech AP, Price AL. Mixed-model association for biobank-scale datasets. *Nat Genet*.  
907 2018 Jul; 50(7):906–908. <https://www.nature.com/articles/s41588-018-0144-6>, doi: 10.1038/s41588-018-0144-6.
- 909 Loh PR, Tucker G, Bulik-Sullivan BK, Vilhjálmsson BJ, Finucane HK, Salem RM, Chasman DI, Ridker PM, Neale  
910 BM, Berger B, Patterson N, Price AL. Efficient Bayesian mixed-model analysis increases association power in  
911 large cohorts. *Nat Genet*. 2015 Mar; 47(3):284–290. doi: 10.1038/ng.3190.
- 912 Mahajan A, Spracklen CN, Zhang W, Ng MCY, Petty LE, Kitajima H, Yu GZ, Rüeger S, Speidel L, Kim YJ, Horikoshi  
913 M, Mercader JM, Taliun D, Moon S, Kwak SH, Robertson NR, Rayner NW, Loh M, Kim BJ, Chiou J, et al. Multi-  
914 ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and  
915 translation. *Nat Genet*. 2022 May; 54(5):560–572. doi: 10.1038/s41588-022-01058-3.
- 916 Malécot G. Mathématiques de l'hérédité. Paris: Masson et Cie; 1948. <http://agris.fao.org/agris-search/search.do?recordID=US201300403470>.
- 918 Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in  
919 genome-wide association studies. *Bioinformatics*. 2010 Nov; 26(22):2867–2873. <https://academic.oup.com/bioinformatics/article/26/22/2867/228512>, doi: 10.1093/bioinformatics/btq559.
- 921 Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, Daly MJ, Bustamante CD, Kenny EE. Human  
922 Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am J Hum Genet*. 2017  
923 Apr; 100(4):635–649. doi: 10.1016/j.ajhg.2017.03.004.
- 924 Martin AR, Lin M, Granka JM, Myrick JW, Liu X, Sockell A, Atkinson EG, Wereley CJ, Möller M, Sandhu MS,  
925 Kingsley DM, Hoal EG, Liu X, Daly MJ, Feldman MW, Gignoux CR, Bustamante CD, Henn BM. An Unex-  
926 pectedly Complex Architecture for Skin Pigmentation in Africans. *Cell*. 2017 Nov; 171(6):1340–1353.e14.  
927 <https://www.sciencedirect.com/science/article/pii/S0092867417313247>, doi: 10.1016/j.cell.2017.11.015.
- 928 Matoba N, Akiyama M, Ishigaki K, Kanai M, Takahashi A, Momozawa Y, Ikegawa S, Ikeda M, Iwata N, Hirata M,  
929 Matsuda K, Murakami Y, Kubo M, Kamatani Y, Okada Y. GWAS of 165,084 Japanese individuals identified  
930 nine loci associated with dietary habits. *Nat Hum Behav*. 2020 Mar; 4(3):308–316. <https://www.nature.com/articles/s41562-019-0805-1>, doi: 10.1038/s41562-019-0805-1.
- 932 Mbatchou J, Barnard L, Backman J, Marcketta A, Kosmicki JA, Ziyatdinov A, Benner C, O'Dushlaine C, Barber  
933 M, Boutkov B, Habegger L, Ferreira M, Baras A, Reid J, Abecasis G, Maxwell E, Marchini J. Computationally  
934 efficient whole-genome regression for quantitative and binary traits. *Nat Genet*. 2021 Jul; 53(7):1097–1103.  
935 <https://www.nature.com/articles/s41588-021-00870-7>, doi: 10.1038/s41588-021-00870-7.

- 936 **McVean G.** A genealogical interpretation of principal components analysis. PLoS Genet. 2009 Oct; 5(10):e1000686. doi: [10.1371/journal.pgen.1000686](https://doi.org/10.1371/journal.pgen.1000686).
- 938 **Medina-Gomez C**, Felix JF, Estrada K, Peters MJ, Herrera L, Kruithof CJ, Duijts L, Hofman A, van Duijn CM, Uit-  
939 terlinden AG, Jaddoe VVW, Rivadeneira F. Challenges in conducting genome-wide association studies in  
940 highly admixed multi-ethnic populations: the Generation R Study. Eur J Epidemiol. 2015 Apr; 30(4):317–330.  
941 <https://doi.org/10.1007/s10654-015-9998-4>, doi: 10.1007/s10654-015-9998-4.
- 942 **Mogil LS**, Andaleon A, Badalamenti A, Dickinson SP, Guo X, Rotter JL, Johnson WC, Im HK, Liu Y, Wheeler  
943 HE. Genetic architecture of gene expression traits across diverse populations. PLOS Genetics. 2018  
944 Aug; 14(8):e1007586. <https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1007586>, doi:  
945 [10.1371/journal.pgen.1007586](https://doi.org/10.1371/journal.pgen.1007586).
- 946 **Mullen KM**, Stokkum IHMv, nnls: The Lawson-Hanson algorithm for non-negative least squares (NNLS). The  
947 Comprehensive R Archive Network; 2012. <https://CRAN.R-project.org/package=nnls>.
- 948 **Novembre J**, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR,  
949 Stephens M, Bustamante CD. Genes mirror geography within Europe. Nature. 2008 Nov; 456(7218):98–101.  
950 doi: 10.1038/nature07331.
- 951 **Ochoa A**, Storey JD. New kinship and FST estimates reveal higher levels of differentiation in the global human  
952 population. bioRxiv; 2019. <https://www.biorxiv.org/content/10.1101/653279v1>, doi: 10.1101/653279.
- 953 **Ochoa A**, Storey JD. Estimating FST and kinship for arbitrary population structures. PLoS Genet. 2021 Jan;  
954 17(1):e1009241. doi: [10.1371/journal.pgen.1009241](https://doi.org/10.1371/journal.pgen.1009241).
- 955 **O'Connor LJ**, Schoech AP, Hormozdiari F, Gazal S, Patterson N, Price AL. Extreme Polygenicity of Complex  
956 Traits Is Explained by Negative Selection. The American Journal of Human Genetics. 2019 Aug; 0(0). [https://www.cell.com/ajhg/abstract/S0002-9297\(19\)30266-6](https://www.cell.com/ajhg/abstract/S0002-9297(19)30266-6), doi: [10.1016/j.ajhg.2019.07.003](https://doi.org/10.1016/j.ajhg.2019.07.003).
- 958 **Paradis E**, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioin-  
959 formatics. 2019 Feb; 35(3):526–528. doi: 10.1093/bioinformatics/bty633.
- 960 **Park JH**, Gail MH, Weinberg CR, Carroll RJ, Chung CC, Wang Z, Chanock SJ, Fraumeni JF, Chatterjee N. Dis-  
961 tribution of allele frequencies and effect sizes and their interrelationships for common genetic suscepti-  
962 bility variants. PNAS. 2011 Nov; 108(44):18026–18031. <https://www.pnas.org/content/108/44/18026>, doi:  
963 [10.1073/pnas.1114759108](https://doi.org/10.1073/pnas.1114759108).
- 964 **Park JH**, Wacholder S, Gail MH, Peters U, Jacobs KB, Chanock SJ, Chatterjee N. Estimation of effect size distribu-  
965 tion from genome-wide association studies and implications for future discoveries. Nature Genetics. 2010  
966 Jul; 42(7):570–575. <https://www.nature.com/articles/ng.610>, doi: [10.1038/ng.610](https://doi.org/10.1038/ng.610).
- 967 **Patterson N**, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. Ancient  
968 admixture in human history. Genetics. 2012 Nov; 192(3):1065–1093. doi: [10.1534/genetics.112.145037](https://doi.org/10.1534/genetics.112.145037).
- 969 **Patterson N**, Price AL, Reich D. Population Structure and Eigenanalysis. PLoS Genet. 2006 Dec; 2(12):e190.  
970 <http://dx.plos.org/10.1371/journal.pgen.0020190>, doi: [10.1371/journal.pgen.0020190](https://doi.org/10.1371/journal.pgen.0020190).
- 971 **Peterson RE**, Kuchenbaecker K, Walters RK, Chen CY, Popejoy AB, Periyasamy S, Lam M, Iyegbe C, Strawbridge  
972 RJ, Brick L, Carey CE, Martin AR, Meyers JL, Su J, Chen J, Edwards AC, Kalungi A, Koen N, Majara L, Schwarz E,  
973 et al. Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls,  
974 and Recommendations. Cell. 2019 Oct; 179(3):589–603. [https://www.cell.com/cell/abstract/S0092-8674\(19\)31002-5](https://www.cell.com/cell/abstract/S0092-8674(19)31002-5), doi: [10.1016/j.cell.2019.08.051](https://doi.org/10.1016/j.cell.2019.08.051).
- 976 **Price AL**, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis cor-  
977 rects for stratification in genome-wide association studies. Nat Genet. 2006 Aug; 38(8):904–909. doi:  
978 [10.1038/ng1847](https://doi.org/10.1038/ng1847).
- 979 **Price AL**, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide asso-  
980 ciation studies. Nat Rev Genet. 2010 Jul; 11(7):459–463. doi: [10.1038/nrg2813](https://doi.org/10.1038/nrg2813).
- 981 **Price AL**, Zaitlen NA, Reich D, Patterson N. Response to Sul and Eskin. Nature Reviews Genetics. 2013 Apr;  
982 14(4):300. <https://www.nature.com/articles/nrg2813-c2>, doi: [10.1038/nrg2813-c2](https://doi.org/10.1038/nrg2813-c2).
- 983 **Pritchard JK**, Stephens M, Rosenberg NA, Donnelly P. Association Mapping in Structured Populations. The  
984 American Journal of Human Genetics. 2000 Jul; 67(1):170–181. <http://www.sciencedirect.com/science/article/pii/S0002929707624422>, doi: [10.1086/302959](https://doi.org/10.1086/302959).

- 986 **Privé F**, Luu K, Blum MGB, McGrath JJ, Vilhjálmsdóttir BJ. Efficient toolkit implementing best practices for principal  
987 component analysis of population genetic data. *Bioinformatics*. 2020 Aug; 36(16):4449–4457. <https://doi.org/10.1093/bioinformatics/btaa520>, doi: 10.1093/bioinformatics/btaa520.
- 989 **Qian J**, Tanigawa Y, Du W, Aguirre M, Chang C, Tibshirani R, Rivas MA, Hastie T. A fast and scalable framework  
990 for large-scale and ultrahigh-dimensional sparse regression with application to the UK Biobank. *PLoS Genet.*  
991 2020 Oct; 16(10):e1009141. doi: 10.1371/journal.pgen.1009141.
- 992 **Rakitsch B**, Lippert C, Stegle O, Borgwardt K. A Lasso multi-marker mixed model for association mapping  
993 with population structure correction. *Bioinformatics*. 2013 Jan; 29(2):206–214. doi: 10.1093/bioinformatics/bts669.
- 995 **Roselli C**, Chaffin MD, Weng LC, Aeschbacher S, Ahlberg G, Albert CM, Almgren P, Alonso A, Anderson CD,  
996 Aragam KG, Arking DE, Barnard J, Bartz TM, Benjamin EJ, Bihlmeyer NA, Bis JC, Bloom HL, Boerwinkle E, Bot-  
997 ttinger EB, Brody JA, et al. Multi-ethnic genome-wide association study for atrial fibrillation. *Nature Genetics*.  
998 2018 Sep; 50(9):1225. <https://www.nature.com/articles/s41588-018-0133-9>, doi: 10.1038/s41588-018-0133-9.
- 999 **Rosenberg NA**, Huang L, Jewett EM, Szpiech ZA, Jankovic I, Boehnke M. Genome-wide association studies in  
1000 diverse populations. *Nat Rev Genet.* 2010 May; 11(5):356–366. <https://www.nature.com/articles/nrg2760>, doi:  
1001 10.1038/nrg2760.
- 1002 **Rosenberg NA**, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW. Genetic Structure  
1003 of Human Populations. *Science*. 2002 Dec; 298(5602):2381–2385. <http://www.sciencemag.org/content/298/5602/2381>, doi: 10.1126/science.1078311.
- 1005 **Schlauch D**, Fier H, Lange C. Identification of genetic outliers due to sub-structure and cryptic relationships.  
1006 *Bioinformatics*. 2017 Jul; 33(13):1972–1979. <https://academic.oup.com/bioinformatics/article/33/13/1972/3045026>, doi: 10.1093/bioinformatics/btx109.
- 1008 **Shchur V**, Nielsen R. On the number of siblings and p-th cousins in a large population sample. *J Math Biol.*  
1009 2018 Nov; 77(5):1279–1298. doi: 10.1007/s00285-018-1252-8.
- 1010 **Simonin-Wilmer I**, Orozco-del Pino P, Bishop DT, Iles MM, Robles-Espinoza CD. An Overview of Strategies for  
1011 Detecting Genotype-Phenotype Associations Across Ancestrally Diverse Populations. *Frontiers in Genetics*.  
1012 2021; 12. <https://www.frontiersin.org/articles/10.3389/fgene.2021.703901>.
- 1013 **Simons YB**, Bullaughey K, Hudson RR, Sella G. A population genetic interpretation of GWAS findings for human  
1014 quantitative traits. *PLOS Biology*. 2018 Mar; 16(3):e2002985. <http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.2002985>, doi: 10.1371/journal.pbio.2002985.
- 1016 **Skoglund P**, Posth C, Sirak K, Spriggs M, Valentín F, Bedford S, Clark GR, Reepmeyer C, Petchey F, Fernandes D,  
1017 Fu Q, Harney E, Lipson M, Mallick S, Novak M, Rohland N, Stewardson K, Abdullah S, Cox MP, Friedlaender  
1018 FR, et al. Genomic insights into the peopling of the Southwest Pacific. *Nature*. 2016 Oct; 538(7626):510–513.  
1019 doi: 10.1038/nature19844.
- 1020 **Sokal RR**, Michener CD. A statistical method for evaluating systematic relationships. *Univ Kansas, Sci Bull.*  
1021 1958; 38:1409–1438.
- 1022 **Song M**, Hao W, Storey JD. Testing for genetic associations in arbitrarily structured populations. *Nat Genet.*  
1023 2015 May; 47(5):550–554. doi: 10.1038/ng.3244.
- 1024 **Speed D**, Hemani G, Johnson MR, Balding DJ. Improved heritability estimation from genome-wide SNPs. *Am J  
1025 Hum Genet.* 2012 Dec; 91(6):1011–1021. doi: 10.1016/j.ajhg.2012.10.010.
- 1026 **Storey JD**. The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann Statist.* 2003 Dec;  
1027 31(6):2013–2035. <http://projecteuclid.org/euclid-aos/1074290335>, doi: 10.1214/aos/1074290335.
- 1028 **Storey JD**, Tibshirani R. Statistical significance for genomewide studies. *Proceedings of the National Academy  
1029 of Sciences of the United States of America*. 2003; 100(16):9440–9445. <http://www.pnas.org/content/100/16/9440.abstract>, doi: 10.1073/pnas.1530509100.
- 1031 **Sul JH**, Eskin E. Mixed models can correct for population structure for genomic regions under selection. *Nature  
1032 Reviews Genetics*. 2013 Apr; 14(4):300. <https://www.nature.com/articles/nrg2813-c1>, doi: 10.1038/nrg2813-c1.
- 1033 **Sul JH**, Martin LS, Eskin E. Population structure in genetic studies: Confounding factors and mixed models.  
1034 *PLoS Genet.* 2018; 14(12):e1007309. doi: 10.1371/journal.pgen.1007309.

- 1035 Sun G, Zhu C, Kramer MH, Yang SS, Song W, Piepho HP, Yu J. Variation explained in mixed-model asso-  
1036 ciation mapping. *Heredity*. 2010 Oct; 105(4):333–340. <https://www.nature.com/articles/hdy201011>, doi:  
1037 10.1038/hdy.2010.11.
- 1038 Svishcheva GR, Axenovich TI, Belonogova NM, van Duijn CM, Aulchenko YS. Rapid variance components-based  
1039 method for whole-genome association analysis. *Nat Genet*. 2012 Oct; 44(10):1166–1170. <https://www.nature.com/articles/ng.2410>, doi: 10.1038/ng.2410.
- 1041 Thornton T, McPeek MS. ROADTRIPS: case-control association testing with partially or completely  
1042 unknown population and pedigree structure. *Am J Hum Genet*. 2010 Feb; 86(2):172–184. doi:  
1043 10.1016/j.ajhg.2010.01.001.
- 1044 Tucker G, Price AL, Berger B. Improving the Power of GWAS and Avoiding Confounding from Population Stratification with PC-Select. *Genetics*. 2014 Jul; 197(3):1045–1049. <http://www.genetics.org/content/197/3/1045>,  
1045 doi: 10.1534/genetics.114.164285.
- 1047 Vilhjálmsson BJ, Nordborg M. The nature of confounding in genome-wide association studies. *Nat Rev Genet*.  
1048 2013 Jan; 14(1):1–2. <https://www.nature.com/articles/nrg3382>, doi: 10.1038/nrg3382.
- 1049 Voight BF, Pritchard JK. Confounding from Cryptic Relatedness in Case-Control Association Studies. *PLOS  
1050 Genetics*. 2005 Sep; 1(3):e32. <https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.0010032>,  
1051 doi: 10.1371/journal.pgen.0010032.
- 1052 Wang H, Aragam B, Xing EP. Trade-offs of Linear Mixed Models in Genome-Wide Association Studies. *Journal  
1053 of Computational Biology*. 2022 Mar; 29(3):233–242. <https://www.liebertpub.com/doi/full/10.1089/cmb.2021.0157>,  
1054 doi: 10.1089/cmb.2021.0157.
- 1055 Wojcik GL, Graff M, Nishimura KK, Tao R, Haessler J, Gignoux CR, Highland HM, Patel YM, Sorokin EP, Avery  
1056 CL, Belbin GM, Bien SA, Cheng I, Cullina S, Hodonsky CJ, Hu Y, Huckins LM, Jeff J, Justice AE, Kocarnik JM, et al.  
1057 Genetic analyses of diverse populations improves discovery for complex traits. *Nature*. 2019; 570(7762):514–  
1058 518. doi: 10.1038/s41586-019-1310-4.
- 1059 Wright S. The Genetical Structure of Populations. *Annals of Eugenics*. 1949; 15(1):323–354. <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-1809.1949.tb02451.x>, doi: 10.1111/j.1469-1809.1949.tb02451.x.
- 1061 Wu C, DeWan A, Hoh J, Wang Z. A comparison of association methods correcting for population stratification  
1062 in case-control studies. *Ann Hum Genet*. 2011 May; 75(3):418–427. doi: 10.1111/j.1469-1809.2010.00639.x.
- 1063 Xu H, Guan Y. Detecting Local Haplotype Sharing and Haplotype Association. *Genetics*. 2014 Jul; 197(3):823–838.  
1064 <http://www.genetics.org/content/197/3/823>, doi: 10.1534/genetics.114.164814.
- 1065 Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum  
1066 Genet*. 2011 Jan; 88(1):76–82. doi: 10.1016/j.ajhg.2010.11.011.
- 1067 Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in the application of mixed-  
1068 model association methods. *Nat Genet*. 2014 Feb; 46(2):100–106. doi: 10.1038/ng.2876.
- 1069 Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB,  
1070 Kresovich S, Buckler ES. A unified mixed-model method for association mapping that accounts for multiple  
1071 levels of relatedness. *Nat Genet*. 2006 Feb; 38(2):203–208. doi: 10.1038/ng1702.
- 1072 Zaidi AA, Mathieson I. Demographic history mediates the effect of stratification on polygenic scores. *eLife*.  
1073 2020 Nov; 9:e61548. <https://doi.org/10.7554/eLife.61548>, doi: 10.7554/eLife.61548.
- 1074 Zeng J, Vlaming R, Wu Y, Robinson MR, Lloyd-Jones LR, Yengo L, Yap CX, Xue A, Sidorenko J, McRae AF, Powell  
1075 JE, Montgomery GW, Metspalu A, Esko T, Gibson G, Wray NR, Visscher PM, Yang J. Signatures of negative  
1076 selection in the genetic architecture of human complex traits. *Nature Genetics*. 2018 May; 50(5):746–753.  
1077 <https://www.nature.com/articles/s41588-018-0101-4>, doi: 10.1038/s41588-018-0101-4.
- 1078 Zhang S, Zhu X, Zhao H. On a semiparametric test to detect associations between quantitative traits and  
1079 candidate genes using unrelated individuals. *Genetic Epidemiology*. 2003; 24(1):44–56. <https://onlinelibrary.wiley.com/doi/abs/10.1002/gepi.10196>, doi: 10.1002/gepi.10196.
- 1081 Zhang Y, Pan W. Principal Component Regression and Linear Mixed Model in Association Analysis of Structured  
1082 Samples: Competitors or Complements? *Genetic Epidemiology*. 2015; 39(3):149–155. <https://onlinelibrary.wiley.com/doi/abs/10.1002/gepi.21879>, doi: 10.1002/gepi.21879.

- 1084** **Zhang Z**, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordovas JM, Buckler ES.  
**1085** Mixed linear model approach adapted for genome-wide association studies. *Nat Genet*. 2010 Apr; 42(4):355–  
**1086** 360. <https://www.nature.com/articles/ng.546>, doi: 10.1038/ng.546.
- 1087** **Zhao K**, Aranzana MJ, Kim S, Lister C, Shindo C, Tang C, Toomajian C, Zheng H, Dean C, Marjoram P, Nordborg M. An Arabidopsis Example of Association Mapping in Structured Samples. *PLOS Genetics*. 2007  
**1088** Jan; 3(1):e4. <https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.0030004>, doi: 10.1371/journal.pgen.0030004.
- 1091** **Zheng X**, Weir BS. Eigenanalysis of SNP data with an identity by descent interpretation. *Theor Popul Biol*. 2016  
**1092** Feb; 107:65–76. doi: 10.1016/j.tpb.2015.09.004.
- 1093** **Zhong Y**, Perera MA, Gamazon ER. On Using Local Ancestry to Characterize the Genetic Architecture of Hu-  
**1094** man Traits: Genetic Regulation of Gene Expression in Multiethnic or Admixed Populations. *The American  
1095 Journal of Human Genetics*. 2019 Jun; 104(6):1097–1115. <https://www.sciencedirect.com/science/article/pii/S0002929719301557>, doi: 10.1016/j.ajhg.2019.04.009.
- 1097** **Zhou Q**, Zhao L, Guan Y. Strong Selection at MHC in Mexicans since Admixture. *PLoS Genet*. 2016 Feb;  
**1098** 12(2):e1005847. doi: 10.1371/journal.pgen.1005847.
- 1099** **Zhou W**, Nielsen JB, Fritzsche LG, Dey R, Gabrielsen ME, Wolford BN, LeFaive J, VandeHaar P, Gagliano SA, Gifford  
**1100** A, Bastarache LA, Wei WQ, Denny JC, Lin M, Hveem K, Kang HM, Abecasis GR, Willer CJ, Lee S. Efficiently  
**1101** controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat  
Genet*. 2018 Sep; 50(9):1335–1341. doi: 10.1038/s41588-018-0184-y.
- 1103** **Zhou X**, Stephens M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet*. 2012  
**1104** Jun; 44(7):821–824. doi: 10.1038/ng.2310.
- 1105** **Zhu C**, Yu J. Nonmetric Multidimensional Scaling Corrects for Population Structure in Association Mapping With  
**1106** Different Sample Types. *Genetics*. 2009 Jul; 182(3):875–888. <https://doi.org/10.1534/genetics.108.098863>, doi:  
**1107** 10.1534/genetics.108.098863.

1108      **Appendix 1**

1110  
**1111 Fitting ancestral allele frequency distribution to real data**  
1112

1113 We calculated  $\hat{p}_i^T$  distributions of each real dataset. However, population structure increases  
1114 the variance of these sample  $\hat{p}_i^T$  relative to the true  $p_i^T$  (*Ochoa and Storey, 2021*). We present  
1115 a new algorithm for constructing a new distribution based on the input data but with the  
1116 lower variance of the true ancestral distribution. Suppose the  $p_i^T$  distribution over loci  $i$   
1117 satisfies  $E[p_i^T] = \frac{1}{2}$  and  $\text{Var}(p_i^T) = V^T$ . The sample allele frequency  $\hat{p}_i^T$ , conditioned on  $p_i^T$ ,  
1118 satisfies

1119       $E[\hat{p}_i^T | p_i^T] = p_i^T, \quad \text{Var}(\hat{p}_i^T | p_i^T) = p_i^T(1 - p_i^T)\bar{\varphi}^T,$

1120 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 (*Ochoa and Storey, 2021*).  
1121 The unconditional moments of  $\hat{p}_i^T$  follow from the laws of total expectation and variance:  
1122  $E[\hat{p}_i^T] = \frac{1}{2}$  and

1123       $W^T = \text{Var}(\hat{p}_i^T) = \bar{\varphi}^T \frac{1}{4} + (1 - \bar{\varphi}^T)V^T.$

1124 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  
1125 with the original, lower variance of

1127       $V^T = \frac{W^T - \frac{1}{4}\bar{\varphi}^T}{1 - \bar{\varphi}^T}. \quad (9)$   
1128

1129 We use the unbiased estimator  $\hat{W}^T = \frac{1}{m} \sum_{i=1}^m \left( \hat{p}_i^T - \frac{1}{2} \right)^2$ , while  $\bar{\varphi}^T$  is calculated from the tree  
1130 parameters: the subpopulation coancestry matrix (**Equation 7**), expanded from subpopula-  
1131 tions to individuals, the diagonal converted to kinship (reversing **Equation 8**), and the matrix  
1132 averaged. However, since our model ignores the MAF filters imposed in our simulations,  $\bar{\varphi}^T$   
1133 was adjusted. For Human Origins the true model  $\bar{\varphi}^T$  of 0.143 was used. For 1000 Genomes  
1134 and HGDP the true  $\bar{\varphi}^T$  are 0.126 and 0.124, respectively, but 0.4 for both produced a better  
1135 fit.  
1136

1137      Lastly, we construct new allele frequencies,

1138       $p^* = w\hat{p}_i^T + (1 - w)q,$

1139 by a weighted average of  $\hat{p}_i^T$  and  $q \in (0, 1)$  drawn independently from a different distribution.  
1140  $E[q] = \frac{1}{2}$  is required to have  $E[p^*] = \frac{1}{2}$ . The resulting variance is

1141       $\text{Var}(p^*) = w^2 W^T + (1 - w)^2 \text{Var}(q),$

1142 which we equate to the desired  $V^T$  (**Equation 9**) and solve for  $w$ . For simplicity, we also set  
1143  $\text{Var}(q) = V^T$ , which is achieved with:

1144       $q \sim \text{Beta}\left(\frac{1}{2}\left(\frac{1}{4V^T} - 1\right), \frac{1}{2}\left(\frac{1}{4V^T} - 1\right)\right).$

1145 Although  $w = 0$  yields  $\text{Var}(p^*) = V^T$ , we use the second root of the quadratic equation to use  
1146  $\hat{p}_i^T$ :

1147       $w = \frac{2V^T}{W^T + V^T}.$

1156 **Appendix 2**

1157 **Comparisons between SRMSD<sub>p</sub>, AUC<sub>PR</sub>, and evaluation measures from the**  
 1158 **literature**

1159 The inflation factor  $\lambda$

1160 Test statistic inflation has been used to measure model calibration (*Astle and Balding, 2009*;  
 1161 *Price et al., 2010*). The inflation factor  $\lambda$  is defined as the median  $\chi^2$  association statistic  
 1162 divided by theoretical median under the null hypothesis (*Devlin and Roeder, 1999*). To com-  
 1163 pare p-values from non- $\chi^2$  tests (such as t-statistics),  $\lambda$  can be calculated from p-values using

$$1164 \quad 1165 \quad \lambda = \frac{F^{-1}(1 - p_{\text{median}})}{F^{-1}(1 - u_{\text{median}})},$$

1166 where  $p_{\text{median}}$  is the median observed p-value (including causal loci),  $u_{\text{median}} = \frac{1}{2}$  is its null  
 1167 expectation, and  $F$  is the  $\chi^2$  cumulative density function ( $F^{-1}$  is the quantile function).

1168 To compare  $\lambda$  and SRMSD<sub>p</sub> directly, for simplicity assume that all p-values are null. In  
 1169 this case, calibrated p-values give  $\lambda = 1$  and SRMSD<sub>p</sub> = 0. However, non-uniform p-values  
 1170 with the expected median, such as from genomic control (*Devlin and Roeder, 1999*), result in  
 1171  $\lambda = 1$ , but SRMSD<sub>p</sub> ≠ 0 except for uniform p-values, a key flaw of  $\lambda$  that SRMSD<sub>p</sub> overcomes.  
 1172 Inflated statistics (anti-conservative p-values) give  $\lambda > 1$  and SRMSD<sub>p</sub> > 0. Deflated statistics  
 1173 (conservative p-values) give  $\lambda < 1$  and SRMSD<sub>p</sub> < 0. Thus,  $\lambda \neq 1$  always implies SRMSD<sub>p</sub> ≠ 0  
 1174 (where  $\lambda - 1$  and SRMSD<sub>p</sub> have the same sign), but not the other way around. Overall,  $\lambda$  de-  
 1175 pends only on the median p-value, while SRMSD<sub>p</sub> uses the complete distribution. However,  
 1176 SRMSD<sub>p</sub> requires knowing which loci are null, so unlike  $\lambda$  it is only applicable to simulated  
 1177 traits.

1178 **Empirical comparison of SRMSD<sub>p</sub> and  $\lambda$**

1179 There is a near one-to-one correspondence between  $\lambda$  and SRMSD<sub>p</sub> in our data (**Figure 2—**  
 1180 **figure Supplement 1**). PCA tended to be inflated ( $\lambda > 1$  and SRMSD<sub>p</sub> > 0) whereas LMM  
 1181 tended to be deflated ( $\lambda < 1$  and SRMSD<sub>p</sub> < 0), otherwise the data for both models fall on  
 1182 the same contiguous curve. We fit a sigmoidal function to this data,

$$1183 \quad 1184 \quad \text{SRMSD}_p(\lambda) = a \frac{\lambda^b - 1}{\lambda^b + 1}, \quad (10)$$

1185 which for  $a, b > 0$  satisfies  $\text{SRMSD}_p(\lambda = 1) = 0$  and reflects  $\log(\lambda)$  about zero ( $\lambda = 1$ ):

$$1186 \quad \text{SRMSD}_p(\log(\lambda) = -x) = -\text{SRMSD}_p(\log(\lambda) = x).$$

1187 We fit this model to  $\lambda > 1$  only since it was less noisy and of greater interest, and obtained  
 1188 the curve shown in **Figure 2—figure Supplement 1** with  $a = 0.564$  and  $b = 0.619$ . The value  $\lambda =$   
 1189 1.05, a common threshold for benign inflation (*Price et al., 2010*), corresponds to  $\text{SRMSD}_p =$   
 1190 0.0085 according to **Equation 10**. Conversely,  $\text{SRMSD}_p = 0.01$ , serving as a simpler rule of  
 1191 thumb, corresponds to  $\lambda = 1.06$ .

1192 **Type I error rate**

1193 The type I error rate is the proportion of null p-values with  $p \leq t$ . Calibrated p-values have  
 1194 type I error rate near  $t$ , which may be evaluated with a binomial test. This measure may  
 1195 give different results for different  $t$ , for example be significantly miscalibrated only for large  
 1196  $t$  (due to lack of power for smaller  $t$ ), and it requires large simulations to estimate well as it  
 1197 depends on the tail of the distribution. In contrast,  $\text{SRMSD}_p$  uses the entire distribution so it  
 1198 is easier to estimate,  $\text{SRMSD}_p = 0$  guarantees calibrated type I error rates at all  $t$ , while large

|SRMSD<sub>p</sub>| indicates incorrect type I errors for a range of  $t$ . Empirically, we find the expected agreement and monotonic relationship between SRMSD<sub>p</sub> and type I error rate (**Figure 2—figure Supplement 2**).

#### Statistical power and comparison to AUC<sub>PR</sub>

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

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

$F(t)$  is a cumulative function, so it is monotonically increasing and has an inverse. Like type I error control, power may rank models differently depending on  $t$ , and it is also harder to estimate than AUC<sub>PR</sub> because power depends on the tail of the distribution.

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

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

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

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

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

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

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

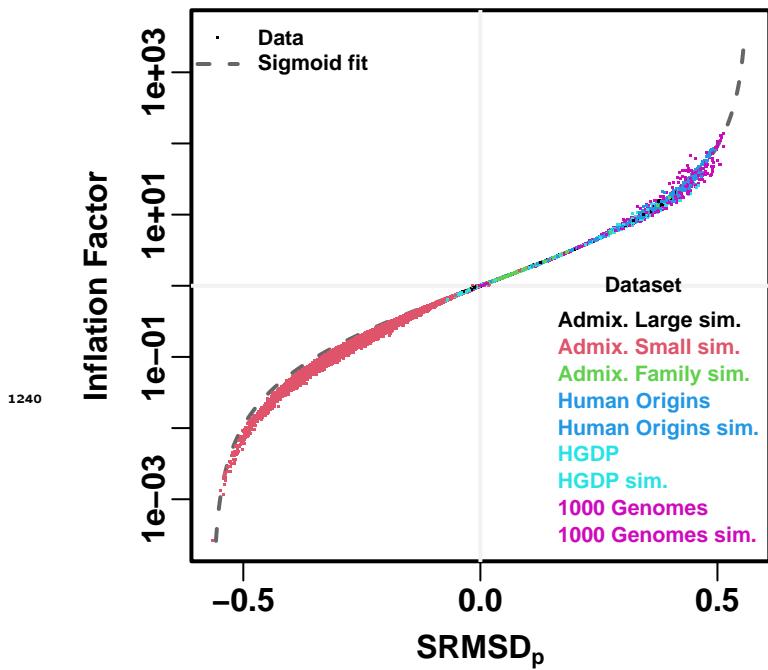
Noting that  $t = F^{-1}(\text{Recall})$ , precision can be written as a function of recall, the power function, and constants:

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

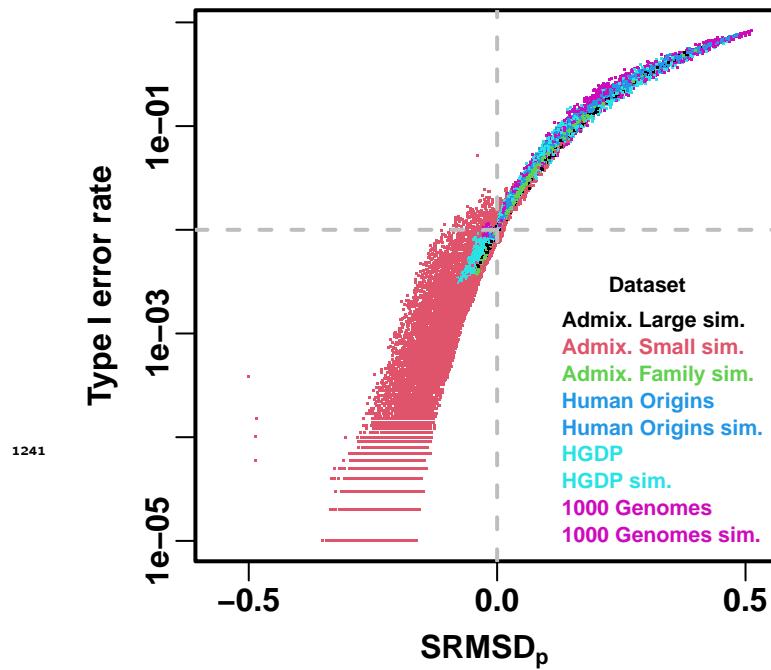
This last form leads most clearly to  $\text{AUC}_{\text{PR}} = \int_0^1 \text{Precision}(\text{Recall}) d\text{Recall}$ .

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

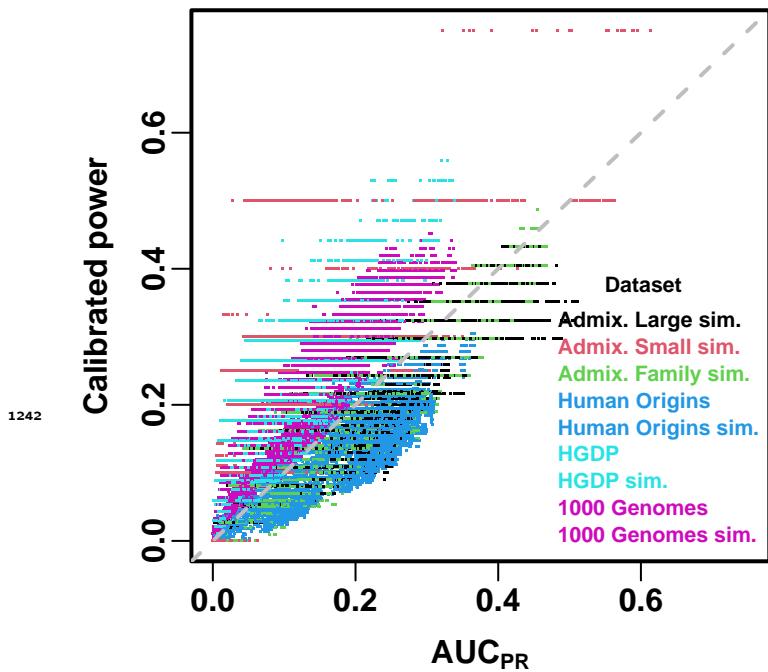
Empirically, we find the predicted positive correlation between AUC<sub>PR</sub> and calibrated power (**Figure 2—figure Supplement 3**). The correlation is clear when considered separately per dataset, but the slope varies per dataset, which is expected because the proportion of alternative cases  $\pi_1$  varies per dataset.



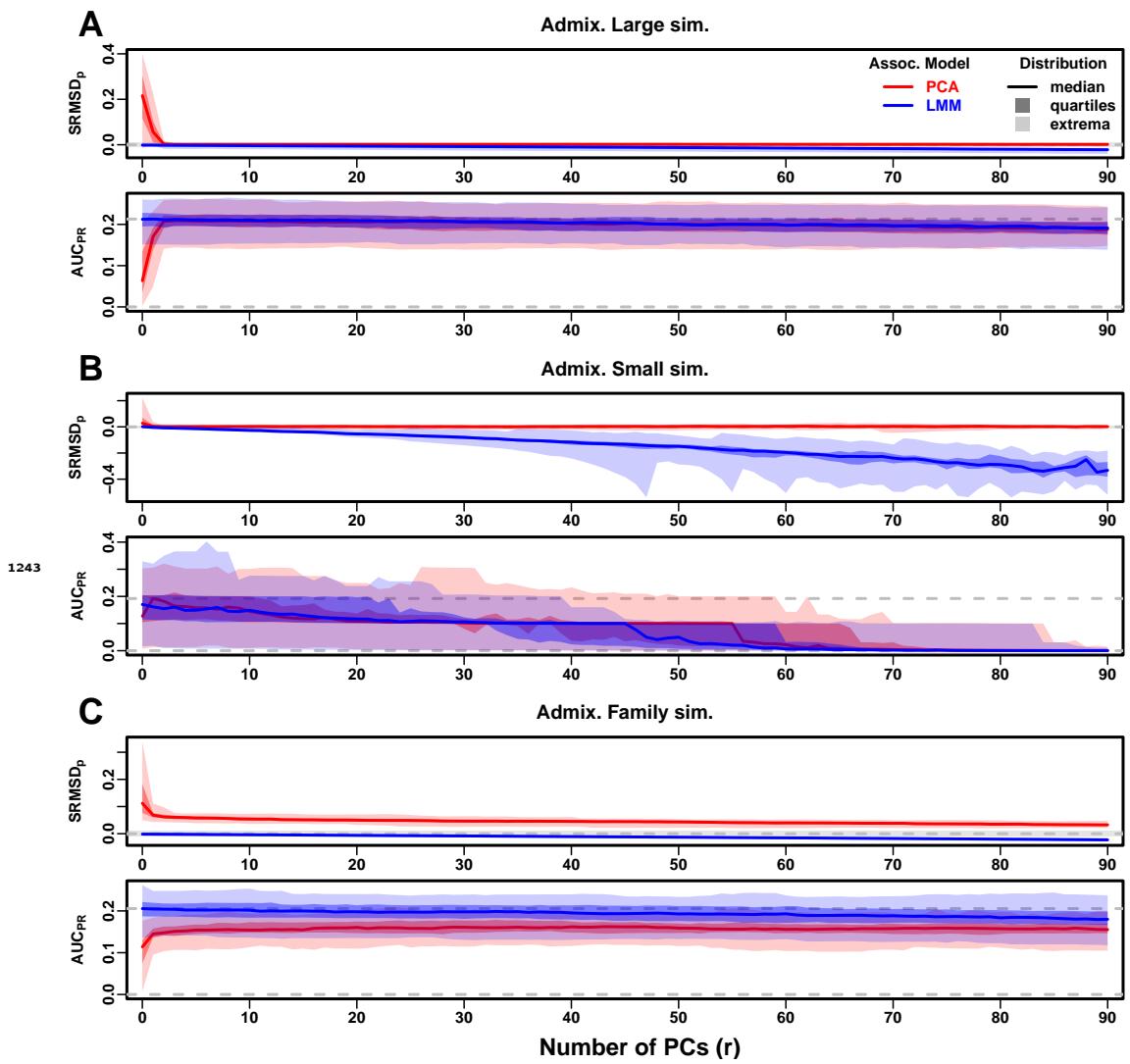
**Figure 2—figure supplement 1.** Comparison between SRMSD<sub>p</sub> and inflation factor. Each point is a pair of statistics for one replicate, one association model (PCA or LMM with some number of PCs  $r$ ), one trait model (FES vs RC, all heritability/environments tested), and one dataset (color coded by dataset). Note log y-axis. The sigmoidal curve in [Equation 10](#) is fit to the data.



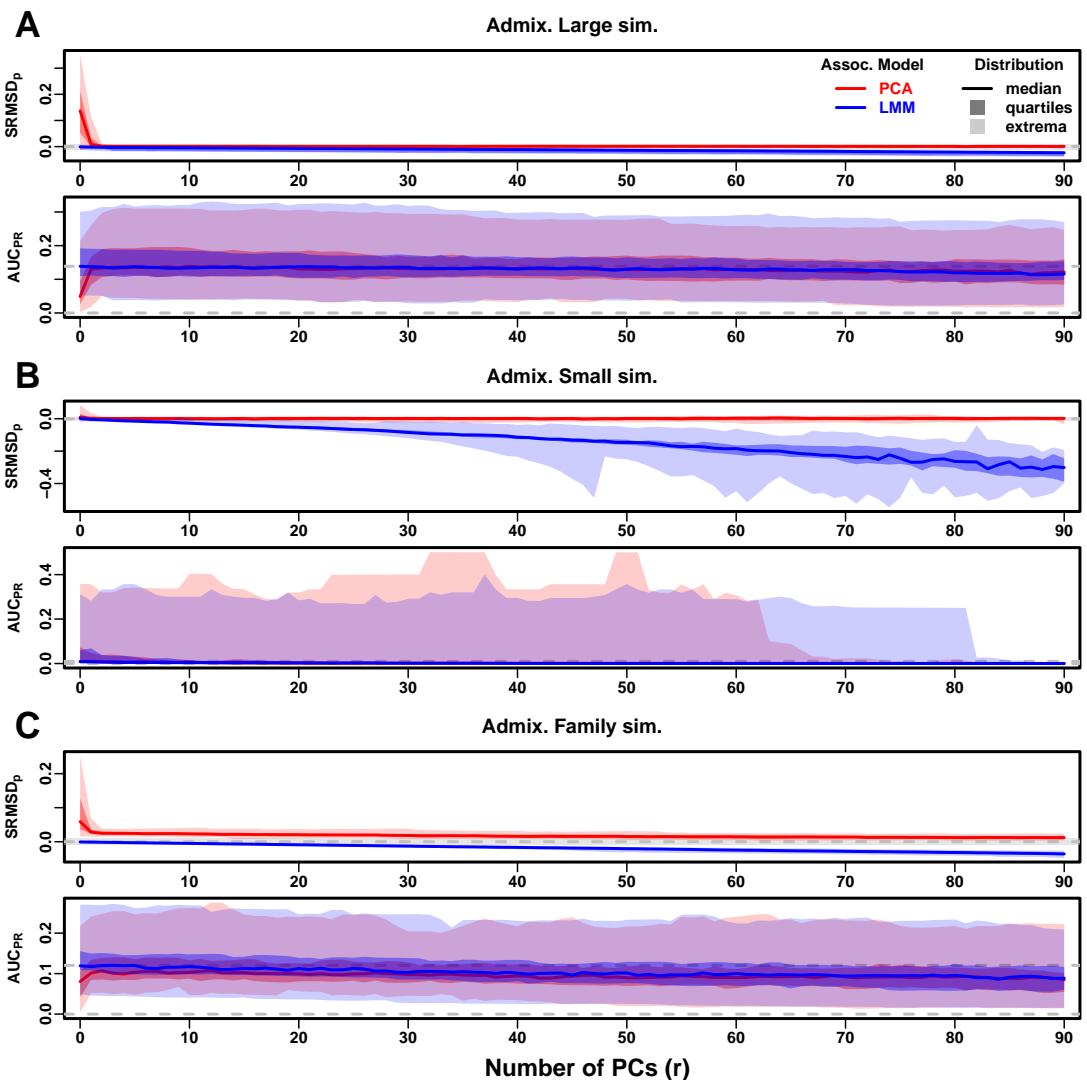
**Figure 2—figure supplement 2.** Comparison between SRMSD<sub>p</sub> and type I error rate. Type I error rate calculated at a p-value threshold of 1e-2 (horizontal dashed gray line). Thus, a calibrated model has a type I error rate of 1e-2 and SRMSD<sub>p</sub> = 0 (where the dashed lines meet). As expected, increased type I error rates correspond to SRMSD<sub>p</sub> > 0, while reduced type I error rates correspond to SRMSD<sub>p</sub> < 0. Each point is a pair of statistics for one replicate, one association model (PCA or LMM with some number of PCs  $r$ ), one trait model (FES vs RC, all heritability/environments tested), and one dataset (color coded by dataset). Note log y-axis.



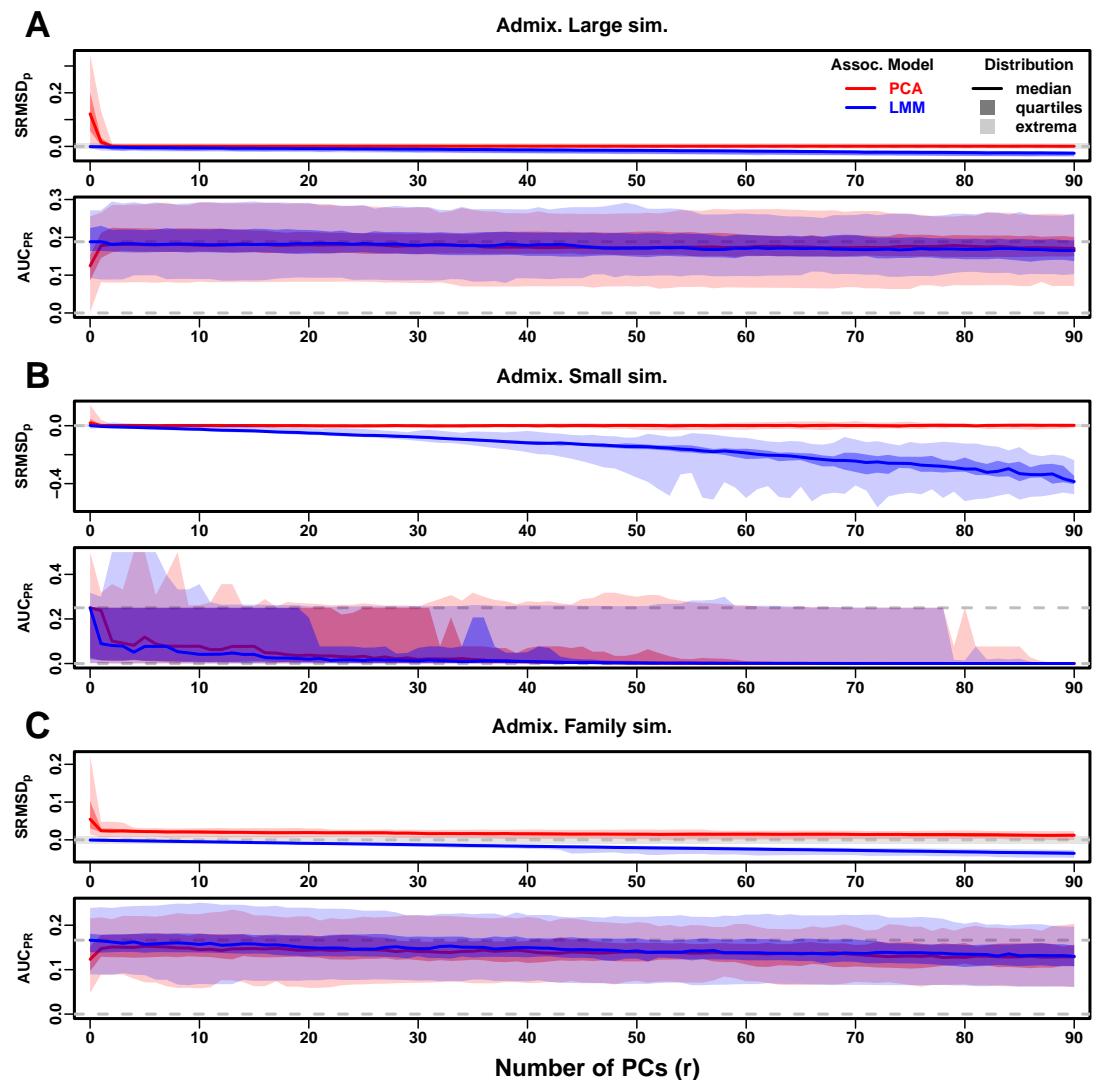
**Figure 2—figure supplement 3.** Comparison between AUC<sub>PR</sub> and calibrated power. Calibrated power is power calculated at an empirical type I error threshold of 1e-4. Each point is a pair of statistics for one replicate, one association model (PCA or LMM with some number of PCs  $r$ ), one trait model (FES vs RC, all heritability/environments tested), and one dataset (color coded by dataset). Gray dashed line is  $y = x$  line.



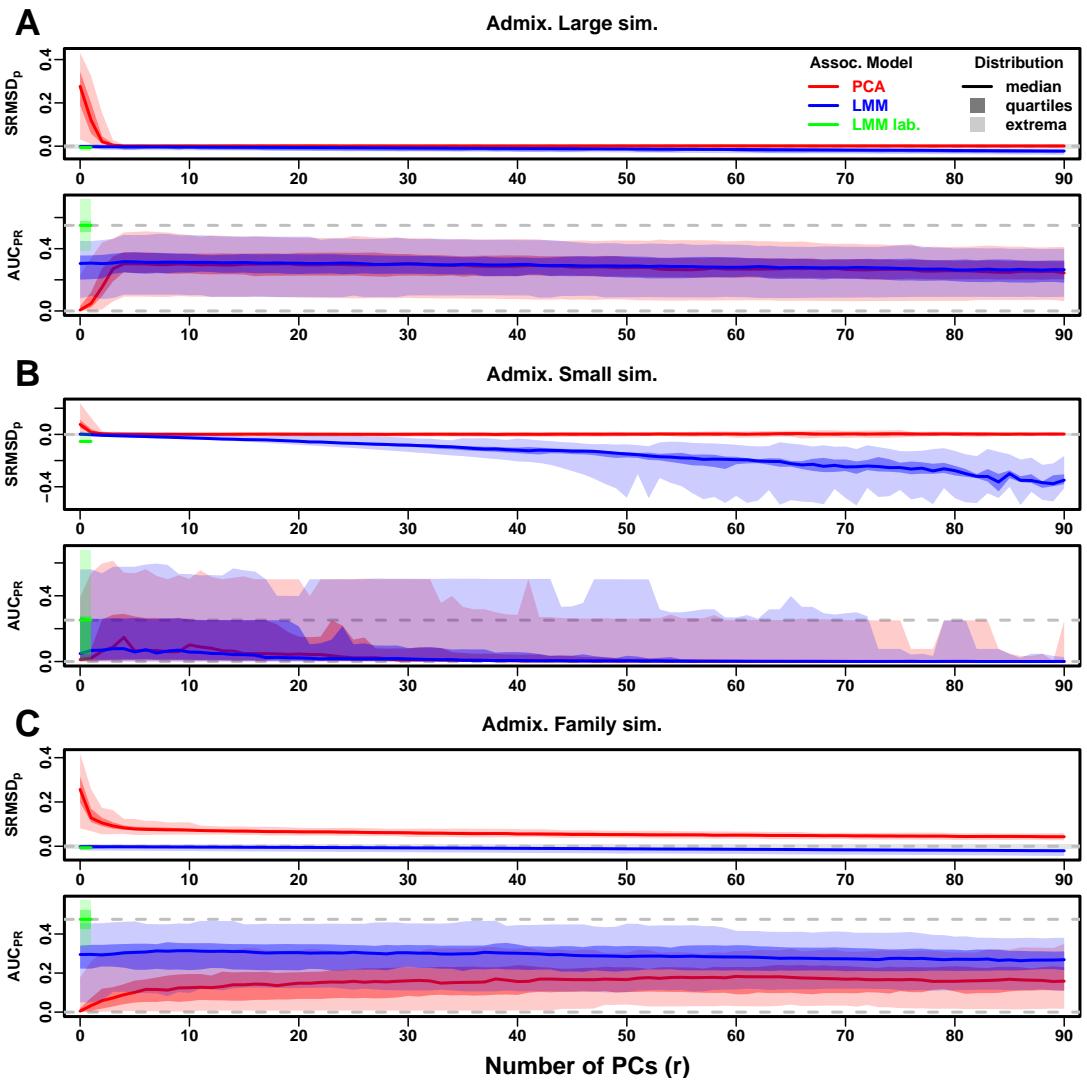
**Figure 3—figure supplement 1.** Evaluations in admixture simulations with RC traits, high heritability.



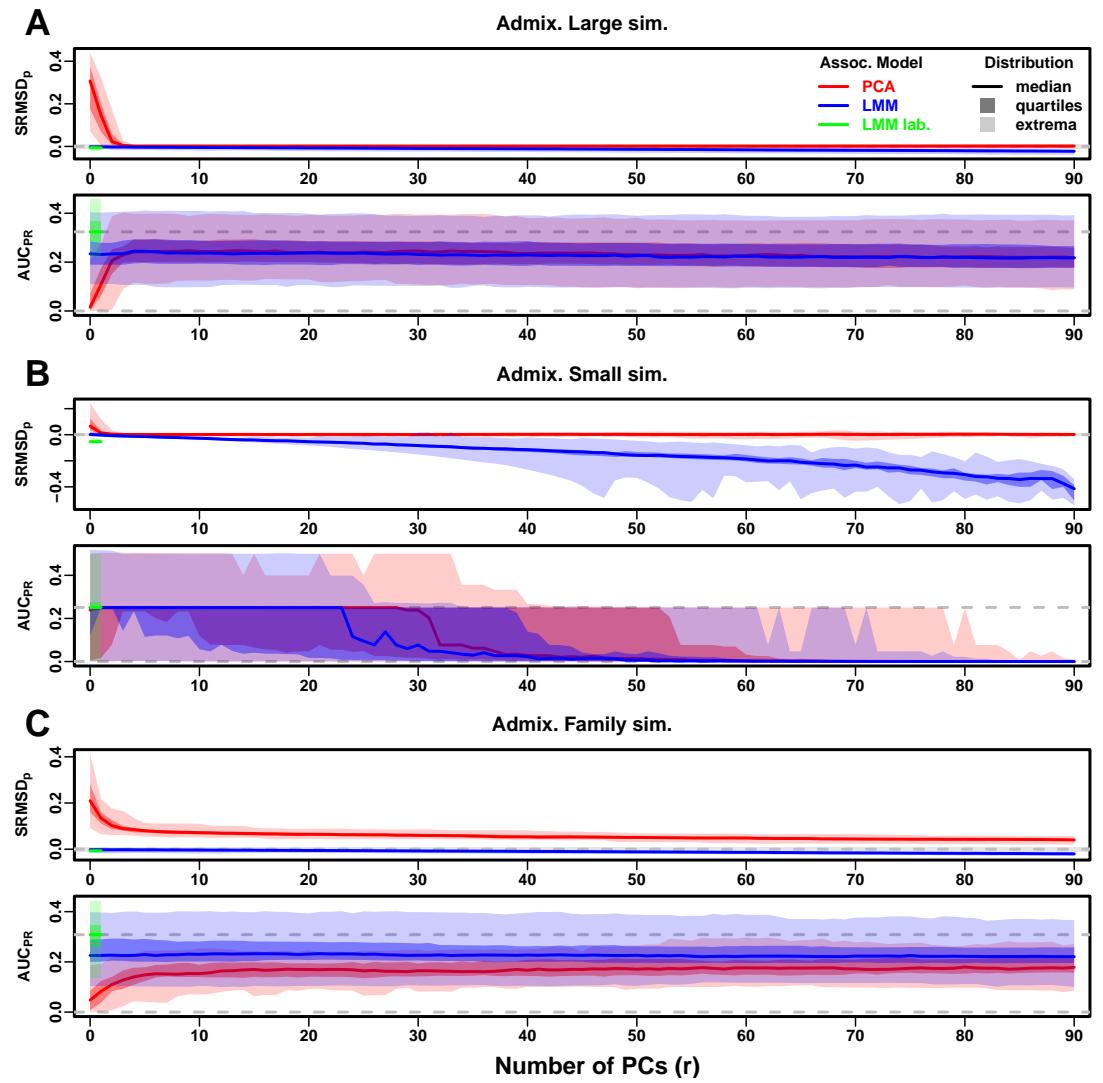
**Figure 3—figure supplement 2.** Evaluations in admixture simulations with FES traits, low heritability.



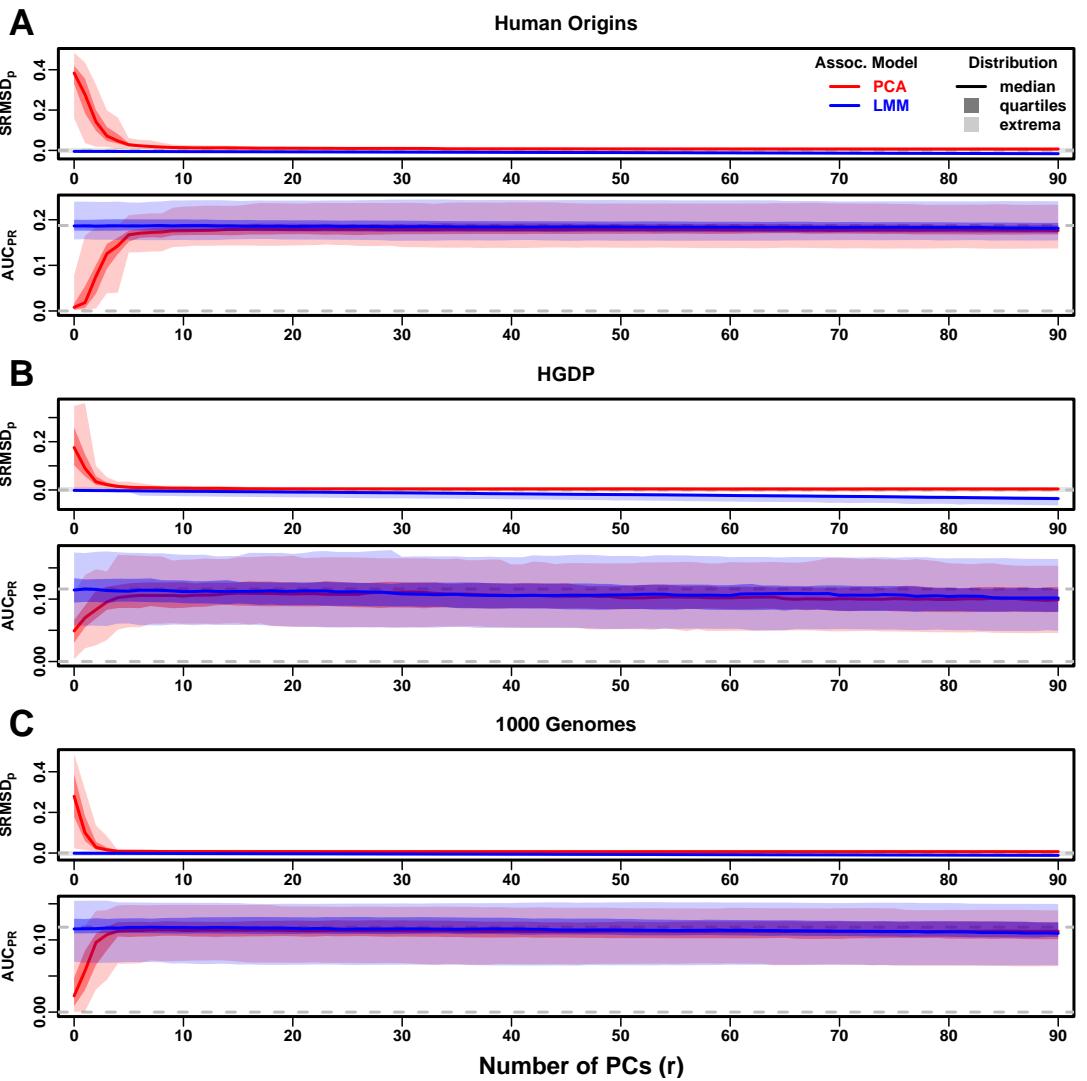
**Figure 3—figure supplement 3.** Evaluations in admixture simulations with RC traits, low heritability.



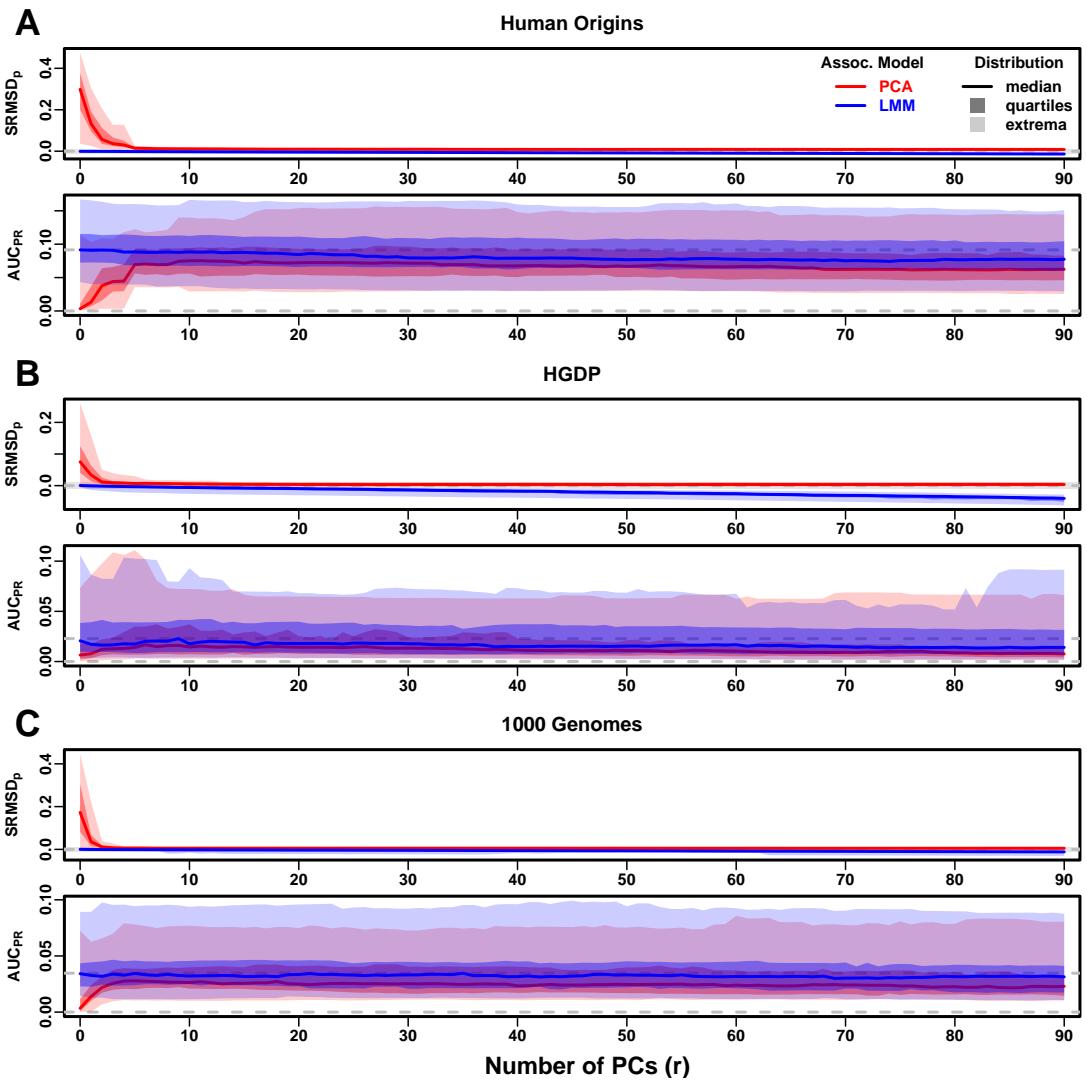
**Figure 3—figure supplement 4.** Evaluations in admixture simulations with FES traits, environment. “LMM lab.” was only tested with  $r = 0$ .



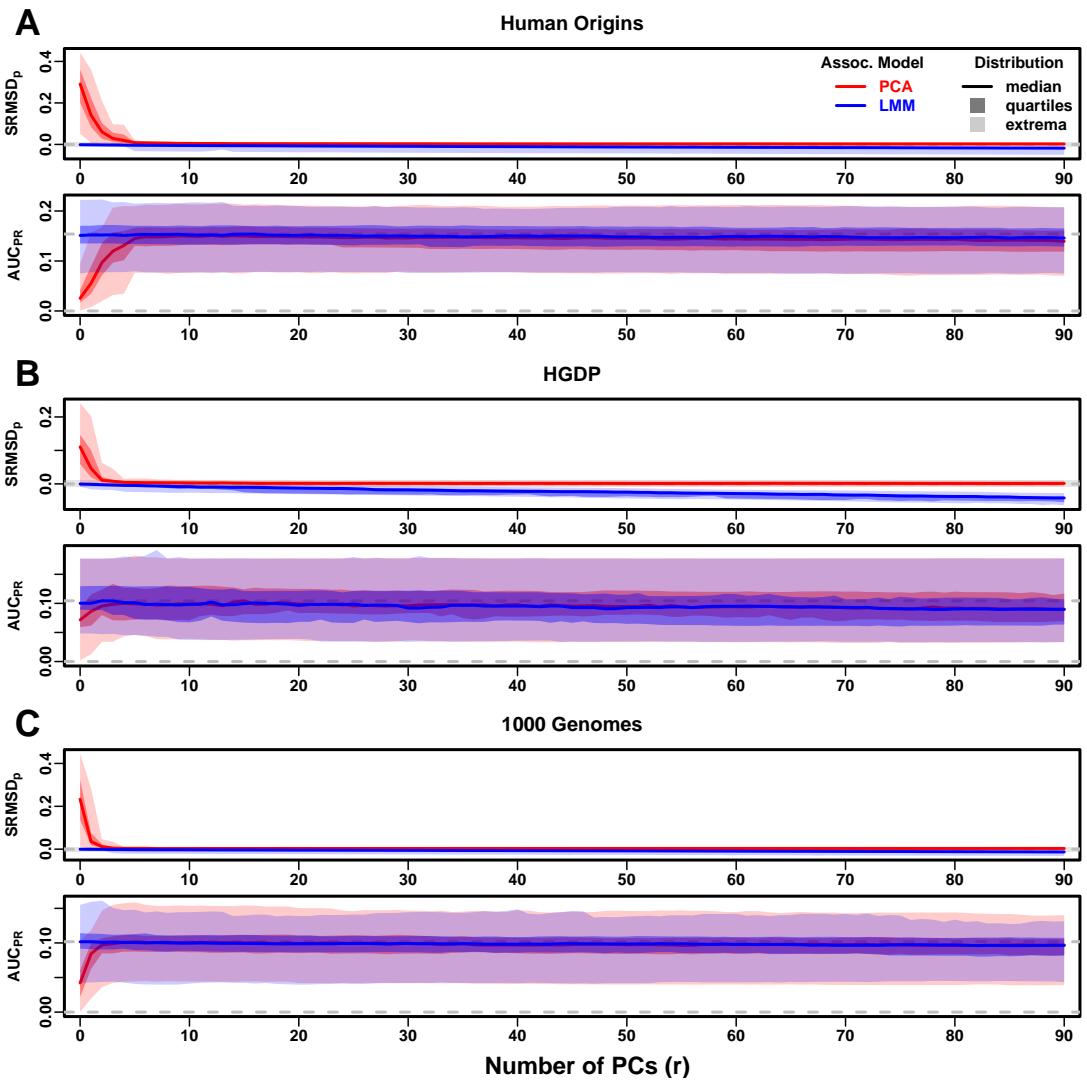
**Figure 3—figure supplement 5.** Evaluations in admixture simulations with RC traits, environment. “LMM lab.” was only tested with  $r = 0$ .



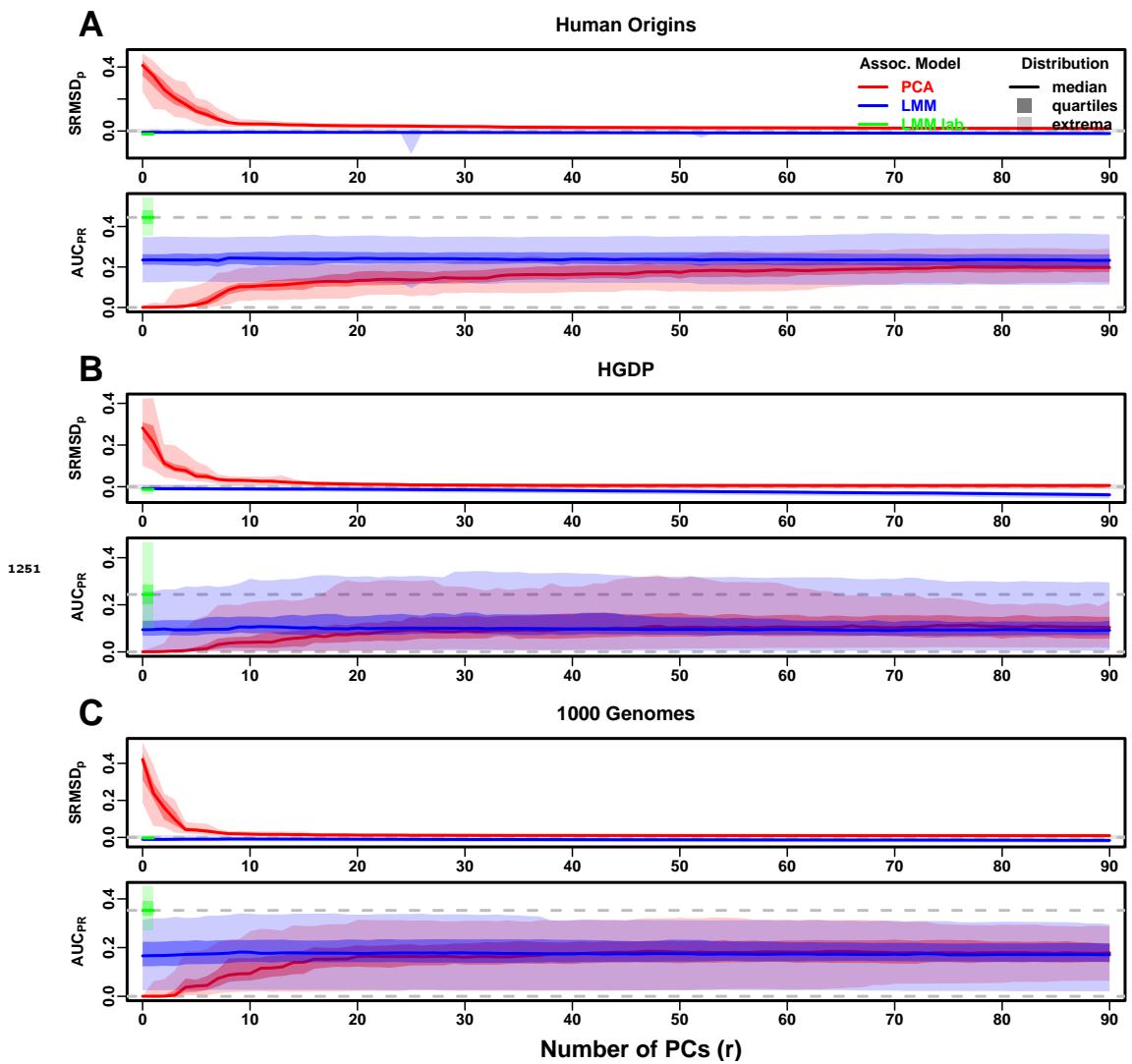
**Figure 4—figure supplement 1.** Evaluations in real human genotype datasets with RC traits, high heritability.



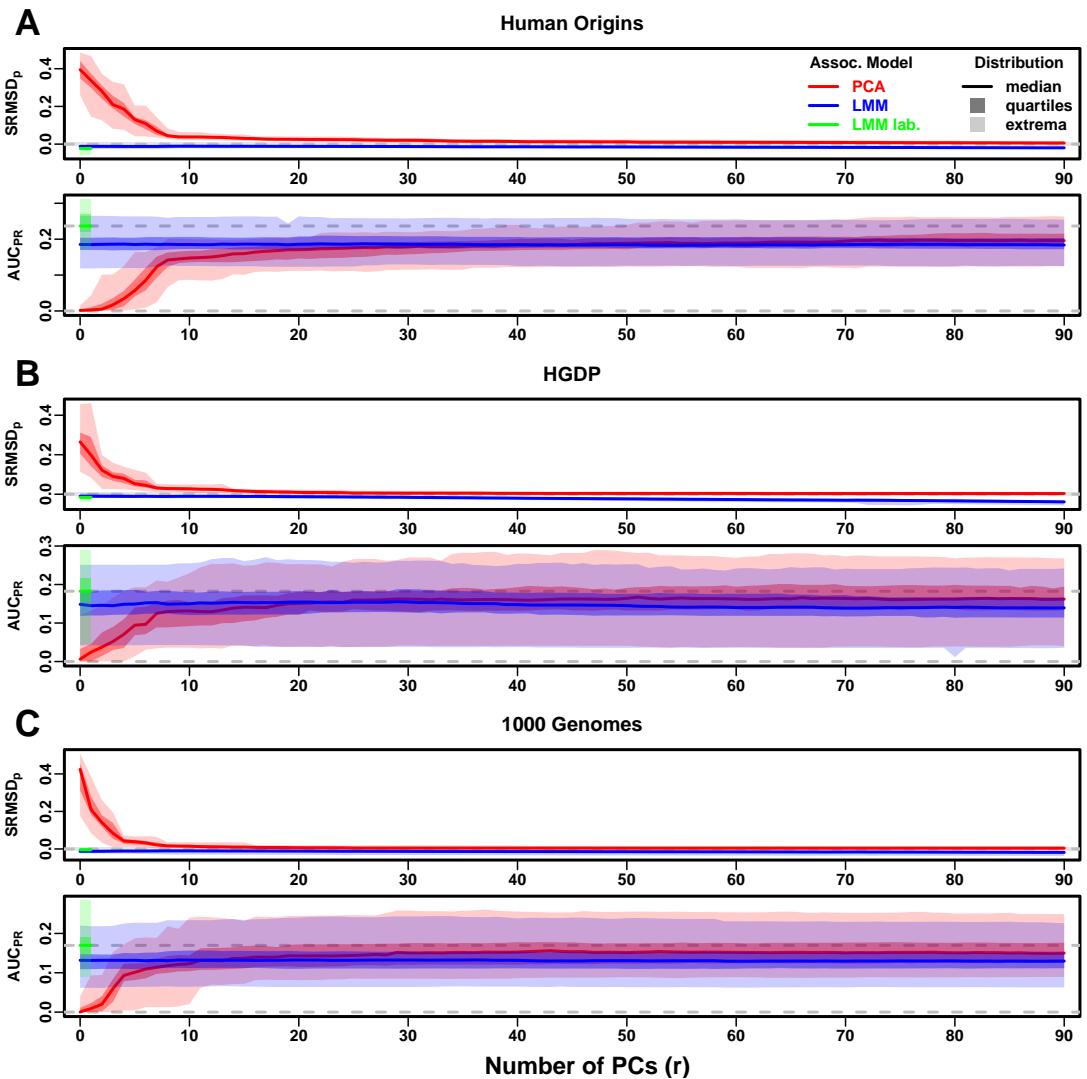
**Figure 4—figure supplement 2.** Evaluations in real human genotype datasets with FES traits, low heritability.



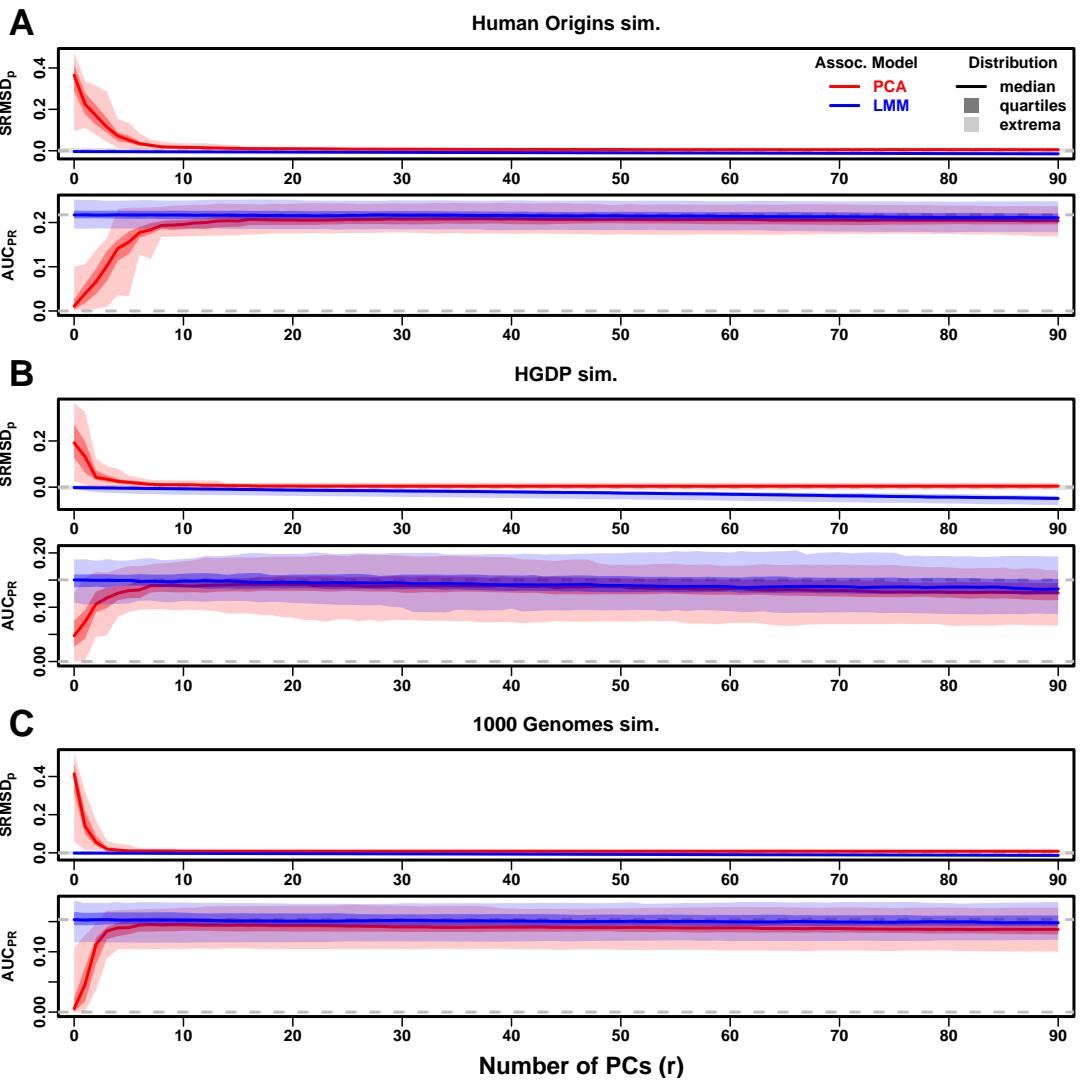
**Figure 4—figure supplement 3.** Evaluations in real human genotype datasets with RC traits, low heritability.



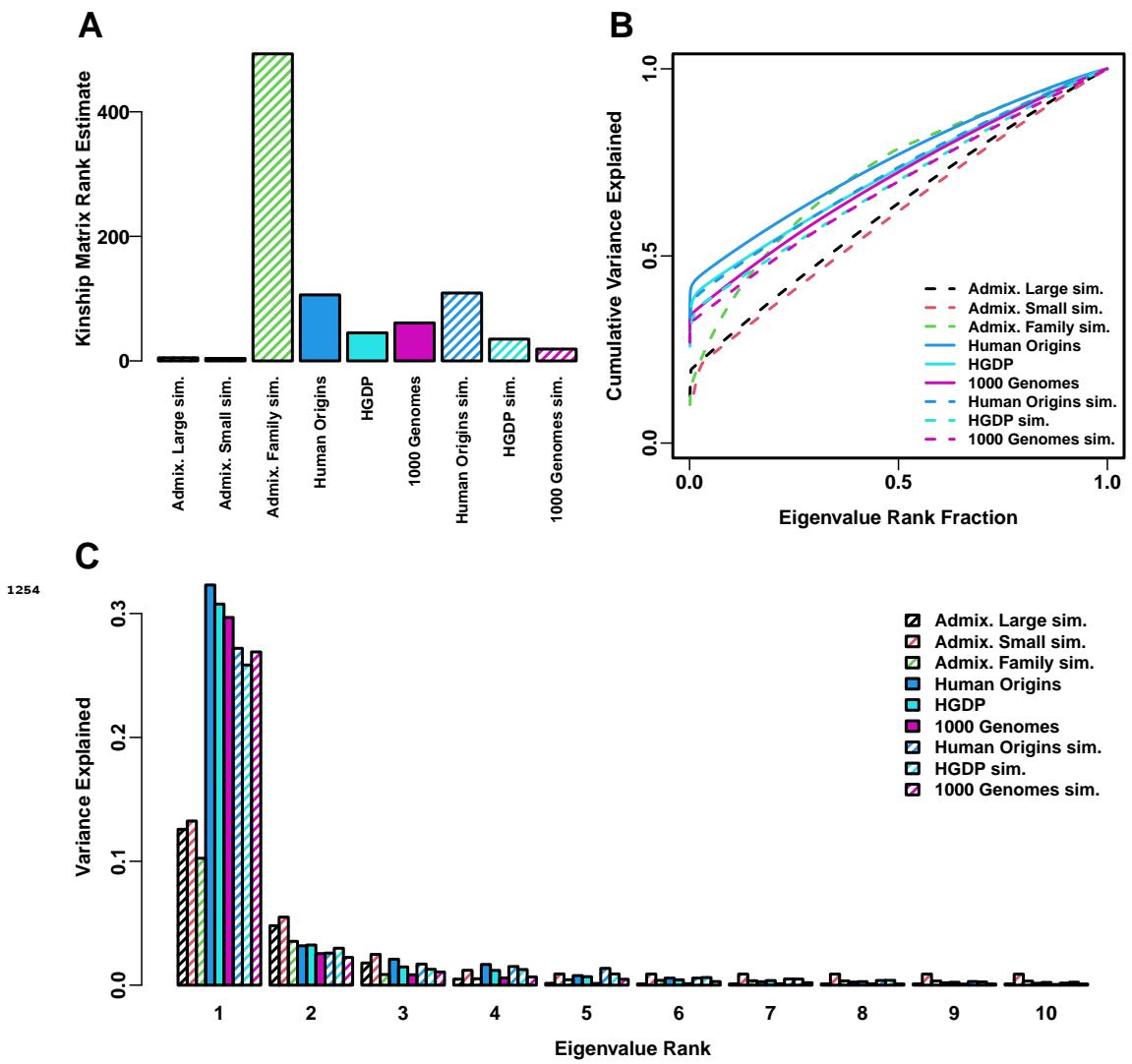
**Figure 4—figure supplement 4.** Evaluations in real human genotype datasets with FES traits, environment. “LMM lab.” was only tested with  $r = 0$ .



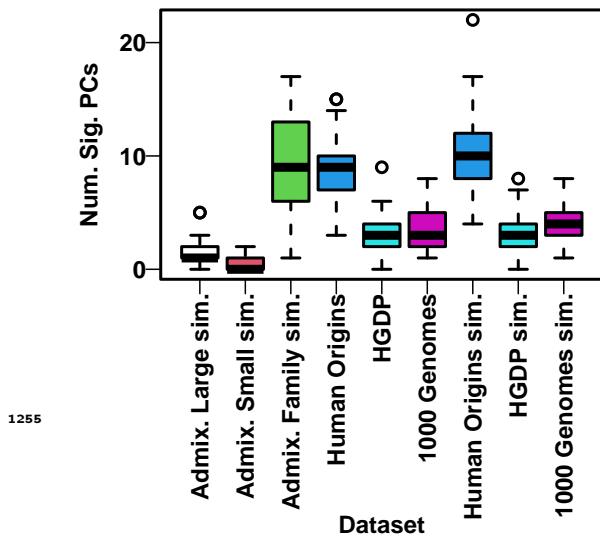
**Figure 4—figure supplement 5.** Evaluations in real human genotype datasets with RC traits, environment. “LMM lab.” was only tested with  $r = 0$ .



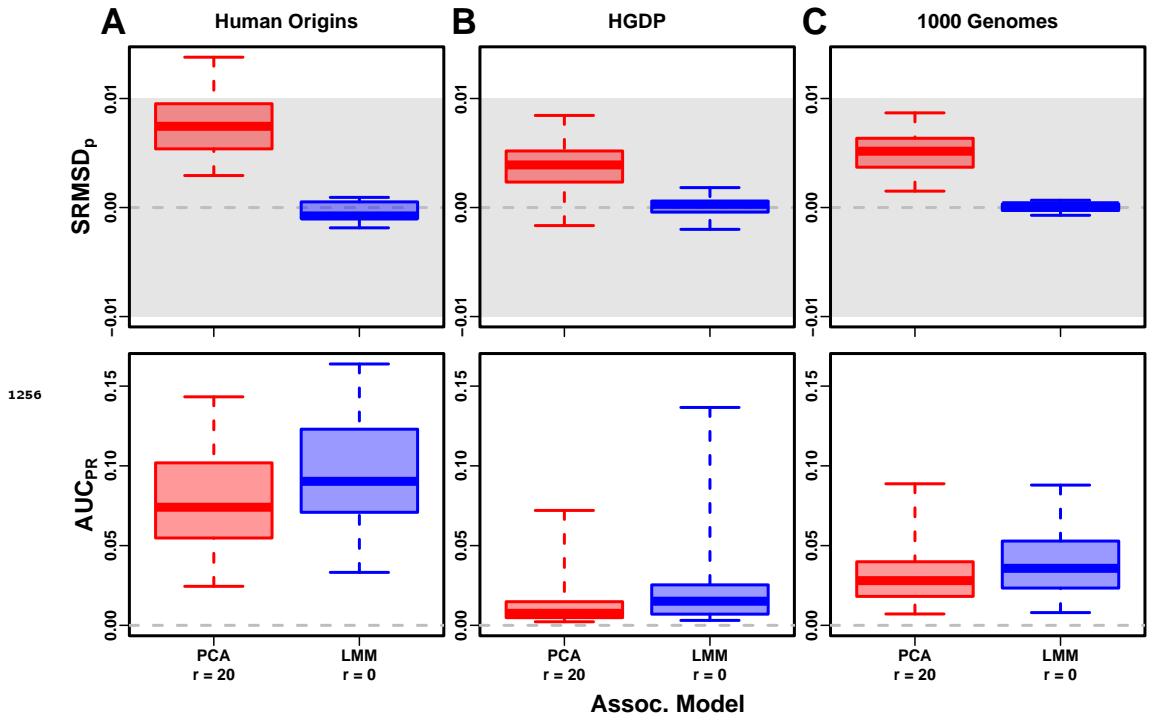
**Figure 5—figure supplement 1.** Evaluations in subpopulation tree simulations fit to human data with RC traits, high heritability.



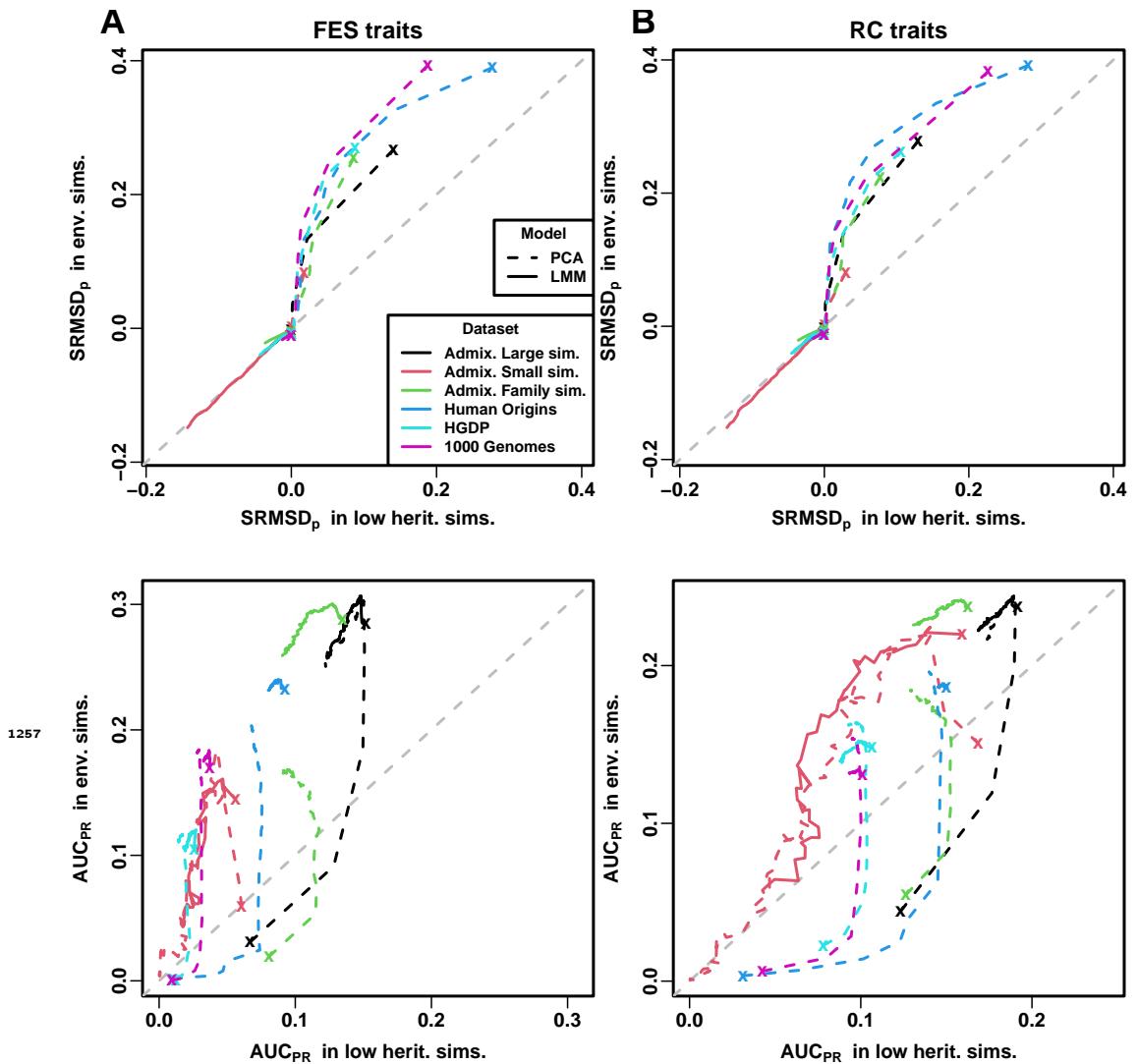
**Figure 6—figure supplement 1.** Estimated relatedness dimensions of datasets. **A.** Kinship matrix rank estimated with the Tracy-Widom test with  $p < 0.01$ . **B.** Cumulative variance explained versus eigenvalue rank fraction. **C.** Variance explained by first 10 eigenvalues.



**Figure 6—figure supplement 2.** Number of PCs significantly associated with traits. PCs are tested using an ordinary linear regression sequentially, with the  $k$ th PC tested conditionally on the previous  $k - 1$  PCs and the intercept. Q-values are estimated from the 90 p-values (one for each PC in a given dataset and replicate) using the R package `qvalue` assuming  $\pi_0 = 1$  (necessary since the default  $\pi_0$  estimates were unreliable for such small numbers of p-values and occasionally produced errors), and an FDR threshold of 0.05 is used to determine the number of significant PCs. Distribution per dataset is over its 50 replicates. Shown are results for FES traits with  $h^2 = 0.8$  (the results for RC were very similar, not shown).



**Figure 7—figure supplement 1.** Evaluation in real datasets excluding 4th degree relatives, FES traits, low heritability.



**Figure 8—figure supplement 1.** Comparison of performance in low heritability vs environment simulations. Each curve traces as the number of PCs  $r$  is increased from  $r = 0$  (marked with an "x") until  $r = 90$  (unmarked end), on one axis is the mean value over replicates of either  $\text{SRMSD}_p$  or  $\text{AUC}_{\text{PR}}$ , for low heritability simulations on the x-axis and environment simulations on the y-axis. Each curve corresponds to one dataset (color) and association model (solid or dashed line type). Columns: **A.** FES and **B.** RC traits show similar results. First row shows that for PCA curves (dashed),  $\text{SRMSD}_p$  is higher (worse) in environment simulations for low  $r$ , but becomes equal in both simulations once  $r$  is sufficiently large; for LMM curves (solid),  $\text{SRMSD}_p$  is equal in both simulations for all  $r$ , all datasets. Second row shows that for PCA,  $\text{AUC}_{\text{PR}}$  is higher (better) in low heritability simulations for low  $r$ , but becomes higher in environment simulations once  $r$  is sufficiently large; for LMM, performance is better in environment simulations for all  $r$ , all datasets.