Limitations of principal components in quantitative genetic association models

Yiqi Yao[1], Alejandro Ochoa[1,2,*]

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

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

[*] Corresponding author: alejandro.ochoa@duke.edu

Abstract

Modern genetic association studies require modeling population structure and family relatedness in order to calculate correct association statistics. Principal Components Analysis (PCA) is an efficient, flexible, and one of the most common approaches for modeling this population structure, but nowadays the Linear Mixed-Effects Model (LMM) is believed by many to be a superior model. Remarkably, previous PCA evaluations have been limited, for example, by not varying the number of principal components (PCs), by simulating unrealistically simple population structures, and by not always measuring both type-I error control and predictive power. The use of LMMs with PCs has also been proposed, but evidence of effectiveness is lacking. In this work, we thoroughly evaluate both PCA and an LMM with varying number of PCs in various realistic genotype and quantitative trait simulation scenarios, including admixture together with family structure, the use of large real human datasets with complex population structures (1000 Genomes Project, the Human Genome Diversity Panel, and Human Origins), and simulations from trees fit to each of these real datasets. Trait simulation permits us to measure both null p-value uniformity and the area under the precision-recall curves. We find that LMM without PCs has the best average performance in all cases. Moreover, PCA is vastly outperformed by LMM in the family simulation, in all of the real human datasets, and the tree simulations. The considerably larger gaps in PCA to LMM performance for the real human datasets compared to the corresponding tree simulations suggests that there is high-dimensional, family-like structure

in these human datasets (beyond the relatedness at the subpopulation level) that PCA is not modeling adequately. Overall, this work better characterizes the limitations of principal components in modeling the complex relatedness structures present in simulated and real multiethnic human data.

1 Introduction

The goal of a genetic association study is to identify loci whose genotype variation is significantly correlated to given trait. An important, implicit assumption made by classical association tests is that, under the null hypothesis, genotypes are unstructured: drawn independently from a common allele frequency. However, this assumption does not hold for structured populations, which includes multiethnic cohorts and admixed individuals, and for family data. When naive approaches are incorrectly applied to structured populations or family data, association statistics (such as chi-squared) become inflated relative to the null expectation, resulting in greater numbers of false positives than expected and loss of power (Devlin and Roeder, 1999; Voight and Pritchard, 2005; Astle and Balding, 2009). Therefore, many specialized approaches have been developed for genetic association in structured data. Here we focus on extensively evaluating the two most popular association models: principal components analysis (PCA) and linear mixed-effects models (LMM).

Overall, many approaches for conducting genetic association studies with structured populations involve modeling the population structure via covariates. Such covariates may be inferred ancestry proportions (Pritchard et al., 2000b) or transformations of these. PCA represents the most common of these variants nowadays, in which the top eigenvectors of the kinship matrix are used to model the population structure (Zhang et al., 2003; Price et al., 2006; Bouaziz et al., 2011). These top eigenvectors are commonly referred to as Principal Components (PCs) in the genetics literature (the convention we adopt here; Patterson et al., 2006), but it is worth noting that in other fields the PCs would instead denote the projections of the data (genotypes) onto the eigenvectors (Jolliffe, 2002). PCs map to ancestry (*e.g.*, Zhou et al., 2016), and they work as well as ancestry in association studies but are estimated more easily (Patterson et al., 2006; Zhao et al., 2007; Bouaziz et al., 2011). An additional strength of PCA is its simplicity, which as covariates can be readily integrated

into more complex models, such as haplotype association (Xu and Guan, 2014). However, PCA fundamentally assumes that relatedness is low-dimensional, which may limit its accuracy in some cases. PCA is known to be inadequate for data containing family structure (Patterson et al., 2006; Thornton and McPeek, 2010; Price et al., 2010), which can be called "cryptic relatedness" when it is unknown to the researchers, but no other specific troublesome relatedness scenarios have been confidently identified. Recent work has focused on developing variants of the PCA algorithm that scale better for large datasets (Lee et al., 2012; Abraham and Inouye, 2014; Galinsky et al., 2016; Abraham et al., 2017; Agrawal et al., 2020). PCA remains a popular and powerful approach for association studies (Wojcik et al., 2019).

The other dominant approach for genetic association studies under population structure is the LMM, in which population structure is a random effect drawn from a covariance model parametrized by the kinship matrix. Unlike PCA, LMM does not assume that relatedness is low-dimensional, and explicitly models family structure via the kinship matrix. Interestingly, LMM and PCA share deep connections (Astle and Balding, 2009; Janss et al., 2012; Hoffman, 2013), which suggest that both models ought to perform similarly under certain conditions such as low-dimensional relatedness. However, many previous simulation studies have found that LMM outperforms PCA (Zhao et al., 2007; Astle and Balding, 2009; Kang et al., 2010; Wu et al., 2011; Song et al., 2015). Other studies find that PCA was less inflated and/or controlled type-I errors better than LMM in certain hypothetical settings, including unusually differentiated markers (Price et al., 2010; Wu et al., 2011), and ??? (Wang et al., 2013), which are an artificial scenario not based on a population genetics model, and are believed to be unusual (Sul and Eskin, 2013). Moreover, various explanations for why LMM might outperforms PCA are vague and have not been tested directly (Price et al., 2010; Sul and Eskin, 2013; Price et al., 2013; Hoffman, 2013). Since LMMs tend to be considerably slower than PCA, it is important to understand when their difference in power or accuracy is outweighed by their difference in runtime.

[TODO haven't incorporated fully maybe: (Thornton and McPeek, 2010)]

The PCA approach has been compared to other approaches previously. However, all of these studies have important limitations, for the most part due to PCA being treated as a competitor

rather than a model worthy of exploring more fully. For example, although there are methods for selecting the numbers of PCs (Patterson et al., 2006), most evaluations either admit to selecting 10 because it has long been the default and it performs well enough, regardless of the dataset in question (Epstein et al., 2007; Li and Yu, 2008; Astle and Balding, 2009; Li et al., 2010; Wu et al., 2011), or otherwise test only one number of PCs, often without justification (Zhang et al., 2003; Kimmel et al., 2007; Zhao et al., 2007; Zhang et al., 2008; Price et al., 2010; Bouaziz et al., 2011; Hoffman, 2013; Wang et al., 2013; Tucker et al., 2014; Yang et al., 2014; Song et al., 2015; Sul et al., 2018). Conversely, only a few studies consider a (small) set of numbers of PCs, where they show remarkable robustness to this choice (Price et al., 2006; Kang et al., 2010; Wojcik et al., 2019). Moreover, most of these evaluations considered simulated data with only $K = 2$ independent subpopulations or admixture from only two subpopulations (exceptions are Astle and Balding (2009) with $K = 3$, and Wu et al. (2011) and Wang et al. (2013) with $K = 4$), although worldwide human population structure is expected to have a larger dimensionality of at least $K = 9$ (Wojcik et al., 2019). Similarly, only two evaluations simulated data from a family pedigree: Price et al. (2010) included sibling pairs, and Thornton and McPeek (2010) included parents, siblings and uncles/aunts. Some studies include evaluations involving real data that featured known or cryptic relatedness, but these analyses did not measure type-I error rates or power calculations, most of which settled for measuring test statistic inflation. Lastly, many of the earlier evaluations employed case-control simulations exclusively (as opposed to quantitative traits as we do here), were based on very small real or simulated datasets relative to today's standards, did not include any LMMs in their evaluations, and often did not measure both type-I error rates and power (or one of their proxies).

Table 1: **Summary of previous evaluations in the literature.**

Publication	n^{a}	m^{a}	K	r	Scen	Sim	F_{ST}	Real	Trait	Causal	Inf	Power	Reps	LMM
Zhang et al., 2003	150	300	4	1	3	IAF	?	AF,N	Q	Y	T	Y	250	N
?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
Thornton and McPeek, 2010	620	100,000	3	10?	6	IAF	0.01	Sc	CC	Y	T	Y	1	N
Price et al., 2010	1,000	100,000	2	1	4	IF	0.01	N	CC	N	I	N	1	YG
Wu et al., 2011	4,000	100,000	2-4	10	5	IA	0.01	N	CC	Y	T	Y	10	YG
This work	2,922	1,185,208	10-243	0-90	9	ATF	0.10	YN	Q	Y	R	Y	50	YG

[a]Max sizes

In this work, we study the performance of the PCA and LMM association models, characterizing their behavior under various numbers of PCs (which are included as fixed covariates in both PCA and LMM). Our evaluation is based on six genotype simulations and three real genotype datasets. The first three simulations consist of an admixture model with $K = 10$ ancestral subpopulations, but which differ in sample size and whether they also feature family structure or not. The real datasets are the 1000 Genomes Project (Consortium, 2010; 1000 Genomes Project Consortium et al., 2012), the Human Genome Diversity Panel (HGDP) (Cann et al., 2002; Rosenberg et al., 2002; Bergström et al., 2020), and Human Origins (Patterson et al., 2012; Lazaridis et al., 2014; Lazaridis et al., 2016; Skoglund et al., 2016). Both 1000 Genomes and HGDP are whole-genome sequencing datasets, whereas Human Origins is based on a genotyping array. Human Origins features the greatest geographical coverage, HGDP is in an intermediate position, and 1000 Genomes features the fewest sampled locations. The last three simulations aim to approximately match each of the real datasets by fitting trees and drawing from the empirical allele frequency distirbutions, to determine whether those features alone recapitulate the observations on the real data or not. In all cases we simulate from two trait models: one with fixed effect sizes (regression coefficients roughly inverse to allele frequency) that approximates estimates in real data (Park et al., 2011) and corresponds to high pleiotropy and strong balancing selection (Simons et al., 2018), which are appropriate assumptions for diseases; and one with random coefficients (independent of allele frequency) that corresponds to neutral traits (Simons et al., 2018). Our evaluation directly measures the uniformity of null p-values (required for accurate type-I error control and FDR control; Storey, 2003; Storey and Tibshirani, 2003) and predictive power by calculating the area under precision-recall curves. Across all tests LMM without PCs consistently performs best. However, in our admixture simulations PCA matches LMM performance when enough PCs are used and there are no close relatives in the study. In reasonably large studies PCA is robust to including far beyond the optimal number of PCs. However, for smaller studies (100 individuals) there is a pronounced loss of power when the number of PCs exceeds the optimal choice. LMMs greatly outperforms PCA in the admixed family simulation, as expected (Patterson et al., 2006; Price et al., 2010). Remarkably, LMM outperforms PCA in all of the real datasets by vast margins. The final three simulations approximately recapitulate both the

complex tree-branching structure and skewed minor allele frequency distributions of the real human data, but recapitulate only part of the gap in PCA to LMM performance, suggesting that additional family-like structure in the real data is the source of the difference in performance. All together, we find that LMMs without PCs are generally preferable, and provide novel simulation and evaluation approaches to measure the performance of these and other genetic association approaches.

2 Results

The success of our investigation hinges on simulating a variety of population structures and quantitative trait models, introduced first, which have the goal of capturing all the essential features present in genetically diverse human studies. Then we summarize the evaluation methods and present the results.

2.1 Overview of genotype simulations and real datasets

We simulated genotypes from six population structure scenarios to cover various features of interest. We also utilized three real genotype datasets. We will introduce them here in sets of three, as they appear in the rest of our results. All simulated and real genotype datasets are summarized in Table 2. The population structures are also conveniently visualized in Fig. 1 using `popkin` to estimate kinship matrices without bias (Ochoa and Storey, 2021).

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

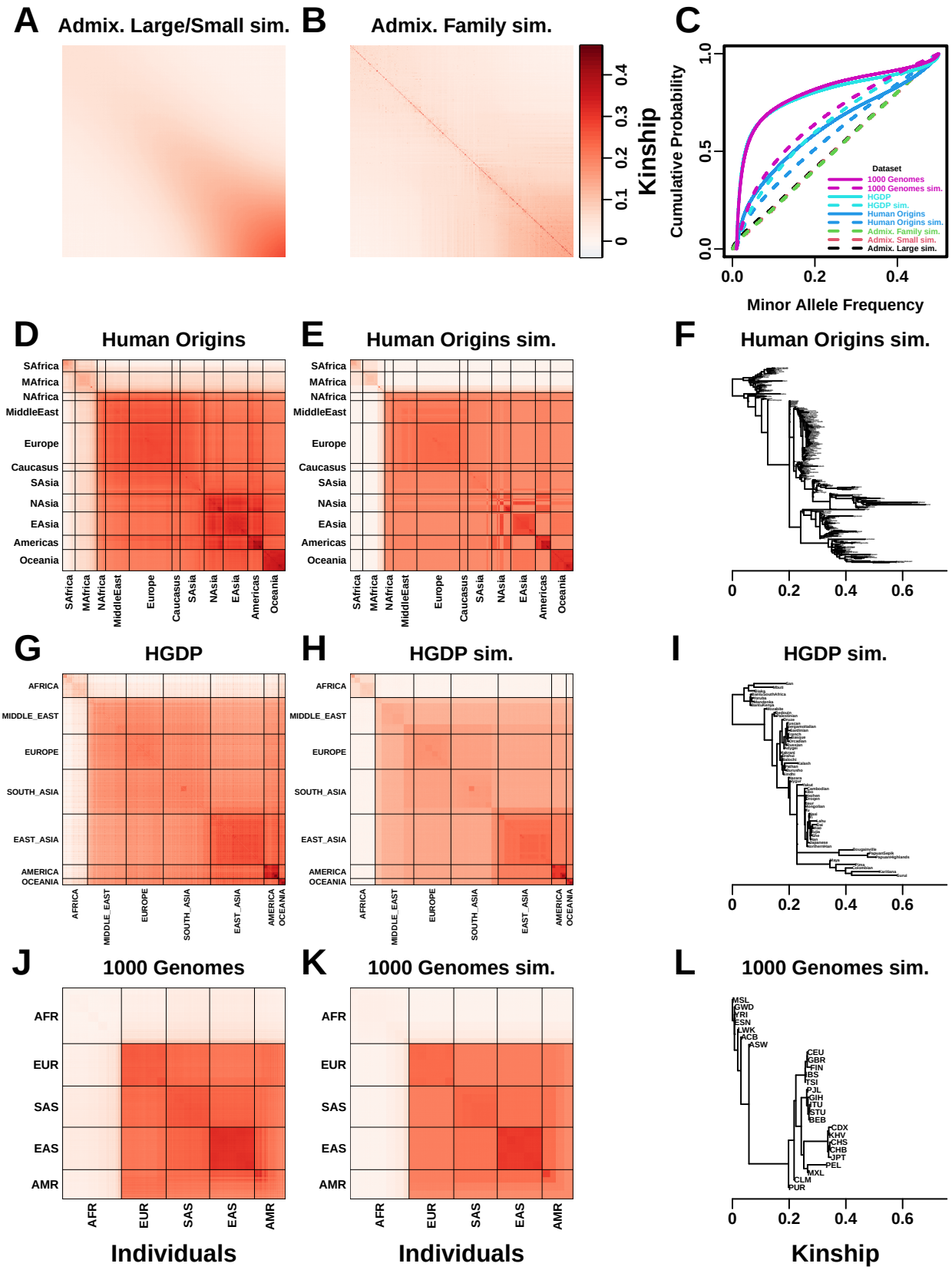
Dataset	Type	Loci (m)	Ind. (n)	Subpops.[a] (K)	Causal loci[b] (m_1)	F_{ST} [c]
Admix. Large sim.	Admix.	100,000	1000	10	100	0.1
Admix. Small sim.	Admix.	100,000	100	10	10	0.1
Admix. Family sim.	Admix.+Pedig.	100,000	1000	10	100	0.1
Human Origins	Real	190,394	2922	11-243	292	0.28
HGDP	Real	924,421	929	7-54	93	0.21
1000 Genomes	Real	1,185,208	2504	5-26	250	0.19
Human Origins sim.	Tree	100,000	2922	243	292	0.23
HGDP sim.	Tree	100,000	929	54	93	0.22
1000 Genomes sim.	Tree	100,000	2504	26	250	0.21

[a]For admixed family, ignores the increase in dimensionality due to 20 generation pedigree structure. For real datasets, lower range is continental subpopulations, upper range is number of fine-grained subpopulations.
[b]$m_1 = n/10$ in all cases to balance power across dataset.
[c]Model parameter for simulations, estimated value on real datasets.

The first set of three simulated genotypes are based on an admixture model from $K = 10$ subpopulations (Fig. 1A) (Ochoa and Storey, 2021; Gopalan et al., 2016; Cabreros and Storey, 2019). The "large" version of this simulation, with 1000 individuals, illustrates the asymptotic performance of the association models (in this sense, it is merely large enough). In contrast, the "small" sample size simulation, with just 100 individuals, illustrates cases of overfitting when using large numbers of PCs. For PCA, the theoretically ideal choice for the number of PCs in this simulation is $K - 1 = 9$ (the rank of the population structure, K, minus the rank of the intercept term, 1). The third starts from an admixed founder population and simulates a 20-generation random pedigree with assortative mating, resulting in a very complex joint family and ancestry structure in the last generation (Fig. 1B).

The second set of three are the real human datasets: Human Origins (Fig. 1D), HGDP (Fig. 1G), and 1000 Genomes (Fig. 1J). All of these represent global human diversity, some in greater resolution than others, making them of great interest as representatives of proposed multiethnic studies. These three datasets had loci filtered to avoid linkage disequilibrium, which both simplified our evaluation and reduced dataset sizes enough to make our large-scale evaluations feasible. Lastly, since real data are enriched for extremely rare variants for which PCA and LMM have no power to detect associations, loci with minor allele frequency under 0.01 were removed. Still, all real dataset remain enriched for lower-than-uniform minor allele frequencies (Fig. 1C).

Figure 1 *(following page)*: **Population structures of simulated and real human genotype datasets.** First two columns are population kinship matrices estimated with `popkin`: Every individual is placed along both x- and y-axes, and the population kinship value of every pair of individuals is visualized as color, where lighter is values closer to zero, and darker red are higher values. Diagonal is inbreeding values. Individuals are divided into continental subpopulations in real datasets. **A.** Admixture scenario, shared by 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 of all datasets. Real datasets and tree simulations had MAF ≥ 0.01 filter. **D.** Human Origins is an array dataset from a large diversity of humans from around the world. **G.** Human Genome Diversity Panel (HGDP) is a WGS dataset from native populations around the world. **J.** 1000 Genomes Project is a WGS dataset sampling cosmopolitan populations around the world. **F,I,L.** Trees between subpopulations fit to real data, used to draw genotypes in simulations. **E,H,K.** Simulations from trees fit to the real data recapitulate well the covariance structure (population kinship) at the subpopulation level.


A
Admix. Large/Small sim.
B
Admix. Family sim.
Kinship
0
0.1
0.2
0.3
0.4
C
Cumulative Probability
0.0
0.5
1.0
0.0
0.2
0.4
Minor Allele Frequency
Dataset
1000 Genomes
1000 Genomes sim.
HGDP
HGDP sim.
Human Origins
Human Origins sim.
Admix. Family sim.
Admix. Small sim.
Admix. Large sim.
D
Human Origins
E
Human Origins sim.
SAfrica
MAfrica
NAfrica
MiddleEast
Europe
Caucasus
SAsia
NAsia
EAsia
Americas
Oceania
F
Human Origins sim.
0
0.2
0.4
0.6
G
HGDP
H
HGDP sim.
AFRICA
MIDDLE_EAST
EUROPE
SOUTH_ASIA
EAST_ASIA
AMERICA
OCEANIA
I
HGDP sim.
0
0.2
0.4
0.6
J
1000 Genomes
K
1000 Genomes sim.
AFR
EUR
SAS
EAS
AMR
Individuals
Individuals
L
1000 Genomes sim.
MSL
GWD
YRI
ESN
LWK
ACB
ASW
CEU
GBR
FIN
IBS
TSI
PJL
GIH
ITU
STU
BEB
CDX
CHV
CHS
CHB
JPT
PEL
MXL
CLM
PUR
0
0.2
0.4
0.6
Kinship

The last set of three are tree-based simulations based on each of the real human datasets. In each case, a tree was fit to the kinship matrix of each dataset averaged over subpopulations (Fig. 1F,I,L), and this tree was used to draw genotypes. The empirical allele frequency distribution of each dataset was transformed to serve as the ancestral allele frequency distribution of the corresponding simulation, to mimic the skew for smaller minor allele frequencies observed in the real datasets (Fig. 1C). Overall, although these simulations do not match the real data exactly (real admixture is not fit well by these trees), fits are remarkably good (Fig. 1E,H,K), resulting in data of comparable covariance structure and scale of differentiation. By design, these subpopulation tree simulations exclude both admixture and relatedness more fine-grained than the subpopulation level. In particular, the potential family structure present in the real dataset is absent in these simulations.

Replicates consisted of as much newly-drawn data as possible. For the simulated genotype datasets, each replicate drew a new genotype matrix (from the same structure model of the scenario). The admixed family simulation additionally drew a new random pedigree for each replicate. For each real dataset, the given genotype matrix is used in every replicate.

2.2 Overview of trait simulation models

We performed all of our tests using two additive quantitative trait models, which we call *fixed effect sizes* and *random coefficients*, respectively. Starting from a given real or simulated genome, both trait simulations pick a given number of random loci to serve as causal loci, but their coefficients (in the assumed linear model) are constructed in two different ways. Complete simulation details are found in the Methods.

The *fixed effect sizes* simulation selects coefficients β_i such that the effect size $2\beta_i^2 p_i^T(1-p_i^T)$ have the same value at every locus i, where p_i^T is the ancestral allele frequency of the simulation. This corresponds with a rough inverse relationship between coefficient and minor allele frequency, which arises under one evolutionary extreme of strong balancing selection (Simons et al., 2018) and has been observed to hold roughly in meta-analysis across several diseases (Park et al., 2011). For these reasons, the results presented in the main figures focus on this trait model, as it more closely resembles disease data.

The *random coefficients* simulation selects random coefficients independently of allele frequency. This corresponds to the other evolutionary extreme, namely neutrality (Simons et al., 2018). Effect size distributions in this simulation are wider, which reduces association power, but overall recapitulates our conclusions from the fixed effect sizes simulation.

2.3 Overview of evaluations

Since our quantitative traits are simulated, true causal loci are known, permitting exact identification of true positives, false positives, and false negatives. We employ two complementary summary measures: (1) SRMSD_p (p-value signed root mean square deviation) measures null p-value uniformity and relates to the accuracy of type-I error control across thresholds (closer to zero is better), and (2) AUC_{PR} (precision-recall area under the curve) measures predictive power (higher is better). The SRMSD_p and AUC_{PR} measures are fully described in the Methods and illustrated in Fig. 2. The SRMSD_p measure is a more robust alternative to the common inflation factor and type-I error measures; a detailed comparison is presented at the end of the results, where we found that $\text{SRMSD}_p > 0.01$ corresponds to an inflation factor > 1.06, and thus evidence of inflation. The AUC_{PR} measure is a more robust alternative to statistical power calculations, which are not meaningful when p-values are not accurate (as is often the case in this investigation). Reducing the complexity of null p-value distributions and precision-recall curves to two scalars is crucial for our extensive evaluations, which consider 0-90 numbers of PCs and 50 replicates for each case.

The overall goal is to characterize the performance of two association models: PCA and LMM. Each of PCA and LMM was evaluated in each dataset while including a number r of PCs as fixed covariates, in both cases varying r between 0 and 90. We determined which value of r was optimal (in terms of SRMSD_p and AUC_{PR} separately) for each of PCA and LMM separately, in each dataset, and lastly compared overall performance per dataset across the best PCA and LMM cases (with their optimal r values). Our overall statistical evaluation is presented in Table 3 and will be summarized first, followed by detailed evaluations in each datasets in the rest of the results.

We first describe the results for null p-value uniformity ($|\text{SRMSD}_p|$; Table 3). Only here the sign of SRMSD_p was ignored, so smaller is better and Wilcoxon paired 1-tailed tests were used

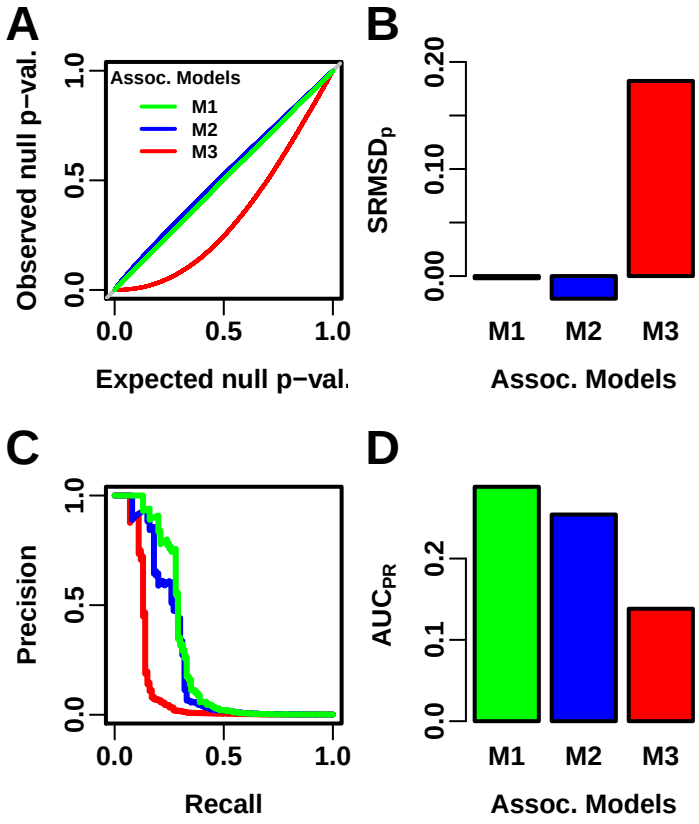


Figure 2: **Illustration of $\mathbf{SRMSD}_p$ and $\mathbf{AUC}_{\mathbf{PR}}$ evaluation measures.** Three archetypal models (M1, M2, M3) illustrate our two complementary measures. M1 is an ideal model that performs best overall, M2 overcorrects for population structure so it incurs a small performance penalty, and M3 does not correct for population structure so it performs most poorly. **A.** Probability-probability plot of the subset of p-values testing "null" (non-causal) loci. M1 has uniform null p-values as desired (overlaps $y = x$). M2/M3 have null p-values larger/smaller than expected. **B.** The SRMSD_p (p-value Signed Root Mean Square Deviation) summarizes null p-value accuracy using a scaled Euclidean distance between the observed null p-values and their uniform expectation, with a negative sign if the median is larger than expected (closer to zero is better). **C.** Precision-Recall plot assesses classification prediction performance across significance thresholds, without assuming that p-values are accurate (only locus ranks matter; higher is better). **D.** The AUC_{PR} (Precision-Recall Area Under the Curve) summarizes predictive performance (higher is better).

to determine whether a suboptimal distribution was significantly different. For PCA, the optimal number of PCs r is typically very large across all datasets (up to $r = 90$, which was the largest value tested), but we found that much smaller "min" r values often performed as well (numbers in parentheses in Table 3 are the smallest r whose $|\text{SRMSD}_p|$ distributions were not significantly different from the distribution of the r with the smallest mean $|\text{SRMSD}_p|$). However, even the min r values for PCA tended to be large on the family simulations and the real datasets, compared to the admixture and tree simulations. In most cases both the best r and the min r had a mean

Table 3: **Overview of PCA and LMM evaluation results**

	Metric:	$\lvert\text{SRMSD}_p\rvert$			AUC_{PR}		
		Best (min[b]) PCs			Best (min[b]) PCs		
Dataset	Trait model[a]	PCA	LMM	Best[c]	PCA	LMM	Best[c]
Admix. Large sim.	FES	84* (3*)	0*	tie	3	3 (0)	LMM
Admix. Small sim.	FES	4* (2*)	0*	LMM	4 (1)	0	LMM
Admix. Family sim.	FES	90 (87)	0*	LMM	83 (34)	0	LMM
Human Origins	FES	90 (87)	0*	LMM	34 (9)	1 (0)	LMM
HGDP	FES	90* (31*)	0*	LMM	17 (15)	1 (0)	LMM
1000 Genomes	FES	51 (34)	0*	LMM	9 (8)	1 (0)	LMM
Human Origins sim.	FES	89* (84*)	0*	PCA	45 (32)	0	LMM
HGDP sim.	FES	43* (16*)	0*	tie	17 (7)	2 (0)	LMM
1000 Genomes sim.	FES	55* (15*)	0*	tie	17 (8)	6 (0)	LMM
Admix. Large sim.	RC	89* (3*)	0*	tie	3	2 (0)	LMM
Admix. Small sim.	RC	8* (2*)	0*	tie (LMM)	1 (0)	0	LMM
Admix. Family sim.	RC	90 (88)	0*	LMM	74 (28)	0	LMM
Human Origins	RC	89* (79*)	0*	LMM	34 (18)	5 (0)	LMM
HGDP	RC	57* (26*)	0*	LMM	19 (8)	3 (0)	LMM
1000 Genomes	RC	68* (39*)	0*	LMM	11 (6)	10 (0)	LMM
Human Origins sim.	RC	90* (77*)	0*	PCA	53 (32)	0	LMM
HGDP sim.	RC	90* (17*)	0*	tie	21 (12)	1 (0)	LMM
1000 Genomes sim.	RC	60* (41*)	0*	LMM	10 (6)	6 (1)	LMM

[a]FES: Fixed Effect Sizes, RC: Random Coefficients.
[b]Smallest r (number of PCs) whose distribution ($|\text{SRMSD}_p|$ or AUC_{PR}) was not significantly different (Wilcoxon paired 1-tailed $p > 0.01$) from the r with best mean value (if any).
[c]Tie if distributions ($|\text{SRMSD}_p|$ or AUC_{PR}) of best PCA and LMM version (previous two columns) did not differ significantly (Wilcoxon paired 1-tailed $p > 0.01$). Result was always the same whether "best" or "min" (in parenthesis) cases were compared, except in one case (in parenthesis).
*r for which mean $|\text{SRMSD}_p| < 0.01$ ($|\text{SRMSD}_p|$ columns only).

$|\text{SRMSD}_p| < 0.01$ (marked with asterisks), indicating small enough effect sizes to consider those null p-value distributions effectively uniform. Mean $|\text{SRMSD}_p| > 0.01$ cases for PCA also tended to be observed on the family simulations and real datasets. In contrast, for LMM $r = 0$ (no PCs) was always the optimal choice (always resulted in the minimum mean $|\text{SRMSD}_p|$), and in those cases we also always had mean $|\text{SRMSD}_p| < 0.01$. Lastly, we compared the $|\text{SRMSD}_p|$ distributions between PCA and LMM, each with their best r, resulting in LMM besting often or in statistical ties, whereas PCA was best in the Human Origins simulations only.

Next we turn to the predictive power evaluations (AUC_{PR}; Table 3). For PCA, the best r for AUC_{PR} was always smaller than the best r for $|\text{SRMSD}_p|$, and also for the respective "min" r comparisons (smallest r which is not significantly different in AUC_{PR} distribution from the best r). So for PCA there is often a tradeoff between the desire for accurate p-values versus maximizing power. For LMM there is no such tradeoff, as $r = 0$ (no PCs) resulted in AUC_{PR} distributions not significantly different from the best r in all tests except one (in the 1000 Genomes simulation with the random coefficients trait model, the min r was 1). Lastly, LMM with its best r always had significantly greater AUC_{PR} distributions than PCA with its best r.

2.4 Evaluations in admixture simulations

Now we look more closely at the results of every individual evaluation. The measured SRMSD_p and AUC_{PR} distributions, for each PCA and LMM and for each value of the number of PCs r, for the first three admixture simulations and the *fixed effect size* trait simulation, are in Fig. 3. We repeated the evaluation with traits simulated from the *random coefficients* model as well, which gave qualitatively similar results (Fig. S1).

The large admixture simulation has $n = 1,000$ individuals, and differs from previous admixture evaluations in featuring a larger number of ancestral populations ($K = 10$) and more differentiation ($F_{\text{ST}} = 0.1$ for the admixed individuals). Admixture is structured over a one-dimensional geography (Ochoa and Storey, 2021). The SRMSD_p of PCA is largest when $r = 0$ (no PCs) and decreases rapidly to zero at $r = 3$, where it stays for up to $r = 90$ (Fig. 3A). Thus, PCA gives effectively accurate p-values for all $r \geq 3$, which is surprisingly smaller than the theoretical optimum we

expected for this simulation of $r = K - 1 = 9$. In contrast, the SRMSD_p distribution for LMM starts near zero for $r = 0$, and as r increases moves away from zero in the negative direction (null test statistics are deflated rather than inflated, so p-values become conservative). The AUC_{PR} distribution of PCA is similarly worst at $r = 0$, increases rapidly and peaks at $r = 3$, then decreases slowly for $r > 3$. Similarly, the AUC_{PR} distribution for LMM starts near its maximum at $r = 0$, and decreases overall for larger r. Although the AUC_{PR} distributions for LMM and PCA overlap considerably for each r, LMM with $r = 0$ has significantly greater AUC_{PR} values than PCA with $r = 3$ (Table 3). However, qualitatively PCA closely matches LMM in performance in this simulation. It is also remarkable how robust both LMM and PCA are to greatly mispecifying r.

The previous robustness to large r led us to consider smaller sample sizes. Our expectation is that a model with large numbers of parameters r should overfit more as r increases, and particularly as r approaches the sample size n (number of individuals). Rather than increase r beyond 90, which is not done in practice, we reduce to $n = 100$ individuals, which is small for typical association studies but may occur in studies of rare diseases, for pilot studies, or be due to low budgets or other constraints. To compensate for the loss of power due to reducing n, we also reduce the number of causal loci from 100 before to $m_1 = 10$, (in all cases a fixed ratio of $n/m_1 = 10$) which increases the magnitude of each individual causal coefficient per locus. As expected, here we found rapid decreases in performance for both PCA and LMM as r increases, with optimal performance attained near $r = 1$ for PCA and $r = 0$ for LMM (Fig. 3B). However, the shift to negative SRMSD_p values for LMM as r increases is much greater in this case. The evaluation declares LMM with $r = 0$ significantly better than PCA ($r = 1$ to 4) in both metrics (Table 3), although qualitatively there is a negligible difference between the two models.

The last of the first three simulations adds a 20-generation random family to our admixture simulation. Previous work has reported, in limited settings, that PCA performs poorly in the presence of family structure, so it is important to establish the detailed behavior of PCA and LMM in this setting as r is varied for both. Since the LMM is formulated in terms of kinship matrices, it is expected to perform better here. Only the last generation is studied for association, which contains numerous siblings, first cousins, etc. The initial population structure due to admixture is preserved

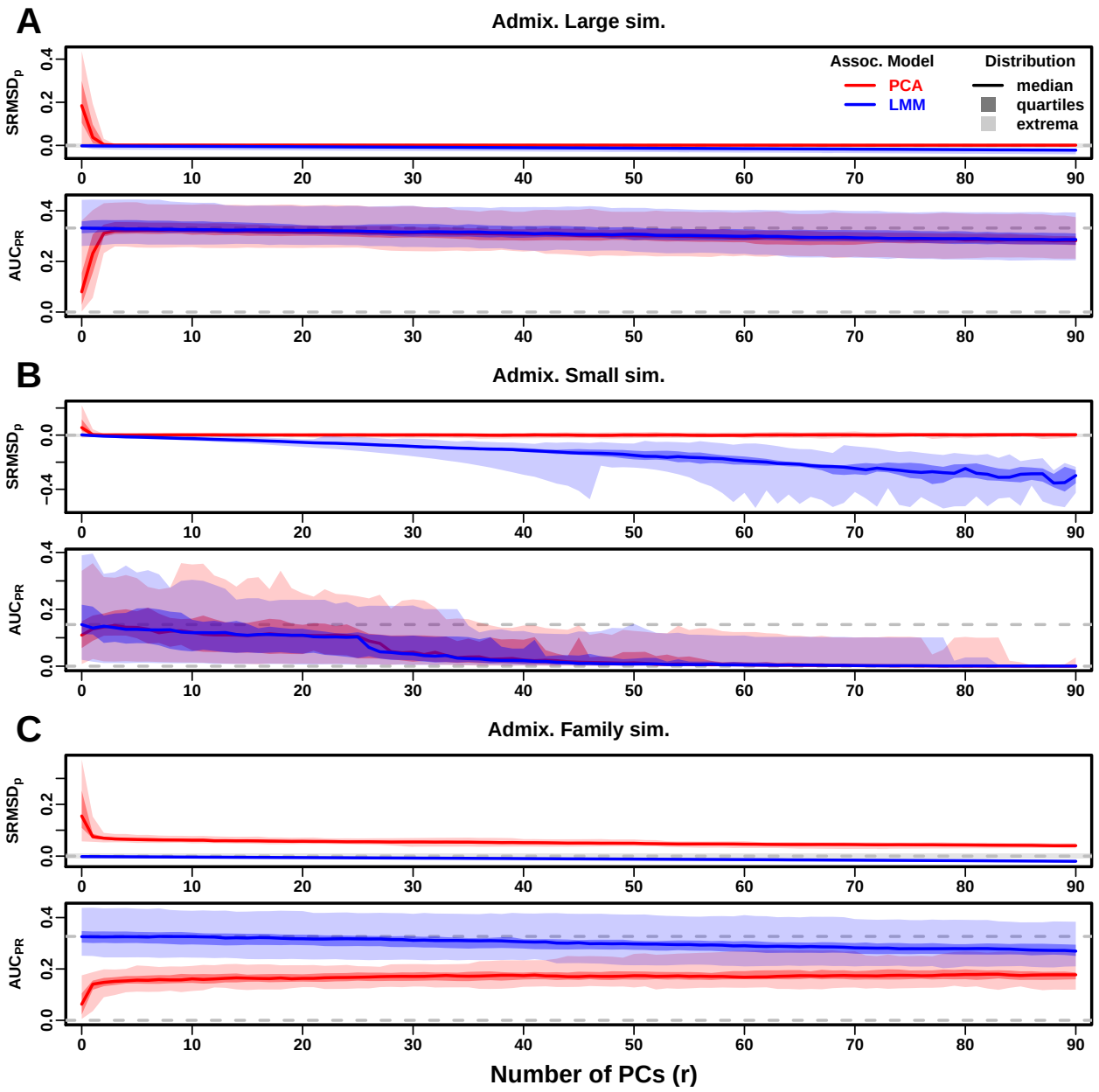


Figure 3: **Evaluations in admixture simulations.** Traits simulated from *fixed effect sizes* model. PCA and LMM approaches are tested under varying number of PCs ($r \in \{0, ..., 90\}$ on x-axis), with the distributions (y-axis) of SRMSD_p (top subpanel) and AUC_{PR} (bottom subpanel) for 50 replicates. Best performance is zero SRMSD_p and large AUC_{PR}. Zero values and maximum median AUC_{PR} values are marked with horizontal gray dashed lines, and the $|\text{SRMSD}_p| < 0.01$ band is marked with a light gray area. LMM always performs best when $r = 0$, and PCA performs best when r is between 1-4. **A.** The large simulation has $n = 1,000$ individuals. **B.** The small simulation has $n = 100$ individuals, helps illustrate overfitting for large r. **C.** The family simulation has $n = 1,000$ individuals from a family with admixed founders and large numbers of pairs of sibling, first/second cousins, etc, from a realistic random 20-generation pedigree. Here PCA performs poorly compared to LMM: $\text{SRMSD}_p > 0$ for all r, and a large gap in AUC_{PR}.

across the generations by strongly biasing mating pairs for proximity over the one-dimensional geography, which results in an indirect ancestry-biased assortative mating (see Methods). Our evaluation reveals a sizable gap in both metrics between LMM and PCA across all values of r (Fig. 3C). Thus, although LMM again performs best with $r = 0$ and achieves mean $|\text{SRMSD}_p| < 0.01$, PCA does not achieve zero SRMSD_p at any r value (all p-values are strongly anti-conservative), and the best mean AUC_{PR} value across r for PCA is worse than the worst mean AUC_{PR} value for LMM. Thus, LMM is conclusively superior to PCA, and the only adequate model, when there is family structure.

2.5 Evaluations in real human genotype datasets

We were next interested in recapitulating our previous results using real human genotype data, which will give the most relevant results in practice, and which differs from our simulations in many ways, including marginal allele frequency distributions, potentially more complex population structures with greater differentiation, correlated and more numerous ancestral subpopulations, as well as the potential presence of cryptic family relatedness. We chose three non-redundant datasets that span global human diversity and include both array and WGS genotyping technologies. Loci in high linkage disequilibrium were removed to simplify our evaluation, and traits were simulated from these genotypes and each of the two trait models: fixed effect sizes (Fig. 4) and random coefficients (Fig. S2).

Among the real datasets, Human Origins has the greatest number and diversity of subpopulations. The SRMSD_p and AUC_{PR} distributions in this dataset and the fixed effect sizes trait model (Fig. 4A) most resemble those from the family simulation (Fig. 3C), which is surprising since close relatives are excluded from this data. In particular, while LMM with $r = 0$ again performed optimally (both metrics) and satisfies mean $|\text{SRMSD}_p| < 0.01$, PCA maintained mean $\text{SRMSD}_p > 0$ for all r values and its AUC_{PR} values were all strictly smaller than even the worst AUC_{PR} values of LMM at any r.

The HGDP dataset has the fewest individuals among real datasets, which are a subset of Human Origins individuals that was recently genotyped with WGS, so it contains many more loci and has

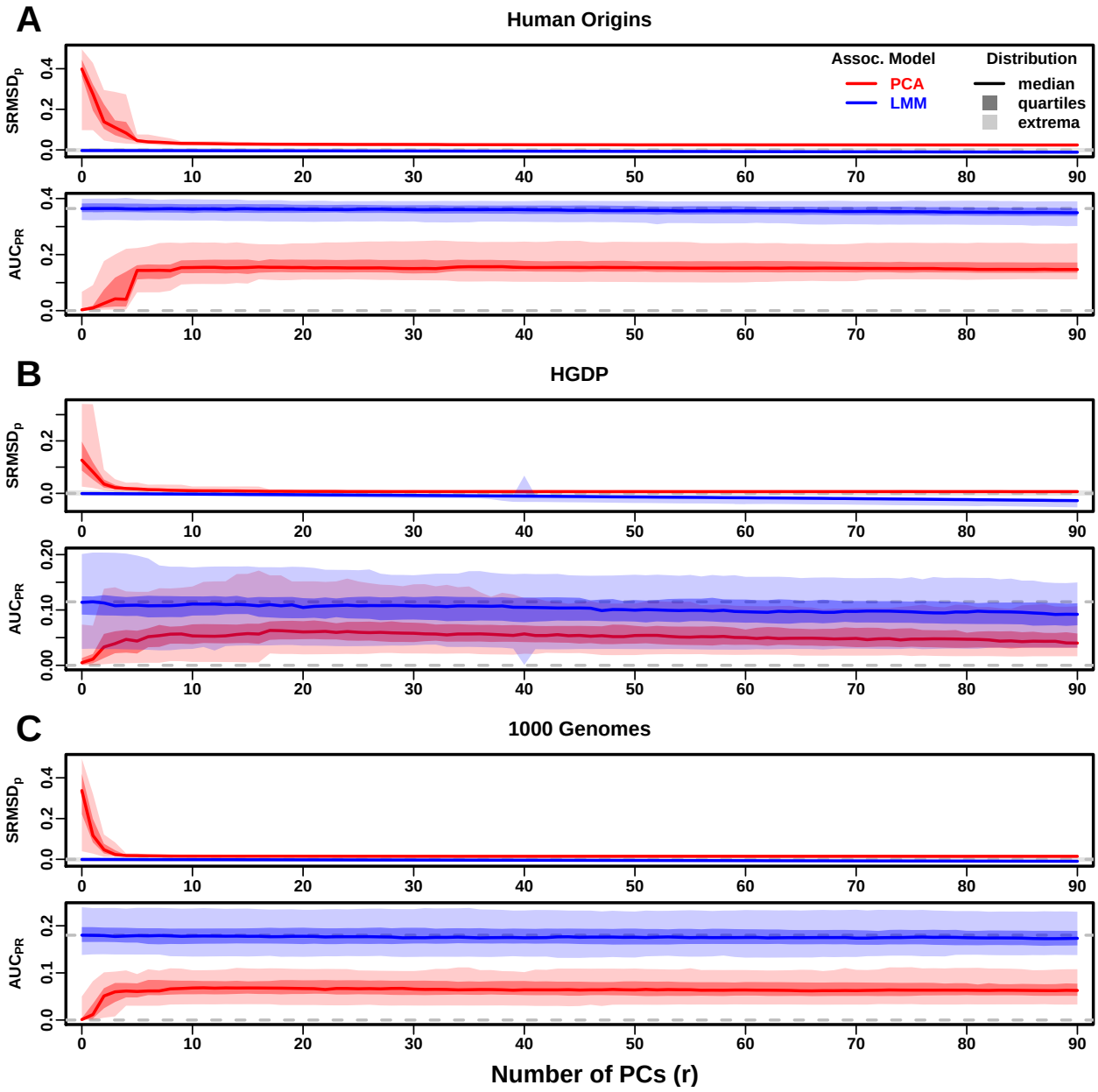


Figure 4: **Evaluations in real human genotype datasets.** Traits simulated from *fixed effect sizes* model. Same setup as Fig. 3, see that for details. These datasets strongly favor LMM with $r = 0$ PCs over PCA, resulting in curves that resemble the previous admixed family simulation, even though these datasets excluded known family members. **A.** The Human Origins dataset. **B.** The Human Genome Diversity Panel (HGDP) dataset. **C.** The 1000 Genomes Project dataset.

more rarer variants. The SRMSD_p and AUC_{PR} distributions (Fig. 4B) are intermediate between the admixture and family simulations. In particular, here both LMM ($r = 0$) and PCA ($r \geq 31$) achieve mean $|\text{SRMSD}_p| < 0.01$, so null p-values will be accurate in both association models. However, there is a sizable mean AUC_{PR} gap between LMM, which performed best across all values of r, and PCA. Maximum AUC_{PR} values were lowest in HGDP compared to the two other human datasets.

1000 Genomes has the fewest subpopulations, but an intermediate number of individuals, compared to the other two real datasets, and like HGDP is also WGS. Thus, although this dataset is expected to have the simplest population structure among the real datasets, we were surprised to find SRMSD_p and AUC_{PR} distributions (Fig. 4C) that again resemble those of our earlier family simulation, with mean $|\text{SRMSD}_p| < 0.01$ for LMM only and large AUC_{PR} gaps between LMM and PCA.

The previous results for the real datasets focused on traits drawn from the fixed effect sizes (FES) model. In this case the results are qualitatively very different for traits drawn from the random coefficients (RC) model (Fig. S2). The key difference is that AUC_{PR} gaps between LMM and PCA, which were very large in FES, are much smaller in RC. Maximum AUC_{PR} were smaller in RC compared to FES in two of the three datasets. SRMSD_p distributions are practically the same in RC versus FES. Nevertheless, our overall statistical evaluations declare LMM with $r = 0$ superior to PCA in both RC and FES traits (Table 3).

2.6 Evaluations in tree simulations fit to human data

To better understand what features of the real datasets lead to the large differences in performance between LMM and PCA, we performed additional simulations. In particular, human subpopulations are related roughly by a tree, which induces the strongest correlations by magnitude (Fig. 1), so we wanted to determine if this tree structure alone could recapitulate our previous results. Thus, we fit trees to each human dataset and verified that the kinship matrices of the simulations were a rough match to those of the real datasets, as desired. The second feature included in these simulations (absent in the original admixture simulations) is a non-uniform ancestral allele frequency distribution, which recapitulated some of the skew for smaller minor allele frequencies of the real

datasets (Fig. 1C).

The SRMSD_p and AUC_{PR} distributions for the tree simulations fit to the human data (Fig. 5) resembled our previous admixture simulation much more than either the family simulation (Fig. 3) or real data results (Fig. 4). In all three simulations, both LMM with $r = 0$ and PCA (various r) achieve mean $|\text{SRMSD}_p| < 0.01$, and in two out of the three cases both association models (with their best r) were tied (not significantly different; Table 3). The AUC_{PR} distributions of both LMM and PCA track closely as r is varied, although there is a small gap in performance that results in LMM ($r = 0$) besting PCA in all three simulations. Lastly, the results are qualitatively similar for traits drawn from the random coefficients model (Fig. S3 and Table 3). Overall, the tree simulations do not recapitulate the large LMM advantage over PCA observed in the previous real human data results.

2.7 Estimated eigenvalues do not explain PCA performance

A first-principles hypothesis for why PCA performs well in some datasets and not in others is their differences in dimensionality, since PCA assumes a low-dimensional genetic structure whereas LMM can model high-dimensional genetic structures. We applied the Tracy-Widom statistical test (Patterson et al., 2006) with $p < 0.01$ to determine the number of significant principal components in each dataset, which estimates the rank of the kinship matrix (i.e., its dimensionality). These kinship ranks (Fig. 6A) slightly underestimated the true dimensionality of our simulations (Table 2). However, rank estimates agree that the admixed family simulation has a greater rank than the admixture-only simulation, and that the real datasets have a much greater rank than the admixture simulations and (to a lesser extent) than the tree simulations that were fit to each real dataset (Fig. 6A). However, these estimated ranks do not differentiate datasets where PCA performs well from those where performance was poor. For example, the ranks of the Human Origins and HGDP tree simulations, were PCA performed relatively well (Fig. 5), are much larger than the rank of the family simulation, where PCA performed very poorly (Fig. 3). Moreover, the 1000 Genomes rank estimate is lower than 90, yet PCA performed poorly for all $r \geq 90$ numbers of PCs tested (Fig. 4). Performance might depend on the ratio of the matrix rank to its dimension (the number

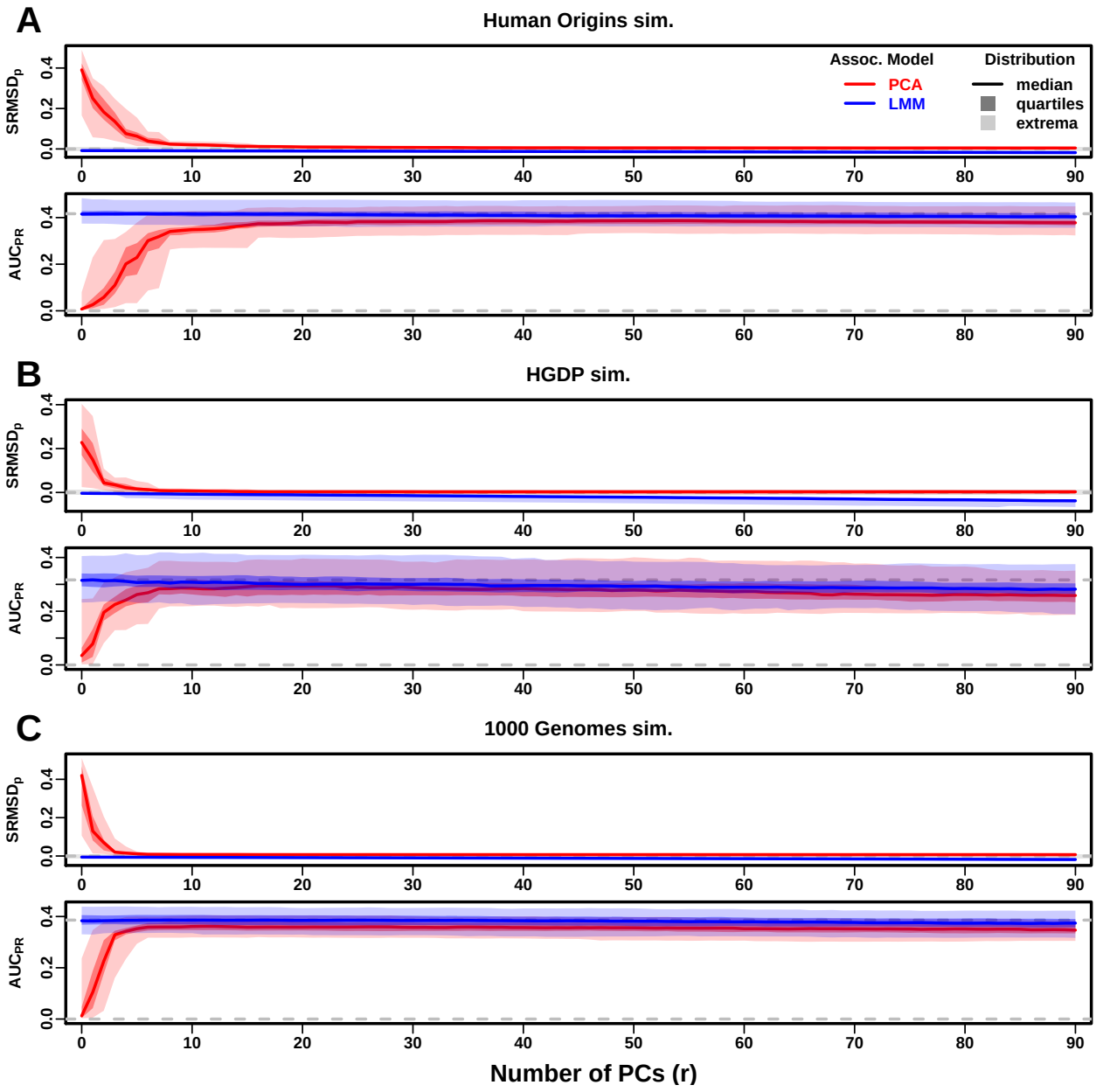


Figure 5: **Evaluations in tree simulations fit to human data.** Traits simulated from *fixed effect sizes* model. Same setup as Fig. 3, see that for details. These tree simulations, which exclude within-subpopulation structure by design, do not explain the large gaps in LMM-PCA performance observed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

of individuals) or a more complicated formula, but the datasets tested do not differ greatly in dimensions (at most 3x, excluding the small simulation), while our analysis does not reveal an obvious line that separates these dataset by PCA performance.

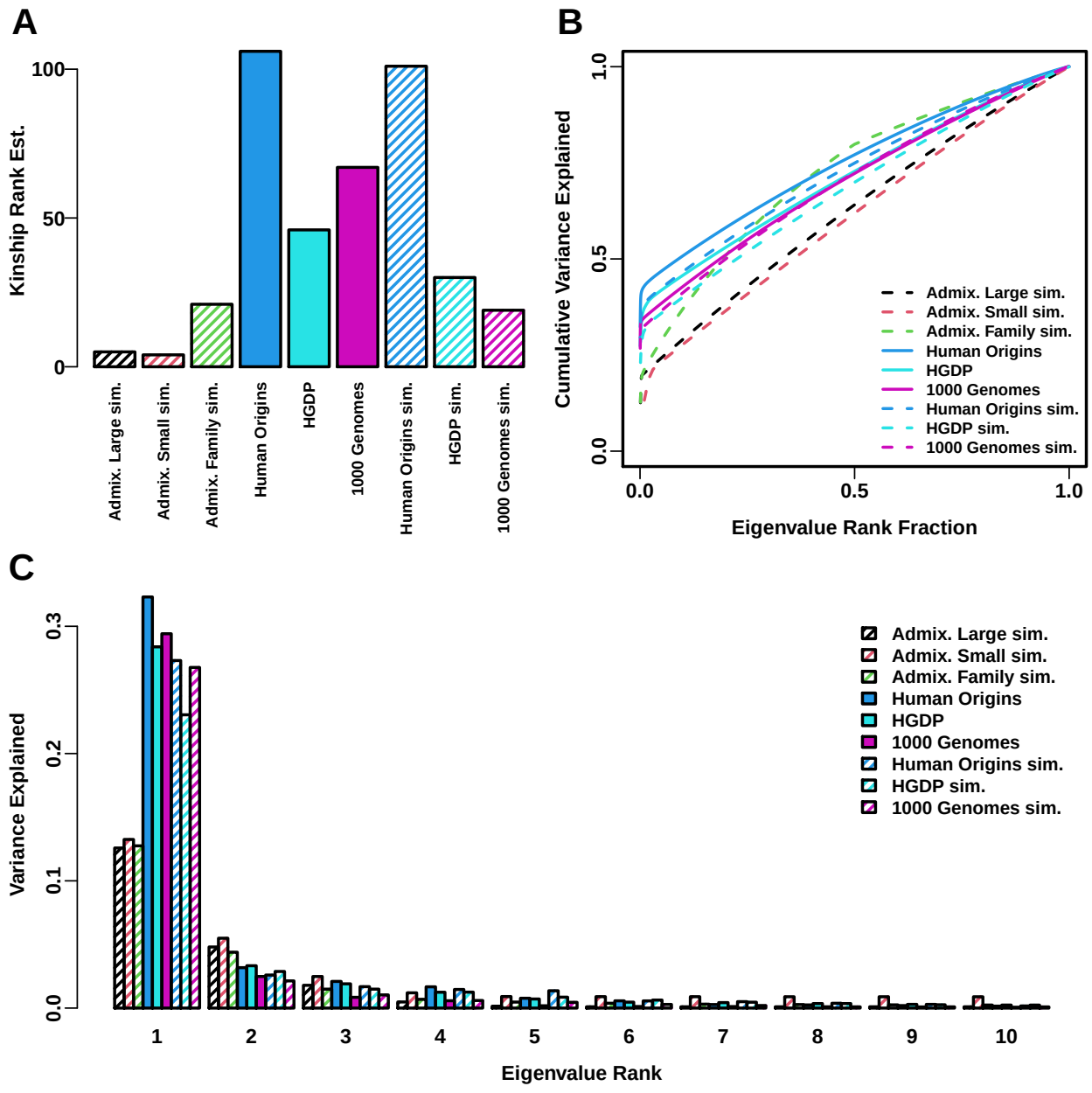


Figure 6: **Estimated dimensionality of datasets. A.** Kinship matrix ranks estimated with the Tracy-Widom test with $p < 0.01$. **B.** Cumulative variance explained versus eigenvalue rank fraction (i/n where i is rank, n is number of eigenvalues/individuals). **C.** Variance explained by first 10 eigenvalues.

We also compared eigenvalues across datasets, expressed as variance explained (each eigenvalue divided by the sum of eigenvalues) to facilitate comparisons across datasets. The top eigenvalue explained a proportion of variance roughly proportional to F_{ST} (Table 2), but the rest of the top 10 eigenvalues show no large differences between datasets (Fig. 6C), except the small admixture simulation had larger variances explained per eigenvalue (as expected since it has fewer of them). We also visualized all eigenvalues (beyond the top 10), computing their cumulative variance explained distributions versus their eigenvalue rank fraction (normalized to account for dataset sample size differences). Each dataset has a different starting point, but all increase almost linearly from there until they reach 1, except for the admixed family simulation, which has much greater variance explained by mid-rank eigenvalues (Fig. 6B). However, there are again no obvious clues separating datasets where PCA performed poorly (such as the real datasets) from those where it performed relatively well (such as the corresponding tree simulations).

2.8 Comparison between SRMSD$_p$ and inflation factor

Now that our main evaluation of the PCA and LMM models is concluded, we switch gears to a comparison of our new SRMSD$_p$ measure and the more traditional inflation factor common in the field. The inflation factor λ measures test statistic inflation, which is a way to measure total or residual population structure (Price et al., 2010). We measured both measures in all of our evaluations, which enables us to discover a correspondence which we fit to all of our data.

Across our tests, we computed inflation factors by mapping median p-values back to their χ^2 quantiles and comparing those values to their expectations under the null hypothesis (see Methods). Remarkably, there appears to be a near one-to-one correspondence between these inflation factors λ and our SRMSD$_p$ statistics (Fig. 7). The curves for PCA and LMM models differ in sign: PCA tended to be inflated ($\lambda > 1$ and SRMSD$_p > 0$) whereas LMM tended to be deflated ($\lambda < 1$ and SRMSD$_p < 0$; not shown). Other than that, it appears that the statistics for both PCA and LMM fall on the same contiguous curve. One immediate advantage of SRMSD$_p$ over λ is that the former does not take on as large values as the latter, which we had to plot on a log scale in Fig. 7.

We fit a sigmoidal curve to this data in order to further characterize the connection between λ

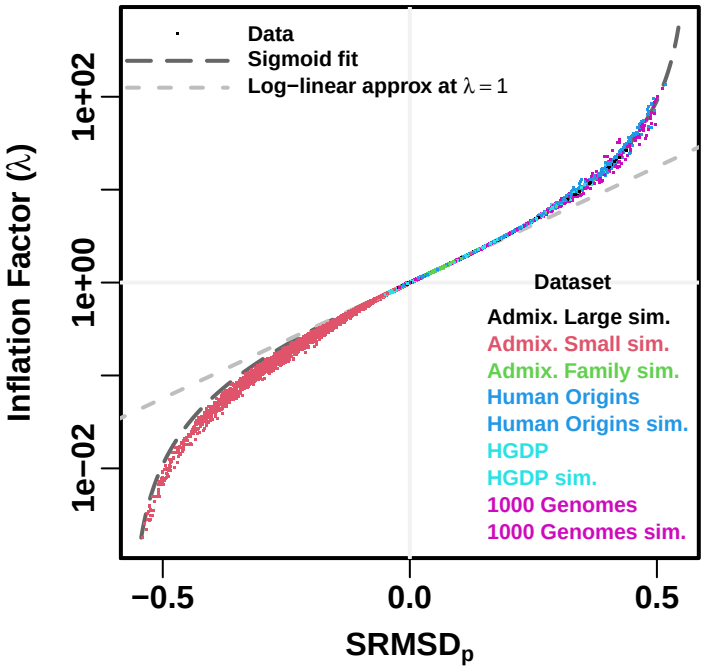


Figure 7: **Comparison between SRMSD$_p$ and inflation factor.** These statistics were pooled from every number of PCs r of both PCA and LMM models, both trait models, and every dataset (color coded by dataset). Note y-axis (λ) is on a log scale, while x-axis (SRMSD$_p$) is linear scale. The sigmoidal curve corresponds to Eq. (1), and the log-linear approximation to Eq. (2).

and SRMSD_p. The family of functions we considered was

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

which for all values of the parameters $a > 0$ and $b > 0$ satisfies $\text{SRMSD}_p(\lambda = 1) = 0$ and reflects $\log(\lambda)$ about zero ($\lambda = 1$) as desired from visual inspection, namely that

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

We fit this model to the upper portion of the data only ($\lambda > 1$), since it was less noisy and of greater interest, and obtained the curve shown in Fig. 7 with parameters $a = 0.566$ and $b = 0.616$. Using this model, we also produced a log-linear approximation based on its Taylor series with respect to $x = \log(\lambda)$ about $x = 0$, resulting in

$$\text{SRMSD}_p(\lambda) \approx \frac{ab}{2}\log(\lambda), \tag{2}$$

which is also shown in Fig. 7. The value $\lambda = 1.05$, a threshold typically used to determine that there is no inflation (Price et al., 2010), corresponds to $\text{SRMSD}_p = 0.0085$ according to both Eqs. (1) and (2). Conversely, $\text{SRMSD}_p = 0.01$, serving as a simpler rule of thumb, corresponds to $\lambda = 1.06$ according to both formulas too.

3 Discussion

Our evaluations conclusively determined that LMM without PCs performs better than PCA (for any number of PCs) across all scenarios, including all real and simulated genotypes and two trait simulation models. Although the addition of a few PCs does not greatly hurt the performance of LMM (except when sample sizes are very small), such additions never resulted in significantly improved perfomance either (barring one marginally significant case with a small effect size; Table 3), which contradict some previous limited observations (Zhao et al., 2007; Price et al., 2010). Our findings make sense since the PCs are the eigenvectors of the kinship matrix used to model the

random effects, so including both is redundant.

Previous work also suggested that PCA can outperform LMM when [TODO: harmonize with intro] there are loci under selection or otherwise highly differentiated (Price et al., 2010; Wu et al., 2011; Yang et al., 2014). Our evaluations on real human data, which presumably contain such loci in realistic proportions, if they are realistic scenarios at all, do not replicate those observations. However, the probable presence of cryptic relatedness on all of these datasets (see below), which favors LMM, may obscure the expected effect. Therefore, while we are not able to completely dismiss this potential PCA advantage, it probably plays a minor role in human studies.

Relative to LMM, the behavior of PCA fell into two extremes. When PCA performed well, there was a (typically small) number of PCs that resulted in both near zero mean SRMSD_p and mean AUC_{PR} near that of LMM without PCs. Conversely, when PCA performed poorly, no choice for the number of PCs led to either acceptably low SRMSD_p or acceptably large AUC_{PR}. PCA performed well in the admixture simulations (without families, both trait models), real human genotypes with random coefficients traits, and, to a lesser extent, the tree simulations (both trait models). Conversely, PCA performed poorly in the admixed family simulation (both trait models) and the real human genotypes with fixed effect sizes traits.

PCA makes an inherent assumption that genetic structure is low-dimensional, whereas LMM can handle high-dimensional structures too. Thus, PCA performs well in the admixture simulation, which is explicitly low-dimensional (see Methods), as well as in tree simulations with few nodes or few long branches, such as the trees we fit to the real human data, where a low-dimensional approximation is sufficient. Conversely, PCA performs poorly under family structure because its kinship matrix is high-dimensional. One theoretical inconvenience is that true kinship matrices are always full rank: for example, an unstructured population where all individuals are equally unrelated and outbred has a kinship matrix of $\mathbf{I}/2$, whose eigenvalues are all equal to $1/2$. Nevertheless, population structure induces a more unbalanced eigenvalue distribution with a few very large eigenvalues (Fig. 6), so we may define dimensionality in practice as the number of eigenvalues that exceed some small threshold. However, evaluating the dimensionality of real datasets is challenging because estimated kinship/covariance matrices result in noisy eigenvalues with skewed distributions. We used

the Tracy-Widom test to estimate dimensionality (Patterson et al., 2006), which gives estimates coherent with the simulation models, although it slightly underestimated their dimensionality (as expected since some non-zero eigenvalues are too small to be significant at the given sample size, and are potentially also less important in practice as evident in our evaluations, as PCA often also performed best with many fewer PCs than the true simulation rank). This analysis confirms that there is considerable structure in real human datasets, which have ranks that exceed the typical 10 PC used in practice and partly explains why LMM performs much better in those cases. However, estimated eigenvalues and kinship matrix ranks by themselves do not fully explain when PCA will perform unacceptably poorly. An additional complication that our evaluations reveal is that the model relating the trait to the genotypes also determines the relative performance of PCA, so genotype-based eigenvalues alone cannot tell the full story.

The real human genotype results, which are the most relevant in practice, suggests that PCA is at best underpowered relative to LMMs, and at worst produces inflated statistics regardless of the numbers of PCs included. Among our simulations, such poor performance with the same features was observed only in the admixed family simulation, so our hypothesis is that cryptic relatedness explains our observations. The very large differences in rank between each real dataset and its tree simulation, particularly for 1000 Genomes and HGDP, support our hypothesis that there is considerable high-dimensional relatedness not captured by the tree model. Admixture is not modeled in our tree simulations, but our other admixture simulations concluded that this feature by itself is not problematic for PCA, so its exclusion should not affect PCA performance. Therefore, by elimination, our analysis again points to cryptic relatedness as the culprit for poor PCA performance in real datasets. Although each of those real human studies excluded individuals known to be related, reanalysis of 1000 Genomes has confirmed the presence of hundreds of close relative pairs (a few as close as siblings or parent-children; Gazal et al., 2015; Al-Khudhair et al., 2015; Fedorova et al., 2016; Schlauch et al., 2017). However, it is unclear which scenarios lead to worse PCA performance, for example, between a few highly related individuals (which are easy to identify and remove) versus a large number of more distantly related pairs. Nevertheless, cryptic relatedness is expected to be prevalent in any large human dataset (Henn et al., 2012; Shchur and

Nielsen, 2018). In view of this evidence, it appears that the challenges of cryptic relatedness are exacerbated when rare variants have large coefficients, as they do in our fixed-effect-sizes trait model. Thus, the high-dimensionality induced by cryptic relatedness appears to be the key challenge for PCA-based association in modern datasets that is readily overcome by LMM.

Minor conclusions follow. Our extensive evaluation also determined that PCA is robust to using a large number of PCs, often far beyond the optimal choice, which agrees with previous anecdotal observations (Price et al., 2006; Kang et al., 2010). This is in contrast to using too few PCs, for which there is a large performance penalty. The only exception was the small simulation, where choosing just the right number of PCs was critical for performance. Thus, when simulations such as ours are not feasible, it is best to err on the side of including too many PCs rather than risk including too few. In this sense, LMM is also more straightforward as users need not select any parameters such as the number of PCs, which to do right requires additional work or knowledge. Conversely, we found that if an LMM has a large number of covariates relative to its sample size (in our case PCs, though we expect this to generalize) then p-values become too conservative/deflated (SRMSD_p was often negative across datasets when $r = 90$), which is a weakness of LMM's use of the likelihood ratio test and its asymptotic χ^2 distribution, which PCA overcomes with the more accurate t-test. Post-hoc evaluations, or simulations such as ours, remain important in all cases to ensure that statistics are as expected, and may help decide whether to use PCA or LMM in a given setting.

Overall, our results force us to always recommend the use of LMM over PCA. Although PCA offer flexibility and speed compared to LMM, additional work is required to ensure that PCA is adequate, including identifying close relatives for exclusion (lowering sample size and resulting in wasted resources) followed by simulations or other evaluations of the output statistics, and even then, without also running LMM there is no guarantee that PCA performance will be close enough to the improved power of LMM we observed in all of our evaluations. Our findings also suggest that other applications that employ PCA to control for population structure, such as polygenic risk scores (Qian et al., 2020), may enjoy gains in power by instead employing an LMM or some other high-dimensional population structure model capable of jointly modeling population structure and

cryptic relatedness.

4 Models and Methods

4.1 Models for genetic association studies

In this subsection we describe the complex trait model and kinship model that motivates both the PCA and LMM models for genetic association studies, followed by further details regarding the PCA and LMM approaches. The derivations of the PCA and LMM models from the general quantitative trait model are similar to previous presentations (Astle and Balding, 2009; Janss et al., 2012; Hoffman, 2013), but we emphasize the kinship model for random genotypes as being crucial for these connections, and make a clear distinction between the true kinship matrix and its most common estimator, which is biased (Ochoa and Storey, 2021; Ochoa and Storey, 2019).

4.1.1 The complex trait model and PCA approximation

Let $x_{ij} \in \{0, 1, 2\}$ be the genotype at locus i for individual j, which counts the number of reference alleles. Suppose there are n individuals and m loci, $\mathbf{X} = (x_{ij})$ is their $m \times n$ genotype matrix, and $\mathbf{y}$ is the length-n (column) vector which represents trait value for each individual. The approaches we consider are based on the following additive linear model for a quantitative (continuous) trait:

$$\mathbf{y} = \mathbf{1}\alpha + \mathbf{X}^\intercal\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{3}$$

where $\mathbf{1}$ is a length-n vector of ones, α is the scalar intercept coefficient, $\boldsymbol{\beta}$ is the length-m vector of locus coefficients, $\boldsymbol{\epsilon}$ is a length-n vector of residuals, and the $\intercal$ superscript denotes matrix transposition. The residuals are assumed to follow a normal distribution: $\epsilon_j \sim \text{Normal}(0, \sigma^2)$ independently for each individual j, for some residual variance parameter σ^2.

Typically the number of loci m is in the order of millions while the number of individuals n is in the thousands, or in other words, $m \gg n$. Hence, the full model above cannot be fit in this typical case, as there are more parameters ($m + 1$, the length of $\boldsymbol{\beta}$ and α) than datapoints (n, the length

of $\mathbf{y}$) to fit. The PCA model with r PCs approximates the full model fit at a single locus i:

$$\mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta_i + \mathbf{U}_r\boldsymbol{\gamma}_r + \boldsymbol{\epsilon}, \tag{4}$$

where $\mathbf{x}_i$ is the length-n vector of genotypes at locus i only, β_i is the coefficient for that locus, $\mathbf{U}_r$ is an $n \times r$ matrix of PCs, and $\boldsymbol{\gamma}_r$ is the length-r vector of coefficients for the PCs. This approximation is explained by first noticing that the genotype matrix has the following singular value decomposition: $\mathbf{X}^\intercal = \mathbf{U}\mathbf{D}\mathbf{V}^\intercal$, where assuming $n < m$ we have that $\mathbf{U}$ is an $n \times n$ matrix of the left singular vectors of $\mathbf{X}$, $\mathbf{V}$ is an $m \times n$ matrix of its right singular vectors, and $\mathbf{D}$ is an $n \times n$ diagonal matrix of its singular values. Thus, in the full model we have $\mathbf{X}^\intercal\boldsymbol{\beta} = \mathbf{U}\boldsymbol{\gamma}$, where $\boldsymbol{\gamma} = \mathbf{D}\mathbf{V}^\intercal\boldsymbol{\beta}$ is a length-n vector. The approximation consists solely of replacing $\mathbf{U}\boldsymbol{\gamma}$ (the full set of n left singular vectors and their coefficients) with $\mathbf{U}_r\boldsymbol{\gamma}_r$ (the top r singular vectors only, which constitutes the best approximation of rank r). Thus, the extra terms in the PCA model approximate the polygenic effect of the whole genome, and assumes that the locus i being tested does not contribute greatly to this signal.

Statistical significance. The null hypothesis is $\beta_j = 0$ (no association). The null and alternative models are each fit (fitting the coefficients of the multiple regression, where β_j is excluded under the null while it is fit under the alternative). The resulting regression residuals are compared to each other using the t-test, yielding a two-sided p-value. Note that many common PCA implementations trade this t-test for a less accurate χ^2 test, which requires the overall degrees of freedom of the model to be much smaller than the number of individuals.

4.1.2 Kinship model for genotypes

In order to better motivate the most common estimation procedure of PCs for genotype data, and to connect PCA to LMMs, we shall review the kinship model for genotypes. The model states that genotypes are random variables with a mean and covariance structure given by

$$\mathrm{E}[x_{ij}|T] = 2p_i^T, \qquad \mathrm{Cov}(x_{ij}, x_{ik}|T) = 4p_i^T(1 - p_i^T)\varphi_{jk}^T,$$

where T denotes the ancestral population quantities are conditioned upon, p_i^T is the ancestral allele frequency at locus i, and φ_{jk}^T is the kinship coefficient between individuals j and k (Malécot, 1948; Wright, 1951; Jacquard, 1970). Thus, if we standardize the genotype matrix using the true ancestral allele frequencies p_i^T, as

$$\mathbf{X}_S = \left(\frac{x_{ij} - 2p_i^T}{\sqrt{4p_i^T \left(1 - p_i^T\right)}} \right),$$

then this results in a straightforward kinship matrix estimator:

$$\mathrm{E}\left[\frac{1}{m} \mathbf{X}_S^\intercal \mathbf{X}_S \right] = \mathbf{\Phi}^T,$$

where $\mathbf{\Phi}^T = (\varphi_{jk}^T)$ is the $n \times n$ kinship matrix (do not confuse the ancestral population superscript T with the matrix transposition symbol $\intercal$). Replacing the raw genotype matrix $\mathbf{X}$ with the standardized matrix $\mathbf{X}_S$ in the trait model of Eq. (3) results in an equivalent model, as this covariate differs only by a linear transformation. Thus, under the standardized genotype model, the PCs of interest are equal in expectation to the top eigenvectors of the kinship matrix.

4.1.3 Estimation of principal components from genotype data

In practice, the matrix of principal components $\mathbf{U}_r$ in Eq. (4) is calculated from an estimate of the earlier standardized genotype matrix $\mathbf{X}_S$, namely

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

where the true ancestral allele frequency p_i^T is replaced by the estimate $\hat{p}_i^T = \frac{1}{2n} \sum_{j=1}^n x_{ij}$, and results in the kinship estimate

$$\hat{\mathbf{\Phi}}^T = \frac{1}{m} \hat{\mathbf{X}}_S^\intercal \hat{\mathbf{X}}_S. \tag{5}$$

This kinship estimate and minor variants are also employed in LMMs (Yang et al., 2011). This estimator of the kinship matrix is biased, and this bias is different for every individual pair (Ochoa and Storey, 2021; Ochoa and Storey, 2019). However, in regression-based genetic association models

such as PCA and LMM, the existing approach performs as well as when the above estimate is replaced by the true kinship matrix (data not shown). The explanation, briefly, is that the biased expectation of the above estimator differs from the true kinship matrix by a rank-1 update, which is exactly compensated for by the intercept term $\mathbf{1}\alpha$ in Eq. (4). [TODO: cite BIAS GWAS]

4.1.4 Connection between PCs and ancestry proportions

Here we show that genetic association using ancestry proportions as covariates is equivalent to using PCs. We shall assume the following individual-specific admixture model commonly assumed when inferring ancestry proportions (Pritchard et al., 2000a; Falush et al., 2003; Alexander et al., 2009; Gopalan et al., 2016; Cabreros and Storey, 2019). There are K subpopulations and every individual j draws a proportion q_{ju} of its alleles from subpopulation S_u. These ancestry proportions must be non-negative and sum to one for every individual j ($\sum_{u=1}^{K} q_{ju} = 1$ for every j). Each subpopulation S_u has an allele frequency $p_i^{S_u}$ at locus i, and thus the individual-specific allele frequency π_{ij} of individual j at locus i is given by the weighted average of the subpopulation allele frequencies, where the ancestry proportions are the weights:

$$\pi_{ij} = \sum_{u=1}^{K} q_{ju} p_i^{S_u}. \tag{6}$$

Genotypes are constructed by drawing each allele independently from this frequency, or $x_{ij}|\pi_{ij} \sim \text{Binomial}(2, \pi_{ij})$. Thus, the rowspace of the genotype matrix is, in expectation, the same as the rowspace of the individual-specific allele frequency matrix, which by Eq. (6) above is the same as the rowspace of the $n \times K$ admixture proportions matrix $\mathbf{Q} = (q_{ju})$. Therefore, the top K principal components suffice to fully model the rowspace of the genotypes, which only have dimension K. As an intercept term is always included in genetic association studies ($\mathbf{1}\alpha$ in Eq. (4)), and the sum of rows of $\mathbf{Q}$ sums to one, then the rowspace of the combined model has dimension K as well, so only $K - 1$ PCs (plus intercept) are needed to span the rowspace of this admixture model.

4.1.5 Linear mixed-effects model

The LMM is another approximation to the complex trait model in Eq. (3). Excluding additional covariates first, the LMM is

$$\mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta_i + \mathbf{s} + \boldsymbol{\epsilon}, \tag{7}$$

which is like the PCA model in Eq. (4) except that the PC terms $\mathbf{U}_r\boldsymbol{\gamma}_r$ are replaced by the random effect $\mathbf{s}$, which is a length-n vector drawn from (Sul et al., 2018)

$$\mathbf{s} \sim \text{Normal}\left(\mathbf{0}, \sigma_s^2\boldsymbol{\Phi}^T\right),$$

where $\boldsymbol{\Phi}^T$ is the kinship matrix and σ_s^2 is a trait-specific variance scaling factor. This model is derived from treating the standardized genotype matrix $\mathbf{X}_S$ as random rather than fixed, so that the standardized genetic effect $\mathbf{X}_S^\intercal\boldsymbol{\beta}_S$ in Eq. (3) has mean zero and a covariance matrix of

$$\text{Cov}\left(\mathbf{X}_S^\intercal\boldsymbol{\beta}_S\right) = ||\boldsymbol{\beta}_S||^2\boldsymbol{\Phi}^T.$$

The above random effect $\mathbf{s}$ satisfies those equations, where the variance scale equals $\sigma_s^2 = ||\boldsymbol{\beta}_S||^2$. Thus, the PCA approach is the fixed model equivalent of the LMM under the additional approximation that only the top r eigenvectors are used in PCA whereas the LMM uses all eigenvectors.

In practice, the key advantage of LMM over PCA is that it has fewer parameters to fit: ignoring the shared terms in Eq. (4) and Eq. (7), PCA has r parameters to fit (each PC coefficient in the $\boldsymbol{\gamma}$ vector), whereas LMMs only fit one additional parameter, namely σ_s^2. Therefore, PCA is expected to overfit more substantially than LMM—and thus lose power—when r is very large, and especially when the sample size (the number of individuals n) is very small. Statistical significance in LMMs is calculated via a likelihood ratio test, whose test statistic has a asymptotic χ^2 distribution under the null hypothesis.

4.1.6 LMM with PCs

An LMM variant we focus on testing in this work incorporates PCs as fixed covariates. Since PCs are the top eigenvectors of the same kinship matrix estimate used to draw the random effects, then the population structure is essentially modeled twice, which can lead to loss of power when the number of PCs is very large. However, some previous work has found the apparent redundancy of an LMM with PCs beneficial (Zhao et al., 2007; Price et al., 2010). Note that earlier LMM approaches estimated non-redundant kinship and fixed effects covariates: kinship matrices were estimated from pedigrees (thus excluding population structure), and population structure was modeled via admixture proportions rather than PCA (Yu et al., 2006; Zhao et al., 2007).

4.2 Simulations

Here the general notation f_B^A denotes the inbreeding coefficient of a subpopulation A from another subpopulation B that is ancestral to A. Often we use f_A^T where T is an overall ancestral population (ancestral to every subpopulation and/or individual under consideration, such as the most recent common ancestor population). [TODO: introduce more notation here, or do it all earlier?]

4.2.1 Genotype simulation from the admixture model

We consider three admixture simulation scenarios, refered to as Large, Small, and Family. All cases are based on the admixture model described previously (Ochoa and Storey, 2016; Ochoa and Storey, 2021).

The Large and Family simulations have $n = 1,000$ individuals, while Small has $n = 100$. The number of loci in all cases is $m = 100,000$. Individuals are admixed from $K = 10$ intermediate subpopulations. Each subpopulation S_u ($u \in \{1, ..., K\}$) has an inbreeding coefficient $f_{S_u}^T = u\tau$, individual-specific admixture proportions q_{ju} for individual j and intermediate subpopulation S_u arise from a random walk model for the intermediate subpopulations on a 1-dimensional geography with spread σ, where the free parameters τ and σ are fit to result in $F_{\text{ST}} = 0.1$ for the admixed individuals and a bias coefficient of $s = 0.5$, as before (Ochoa and Storey, 2021).

Random allele frequencies and genotypes are drawn from the following hierarchical model:

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

Briefly, allele frequencies p_i^T for the ancestral population T are drawn independently per locus i. Subpopulation allele frequencies $p_i^{S_u}$ are drawn independently for each intermediate subpopulation S_u from the Balding-Nichols distribution, resulting in marginal distributions with mean p_i^T and variance $p_i^T \left(1 - p_i^T\right) f_{S_u}^T$ (Balding and Nichols, 1995). The individual-specific allele frequency π_{ij} at locus i of individual j weigh the intermediate subpopulation allele frequencies $p_i^{S_u}$ according to the admixture proportions q_{ju}, and genotypes are drawn from these admixed frequencies. Loci that are fixed (i where $x_{ij} = 0$ for all j, or $x_{ij} = 2$ for all j) are drawn again from the model, starting from p_i^T, iterating until no loci are fixed.

4.2.2 Genotype simulation from a random admixed family

We simulated a pedigree with admixed founders that features: (1) strict avoidance of close relatives when pairing individuals; (2) strong favoring of pairs that are nearby in their 1-dimensional geography, which helps preserve the population structure across the generations by preferentially pairing individuals with more similar admixture proportions (a form of assortative mating); and (3) many generations, so that a broad distribution of close and distant relatives is present in the data.

Each generation in the pedigree is drawn iteratively for 20 generations. Generation 1 has individuals with genotypes drawn from the Large simulation described earlier, which features admixture. These individuals are ordered by the 1-dimensional geography of the admixture scenario. The local kinship matrix measures the pedigree relatedness; in the first generation, everybody is locally unrelated.

The children of the previous generation serve as the parents in the next generation, which are paired iteratively as follows. From the pool of unpaired individuals, one is picked randomly, and it is paired with the nearest individual that is not a second cousin or closer relative (local kinship must be $< 1/4^3$). If there are individuals that could not be paired (occurs if remaining unpaired individuals are all close relatives), then the current generation is erased and all individuals are paired anew. After all individuals are paired, two children per pair are created, to maintained a fixed population size. Children are reordered by the average coordinate of their parents, preserving the original order when there are ties.

This random pedigree was drawn once and shared across replicates. However, in each replicate the genotypes of the founder population are drawn anew from the admixture model, and genotypes across the generations are selected randomly. Each child draws, independently per locus, a random allele from each of its parents.

4.2.3 Genotype simulation from a tree model

A variant of the earlier admixture simulation model consists of drawing subpopulations allele frequencies from a hierarchical model, parametrized by a tree. The ancestral population T is at the root of the tree, and each node in the tree corresponds to a subpopulation S_w where the nodes are indexed arbitrarily. Each edge has a value $f_{S_w}^{P_w}$ corresponding to the inbreeding coefficient of subpopulation S_w from its parent population, denoted as P_w.

With a tree so defined, allele frequencies are drawn from the root to the tips of the tree iteratively, as a hierarchical or graphical model, particularly a directed acyclic graph. For the root T, allele frequencies p_i^T are drawn from a given distribution constructed to mimic each given real dataset (see below). Now, given the allele frequencies $p_i^{P_w}$ of the parent population P_w (here $P_w = T$ for the first level of the tree), then the child population S_w's allele frequencies are drawn from the following Balding-Nichols distribution:

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

Finally, individuals j in the tip subpopulation S_w have genotypes drawn independently from its

allele frequency:

$$x_{ij}|p_i^{S_w} \sim \text{Binomial}\left(2, p_i^{S_w}\right).$$

Each simulated subpopulation size equals its corresponding real subpopulation size.

To match the real datasets, which had loci with MAF $= \min\left\{\hat{p}_i^T, 1-\hat{p}_i^T\right\} < 0.01$ removed, our simulations had loci equivalently ascertained: loci with MAF < 0.01 are drawn again from the model, starting from drawing a new p_i^T from the input distribution, iterating until no such loci remain.

4.2.4 Fitting tree to data

We developed new methods to fit trees to real data based on estimating kinship using `popkin`. The general approach is divided into these parts: deriving a simple additive estimation model, estimating population-averaged coancestry values, estimating tree topology, and estimating inbreeding edge values for a given tree topology.

Estimation model. A tree with given inbreeding edges gives rise to a specific coancestry matrix, which we will calculate recursively here. Suppose as before that every node in the tree, including root and tip nodes, are indexed as S_w. All coancestry values ϑ_{uv}^T for a pair of subpopulations S_u and S_v given by a tree are total inbreeding values of subpopulations in the tree. In particular, the self-coancestry of S_u equals its total inbreeding coefficient ($\vartheta_{uu}^T = f_{S_u}^T$), and more generally, the coancestry of subpopulations S_u and S_v equals the total inbreeding of the most common ancestor (MRCA) population of those subpopulations:

$$\vartheta_{uv}^T = f_{S_w}^T \qquad \text{for } w \text{ such that} \qquad S_w = \text{MRCA}(S_u, S_v).$$

Since the above S_w is always some node in the tree, we obtain the desired coancestry matrix by calculating the total inbreeding values of every S_w.

Letting P_w denote the parent subpopulation of S_w in the tree, the value of the edge to S_w is an inbreeding coefficient between a child and parent node pair, or $f_{S_w}^{P_w}$ for each w. We will calculate total coancestries (from the ancestral population T, namely $f_{S_w}^T$) for every node S_w as follows. Note

that nodes whose parent is $P_w = T$ are already of this form. We proceed recursively through the tree branches moving outward from the root. Suppose we have already calculated the total coancestry of P_w, which is $f^T_{P_w}$. Then the desired total coancestry of S_w is given by

$$f^T_{S_w} = f^T_{P_w} + f^{P_w}_{S_w}\left(1 - f^T_{P_w}\right),$$

which is a special case of a previous calculation for three nested subpopulations (Ochoa and Storey, 2016). Note that the previous calculation is nearly additive, but instead of adding $f^{P_w}_{S_w}$ to $f^T_{P_w}$ we have to shrink $f^{P_w}_{S_w}$ first by a factor of $\left(1 - f^T_{P_w}\right)$. Denote the additive contribution of the edge to S_w as

$$\delta_w = f^T_{S_w} - f^T_{P_w} = f^{P_w}_{S_w}\left(1 - f^T_{P_w}\right),$$

which we introduce because, as we will see shortly, δ_w can be estimated more readily from a coancestry matrix. Note that $\delta_w \geq 0$ because $0 \leq f^{P_w}_{S_w}, f^T_{P_w} \leq 1$ for every w. Importantly, the inbreeding edge values can be recovered from these additive edges recursively starting from the root, since nodes S_w connected to the root satisfy $f^{P_w}_{S_w} = f^T_{S_w} = \delta_w$, and in the next level we have $f^T_{P_w}$ so we can calculate the desired quantity $f^{P_w}_{S_w}$ and also $f^T_{S_w}$ (for the level after that, if needed):

$$f^{P_w}_{S_w} = \frac{\delta_w}{1 - f^T_{P_w}}, \quad f^T_{S_w} = f^T_{P_w} + \delta_w.$$

The coancestry matrix is given most simply as a sum of the additive contributions δ_w for the nodes that are ancestors of the pair of subpopulations under consideration, or

$$\vartheta^T_{uv} = \sum_w \delta_w I_w(u, v), \tag{8}$$

where the sum goes over all nodes S_w in the tree, and $I_w(u, v)$ is an indicator function equal to 1 if S_w is an ancestor to both S_u and S_v, and 0 otherwise. Note that $I_w(u, v)$ are given solely by the topology of the tree, while δ_w reflect the edge values for that topology. Therefore, given a topology, this form is amenable to estimation of δ_w values by a variant of linear regression, where the $I_w(u, v)$ define the design matrix.

Estimating population-averaged coancestry. The `popkin` algorithm is used to estimate the kinship matrix ($\hat{\varphi}^T_{jk}$) between all individual pairs (j, k) in the data. Coancestry ($\hat{\theta}^T_{jk}$) is estimated from kinship by replacing self-kinship with inbreeding ($\hat{f}^T_j$) along the diagonal:

$$\hat{\theta}^T_{jk} = \begin{cases} \hat{\varphi}^T_{jk} & \text{if} \quad k \neq j, \\ \hat{f}^T_j = 2\hat{\varphi}^T_{jj} - 1 & \text{if} \quad k = j. \end{cases} \tag{9}$$

Subpopulation coancestry values $\hat{\vartheta}^T_{uv}$ between subpopulations S_u and S_v are averages of the individual coancestry values across both subpopulations, or within the subpopulation when $u = v$:

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

Estimating tree topology. Our topology estimation approach is remarkably simple, stemming from the simple, monotonic relationship between node depth and coancestry value due to Eq. (8). The tree topology is estimated with hierarchical clustering using the weighted pair group method with arithmetic mean (WPGMA) algorithm (Sokal and Michener, 1958). The distance function between subpopulations we use is

$$d(S_u, S_v) = \vartheta^T_{\max} - \hat{\vartheta}^T_{uv},$$

where $\vartheta^T_{\max}$ is the maximum of all the $\hat{\vartheta}^T_{uv}$ values. This algorithm recovers the true tree topology when the true coancestry values (ϑ^T_{uv}) are provided, and performs well when $\hat{\vartheta}^T_{uv}$ are noisy estimates from genotypes. However, edge lengths as estimated by hierarchical clustering are incorrect, and are refit in the next step.

Estimating tree edge lengths. Additive edge lengths δ_w are estimated from Eq. (8) from the estimated subpopulation coancestry matrix $\hat{\vartheta}^T_{uv}$, using non-negative least squares linear regression (Lawson and Hanson, 1974), which minimizes the sum of squared residuals to the data while ensuring that every estimated coefficient (δ_w) is non-negative. The desired inbreeding edge values $f^{P_w}_{S_w}$ are then estimated from these δ_w using the recursive algorithm described earlier. To account for small

biases in coancestry estimation, an intercept term δ_0 is fit (with $I_0(u, v) = 1$ for all u, v), and when converting δ_w to $f_{S_w}^{P_w}$ values this is treated as an additional edge from the root of the input topology and the new root. However, as this root edge models estimation bias rather than an evolutionary relationship, it is ignored when simulating allele frequencies from the earlier hierarchical Balding-Nichols distribution.

4.2.5 Fitting ancestral allele distribution to data

We calculated the allele frequency distribution $\hat{p}_i^T$ of each real dataset. However, differentiation increases the variance of $\hat{p}_i^T$ relative to the true ancestral allele frequency p_i^T (Ochoa and Storey, 2021). Here we present a new procedure for constructing an "undifferentiated" distribution of ancestral allele frequencies based on the input data $\hat{p}_i^T$ but which has the lower variance of the true p_i^T distribution.

Model. Suppose we started from an allele frequency distribution p_i^T over all loci i with $\mathrm{E}\left[p_i^T\right] = \frac{1}{2}$ and $\mathrm{Var}\left(p_i^T\right) = V^T$. The sample allele frequency $\hat{p}_i^T$ was previously found to have a conditional mean and variance (treating p_i^T as fixed) of

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

where $\bar{\varphi}^T = \frac{1}{n^2}\sum_{j=1}^n \sum_{k=1}^n \varphi_{jk}^T$ is the mean kinship over all individual (Ochoa and Storey, 2021). The desired moments of the total (unconditional) distribution of $\hat{p}_i^T$ are given by the laws of total expectation and variance:

$$\begin{aligned}
\mathrm{E}\left[\hat{p}_i^T\right] &= \mathrm{E}\left[\mathrm{E}\left[\hat{p}_i^T \middle| p_i^T\right]\right] = \mathrm{E}\left[p_i^T\right] = \frac{1}{2}, \\
W^T = \mathrm{Var}\left(\hat{p}_i^T\right) &= \mathrm{E}\left[\mathrm{Var}\left(\hat{p}_i^T \middle| p_i^T\right)\right] + \mathrm{Var}\left(\mathrm{E}\left[\hat{p}_i^T \middle| p_i^T\right]\right) \\
&= \mathrm{E}\left[p_i^T\left(1 - p_i^T\right)\bar{\varphi}^T\right] + \mathrm{Var}\left(p_i^T\right) \\
&= \bar{\varphi}^T \mathrm{E}\left[p_i^T\right]\left(1 - \mathrm{E}\left[p_i^T\right]\right) + \left(1 - \bar{\varphi}^T\right)\mathrm{Var}\left(p_i^T\right) \\
&= \bar{\varphi}^T \frac{1}{4} + \left(1 - \bar{\varphi}^T\right) V^T.
\end{aligned}$$

Since $V^T \leq \frac{1}{4}$ (a bound that all allele variances satisfy) and $\bar{\varphi}^T \geq 0$, the variance of $\hat{p}_i^T$ is greater: $W^T \geq V^T$. Thus, given W^T and $\bar{\varphi}^T$, the goal is to construct a new distribution with the original, lower variance of

$$V^T = \frac{W^T - \frac{1}{4}\bar{\varphi}^T}{1 - \bar{\varphi}^T}. \tag{10}$$

Estimation of ancestral variance. Given empirical sample allele frequencies $\hat{p}_i^T$, we use a sample estimator for W^T that assumes a known expectation of one half, which is unbiased:

$$\hat{W}^T = \frac{1}{m}\sum_{i=1}^{m}\left(\hat{p}_i^T - \frac{1}{2}\right)^2.$$

Although often $\hat{p}_i^T$ are calculated as minor allele frequencies, which have a sample mean lower than $\frac{1}{2}$, treating the choice of reference allele as random guarantees an expectation of $\frac{1}{2}$ and the above calculation is invariant to the choice of reference allele.

The mean kinship $\bar{\varphi}^T$ should correspond to the simulation parameter, which is calculated from the tree: the subpopulation coancestry matrix is calculated from the tree inbreeding parameters using Eq. (8), which is expanded so that every row corresponds to an individual rather than a subpopulations (individual coancestries are copies of the subpopulation coancestries), and finally the diagonal is converted to kinship (reversing Eq. (9)) and the matrix averaged. However, this variance model ignores the MAF-based locus ascertainment performed in our simulations, which introduces additional biases. We found that greater values of $\bar{\varphi}^T$ (than the true model parameter) resulted in simulations with more accurately specified population structures. For Human Origins, which has the largest numbers of individuals and therefore the least MAF ascertainment bias, the true model $\bar{\varphi}^T$ (0.143) was used. For 1000 Genomes and HGDP the true model values $\bar{\varphi}^T$ are 0.126 and 0.124, respectively, but our internal simulation accuracy evaluations led us to instead use 0.4 for both.

Construction of "undifferentiated" allele frequencies. We construct a new random allele frequency,

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

by averaging the sample allele frequencies $\hat{p}_i^T$ (with known variance W^T) with another frequency

$q \in (0, 1)$ drawn independently from a lower-variance "mixing" distribution (constructed shortly) using some weight w. We require that the mixing distribution have $\mathrm{E}[q] = \frac{1}{2}$, which (since $\mathrm{E}\left[\hat{p}_i^T\right] = \frac{1}{2}$) results in $\mathrm{E}\left[p_i^{T'}\right] = \frac{1}{2}$. Letting $V_{\mathrm{mix}} = \mathrm{Var}(q)$, the output variance is

$$V^{T'} = w^2 W^T + (1 - w)^2 V_{\mathrm{mix}},$$

which we set to the desired V^T in Eq. (10) and solve for w in this quadratic equation. For simplicity and flexibility, we set $V_{\mathrm{mix}} = V^T$, which is achieved with the following symmetric Beta distribution:

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

Although in this case $w = 0$ yields $V^{T'} = V^T$, instead we use the second root of the quadratic equation, which makes use of the input $\hat{p}_i^T$ data:

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

4.2.6 Real human genotype datasets

Three real human genotype datasets were used in our evaluations, from which traits were simulated as described later. These datasets were processed as before (Ochoa and Storey, 2019) (summarized below), except with an additional filter so loci are in approximate linkage equilibrium and rare variants are avoided. This is required to keep our evaluations simple (so loci that are not causal are not correlated to causal loci). All processing was performed with `plink2` (Chang et al., 2015). Each dataset groups individuals in a two-level hierarchy, which we call continental and fine-grained subpopulations, respectively. Final dataset sizes are in Table 2.

Human Origins. We obtained the full (including non-public) Human Origins data by contacting the authors and agreeing to their usage restrictions. The public subset of these data is available at `https://reich.hms.harvard.edu/datasets`. The majority of the genotypes (Lazaridis et al., 2014; Lazaridis et al., 2016) were obtained as one dataset, while the Pacific data (Skoglund et al., 2016) was a separate dataset. Individuals were merged, and only loci in the intersection between

both datasets was considered. Non-autosomal loci were removed. We removed ancient individuals, and individuals from singleton and non-native subpopulations. Loci that were fixed across individuals in the remaining dataset were removed.

Human Genome Diversity Panel (HGDP) dataset. The whole-genome sequencing version of HGDP (Bergström et al., 2020) was downloaded from the Wellcome Sanger Institute FTP site at `ftp://ngs.sanger.ac.uk/production/hgdp/hgdp_wgs.20190516/`. Our analysis was restricted to autosomal biallelic SNP loci.

1000 Genomes Project. The 1000 Genomes Project "Phase 3" integrated call data (Consortium, 2010; 1000 Genomes Project Consortium et al., 2012) was downloaded from `https://www.cog-genomics.org/plink/2.0/resources#1kg_phase3`. This analysis was restricted to autosomal biallelic SNP loci after removing loci with repeated identifiers.

LD prunning. Our evaluations require uncorrelated loci, so that non-causal loci are not correlated to the trait, which would be labeled as false positives. We filtered each dataset with `plink2` using parameters "`-indep-pairwise 1000kb 0.3`", which iteratively removes loci that have a greater than 0.3 correlation coefficient with another locus that is within 1000kb, stopping until no such loci remain.

MAF filters. All real datasets have extremely large numbers of rare variants compared to a uniform distribution. Since the models we are evaluating are not able to detect associations involving rare variants, for simplicity we removed all loci with MAF < 0.01.

4.2.7 Trait Simulation

For a given genotype matrix (simulated or real), simulated complex traits that follow the additive quantitative trait model in Eq. (3) are constructed. To simulate the correct heritability, the true ancestral allele frequencies p_i^T must be known, which are only available for simulations. We extend the procedure to real datasets by utilizing estimated allele frequencies and appropriate bias corrections, which rely on the unbiased kinship estimator `popkin` (Ochoa and Storey, 2021).

All simulations share the following features. The (narrow-sense) heritability of the trait is $h^2 = 0.8$. The non-genetic effects are drawn from $\epsilon_j \sim \text{Normal}(0, 1 - h^2)$ independently for each

individual j. To balance power across datasets with varying numbers of individuals n, the number of causal loci is $m_1 = n/10$. For each replicate, new causal loci are picked randomly from the genome, and new coefficients are drawn or constructed depending on the trait model. The length-m_1 set of causal loci C is drawn from the subset of loci $\{i\}$ that satisfy MAF ≥ 0.01, to avoid simulations with very rare causal variants (in the random coefficients model they are typically undetectable, while in the fixed effect size model they have extremely large coefficients, which may be problematic; either way PCA and LMM are not appropriate inference models for these cases).

Initial coefficients for *fixed effect sizes* model. The effect size of a locus i is defined as $2p_i^T\left(1-p_i^T\right)\beta_i^2$, which is its contribution of the trait variance (Park et al., 2010). Thus, when p_i^T are known, effect sizes of equal magnitude across all causal loci i are obtained by setting the coefficients initially to

$$\beta_i = \frac{1}{\sqrt{2p_i^T\left(1-p_i^T\right)}}.$$

A random sign is then added to each β_i (becomes negative with probability 0.5).

When p_i^T are unknown, we replace the factor containing p_i^T with the following unbiased estimator (Ochoa and Storey, 2021):

$$v_i^T = p_i^T\left(1-p_i^T\right), \qquad \hat{v}_i^T = \frac{\hat{p}_i^T\left(1-\hat{p}_i^T\right)}{1-\bar{\varphi}^T}, \tag{11}$$

where $\bar{\varphi}^T$ is the mean kinship in the data, which is estimated as the mean of the kinship matrix from `popkin`.

Initial coefficients for *random coefficients* model. The coefficients at selected causal loci i are initially drawn independently from $\beta_i \sim \text{Normal}(0,1)$.

Coefficient normalization. All coefficients (both models) are scaled as follows to attain the desired heritability. Under the kinship model, the resulting genetic variance component is given by

$$\sigma_0^2 = \sum_{i \in C} 2v_i^T\beta_i^2,$$

where v_i^T is as in Eq. (11); $\hat{v}_i^T$ is used instead if needed, in which case σ_0^2 is an unbiased estimate of

the total genetic variance. The desired genetic variance of h^2 is therefore obtained by multiplying every β_i by $\frac{h}{\sigma_0}$.

Lastly, the intercept coefficient in Eq. (3) is set to

$$\alpha = -\sum_{i \in C} 2 p_i^T \beta_i,$$

so the trait expectation is zero. When p_i^T are unknown, the above formulation distorts the covariance structure of the trait if $\hat{p}_i^T$ simply replaces p_i^T (for the same reason the standard kinship estimator in Eq. (5) is biased; Ochoa and Storey, 2021), which is avoided with the form

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

4.2.8 Kinship rank estimates

The `popkin` kinship estimates from each dataset (from the first replicate for simulated genotypes; same ones shown in Fig. 1) were used to calcuate the eigenvalues. No individuals were excluded in this analysis. The vector of eigenvalues was passed to the `twstats` binary of the Eigensoft package (Patterson et al., 2006), which returns a table including p-values for each eigenvalue. The estimated kinship rank was the largest eigenvalue rank for which $p < 0.01$ for it and all higher-ranking eigenvalues (p-values did not increase monotonically with eigenvalue rank).

4.3 Evaluation of performance

All of the approaches considered here are evaluated in two orthogonal dimensions. The first one—SRMSD_p—quantifies the extent to which non-causal p-values are uniform, which is a prerequisite for accurate control of the type-I error and successful FDR control. The second measure—the area under the precision-recall curve—quantifies the predictive power of each model, which makes it possible to qualitatively compare the statistical power of each model without having to select a single threshold, and most importantly, overcoming the problem of comparing models that may not have accurate p-values (Bouaziz et al., 2011).

4.3.1 SRMSD$_p$: a measure of p-value uniformity

From their definition, correct p-values for continuous test statistics have a uniform distribution when the null hypothesis holds. This fact is crucial for accurate control of the type-I error, and is a prerequisite for the most common approaches that control the FDR, such as q-values (Storey, 2003; Storey and Tibshirani, 2003). We use the Signed Root Mean Square Deviation (SRMSD) to measure the disagreement between the observed p-value quantiles and the expected uniform quantiles:

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

where $m_0 = m - m_1$ is the number of null (non-causal) loci, here i indexes null loci only, $p_{(i)}$ is the ith ordered null p-value, $u_i = (i - 0.5)/m_0$ is its expectation, p_{median} is the median observed null p-value, $u_{\text{median}} = \frac{1}{2}$ is the median expected null p-value, and sgn is the sign function (in this case 1 if $u_{\text{median}} \geq p_{\text{median}}$, -1 otherwise). Thus, $\text{SRMSD}_p = 0$ corresponds to the best performance (well-calibrated p-values), large positive SRMSD_p indicate anti-conservative p-values, and large negative values are conservative p-values.

One scenario that achieves the maximum SRMSD_p (or worst performance) is when all estimated p-values approach zero, which is what happens to anti-conservative approaches. In that case all of the observed quantiles approach $p_{(i)} = 0$, and then, in the limit as the number of loci goes to infinity, the statistic in this worst-case scenario approaches

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

The same worst-case value is achieved if all p-values approach 1 instead of 0, except for the change in sign.

4.3.2 The inflation factor λ

In previous evaluations, test statistic inflation has been used to measure the success of corrections for population structure (Astle and Balding, 2009; Price et al., 2010). The inflation factor λ is defined as

the median χ^2 association statistic divided by theoretical median under the null hypothesis (Devlin and Roeder, 1999). The inflation factor can be calculated from the median p-value (in this case across all p-values, not just the null ones) using

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

where p_{median} is the median observed p-value, $u_{\text{median}} = \frac{1}{2}$ is the median expected null p-value, and F is the cumulative density function of the χ^2 distribution (F^{-1} is the quantile function). This equation is useful to compare p-values from statistics that have non-χ^2 distributions (linear regression can use the t-test, whose statistics have a t-distribution).

To compare the properties of λ and SRMSD_p directly, for simplicity assume that all p-values are from the null distribution (nearly all usually are in association studies). In this case, when null test statistics have their expected distribution, we get $\lambda = 1$ and $\text{SRMSD}_p = 0$. However, any other null test statistic distribution with the same median results in $\lambda = 1$ as well, but $\text{SRMSD}_p \neq 0$ unless the entire test statistic distribution is as expected; this is the important flaw of λ that SRMSD_p overcomes. In particular, approaches such as genomic control (Devlin and Roeder, 1999) that scale test statistics to artificially result in $\lambda = 1$ will be evaluated fairly using SRMSD_p. The $\lambda > 1$ case always gives $\text{SRMSD}_p > 0$, and corresponds to inflated test statistics (resulting in smaller than expected, or anti-conservative, p-values), which occurs when residual population structure is present. On the other hand, $\lambda < 1$ always gives $\text{SRMSD}_p < 0$, and arises if p-values are larger than expected, or conservative. Thus, $\lambda \neq 1$ always implies $\text{SRMSD}_p \neq 0$, but not the other way around, and SRMSD_p has the same sign as $\lambda - 1$. Overall, the weakness of λ is that it depends only on the median of the distribution, whereas the SRMSD_p makes use of the complete p-value distribution to evaluate its uniformity, which is stricter. The drawback is that SRMSD_p requires knowing which loci are null, so unlike λ, it is only applicable to simulated traits.

4.3.3 The area under the precision-recall curve

Precision and recall are two common measures for evaluating binary classifiers. Let c_i be the the true classification of locus i, where $c_i = 1$ for truly causal loci (if the true $\beta_i \neq 0$, where the alternative

hypothesis holds), and $c_i = 0$ otherwise (null cases). For given test statistics t_i from a model and some threshold t, the model predicts classifications as

$$\hat{c}_i(t) = \begin{cases} 1 & \text{if} \quad t_i \geq t, \\ 0 & \text{otherwise.} \end{cases}$$

Across all loci, the number of true positives (TP), false positives (FP) and false negatives (FN) at the threshold t is given by

$$\begin{aligned} \mathrm{TP}(t) &= \sum_{i=1}^{m} c_i \hat{c}_i(t), \\ \mathrm{FP}(t) &= \sum_{i=1}^{m} (1 - c_i) \hat{c}_i(t), \\ \mathrm{FN}(t) &= \sum_{i=1}^{m} c_i \left(1 - \hat{c}_i(t)\right). \end{aligned}$$

Precision and recall at this threshold are given by

$$\begin{aligned} \mathrm{Precision}(t) &= \frac{\mathrm{TP}(t)}{\mathrm{TP}(t) + \mathrm{FP}(t)} = \frac{\sum_{i=1}^{m} c_i \hat{c}_i(t)}{\sum_{i=1}^{m} \hat{c}_i(t)}, \\ \mathrm{Recall}(t) &= \frac{\mathrm{TP}(t)}{\mathrm{TP}(t) + \mathrm{FN}(t)} = \frac{\sum_{i=1}^{m} c_i \hat{c}_i(t)}{\sum_{i=1}^{m} c_i}. \end{aligned}$$

The precision-recall curve results from calculating the above two values at every threshold t, tracing a curve as recall goes from zero (everything is classified as null) to one (everything is classified as alternative), and the area under this curve is our final measure $\mathrm{AUC}_{\mathrm{PR}}$. A model obtains the maximum $\mathrm{AUC}_{\mathrm{PR}} = 1$ if there is some threshold that classifies all loci perfectly. In contrast, a model that classifies at random (for example, $\hat{c}_i(t) \sim \mathrm{Bernoulli}(p)$ for any p) has an expected precision ($= \mathrm{AUC}_{\mathrm{PR}}$) equal to the overall proportion of alternative cases: $\pi_1 = \frac{m_1}{m} = \frac{1}{m} \sum_{i=1}^{m} c_i$.

4.4 Software

We selected modern software implementing each of the basic PCA and LMM approaches that are the fastest and most robust of their category based on our internal, unpublished benchmarks.

PCA association was performed using `plink2` (Chang et al., 2015). When applied to quantitative traits, the model is a linear regression with covariates that employs the t-test for significance testing. The PCs used for PCA were calculated with `plink2`, which equals the top eigenvectors of Eq. (5). Only loci with MAF ≥ 0.1 were used in calculating the PCs.

LMM association was performed using GCTA (Yang et al., 2011). GCTA also uses the kinship matrix estimator $\hat{\mathbf{\Phi}}$ in Eq. (5), except the diagonal estimates use a slightly different formula (Yang et al., 2011). For LMM with PCs, the PCs were calculated using GCTA from its kinship estimates. When running GCTA with large numbers of PCs in the small admixture simulation, we often encountered errors such as "the information matrix is not invertible", "analysis stopped because more than half of the variance components are constrained", and "Log-likelihood not converged (stop after 100 iteractions)" (sic), in which cases SRMSD_p and AUC_{PR} were treated as missing; these errors were not observed in the other scenarios.

All following R packages are available on the Comprehensive R Archive Network (CRAN).

Our genotype admixture and tree simulations are implemented in the R package `bnpsd` (Ochoa and Storey, 2021). Our tree fitting and simulation implementations, introduced in this work, also make use of the R packages `nnls` for non-negative least squares (Mullen and Stokkum, 2012) and `ape` for general tree data structures and methods (Paradis and Schliep, 2019).

Our random family simulation procedure, introduced in this work, is implemented in the R package `simfam` available at `https://github.com/OchoaLab/simfam`.

Our trait simulation procedure and the AUC_{PR} and SRMSD_p measures, introduced in this work, are implemented in the R package `simtrait` available at `https://github.com/OchoaLab/simtrait`. Our AUC_{PR} function makes use of the R package `PRROC`, which integrates the correct non-linear piecewise function when interpolating between points (Grau et al., 2015).

[TODO: add simfam, simtrait to CRAN!]

Unbiased kinship estimates are obtained with the R package `popkin` (Ochoa and Storey, 2021). The data processing in this work is also uniquely enabled by the R packages `BEDMatrix` (Grueneberg and Campos, 2019) and `genio` (introduced here).

References

1000 Genomes Project Consortium et al. (2012). "An integrated map of genetic variation from 1,092 human genomes". *Nature* 491(7422), pp. 56–65.

Abraham, Gad and Michael Inouye (2014). "Fast Principal Component Analysis of Large-Scale Genome-Wide Data". *PLOS ONE* 9(4), e93766.

Abraham, Gad, Yixuan Qiu, and Michael Inouye (2017). "FlashPCA2: principal component analysis of Biobank-scale genotype datasets". *Bioinformatics* 33(17), pp. 2776–2778.

Agrawal, Aman et al. (2020). "Scalable probabilistic PCA for large-scale genetic variation data". *PLOS Genetics* 16(5). Publisher: Public Library of Science, e1008773.

Al-Khudhair, Ahmed et al. (2015). "Inference of Distant Genetic Relations in Humans Using "1000 Genomes"". *Genome Biology and Evolution* 7(2), pp. 481–492.

Alexander, David H., John Novembre, and Kenneth Lange (2009). "Fast model-based estimation of ancestry in unrelated individuals". *Genome Res.* 19(9), pp. 1655–1664.

Astle, William and David J. Balding (2009). "Population Structure and Cryptic Relatedness in Genetic Association Studies". *Statist. Sci.* 24(4). Mathematical Reviews number (MathSciNet): MR2779337, pp. 451–471.

Balding, D. J. and R. A. Nichols (1995). "A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity". *Genetica* 96(1-2), pp. 3–12.

Bergström, Anders et al. (2020). "Insights into human genetic variation and population history from 929 diverse genomes". *Science* 367(6484).

Bouaziz, Matthieu, Christophe Ambroise, and Mickael Guedj (2011). "Accounting for Population Stratification in Practice: A Comparison of the Main Strategies Dedicated to Genome-Wide Association Studies". *PLOS ONE* 6(12), e28845.

Cabreros, Irineo and John D. Storey (2019). "A Likelihood-Free Estimator of Population Structure Bridging Admixture Models and Principal Components Analysis". *Genetics* 212(4), pp. 1009–1029.

Cann, Howard M. et al. (2002). "A human genome diversity cell line panel". *Science* 296(5566), pp. 261–262.

Chang, Christopher C. et al. (2015). "Second-generation PLINK: rising to the challenge of larger and richer datasets". *GigaScience* 4(1), p. 7.

Consortium, The 1000 Genomes Project (2010). "A map of human genome variation from population-scale sequencing". *Nature* 467(7319), pp. 1061–1073.

Devlin, B. and Kathryn Roeder (1999). "Genomic Control for Association Studies". *Biometrics* 55(4), pp. 997–1004.

Epstein, Michael P., Andrew S. Allen, and Glen A. Satten (2007). "A Simple and Improved Correction for Population Stratification in Case-Control Studies". *The American Journal of Human Genetics* 80(5), pp. 921–930.

Falush, Daniel, Matthew Stephens, and Jonathan K. Pritchard (2003). "Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies". *Genetics* 164(4), pp. 1567–1587.

Fedorova, Larisa et al. (2016). "Atlas of Cryptic Genetic Relatedness Among 1000 Human Genomes". *Genome Biology and Evolution* 8(3), pp. 777–790.

Galinsky, Kevin J. et al. (2016). "Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia". *The American Journal of Human Genetics* 98(3), pp. 456–472.

Gazal, Steven et al. (2015). "High level of inbreeding in final phase of 1000 Genomes Project". *Sci Rep* 5(1). Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Inbreeding;Population genetics Subject_term_id: inbreeding;population-genetics, p. 17453.

Gopalan, Prem et al. (2016). "Scaling probabilistic models of genetic variation to millions of humans". *Nat. Genet.* 48(12), pp. 1587–1590.

Grau, Jan, Ivo Grosse, and Jens Keilwagen (2015). "PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R". *Bioinformatics* 31(15), pp. 2595–2597.

Grueneberg, Alexander and Gustavo de los Campos (2019). "BGData - A Suite of R Packages for Genomic Analysis with Big Data". *G3: Genes, Genomes, Genetics* 9(5). Publisher: G3: Genes, Genomes, Genetics Section: SOFTWARE AND DATA RESOURCES, pp. 1377–1383.

Henn, Brenna M. et al. (2012). "Cryptic Distant Relatives Are Common in Both Isolated and Cosmopolitan Genetic Samples". *PLOS ONE* 7(4). Publisher: Public Library of Science, e34267.

Hoffman, Gabriel E. (2013). "Correcting for population structure and kinship using the linear mixed model: theory and extensions". *PLoS ONE* 8(10), e75707.

Jacquard, Albert (1970). *Structures génétiques des populations.* Paris: Masson et Cie.

Janss, Luc et al. (2012). "Inferences from Genomic Models in Stratified Populations". *Genetics* 192(2), pp. 693–704.

Jolliffe, Ian T. (2002). *Principal Component Analysis*. 2nd ed. New York: Springer-Verlag.

Kang, Hyun Min et al. (2010). "Variance component model to account for sample structure in genome-wide association studies". *Nat. Genet.* 42(4), pp. 348–354.

Kimmel, Gad et al. (2007). "A Randomization Test for Controlling Population Stratification in Whole-Genome Association Studies". *The American Journal of Human Genetics* 81(5), pp. 895–905.

Lawson, Charles L. and R. J. Hanson (1974). "Solving least squares problems prentice-hall". *Englewood Cliffs*.

Lazaridis, Iosif et al. (2014). "Ancient human genomes suggest three ancestral populations for present-day Europeans". *Nature* 513(7518), pp. 409–413.

Lazaridis, Iosif et al. (2016). "Genomic insights into the origin of farming in the ancient Near East". *Nature* 536(7617), pp. 419–424.

Lee, Seokho et al. (2012). "Sparse Principal Component Analysis for Identifying Ancestry-Informative Markers in Genome-Wide Association Studies". *Genetic Epidemiology* 36(4), pp. 293–302.

Li, Mingyao et al. (2010). "Correcting population stratification in genetic association studies using a phylogenetic approach". *Bioinformatics* 26(6), pp. 798–806.

Li, Qizhai and Kai Yu (2008). "Improved correction for population stratification in genome-wide association studies by identifying hidden population structures". *Genetic Epidemiology* 32(3), pp. 215–226.

Malécot, Gustave (1948). *Mathématiques de l'hérédité*. Masson et Cie.

Mullen, Katharine M. and Ivo H. M. van Stokkum (2012). *nnls: The Lawson-Hanson algorithm for non-negative least squares (NNLS)*.

Ochoa, Alejandro and John D. Storey (2016). "F_{ST} and kinship for arbitrary population structures I: Generalized definitions". *bioRxiv* (10.1101/083915). `https://doi.org/10.1101/083915`.

— (2019). "New kinship and F_{ST} estimates reveal higher levels of differentiation in the global human population". *bioRxiv* (10.1101/653279). `https://doi.org/10.1101/653279`.

— (2021). "Estimating FST and kinship for arbitrary population structures". *PLoS Genet* 17(1), e1009241.

Paradis, E. and K. Schliep (2019). "ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R". *Bioinformatics* 35, pp. 526–528.

Park, Ju-Hyun et al. (2010). "Estimation of effect size distribution from genome-wide association studies and implications for future discoveries". *Nature Genetics* 42(7). Number: 7 Publisher: Nature Publishing Group, pp. 570–575.

Park, Ju-Hyun et al. (2011). "Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants". *PNAS* 108(44). Publisher: National Academy of Sciences Section: Biological Sciences, pp. 18026–18031.

Patterson, Nick, Alkes L Price, and David Reich (2006). "Population Structure and Eigenanalysis". *PLoS Genet* 2(12), e190.

Patterson, Nick et al. (2012). "Ancient admixture in human history". *Genetics* 192(3), pp. 1065–1093.

Price, Alkes L. et al. (2006). "Principal components analysis corrects for stratification in genome-wide association studies". *Nat. Genet.* 38(8), pp. 904–909.

Price, Alkes L. et al. (2010). "New approaches to population stratification in genome-wide association studies". *Nature Reviews Genetics* 11(7), pp. 459–463.

Price, Alkes L. et al. (2013). "Response to Sul and Eskin". *Nature Reviews Genetics* 14(4), p. 300.

Pritchard, J. K., M. Stephens, and P. Donnelly (2000a). "Inference of population structure using multilocus genotype data". *Genetics* 155(2), pp. 945–959.

Pritchard, Jonathan K. et al. (2000b). "Association Mapping in Structured Populations". *The American Journal of Human Genetics* 67(1), pp. 170–181.

Qian, Junyang et al. (2020). "A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK Biobank". *PLOS Genetics* 16(10). Publisher: Public Library of Science, e1009141.

Rosenberg, Noah A. et al. (2002). "Genetic Structure of Human Populations". *Science* 298(5602), pp. 2381–2385.

Schlauch, Daniel, Heide Fier, and Christoph Lange (2017). "Identification of genetic outliers due to sub-structure and cryptic relationships". *Bioinformatics* 33(13), pp. 1972–1979.

Shchur, Vladimir and Rasmus Nielsen (2018). "On the number of siblings and p-th cousins in a large population sample". *J Math Biol* 77(5), pp. 1279–1298.

Simons, Yuval B. et al. (2018). "A population genetic interpretation of GWAS findings for human quantitative traits". *PLOS Biology* 16(3), e2002985.

Skoglund, Pontus et al. (2016). "Genomic insights into the peopling of the Southwest Pacific". *Nature* 538(7626), pp. 510–513.

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

Song, Minsun, Wei Hao, and John D. Storey (2015). "Testing for genetic associations in arbitrarily structured populations". *Nat. Genet.* 47(5), pp. 550–554.

Storey, John D. (2003). "The positive false discovery rate: a Bayesian interpretation and the q-value". *Ann. Statist.* 31(6). Mathematical Reviews number (MathSciNet): MR2036398; Zentralblatt MATH identifier: 02067675, pp. 2013–2035.

Storey, John D. and Robert Tibshirani (2003). "Statistical significance for genomewide studies". *Proceedings of the National Academy of Sciences of the United States of America* 100(16), pp. 9440–9445.

Sul, Jae Hoon and Eleazar Eskin (2013). "Mixed models can correct for population structure for genomic regions under selection". *Nature Reviews Genetics* 14(4), p. 300.

Sul, Jae Hoon, Lana S. Martin, and Eleazar Eskin (2018). "Population structure in genetic studies: Confounding factors and mixed models". *PLoS Genet.* 14(12), e1007309.

Thornton, Timothy and Mary Sara McPeek (2010). "ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure". *Am. J. Hum. Genet.* 86(2), pp. 172–184.

Tucker, George, Alkes L. Price, and Bonnie Berger (2014). "Improving the Power of GWAS and Avoiding Confounding from Population Stratification with PC-Select". *Genetics* 197(3), pp. 1045–1049.

Voight, Benjamin F. and Jonathan K. Pritchard (2005). "Confounding from Cryptic Relatedness in Case-Control Association Studies". *PLOS Genetics* 1(3), e32.

Wang, Kai, Xijian Hu, and Yingwei Peng (2013). "An Analytical Comparison of the Principal Component Method and the Mixed Effects Model for Association Studies in the Presence of Cryptic Relatedness and Population Stratification". *HHE* 76(1), pp. 1–9.

Wojcik, Genevieve L. et al. (2019). "Genetic analyses of diverse populations improves discovery for complex traits". *Nature* 570(7762), pp. 514–518.

Wright, S. (1951). "The genetical structure of populations". *Ann Eugen* 15(4), pp. 323–354.

Wu, Chengqing et al. (2011). "A Comparison of Association Methods Correcting for Population Stratification in Case–Control Studies". *Annals of Human Genetics* 75(3), pp. 418–427.

Xu, Hanli and Yongtao Guan (2014). "Detecting Local Haplotype Sharing and Haplotype Association". *Genetics* 197(3), pp. 823–838.

Yang, Jian et al. (2011). "GCTA: a tool for genome-wide complex trait analysis". *Am. J. Hum. Genet.* 88(1), pp. 76–82.

Yang, Jian et al. (2014). "Advantages and pitfalls in the application of mixed-model association methods". *Nat Genet* 46(2), pp. 100–106.

Yu, Jianming et al. (2006). "A unified mixed-model method for association mapping that accounts for multiple levels of relatedness". *Nat. Genet.* 38(2), pp. 203–208.

Zhang, Feng, Yuping Wang, and Hong-Wen Deng (2008). "Comparison of Population-Based Association Study Methods Correcting for Population Stratification". *PLOS ONE* 3(10), e3392.

Zhang, Shuanglin, Xiaofeng Zhu, and Hongyu Zhao (2003). "On a semiparametric test to detect associations between quantitative traits and candidate genes using unrelated individuals". *Genetic Epidemiology* 24(1), pp. 44–56.

Zhao, Keyan et al. (2007). "An Arabidopsis Example of Association Mapping in Structured Samples". *PLOS Genetics* 3(1), e4.

Zhou, Quan, Liang Zhao, and Yongtao Guan (2016). "Strong Selection at MHC in Mexicans since Admixture". *PLoS Genet.* 12(2), e1005847.

S1 Supplementary figures

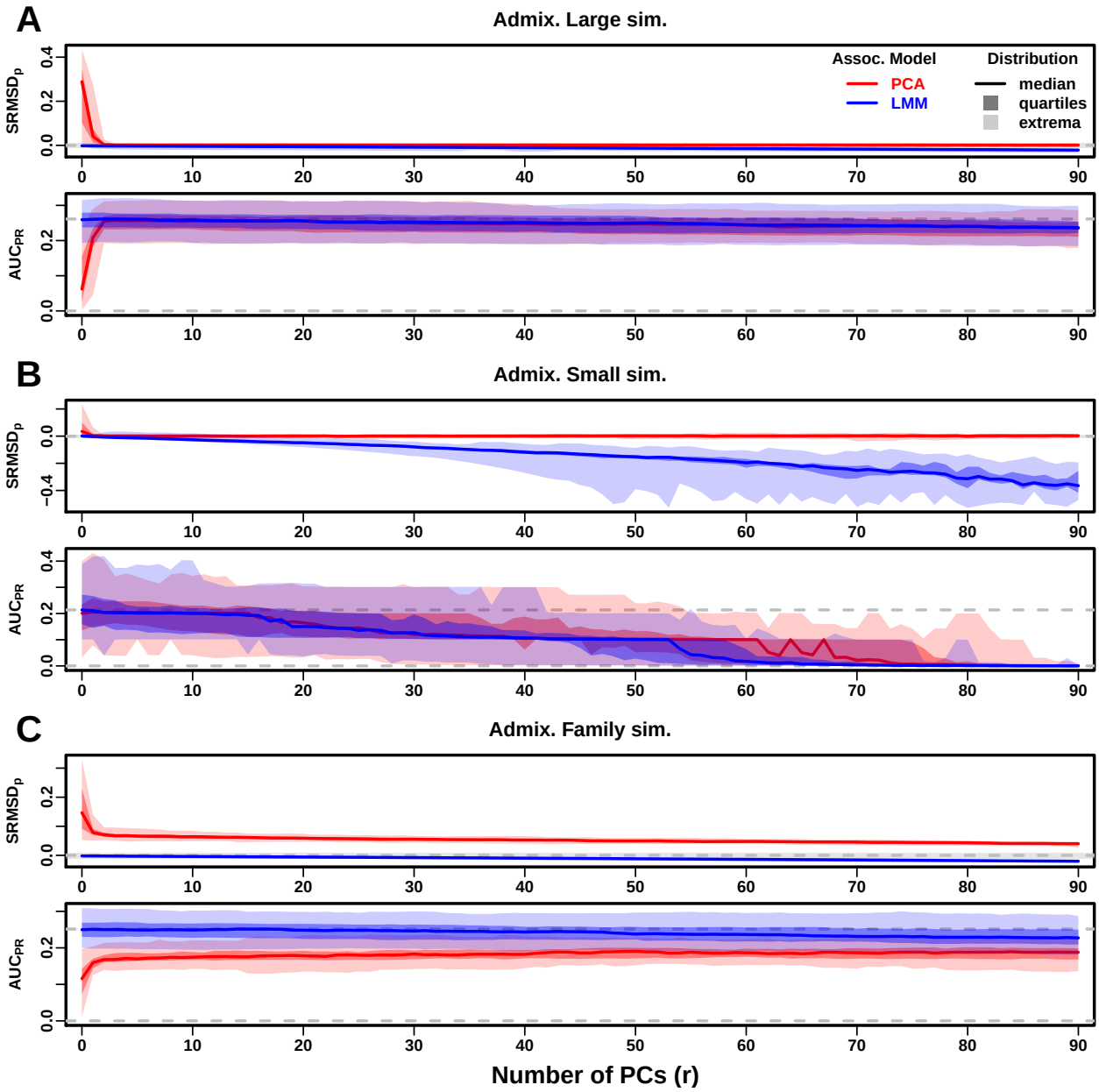


Figure S1: **Evaluations in admixture simulations.** Traits simulated from *random coefficients* model, otherwise the same as Fig. 3.

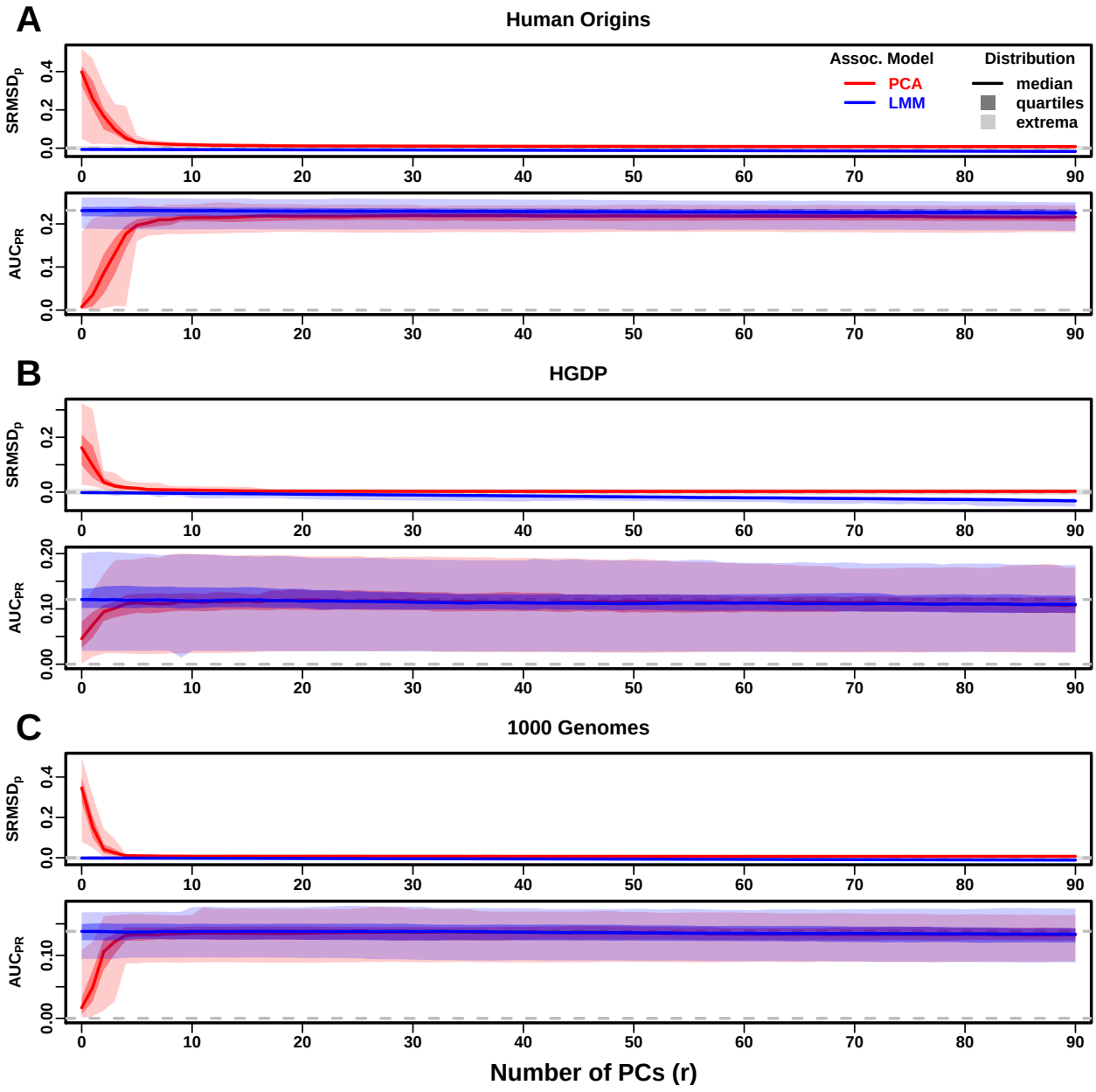


Figure S2: **Evaluations in real human genotype datasets.** Traits simulated from *random coefficients* model, otherwise the same as Fig. 4.

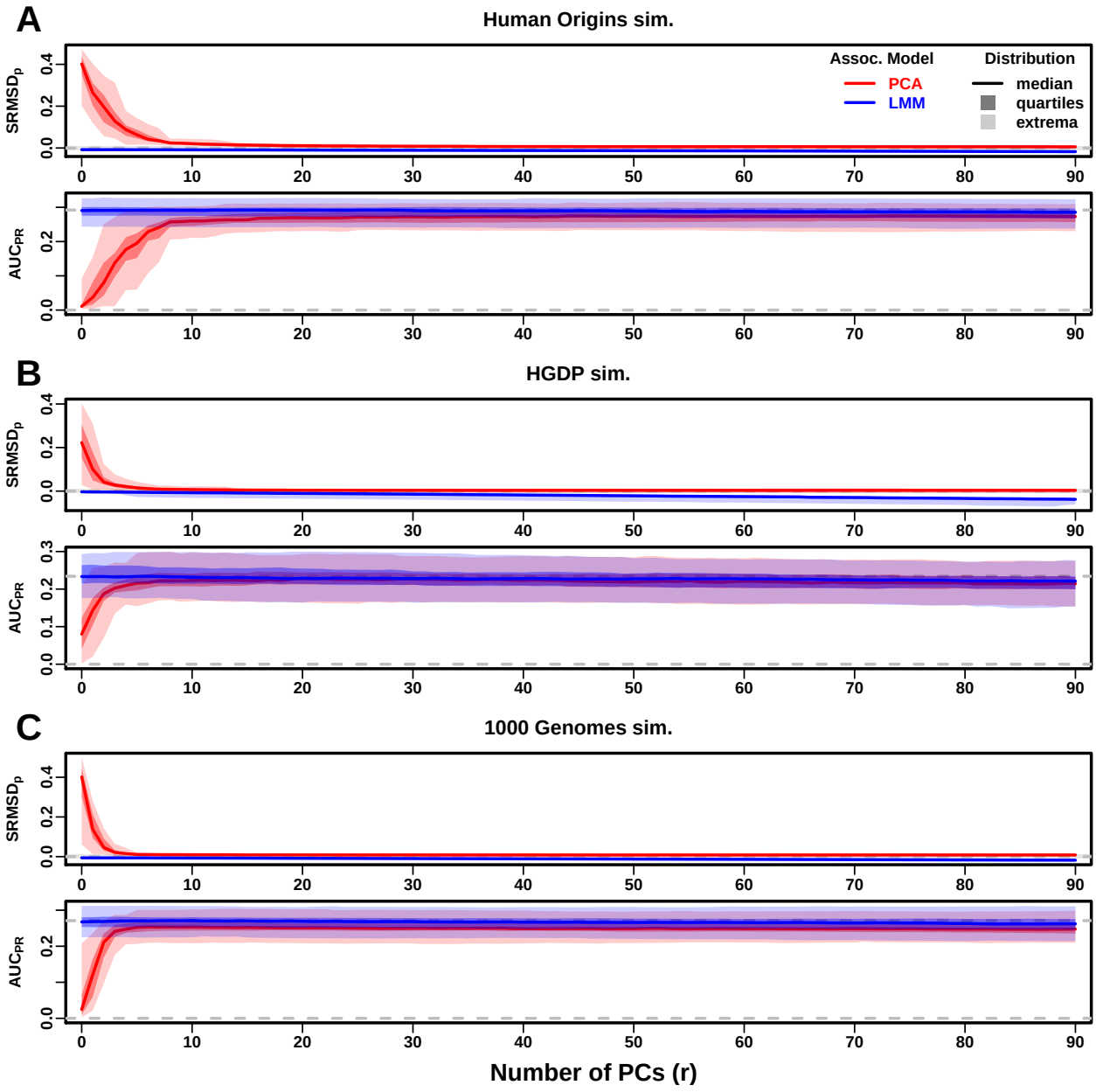


Figure S3: **Evaluations in tree simulations fit to human data.** Traits simulated from *random coefficients* model, otherwise the same as Fig. 5.