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

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 models for population structure, but nowadays the Linear Mixed-Effects Model (LMM) is believed by many to be a superior association 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 not using real genotype data. The use of LMMs with PCs has also been proposed, but evidence of effectiveness is lacking. In this work, we thoroughly evaluate PCA and LMM both with varying number of PCs in various realistic genotype and complex trait simulation scenarios, including admixture together with family structure, large real multiethnic human datasets (1000 Genomes Project, the Human Genome Diversity Panel, and Human Origins), and simulations from trees fit to each real dataset. We find that LMM without PCs performs best in all cases, with the largest effects in the family simulation and all of the real human datasets. We determined that the large gaps in PCA to LMM performance on the real human datasets is due to the high-dimensional family structure from large numbers of distant relatives, and not from the smaller number of highly related individuals present. While it was known that PCA fails on family data, here we report a strong effect on association of cryptic family relatedness in several human datasets that focused on capturing genetic diversity. Overall, this work better characterizes the limitations of principal components compared to mixed

effects models 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).

Many association models for structured populations are generalized linear models that incorporate this structure via covariates, which have included inferred ancestry or admixture 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 population 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 loci onto the eigenvectors (Jolliffe, 2002). PCs map to ancestry (*e.g.*, Alexander et al., 2009; 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; Alexander et al., 2009; 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) and polygenic models (Qian et al., 2020). 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 is 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.

The other dominant association model for structured populations is the LMM, in which this 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. The LMM model assumes a quantitative and polygenic (complex) trait. Interestingly, LMM and PCA share deep connections (Astle and Balding, 2009; Hoffman, 2013), which suggest that both models ought to perform similarly under certain conditions, particularly under low-dimensional relatedness. However, many previous 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 a hypothetical setting, namely unusually differentiated markers (Price et al., 2010; Wu et al., 2011), which as simulated are an artificial scenario not based on a population genetics model, and are otherwise believed to be unusual (Sul and Eskin, 2013; Price et al., 2013). A theoretical analysis claimed estimation biases under flawed assumptions, but LMM outperformed PCA in its simulation study (Wang et al., 2013). Moreover, various explanations for why LMM might outperforms PCA or viceversa 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. For that reason, much effort has been devoted to improving the runtime and scalability of LMMs (Aulchenko et al., 2007; Kang et al., 2008; Kang et al., 2010; Zhang et al., 2010; Lippert et al., 2011; Yang et al., 2011; Listgarten et al., 2012; Zhou and Stephens, 2012; Svishcheva et al., 2012; Loh et al., 2015; Zhou et al., 2018).

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

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

Publication	n[a]	m[a]	K	r	Scen	Sim[b]	F_{ST}	Real[c]	Trait	Inf	Power	Reps	LMM	Best
Zhang et al., 2003	150	300	4	1	3	IAF	?	AF,N	Q	T	Y	250	N	NA
Price et al., 2006	1,000	100,000	2	1-10	12	IA	0.01	N	CC	T	Y	10	N	NA
Yu et al., 2006*	277	1,384	NA	N(3)	6	N	0.12	Y	Q	T	Y	1	Y	NA
Epstein et al., 2007	1,000	111	3	10	2	$\tilde{I}$	0.15	N	CC	T	Y	1	N	NA
Kimmel et al., 2007	2,000	38,864	2-3	10?	2	$\tilde{I}$	NA	hap	CC	T	Y	100	N	NA
Zhao et al., 2007	95	900	NA	8	1	N	NA	Y	Q	Q	Y	1	Y	LMM
Luca et al., 2008	400	24,000	2-9	10?	2	I	0.01	N	CC	T	Y	1	N	NA
Zhang et al., 2008	4000	240	2	10?	1	I	NA	hap	CC	T	Y	1000	N	NA
Li and Yu, 2008	2000	10,000	2-4	10	5	IA	0.03	AF,N	CC	TI	Y	100	N	NA
Kang et al., 2008**	277	140,000	NA	N	3	N	NA	Y	Q	Q	Y	1000	YG?	NA
Astle and Balding, 2009	2,000	10,000	3	10	2	I	0.10	N	CC	Q	ROC	500	YG	Tie
Li et al., 2010	1,000	1,000	2-4	10	3	IA	0.01	AF,N	CC	TI	Y/N	100	N	NA
Kang et al., 2010	5,326	368,177	NA	2-100	2	N	NA	Y	Q+CC	IQ	N	1	YG	LMM
Thornton and McPeek, 2010	620	100,000	3	10?	6	IAF	0.01	Sc	CC	T	Y	1	N	NA
Price et al., 2010	1,000	100,000	2	1	4	IF	0.01	N	CC	I	N	1	YG	L+P
Wu et al., 2011	4,000	100,000	2-4	10	5	IA	0.01	N	CC	T	Y	10	YG	PCA
Bouaziz et al., 2011	?	5,500	1-5	5	6	GF	NA	GF	CC	T	Y	1	N	NA
Liu et al., 2011	1,000	10,000	2-3	10	4	IA	?	hap	Q	TQ	Y,ROC	1000	YG?	Tie
Hoffman, 2013	1,000	45,000	NA	N	2	N	NA	Y	Q	Q	Y	50	YG	NA
Sul and Eskin, 2013	1,000?	100,000?	2?	1?	2	I?	0.01	N	CC?	I	N	1	YG	Tie
Wang et al., 2013	500	NA	1,4	1-4	2	Other	NA	N	Q	N	N	1000	YF	LMM
Tucker et al., 2014	15,633	360,557	2	5	4	I	0.05	YN	Q	I	Y	100	YG	Tie
Yang et al., 2014	105,633	458,560	NA	5	2	N	NA	Y	CC	I	$\tilde{Y}$	1	YG	Tie?
Song et al., 2015	5,000	100,000	2-3	10?	33	IA	$\tilde{0}$.10	PC,N	Q+CC	T	N	100	YG	LMM
Loh et al., 2015	23,294	360,000	NA	10	2	N	NA	Y	Q	ITQ	Y	100	YG	LMM
?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
Sul et al., 2018	5,326	331,475	NA	100	1	N	NA	Y	Q	I	N	10	YG	LMM
This work	2,922	1,185,208	10-243	0-90	18	ATF	0.10	YN	Q	R	Y	50	YG	LMM

[a]Max sizes.
[b]Genotype simulation types.
[c]Evaluations combining real data with simulated data (not just a real data analysis).
*LMM had kinship from SPAGeDi (family-level relatedness), admixture proportions (no PCA).
**No PCA (only admix prop/structured assoc(SA)).

PCA has been compared to other association models, particularly to LMMs. 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; 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) 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).

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 (Astle and Balding, 2009; Hoffman, 2013), then the population structure is essentially modeled twice in an LMM with PCs, 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 (Price et al., 2010; Tucker et al., 2014), while others did not (Liu et al., 2011). ([TODO rephrase] Note that earlier LMM approaches estimated non-redundant kinship and fixed effects covariates: kinship matrices were estimated from pedigrees (thus excluding population structure) or estimated equivalent parameters from genotypes by assuming no inbreeding [TODO: cite SPaGeDi?], and population structure was modeled via admixture proportions rather than PCA (Yu et al., 2006; Zhao et al., 2007).)

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). The last three simulations aim to approximately match each of the real datasets by fitting trees and ancestral 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 and FDR control; Storey, 2003; Storey and Tibshirani, 2003) and classification performance (a robust alternative to power for miscalibrated models) via the area under precision-recall curves (Grau et al., 2015). Across all tests LMM without PCs consistently performs best. However, in our admixture simulations PCA nearly 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. For smaller studies (100 individuals) there is a pronounced loss of power when the number of PCs exceeds the optimal choice. However, 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 a small fraction 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. We confirmed the presence of family structure using the KING-robust estimator (Manichaikul et al., 2010), revealing a small number of highly related individuals greatly outnumbered by more distantly related individuals. Removing up to 4th degree relatives shows equally poor PCA performance as in the full datasets. All together, we find that LMMs without PCs are generally preferable, and present 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 utilized three real genotype datasets and simulated genotypes from six population structure scenarios to cover various features of interest (Table 2). We will introduce them here in sets of three, as they appear in the rest of our results. The population structures are also conveniently visualized in Fig. 1 using `popkin` to estimate population kinship matrices (which capture both family and population relatedness) 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	924,892	929	7-54	93	0.28
1000 Genomes	Real	1,111,266	2504	5-26	250	0.22
Human Origins sim.	Tree	190,394	2922	243	292	0.23
HGDP sim.	Tree	924,892	929	54	93	0.25
1000 Genomes sim.	Tree	1,111,266	2504	26	250	0.21

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

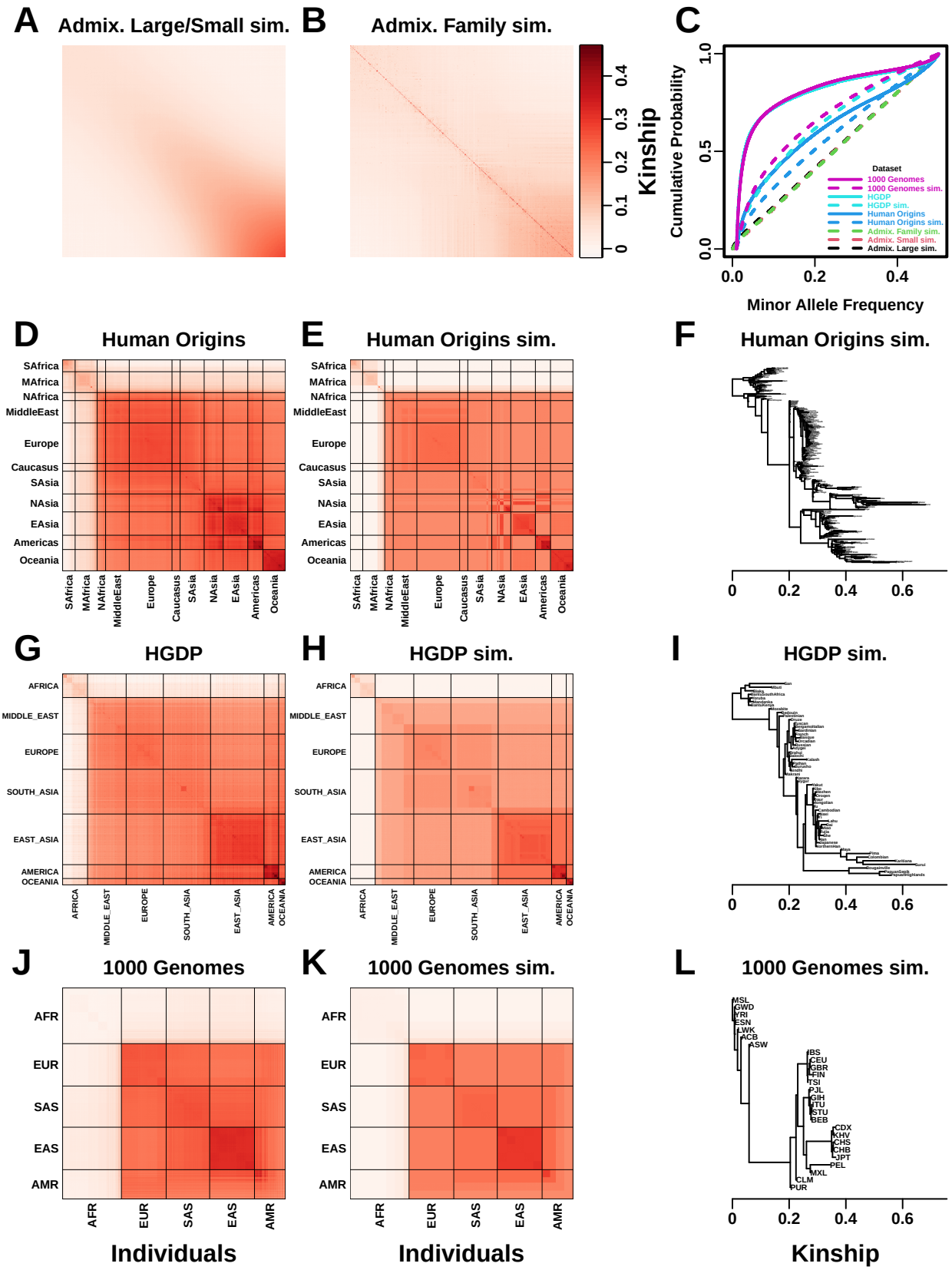
The first set of three simulated genotypes are based on an admixture model from $K = 10$ subpopulations (Fig. 1A) (Ochoa and Storey, 2021; Gopalan et al., 2016; Cabreros and Storey, 2019). The "large" version of this simulation, with 1000 individuals, illustrates the asymptotic performance of the association models. In contrast, the "small" simulation has just 100 individuals and displays overfitting for large numbers of PCs. For PCA, the theoretically ideal 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 simulation starts from an admixed founder population and draws 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. All real dataset are enriched for lower-than-uniform minor allele frequencies, even after excluding rare variants (MAF $< 1\%$; Fig. 1C).

The last set of three are tree-based simulations based on each of the real human datasets. A tree was fit to each kinship matrix 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, fits to the real data result in comparable covariance structures and scale of differentiation (Fig. 1E,H,K). By design, these tree simulations exclude relatedness more fine-grained than the subpopulation level, particularly any family structure present in the real dataset.

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, kinship represented with color (lighter is closer to zero, darker red are higher values). Diagonal shows 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 structure 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
IBS
CEU
GBR
FIN
TSI
PJL
GIH
ITU
STU
BEB
CDX
KHV
CHS
CHB
JPT
PEL
MXL
CLM
PUR
0
0.2
0.4
0.6
Kinship

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 are constructed in two different ways.

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 (T denotes the ancestral population). This corresponds with a rough inverse relationship between coefficient and minor allele frequency, which arises under evolutionary extremes of strong purifying [TODO add ref] and 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 causal locus classification performance (higher is better; Fig. 2). SRMSD_p is a more robust alternative to the common inflation factor λ and type I error measures; we found in our data a near one-to-one correspondence between λ and SRMSD_p, and determined that the threshold $\text{SRMSD}_p > 0.01$ corresponds to $\lambda > 1.06$ (Fig. S1) and thus evidence of inflation close to the rule of thumb of $\lambda > 1.05$ (Price et al., 2010). AUC_{PR} has been used previously to evaluate association models (Rakitsch et al., 2013), reflects

statistical power for calibrated models (see Models and Methods), and in general is a more robust alternative to statistical power for miscalibrated models. 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.

The overall goal is to characterize the performance of two association models: PCA and LMM.

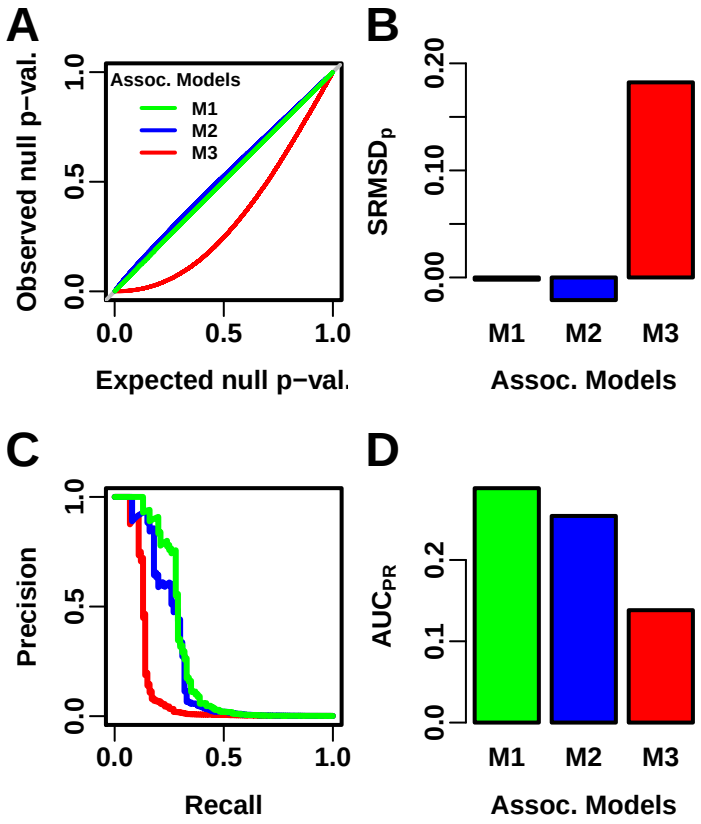


Figure 2: **Illustration of SRMSD$_p$ and AUC$_{PR}$ evaluation measures.** Three archetypal models (M1, M2, M3) illustrate our two complementary measures. M1 is an ideal model that performs best overall, M2 overcorrects for population structure so it incurs a small performance penalty, and M3 does not 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 causal locus classification 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 classification performance and reflects power (higher is better).

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 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 to determine whether a suboptimal distribution was significantly different. For PCA, the optimal number of PCs r is typically 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 $|\text{SRMSD}_p| < 0.01$ (marked with asterisks), whose null p-value distributions effectively uniform. Mean $|\text{SRMSD}_p| > 0.01$ cases for PCA were most common on the family simulation 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, comparing the $|\text{SRMSD}_p|$ distributions between PCA and LMM, each with their best r, resulted in LMM besting often or in statistical ties, whereas PCA was best in the Human Origins simulations only.

Next we turn to classification performance (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). Thus, for PCA there is often a tradeoff between accurate p-values versus classification performance. 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 SRMSD_p and AUC_{PR} distributions 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. S2).

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	87* (34*)	0*	LMM	19 (16)	1 (0)	LMM
1000 Genomes	FES	39 (32)	0*	LMM	8	1 (0)	LMM
Human Origins sim.	FES	90* (80*)	0*	tie	47 (36)	0	LMM
HGDP sim.	FES	43* (20*)	0*	LMM	17 (15)	0	LMM
1000 Genomes sim.	FES	77* (15*)	0*	LMM	16 (6)	2	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	77* (30*)	0*	LMM	19 (13)	3 (0)	tie (LMM)
1000 Genomes	RC	37* (27*)	0*	LMM	19 (4)	9 (2)	LMM
Human Origins sim.	RC	89* (85*)	0*	tie	45 (25)	0	LMM
HGDP sim.	RC	30* (23*)	0*	LMM	18 (15)	5 (0)	LMM
1000 Genomes sim.	RC	90* (16)	0*	LMM	10 (6)	2 (0)	LMM

[a]FES: Fixed Effect Sizes, RC: Random Coefficients.
[b]Smallest r (number of PCs) whose distribution ($\lvert\text{SRMSD}_p\rvert$ 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 ($\lvert\text{SRMSD}_p\rvert$ 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 $\lvert\text{SRMSD}_p\rvert < 0.01$ ($\lvert\text{SRMSD}_p\rvert$ columns only).

The large admixture simulation differs from previous admixture evaluations in featuring a larger number of ancestral populations ($K = 10$) and more differentiation ($F_{\mathrm{ST}} = 0.1$ 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 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, so p-values become conservative). The $\mathrm{AUC}_{\mathrm{PR}}$ distribution of PCA is similarly worst at $r = 0$, increases rapidly and peaks at $r = 3$, then decreases slowly for $r > 3$. Similarly, the $\mathrm{AUC}_{\mathrm{PR}}$ distribution for LMM starts near its maximum at $r = 0$, and decreases overall for larger r. Although the $\mathrm{AUC}_{\mathrm{PR}}$ distributions for LMM and PCA overlap considerably at each r, LMM with $r = 0$ has significantly greater $\mathrm{AUC}_{\mathrm{PR}}$ values than PCA with $r = 3$ (Table 3). However, qualitatively PCA closely matches LMM in performance in this simulation. Both LMM and PCA are robust to extreme values of 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 individuals to $n = 100$, which is small for typical association studies but may occur in studies of rare diseases, pilot studies, or other constraints. To compensate for the loss of power due to reducing n, we also reduce the number of causal loci from 100 before to $m_1 = 10$, (in all cases a fixed ratio of $n/m_1 = 10$) which increases the locus effect sizes. As expected, we found a large decrease 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). LMM attains much larger negative SRMSD_p values than in our other evaluations. While LMM with $r = 0$ is significantly better than PCA ($r = 1$ to 4) in both metrics (Table 3), qualitatively the difference is negligible.

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

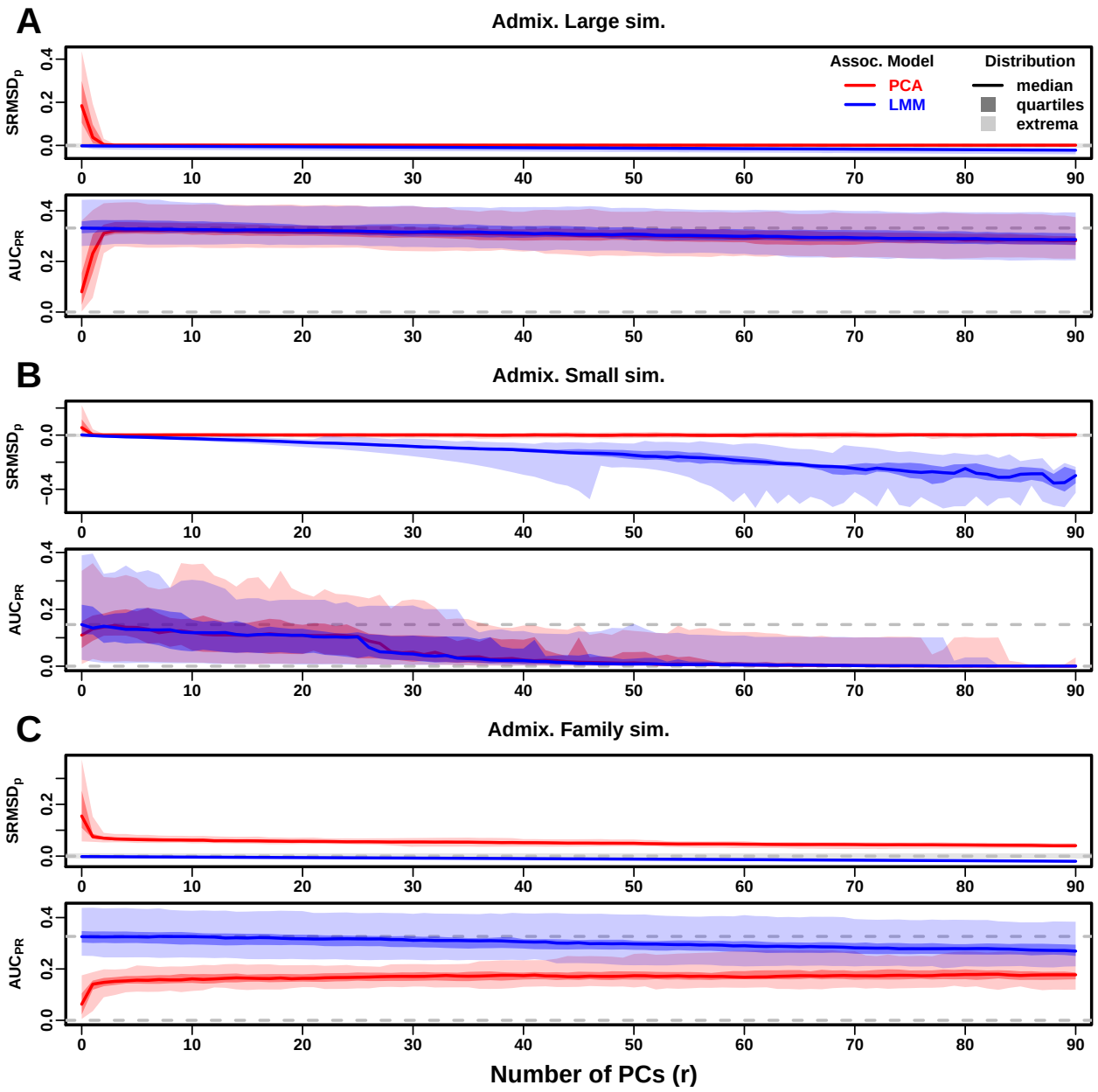


Figure 3: **Evaluations in admixture simulations.** Traits simulated from *fixed effect sizes* model. PCA and LMM approaches are tested with varying number of PCs ($r \in \{0, ..., 90\}$ on x-axis), with the distributions (y-axis) of SRMSD$_p$ (top subpanel) and AUC$_{\text{PR}}$ (bottom subpanel) for 50 replicates. Best performance is zero SRMSD$_p$ and large AUC$_{\text{PR}}$. Zero values and maximum median AUC$_{\text{PR}}$ 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, shows 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: SRMSD$_p > 0$ for all r, and a large gap in AUC$_{\text{PR}}$.

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 LMM is formulated in terms of kinship, 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 across the generations by strongly biasing mating pairs for proximity over the one-dimensional geography, which results in an indirect ancestry-biased assortative mating. Our evaluation reveals a sizable gap in both metrics between LMM and PCA across all values of r (Fig. 3C). LMM again performs best with $r = 0$ and achieves mean $|\text{SRMSD}_p| < 0.01$. However, 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 calibrated model, when there is family structure.

2.5 Evaluations in real human genotype datasets

Next we repeat our evaluations with real human genotype data, which differs from our simulations in allele frequency distributions and more complex population structures with greater differentiation, numerous correlated subpopulations, and potential cryptic family relatedness. We chose three datasets that span global human diversity and include both array and WGS genotyping platforms. 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. S3).

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). 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, but compared to Human Origins it contains many more loci and more low-frequency variants. The SRMSD_p and AUC_{PR}

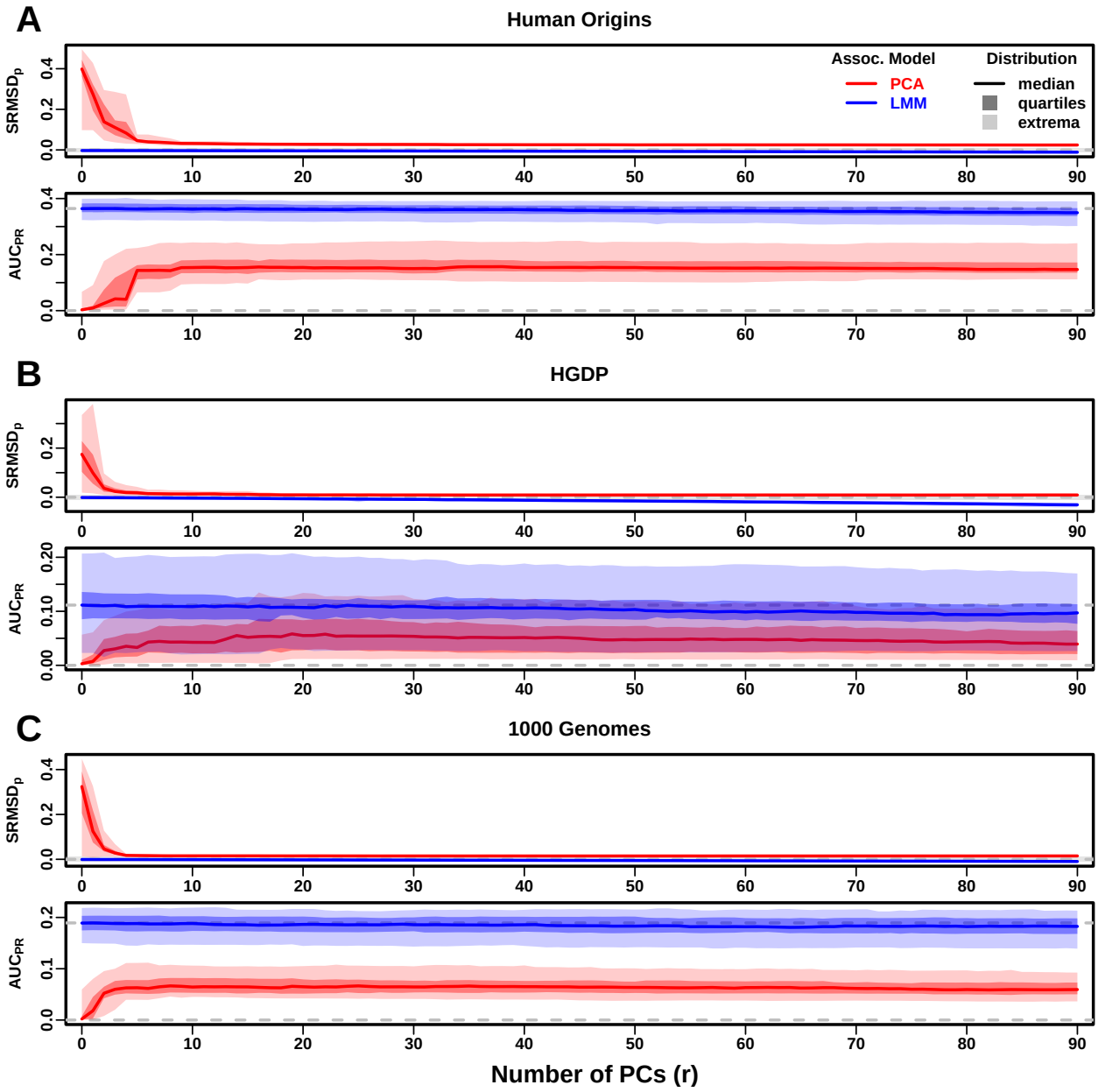


Figure 4: **Evaluations in real human genotype datasets.** Traits simulated from *fixed effect sizes* model. Same setup as Fig. 3, see that for details. These datasets strongly favor LMM with $r = 0$ PCs over PCA, resulting in curves that resemble the previous admixed family simulation, even though these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

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 a large number of individuals, and like HGDP is also WGS. Thus, although this dataset is expected to have the simplest population structure among the real datasets, we 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. S3). 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 (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 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 these tree simulations (Fig. 5) resembled our ad-

mixture simulation more than either the family simulation (Fig. 3) or real data results (Fig. 4). In all three of these 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 not significantly different for $|\text{SRMSD}_p|$ (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. The results are qualitatively similar for the random coefficients trait model (Fig. S4 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. S5A) slightly underestimated the true dimensionality of our simulations (Table 2). However, rank estimates agree that the admixed family simulation has the greatest rank, and that the real datasets have greater ranks than the admixture simulations and, to a lesser extent, their respective tree simulations. However, these estimated ranks do not differentiate datasets where PCA performs well from those where performance was poor, particularly between the real and tree simulations, which span a similar range of ranks. 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 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.

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

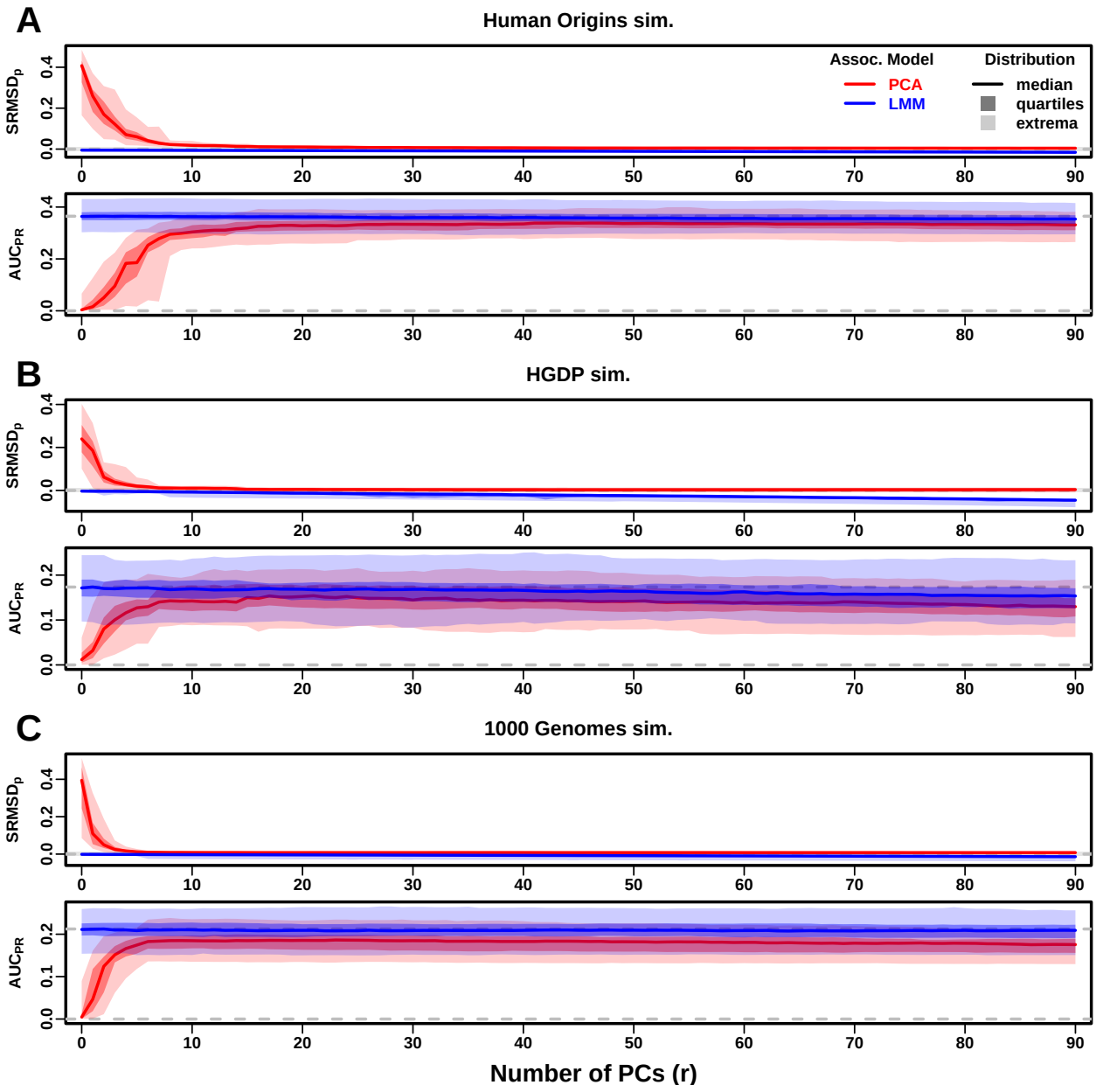


Figure 5: **Evaluations in tree simulations fit to human data.** Traits simulated from *fixed effect sizes* model. Same setup as Fig. 3, see that for details. These tree simulations, which exclude within-subpopulation family structure by design, do not explain the large gaps in LMM-PCA performance observed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

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. S5C), except the small admixture simulation had larger variances explained per eigenvalue (as expected since it has fewer eigenvalues). We also visualized all eigenvalues (beyond the top 10), computing their cumulative variance explained distributions versus their eigenvalue rank fraction (normalized to account for sample size differences). Each dataset has a different starting point, but all increase almost linearly from there until they reach 1, except for the family simulation, which has much greater variance explained by mid-rank eigenvalues (Fig. S5B). 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 Local kinship explains PCA performance

Local kinship, which is relatedness due to family structure ignoring population structure if present, is the presumed cause of the LMM to PCA gap observed in real datasets but not in their tree simulation counterparts. Rather than measure its presence indirectly as increased dimensionality, as attempted in the previous section, here we measure it directly using the KING-robust estimator (Fig. 6). As expected, we observe more larger local kinship values in the real datasets and the family simulation compared to the admixture and tree the simulations. However, this distribution in the real datasets depends on both sample size and number of subpopulations, as locally related pairs are most likely in the same subpopulation and conversely pairs between subpopulations tend to be negative for this estimator. For these reasons, the only comparable curve to each real dataset is their corresponding tree simulation, which matches sample size and subpopulation structure.

In all real datasets we identified highly related individual pairs, defined here as those with kinship values above the 4th degree relative threshold of 0.022 (Manichaikul et al., 2010; Conomos et al., 2016). However, these highly related pairs are vastly outnumbered by pairs who are less related but have greater than zero kinship, which may be inferred stringently as exceeding the maximum values in the corresponding tree simulations, ≈ 0.01, or if we are more permissive in allowing some small fraction of false positives we might reasonably include values above 0.001 (Fig. 6).

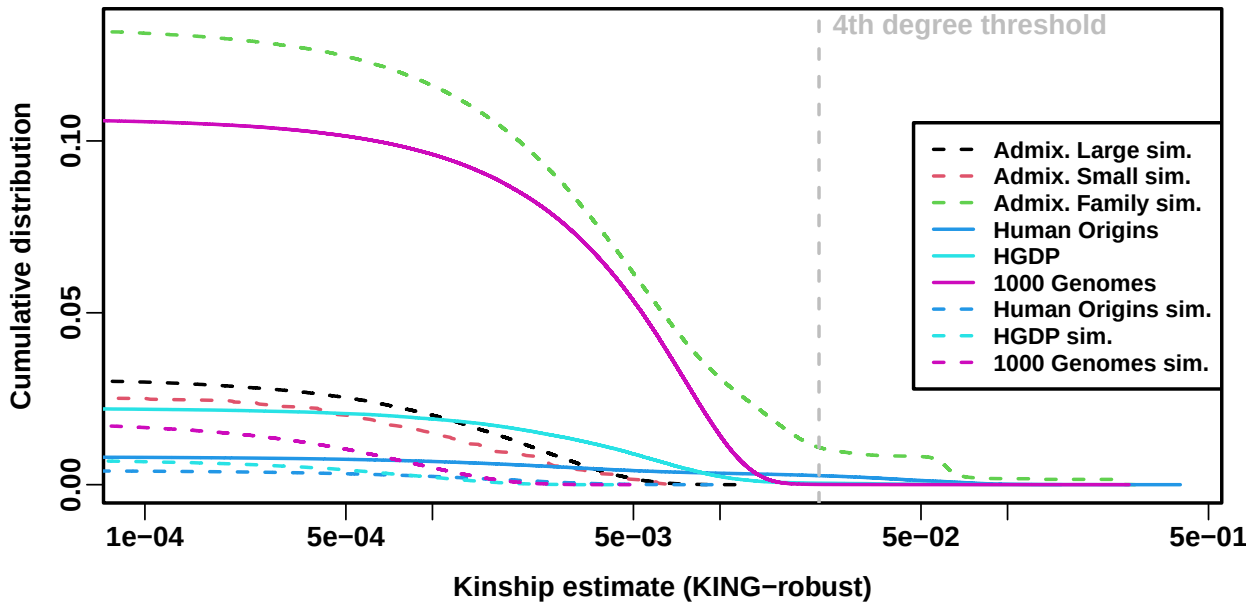


Figure 6: **Local kinship estimate distribution.** Curves are the complementary cumulative distribution of all lower triangular kinship matrix values from the KING-robust estimator. Self kinship is excluded. Note log scale of x-axis; negative estimates are counted in cumulatives but not shown. The majority of values in all datasets are below the 4th degree relative threshold value. Real datasets have greater cumulative distributions at every kinship value compared to their respective tree simulations.

We tested whether removal of the small number of highly related individual pairs improves PCA performance or if the presence of larger numbers of more distantly related pairs are to blame for poor PCA performance. After removing 4th degree relatives, which reduced sample sizes between 5% and 10% (Table S1), we find largely the same results as when all individuals were included. For simplicity a single number of PCs was tested for each model, $r = 0$ for LMM and $r = 20$ for PCA, as these performed well in our earlier evaluation. Similarly, only the fixed effect sizes trait was tested, as that previously showed a large gap in performance between association models. LMM significantly outperformed PCA in all these cases (Wilcoxon paired 1-tailed $p < 0.01$; Fig. 7). Notably, PCA still had miscalibrated p-values in Human Origins and 1000 Genomes ($|\text{SRMSD}_p| > 0.01$). Otherwise, AUC_{PR} and SRMSD_p ranges were similar in this evaluation and the one with all individuals. Therefore, the small number of highly related individual pairs had a negligible effect in PCA performance, so the larger number of more distantly related pairs explain the poor PCA performance compared to LMM in the real datasets.

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 for small sample sizes), such additions never resulted in significantly improved perfomance either (barring one marginally significant case with a small effect size; Table 3), which aggrees with previous observations (Liu et al., 2011) but contradicts others (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 contain such loci in relevant proportions if they exist, do not replicate those observations. However, the presence of cryptic relatedness on all of these datasets, which favors LMM, may obscure the effect. Therefore, while we

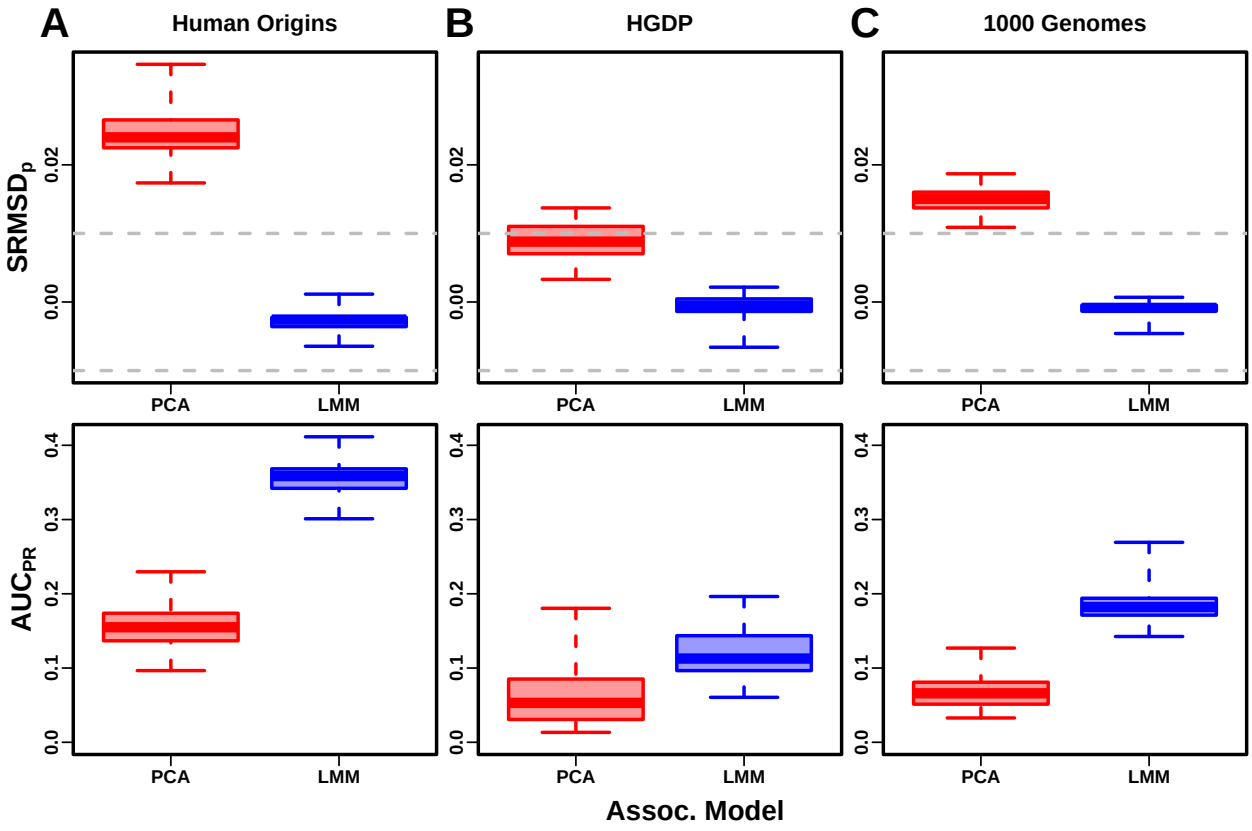


Figure 7: **LMM and PCA performance in real datasets excluding 4th degree relatives.** LMM had $r = 0$ PCs and PCA had $r = 20$, and only fixed effect sizes trait model was tested. Each dataset is a column, rows are metrics. The SRMSD_p cases (first row) have $|\text{SRMSD}_p| < 0.01$ band marked with dashed gray lines.

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 assumes that genetic structure is low-dimensional, whereas LMM can handle high-dimensional structures. Thus, PCA performs well in the admixture simulation, which is explicitly low-dimensional (see Models and Methods), as well as in tree simulations with few nodes with long branches, such as the trees we fit to the real human data, where a low-dimensional approximation suffices. 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. S5), 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 there may be low power to detect small eigenvalues, which were less important in our PCA evaluations as well). Our local kinship analysis confirms that there is considerable cryptic relatedness in all real human datasets, which explains why LMM outperforms PCA there. However, estimated eigenvalues and kinship matrix ranks by themselves do not fully explain when PCA will perform poorly. Further-

more, our evaluations reveal that the trait model 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. Direct local kinship estimation shows that there is considerable family relatedness in the real datasets absent in the corresponding tree simulations. Admixture is not modeled in our tree simulations, but our other admixture simulations concluded that this feature by itself is not problematic for PCA. Reanalysis of 1000 Genomes has identified 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, our evaluations showed that removal of these few highly related individuals does not improve PCA performance sufficiently, revealing the larger number of more distantly related pairs to be PCA's most serious obstacle. Cryptic relatedness is expected to be prevalent in any large human dataset (Henn et al., 2012; Shchur and Nielsen, 2018). Furthermore, our fixed-effect-sizes trait model shows that the challenges of cryptic relatedness are exacerbated when rare variants have large coefficients. Overall, 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 exception was the small simulation, where only a narrow range of PCs performed well. Thus, when using PCA it is best to err on including too many PCs rather than too few. Here LMM is simpler for users since there is no need to choose the number of PCs. However, 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, 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.

Overall, our results lead us to always recommend the use of LMM over PCA for association studies. 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 there is no guarantee that PCA will perform as well as LMM, in terms of both type I error control and power. The presence of large numbers of distant relatives, expected on any real dataset, all but ensures that PCA will perform poorly in practice compared to LMM for association studies. Our findings also suggest that other applications that employ PCA to control for population structure, such as polygenic models (Qian et al., 2020), may enjoy gains in power by instead employing an LMM or some other high-dimensional population structure model capable of modeling both population structure and cryptic relatedness.

4 Models and Methods

4.1 Models for genetic association studies

Here we describe the complex trait and kinship models that motivates both the PCA and LMM models for genetic association studies. The derivations of the PCA and LMM models from the general quantitative trait model are similar to previous presentations (Astle and Balding, 2009; 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 additive

linear model for a quantitative (continuous) trait is:

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

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. For simplicity, non-genetic covariates are not part of this model (or the PCA and LMM counterparts) but are trivial to include without changing any of our theoretical results.

In current datasets $m \gg n$, as there are millions of loci m while most studies have many fewer than a million individuals n. In this case the full model above cannot be fit, 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:

$$\mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta_i + \mathbf{U}_r\boldsymbol{\gamma}_r + \boldsymbol{\epsilon}, \tag{2}$$

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 follows from the singular value decomposition of the genotype matrix: $\mathbf{X}^\intercal = \mathbf{U}\mathbf{D}\mathbf{V}^\intercal$, where $\mathbf{U}$ 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 is 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

To better motivate the most common PC estimator for genotype data, and to connect PCA to LMMs, we shall review the kinship model for genotypes. Here 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 (on which random variables 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, the genotype matrix can be standardized using the true ancestral allele frequencies p_i^T, as

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

which results in a 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. (1) results in an equivalent model, as this covariate differs only by a linear transformation. Thus, starting from standardized genotypes, 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. (2) 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

$$\hat{\mathbf{\Phi}}^T = \frac{1}{m} \hat{\mathbf{X}}_S^\intercal \hat{\mathbf{X}}_S. \tag{3}$$

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).

4.1.4 Connection between PCs and ancestry proportions

Here we show that genetic association using ancestry proportions as covariates is equivalent in expectation to using PCs under the assumptions of the admixture model (*i.e.*, there are no other forms of relatedness), which has been demonstrated empirically before (Alexander et al., 2009; Zhou et al., 2016). 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, so ancestry proportions are non-negative and sum to one for every individual 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 the

mean subpopulation allele frequencies weighted by the ancestry proportions:

$$\pi_{ij} = \sum_{u=1}^{K} q_{ju} p_i^{S_u}. \tag{4}$$

Genotypes are the sum of alleles drawn independently from this frequency, or $x_{ij}|\pi_{ij} \sim \text{Binomial}(2, \pi_{ij})$. Thus, the rowspace of the genotype matrix equals in expectation that of the individual-specific allele frequency matrix, which by Eq. (4) 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. Moreover, since an intercept term is always included in association models ($\mathbf{1}\alpha$ in Eq. (2)), 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. (1). Most LMM implementations support fixed covariates (how we combine LMM with PCs), but for simplicity we exclude them in this presentation of the classical LMM, which is

$$\mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta_i + \mathbf{s} + \boldsymbol{\epsilon}, \tag{5}$$

which is like the PCA model in Eq. (2) 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. (1) 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, PCA is the fixed model equivalent of LMM under the additional approximation that only the top r eigenvectors are used, whereas LMM uses all eigenvectors. A more explicit comparison follows in the next subsection.

Another advantage of LMM over PCA is that it has fewer parameters to fit: ignoring the shared terms in Eq. (2) and Eq. (5), 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 most often employs a likelihood ratio test, whose test statistic has a asymptotic χ^2 distribution under the null hypothesis.

4.1.6 Connection between LMM and PCA

The LMM of Eq. (5) can be written to resemble more greatly the PCA model of Eq. (2) (Astle and Balding, 2009; Hoffman, 2013):

$$\mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta_i + \mathbf{U}\boldsymbol{\gamma}_{\text{LMM}} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\gamma}_{\text{LMM}} = \sigma_s \boldsymbol{\Lambda}^{1/2}\mathbf{r}, \tag{6}$$

where $\mathbf{U}$ is the complete matrix of PCs (all n of them), $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues of the kinship matrix such that its eigendecomposition is $\boldsymbol{\Phi}^T = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\intercal$, and $\mathbf{r} \sim \text{Normal}(\mathbf{0}, \mathbf{I})$ is a standard normal random effect. The connection follows since $\mathbf{s} = \sigma_s \mathbf{U}\boldsymbol{\Lambda}^{1/2}\mathbf{r}$ satisfies

$$\mathbf{s} \sim \text{Normal}\left(\mathbf{0}, \left(\sigma_s \mathbf{U}\boldsymbol{\Lambda}^{1/2}\right)\left(\sigma_s \mathbf{U}\boldsymbol{\Lambda}^{1/2}\right)^\intercal\right) = \text{Normal}(\mathbf{0}, \sigma_s^2 \boldsymbol{\Phi}^T),$$

which itself follows from the affine transformation property of multivariate normal distributions. $\mathbf{U}$ also equals in the limit the right singular vectors of the standardized genotype matrix $\mathbf{X}_S = \mathbf{V}_S\mathbf{D}_S\mathbf{U}_S^\intercal$, which is how we originally motivated PCA, since

$$\frac{1}{m}\mathbf{X}_S^\intercal\mathbf{X}_S = \mathbf{U}_S\mathbf{\Lambda}_S\mathbf{U}_S^\intercal \xrightarrow[m\to\infty]{\text{a.s.}} \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\intercal = \mathbf{\Phi}^T$$

since $\mathbf{V}_S$ is orthonormal, and where $\mathbf{\Lambda}_S = \frac{1}{m}\mathbf{D}_S^2$.

Therefore, when the true kinship matrix is low-dimensional, the LMM fits the same low-dimensional space as PCA would with an appropriate number of PCs, except that the coefficients $\boldsymbol{\gamma}_{\text{LMM}}$ are constrained by the likelihood model, whereas the analogous PCA coefficients $\boldsymbol{\gamma}_r$ are unconstrained parameters. On the other hand, when the kinship matrix is high-dimensional then the LMM can fit this space better than a corresponding PCA model with a small number of PCs, while PCA with a larger number of PCs will instead overfit the data.

4.2 Simulations

Let f_B^A denote 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. The basic admixture model is as described previously (Ochoa and Storey, 2016; Ochoa and Storey, 2021).

Large and Family 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, or ancestries. Each subpopulation S_u ($u \in \{1, ..., K\}$) lies at coordinate u and has an inbreeding coefficient $f_{S_u}^T = u\tau$ for some τ. Ancestry proportions q_{ju} for individual j and subpopulation S_u arise from a random walk model on the given 1-dimensional geography with spread σ, and the free parameters τ and σ are fit to result in $F_{\text{ST}} = 0.1$ and mean kinship $\bar{\theta}^T = 0.5F_{\text{ST}}$ for the admixed

individuals, as before (Ochoa and Storey, 2021).

Random allele frequencies and genotypes are drawn from this hierarchical model:

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

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 with mean p_i^T and variance $p_i^T \left(1 - p_i^T\right) f_{S_u}^T$ (Balding and Nichols, 1995). Lastly, the individual-specific allele frequencies π_{ij} and genotypes x_{ij} are as described earlier (Eq. (4)). Fixed loci (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. Each replicate draws a new genotype matrix starting from new ancestral allele frequencies.

4.2.2 Genotype simulation from random admixed families

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

The 20 generations are drawn iteratively. Generation 1 has individuals with genotypes drawn from the large admixture simulation described earlier. 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 and outbred. Individuals are randomly assigned to male or female with equal probability.

The children of the previous generation serve as the parents in the next generation, paired iteratively as follows. From the pool of available males, one is picked randomly and is paired with the nearest female that is not a second cousin or closer relative (local kinship must be $< 1/4^3$); males that are not pairable are removed from the pool of available males. Pairing stops when there are no more available males or females.

Next, a random number of children per pair is constructed to yield a desired population size n and a minimum family size of $n_m = 1$, as follows. Let n_f the be number of families (paired parents) in the current generation, then the number of additional children (beyond the minimum) is drawn from a Poisson distribution with parameter $n/n_f - n_m$. Although the mean population size is n as desired, the random sample may deviate from this target size. Let δ be the difference between desired and current population sizes. If $\delta > 0$, then δ random families are incremented by 1. If $\delta < 0$, then $|\delta|$ random families with at least $n_m + 1$ children are decremented by 1. If $|\delta|$ exceeds the number of families, all families are incremented or decremented as needed and the process is iterated. Children are assigned sex randomly, and are reordered by the average coordinate of their parents, preserving the original order when there are ties.

A new random pedigree was drawn for each replicate, as well as new genotypes for the founders drawn anew from the admixture model. Genotypes are drawn across the pedigree, children drawing alleles from their parents independently per locus.

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 is a subpopulation S_w, where the nodes are indexed (w) arbitrarily. Each edge between S_w and its parent population P_w has an inbreeding coefficient $f_{S_w}^{P_w}$. Allele frequencies are drawn from the root to the tips of the tree iteratively, as a hierarchical or graphical model, since this tree is 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). Given the allele frequencies $p_i^{P_w}$ of the parent population P_w, 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:

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

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 calculate recursively here. Suppose as before that every node in the tree, including root and tip nodes, are indexed as S_w. Coancestry values ϑ_{uv}^T for a pair of subpopulations S_u and S_v 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 the coancestry of subpopulations S_u and S_v equals the total inbreeding of the most recent common ancestor (MRCA) population of those subpopulations:

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

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

We will calculate total coancestries (from the ancestral population T, namely $f_{S_w}^T$) for every

node S_w recursively through the tree branches from the root, as follows. Recall the value of the edge to S_w from its parent P_w is $f_{S_w}^{P_w}$, which is given. Nodes whose parent is $P_w = T$ are already of the desired form. If $f_{P_w}^T$ has been calculated, then $f_{S_w}^T$ is given by

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

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 $f_{S_w}^{P_w}$ to $f_{P_w}^T$ we have to shrink $f_{S_w}^{P_w}$ first by a factor of $\left(1 - f_{P_w}^T\right)$. Define the additive contribution of the edge to S_w as

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

These $\delta_w \geq 0$ because $0 \leq f_{S_w}^{P_w}, f_{P_w}^T \leq 1$ for every w. Importantly, the inbreeding edge values can be recovered from these additive edges recursively from the root, since nodes S_w connected to the root satisfy $f_{S_w}^{P_w} = f_{S_w}^T = \delta_w$, and given $f_{P_w}^T$ we can calculate $f_{S_w}^{P_w}$ and $f_{S_w}^T$ using

$$f_{S_w}^{P_w} = \frac{\delta_w}{1 - f_{P_w}^T}, \quad f_{S_w}^T = f_{P_w}^T + \delta_w.$$

The coancestry values are simplified as a sum of additive contributions δ_w for the nodes that are ancestors of the given pair of subpopulations,

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

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 by the tree topology, while δ_w reflect the edge values. Therefore, given a topology, δ_w can be estimated by non-negative linear regression (see below), where $I_w(u, v)$ define the design matrix.

Estimating population-averaged coancestry. Kinship ($\hat{\varphi}_{jk}^T$) is estimated using `popkin` (Ochoa and Storey, 2021). Coancestry ($\hat{\theta}_{jk}^T$) is estimated from kinship by replacing self-kinship with

inbreeding ($\hat{f}_j^T$) along the diagonal:

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

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

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

Estimating tree topology. Our topology estimation approach is remarkably simple, stemming from the monotonic relationship between node depth and coancestry from Eq. (7). 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 is

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

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

Estimating tree edge lengths. Additive edge lengths δ_w are estimated from Eq. (7) from the estimated subpopulation coancestry matrix $\hat{\vartheta}_{uv}^T$, 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 $f_{S_w}^{P_w}$ 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, δ_0 is ignored when drawing allele frequencies from the estimated tree.

4.2.5 Fitting ancestral allele frequency 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 the p_i^T distribution over all loci i satisfies $\mathrm{E}\left[p_i^T\right] = \frac{1}{2}$ (it is symmetric about 0.5) 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 $V^T \leq \frac{1}{4}$ 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

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

Estimation of ancestral variance. Given empirical sample allele frequencies $\hat{p}_i^T$, we use

an unbiased sample estimator for W^T that assumes a known expectation of one half, which is appropriate treating the choice of reference allele as random:

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

The mean kinship $\bar{\varphi}^T$ is calculated from the tree simulation parameters: the subpopulation coancestry matrix calculated using Eq. (7), expanded so rows and columns corresponds to individuals rather than subpopulations, the diagonal is converted to kinship (reversing Eq. (8)), and the matrix averaged. However, our 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 model parameter resulted in simulations with more accurately specified population structures. For Human Origins the true model $\bar{\varphi}^T$ of 0.143 was used. For 1000 Genomes and HGDP the true model values $\bar{\varphi}^T$ are 0.126 and 0.124, respectively, but tests showed that 0.4 was better 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 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 equate to the desired V^T in Eq. (9) and solve for w in this quadratic equation. For simplicity, we set $V_{\mathrm{mix}} = V^T$, which is achieved with the following 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 $w = 0$ yields $V^{T'} = V^T$, we use the second root of the quadratic equation to use the input

$\hat{p}_i^T$ data:

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

4.2.6 Real human genotype datasets

The three 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 removed. 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 Pacific data (Skoglund et al., 2016) was obtained as a separate dataset from the rest (Lazaridis et al., 2014; Lazaridis et al., 2016), and datasets were merged using the intersection of loci. We removed ancient individuals, and individuals from singleton and non-native subpopulations. Non-autosomal loci 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 with filter "PASS".

1000 Genomes Project. The recent high-coverage NYGC version of the 1000 Genomes Project (Fairley et al., 2020) was downloaded from `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20190425_NYGC_GATK/`. Our analysis was restricted to autosomal biallelic SNP loci with filter "PASS".

LD prunning. Our evaluations require uncorrelated loci, so that non-causal loci are not correlated to the trait and 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.

Close relative removal. For the evaluation with close relatives removed, each dataset was filtered with `plink2` using the option "`--king-cutoff`" with cutoff 0.02209709 ($= 2^{-11/2}$) for removing up to 4th degree relatives using the KING-robust local kinship estimator (Manichaikul et al., 2010). After removing these individuals, an MAF filter of 0.01 is applied again (Table S1).

4.2.7 Trait Simulation

Simulated complex traits are constructed from the additive quantitative trait model in Eq. (1) from a given genotype matrix (simulated or real). To simulate the correct heritability, true ancestral allele frequencies p_i^T are required, which are only available for simulated genotypes. We extend the procedure to real genotypes and estimated allele frequencies $\hat{p}_i^T$ by implementing novel 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 $h^2 = 0.8$. Non-genetic effects are drawn independently: $\epsilon_j \sim \text{Normal}(0, 1 - h^2)$. 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 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 with MAF ≥ 0.01, to avoid simulations with very rare causal variants (neither PCA nor LMM are appropriate inference models for rare variants).

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

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

A subset of causal coefficients is randomly selected, each with probability 0.5, and multiplied by -1.

For unknown p_i^T, we replace v_i^T with the following unbiased estimator (Ochoa and Storey, 2021):

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

where $\bar{\varphi}^T$ is the mean kinship estimated with `popkin`.

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

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

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

where $\hat{v}_i^T$ is used instead of v_i^T if needed, in which case σ_0^2 is an unbiased estimate of the total genetic variance. The desired genetic variance h^2 is obtained by multiplying every β_i by $\frac{h}{\sigma_0}$.

Intercept coefficient. For known p_i^T, the intercept coefficient in Eq. (1) 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. (3) 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 calculate the eigenvalues, without excluding individuals. The vector of eigenvalues was passed to `twstats` 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 (note 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 uniformity of non-causal p-values, which is a prerequisite for type I error and FDR control. The second measure, AUC_{PR}, quantifies causal locus classification performance of each model and reflects power, while making it possible to rank miscalibrated models fairly. We also define and contrast to related performance measures from the literature.

4.3.1 SRMSD_p: a measure of p-value uniformity

From their definition, p-values for continuous test statistics have a uniform distribution when the null hypothesis holds. This fact is crucial for type I error and FDR control by common approaches such as q-values (Storey, 2003; Storey and Tibshirani, 2003). We use the Signed Root Mean Square Deviation (SRMSD) to measure the difference 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 its expectation, and sgn is the sign function (1 if $u_{\text{median}} \geq p_{\text{median}}$, -1 otherwise). Thus, $\text{SRMSD}_p = 0$ corresponds to the best performance (calibrated p-values), $\text{SRMSD}_p > 0$ indicate anti-conservative p-values, and $\text{SRMSD}_p < 0$ are conservative p-values. The maximum SRMSD_p is achieved when all p-values are zero (the limit of anti-conservative p-values), which for infinite loci approaches

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

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

4.3.2 The inflation factor λ

Test statistic inflation has been used to measure successful population structure modeling (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 (across all p-values, not the null subset) 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 its null expectation, and F is the χ^2 cumulative density function (F^{-1} is the quantile function). This equation is useful to compare p-values from statistics that have non-χ^2 distributions (such as t statistics).

To compare λ and SRMSD_p directly, for simplicity assume that all p-values are null. In this case, calibrated p-values give $\lambda = 1$ and $\text{SRMSD}_p = 0$. However, misspecified null test statistic distributions with the expected median (such as genomic control; Devlin and Roeder (1999)) result in $\lambda = 1$, but $\text{SRMSD}_p \neq 0$ except for the expected distribution; this is the important flaw of λ that SRMSD_p overcomes. Inflated statistics (anti-conservative p-values) give $\lambda > 1$ and $\text{SRMSD}_p > 0$. Deflated statistics (conservative p-values) give $\lambda < 1$ and $\text{SRMSD}_p < 0$. Thus, $\lambda \neq 1$ always implies $\text{SRMSD}_p \neq 0$ (and in that case $\lambda - 1$ and SRMSD_p have the same sign), but not the other way around. Overall, λ depends only on the median of the statistic distribution, which SRMSD_p improves upon by making use of the complete distribution. However, SRMSD_p requires knowing which loci are null, so unlike λ it is only applicable to simulated traits.

4.3.3 Comparison between SRMSD_p and inflation factor λ

One advantage of SRMSD_p is that its range is bounded, while λ is unbounded and on a log scale. There is a near one-to-one correspondence between λ and our SRMSD_p statistics (Fig. S1). PCA tended to be inflated ($\lambda > 1$ and $\text{SRMSD}_p > 0$) whereas LMM tended to be deflated ($\lambda < 1$ and $\text{SRMSD}_p < 0$), otherwise the data for both models fall on the same contiguous curve. We fit the

following sigmoidal function to this data,

$$\mathrm{SRMSD}_p(\lambda) = a\frac{\lambda^b - 1}{\lambda^b + 1}, \tag{10}$$

which for all $a, b > 0$ satisfies $\mathrm{SRMSD}_p(\lambda = 1) = 0$ and reflects $\log(\lambda)$ about zero ($\lambda = 1$) as desired, namely that

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

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. S1 with parameters $a = 0.566$ and $b = 0.616$. The value $\lambda = 1.05$, a common threshold for benign inflation (Price et al., 2010), corresponds to $\mathrm{SRMSD}_p = 0.0085$ according to Eq. (10). Conversely, $\mathrm{SRMSD}_p = 0.01$, serving as a simpler rule of thumb, corresponds to $\lambda = 1.06$.

4.3.4 Type I error rate

The type I error rate is the proportion of null p-values below a given p-value threshold t. Calibrated p-values result in a type I error rate near t, which may be evaluated for significance using a binomial test. This measure depends on the choice of threshold t and can give rise to misleading results, as a model may appear calibrated for some t but not for others, or may be significantly miscalibrated only for sufficiently large t (due to lack of power for smaller t). In contrast, $\mathrm{SRMSD}_p = 0$ guarantees calibrated type I error rates at all thresholds t, while large $|\mathrm{SRMSD}_p|$ indicates incorrect type I errors for at least some thresholds t.

4.3.5 $\mathrm{AUC}_{\mathrm{PR}}$: the area under the precision-recall curve

Precision and recall are standard performance measures for binary classifiers that does not require calibrated p-values (Grau et al., 2015). Let c_i be the the true classification of locus i: $c_i = 1$ for causal loci ($\beta_i \neq 0$), $c_i = 0$ otherwise. 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 loci, the number of true positives (TP), false positives (FP) and false negatives (FN) at threshold t is

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

Precision and recall at this threshold are given by

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

Precision and Recall trace a curve as the threshold t is varied, and the area under this curve is $\mathrm{AUC_{PR}}$. A model obtains the maximum $\mathrm{AUC_{PR}} = 1$ if there is some threshold that classifies all loci perfectly. In contrast, the worst models, which classify at random, have an expected precision ($= \mathrm{AUC_{PR}}$) equal to the overall proportion of alternative cases: $\pi_1 = \frac{m_1}{m} = \frac{1}{m} \sum_{i=1}^{m} c_i$.

4.3.6 Statistical power and comparison to $\mathbf{AUC_{PR}}$

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

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

Note that $F(t)$ is a cumulative function, so it is monotonically increasing with t and has an inverse.

Power is hard to interpret when p-values are not calibrated. However, to establish a clear connection to $\mathrm{AUC}_{\mathrm{PR}}$, assume calibrated p-values so the distribution under the null hypothesis H_0 is uniform: $\Pr(p < t|H_0) = t$. The numbers of true positives, false positives, and false negatives at t are therefore

$$\begin{aligned}
\mathrm{TP}(t) &= m\pi_1 F(t), \\
\mathrm{FP}(t) &= m\pi_0 t, \\
\mathrm{FN}(t) &= m\pi_1 (1 - F(t)),
\end{aligned}$$

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

$$\begin{aligned}
\mathrm{Precision}(t) &= \frac{\pi_1 F(t)}{\pi_1 F(t) + \pi_0 t}, \\
\mathrm{Recall}(t) &= F(t).
\end{aligned}$$

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

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

Finally, the area under the curve equals $\mathrm{AUC}_{\mathrm{PR}} = \int_0^1 \mathrm{Precision}(\mathrm{Recall}) d\mathrm{Recall}$.

Now lets consider a simple yet common case in which model A is uniformly more powerful than model B, so that their power functions satisfy $F_A(t) \geq F_B(t)$ for every t. It follows that the inverse functions satisfy $F_A^{-1}(\mathrm{Recall}) \leq F_B^{-1}(\mathrm{Recall})$ for every recall value. This ensures that the precision of model A is greater or equal than the precision of model B at every recall value, and therefore $\mathrm{AUC}_{\mathrm{PR}}$ is greater or equal for A than B.

4.4 Software

We selected fast and robust software implementing the basic PCA and LMM models based on internal benchmarks.

PCA association was performed using `plink2` (Chang et al., 2015). The quantitative trait model is a linear regression with covariates, with significance assessed using the t-test. PCs were calculated with `plink2`, which equals the top eigenvectors of Eq. (3) after removing loci with MAF < 0.1. KING-robust local kinship estimates and removal of related individuals were also performed with `plink2`.

LMM association was performed using GCTA (Yang et al., 2011). GCTA also estimates kinship using Eq. (3), except self-kinship uses a different formula (Yang et al., 2011). PCs were calculated using GCTA from its kinship estimate. When running GCTA with large numbers of PCs in the Small simulation, we encountered convergence and singularity errors in some replicates, where 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`.

Our trait simulation procedure and the AUC_{PR} and SRMSD_p measures, all introduced in this work, are implemented in the R package `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).

Unbiased population 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).

Complete code reproducing our results, data, and this manuscript is available at `https://`

`github.com/OchoaLab/pca-assoc-paper`.

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.

Aulchenko, Yurii S., Dirk-Jan de Koning, and Chris Haley (2007). "Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis". *Genetics* 177(1), pp. 577–585.

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.

Conomos, Matthew P. et al. (2016). "Model-free Estimation of Recent Genetic Relatedness". *The American Journal of Human Genetics* 98(1), pp. 127–148.

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.

Fairley, Susan et al. (2020). "The International Genome Sample Resource (IGSR) collection of open human genomic variation resources". *Nucleic Acids Research* 48(D1), pp. D941–D947.

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.

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

Kang, Hyun Min et al. (2008). "Efficient control of population structure in model organism association mapping". *Genetics* 178(3), pp. 1709–1723.

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.

Lippert, Christoph et al. (2011). "FaST linear mixed models for genome-wide association studies". *Nat. Methods* 8(10), pp. 833–835.

Listgarten, Jennifer et al. (2012). "Improved linear mixed models for genome-wide association studies". *Nat Methods* 9(6), pp. 525–526.

Liu, Nianjun et al. (2011). "Controlling Population Structure in Human Genetic Association Studies with Samples of Unrelated Individuals". *Stat Interface* 4(3), pp. 317–326.

Loh, Po-Ru et al. (2015). "Efficient Bayesian mixed-model analysis increases association power in large cohorts". *Nat. Genet.* 47(3), pp. 284–290.

Luca, Diana et al. (2008). "On the Use of General Control Samples for Genome-wide Association Studies: Genetic Matching Highlights Causal Variants". *The American Journal of Human Genetics* 82(2), pp. 453–463.

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

Manichaikul, Ani et al. (2010). "Robust relationship inference in genome-wide association studies". *Bioinformatics* 26(22), pp. 2867–2873.

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.

— (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.

Rakitsch, Barbara et al. (2013). "A Lasso multi-marker mixed model for association mapping with population structure correction". *Bioinformatics* 29(2), pp. 206–214.

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.

Svishcheva, Gulnara R. et al. (2012). "Rapid variance components–based method for whole-genome association analysis". *Nat Genet* 44(10). Bandiera_abtest: a Cg_type: Nature Research Journals Number: 10 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Computational biology and bioinformatics;Genetic models;Genome-wide association studies Subject_term_id: computational-biology-and-bioinformatics;genetic-models;genome-wide-association-studies, pp. 1166–1170.

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.

Zhang, Zhiwu et al. (2010). "Mixed linear model approach adapted for genome-wide association studies". *Nat Genet* 42(4). Bandiera_abtest: a Cg_type: Nature Research Journals Number: 4 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Genome-wide association studies;Population dynamics;Statistical methods Subject_term_id: genome-wide-association-studies;population-dynamics;statistical-methods, pp. 355–360.

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.

Zhou, Wei et al. (2018). "Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies". *Nat Genet* 50(9). Bandiera_abtest: a Cg_type: Nature Research Journals Number: 9 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Genome-wide association studies;Statistics Subject_term_id: genome-wide-association-studies;statistics, pp. 1335–1341.

Zhou, Xiang and Matthew Stephens (2012). "Genome-wide efficient mixed-model analysis for association studies". *Nat. Genet.* 44(7), pp. 821–824.

S1 Supplementary figures

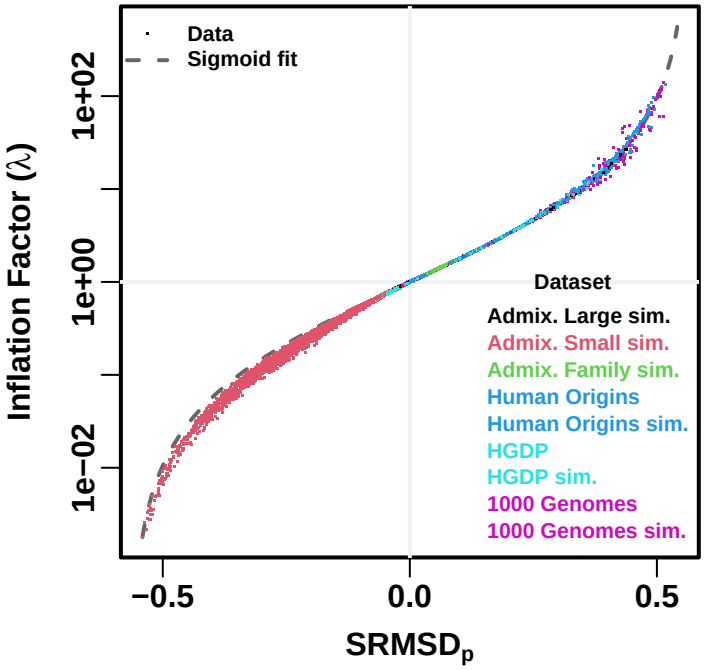


Figure S1: **Comparison between SRMSD$_p$ and inflation factor.** Each point is a pair of statistics for one replicate evaluation for one association model (PCA or LMM with some number of PCs r), one trait model (FES vs RC), and one dataset (color coded by dataset). Note y-axis (λ) is on a log scale,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

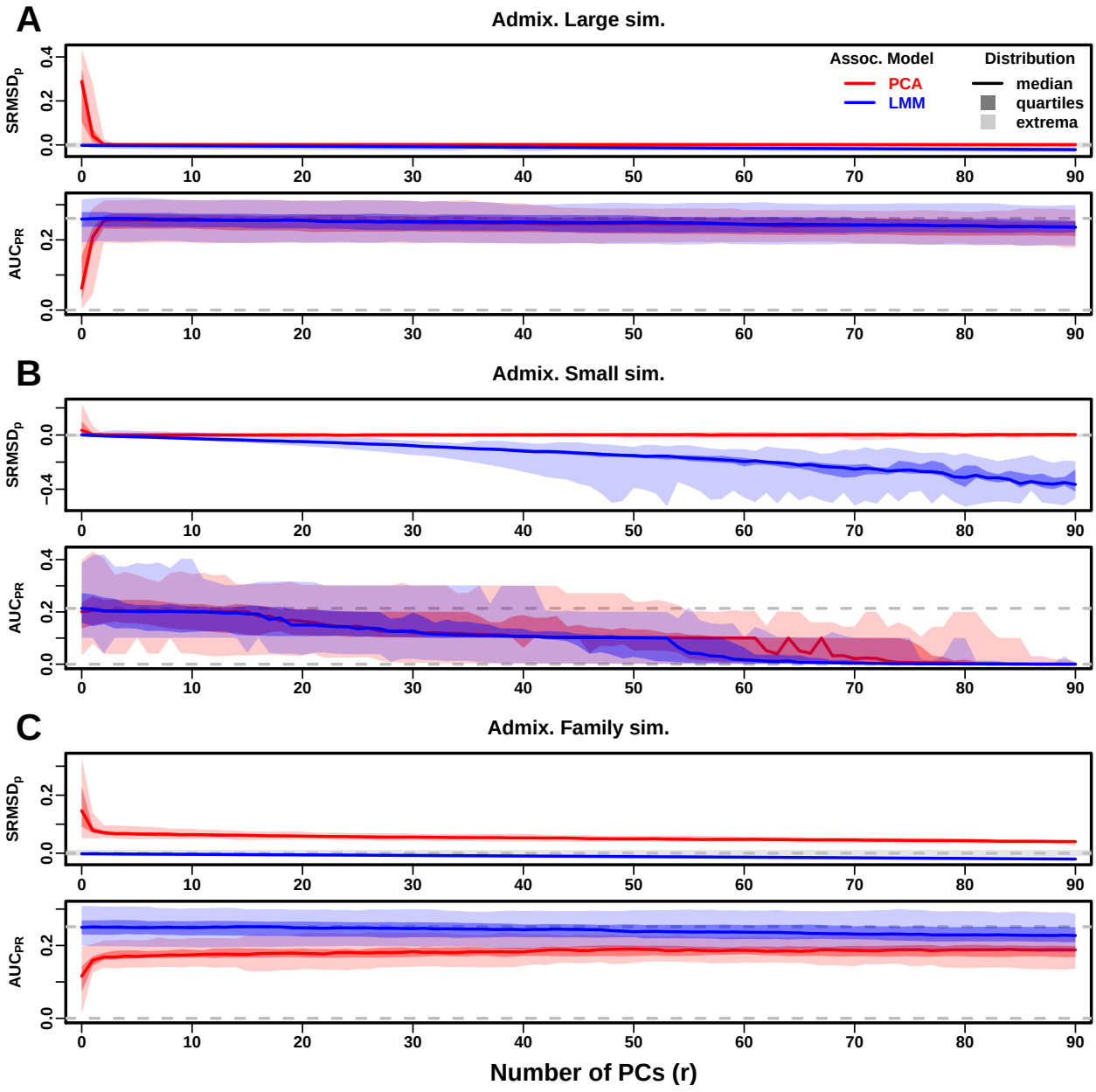


Figure S2: **Evaluations in admixture simulations.** Traits simulated from *random coefficients* model, otherwise the same as Fig. 3.

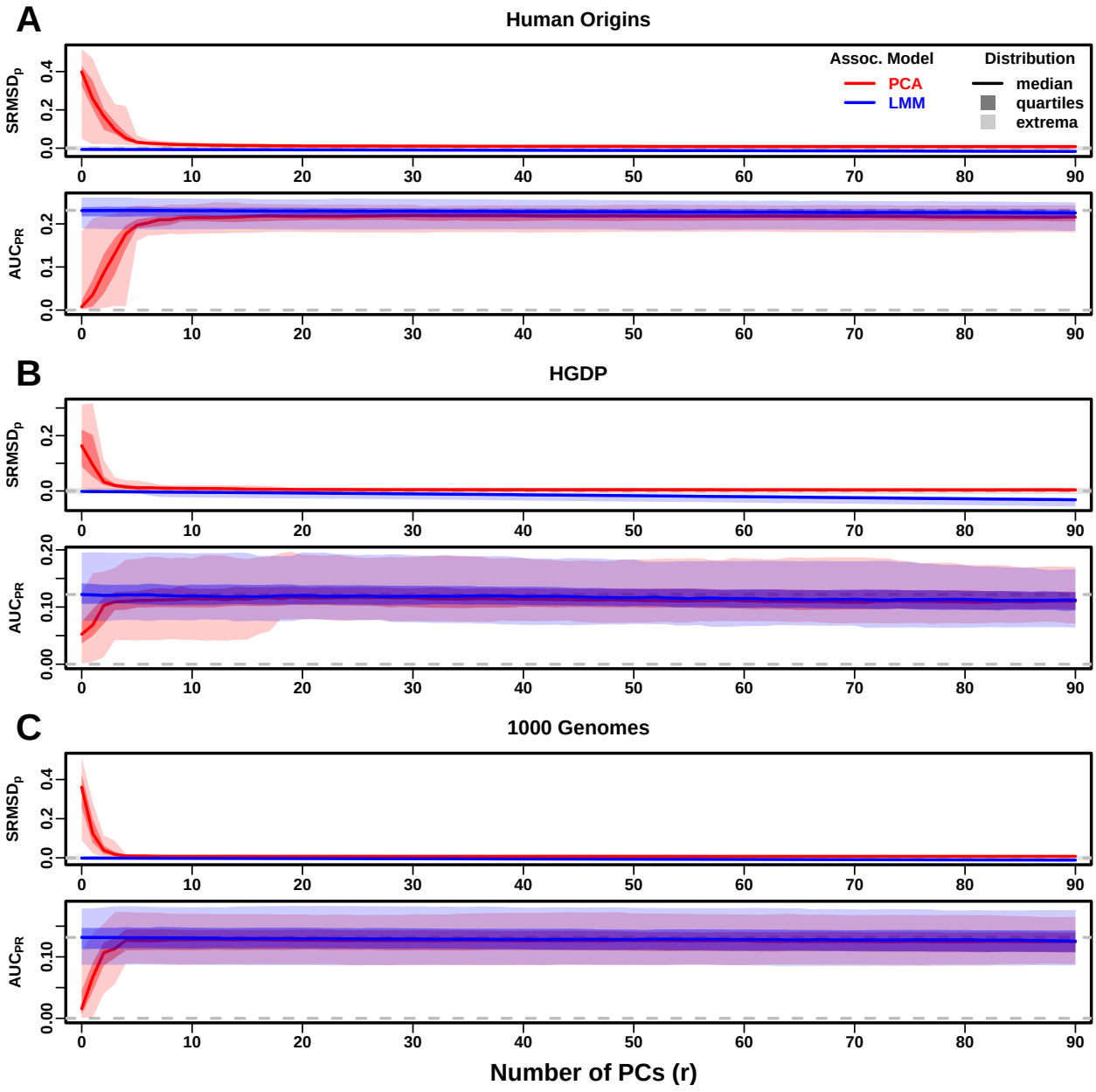


Figure S3: **Evaluations in real human genotype datasets.** Traits simulated from *random coefficients* model, otherwise the same as Fig. 4.

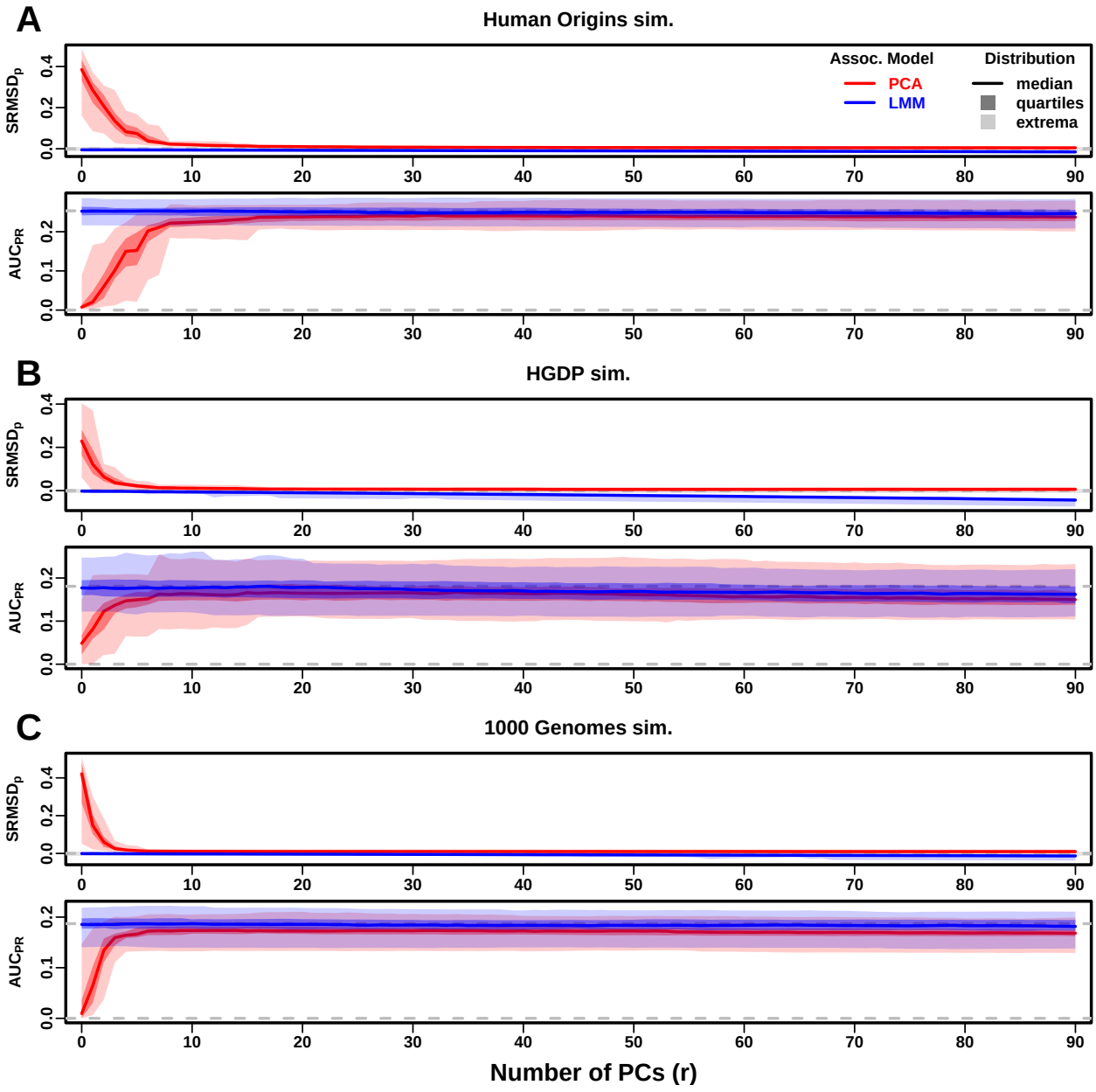


Figure S4: **Evaluations in tree simulations fit to human data.** Traits simulated from *random coefficients* model, otherwise the same as Fig. 5.

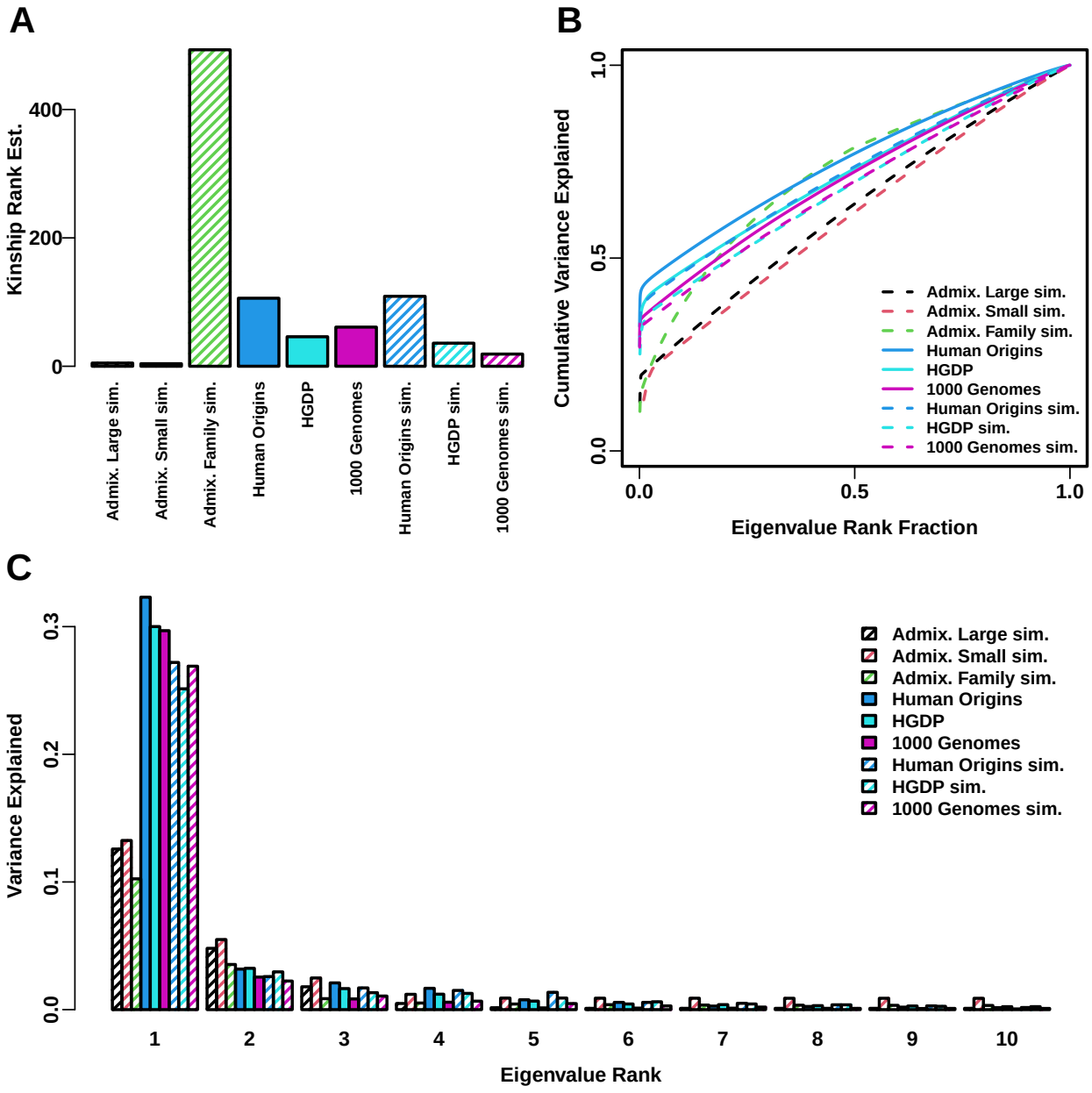


Figure S5: **Estimated dimensionality of datasets. A.** Kinship matrix ranks estimated with the Tracy-Widom test with $p < 0.01$. **B.** Cumulative variance explained versus eigenvalue rank fraction (i/n where i is rank, n is number of eigenvalues). **C.** Variance explained by first 10 eigenvalues.

S2 Supplementary tables

Table S1: **Dataset sizes after 4th degree relative filter.**

Dataset	Loci (m)	Ind. (n)	Ind. removed (%)
Human Origins	189,722	2636	9.8
HGDP	905,838	842	9.4
1000 Genomes	1,097,415	2390	4.6