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, United States; [2]Duke Center for Statistical Genetics and Genomics, Duke University, Durham, United States

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

Editor's evaluation

This is an important paper that presents compelling arguments (based on simulation and comprehensively reviewed background theory) that Linear Mixed Models generally should perform better at correcting for genetic and environmental confounding in GWAS than more commonly used Principal Components methods.

***For correspondence:** alejandro.ochoa@duke.edu

Present address: †BenHealth Consulting, Shanghai, Shanghai, China

Competing interest: See page 27

Funding: See page 27

Preprinted: 27 March 2022
Received: 04 April 2022
Accepted: 04 May 2023
Published: 04 May 2023

Reviewing Editor: Magnus Nordborg, Gregor Mendel Institute, Austria

Introduction

The goal of a genetic association study is to identify loci whose genotype variation is significantly correlated to given trait. Naive association tests assume that genotypes are drawn independently from a common allele frequency. This assumption does not hold for structured populations, which includes multiethnic cohorts and admixed individuals (ancient relatedness), and for family data (recent relatedness; ***Astle and Balding, 2009***). Association studies of admixed and multiethnic cohorts, the focus of this work, are becoming more common, are believed to be more powerful, and are necessary to bring more equity to genetic medicine (***Rosenberg et al., 2010***; ***Hoffman and Dubé, 2013***; ***Coram et al., 2013***; ***Medina-Gomez et al., 2015***; ***Conomos et al., 2016a***; ***Hodonsky et al., 2017***; ***Martin et al., 2017a***; ***Martin et al., 2017b***; ***Hindorff et al., 2018***; ***Hoffmann et al., 2018***; ***Mogil et al., 2018***; ***Roselli et al., 2018***; ***Wojcik et al., 2019***; ***Peterson et al., 2019***; ***Zhong et al., 2019***; ***Hu

et al., 2020; ***Simonin-Wilmer et al., 2021***; ***Kamariza et al., 2021***; ***Lin et al., 2021***; ***Mahajan et al., 2022***; ***Hou et al., 2023a***). When insufficient approaches are applied to data with relatedness, their association statistics are miscalibrated, resulting in excess false positives 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 under relatedness, of which PCA and LMM are the most popular.

Genetic association with PCA consists of including the top eigenvectors of the population kinship matrix as covariates in a generalized linear model (***Zhang et al., 2003***; ***Price et al., 2006***; ***Bouaziz et al., 2011***). These top eigenvectors are a new set of coordinates for individuals that are commonly referred to as PCs in genetics (***Patterson et al., 2006***), the convention adopted here, but in other fields PCs instead denote what in genetics would be the projections of loci onto eigenvectors, which are new independent coordinates for loci (***Jolliffe, 2002***). The direct ancestor of PCA association is structured association, in which inferred ancestry (genetic cluster membership, often corresponding with labels such as “European”, “African”, “Asian”, etc.) or admixture proportions of these ancestries are used as regression covariates (***Pritchard et al., 2000***). These models are deeply connected because PCs map to ancestry empirically (***Alexander et al., 2009***; ***Zhou et al., 2016***) and theoretically (***McVean, 2009***; ***Zheng and Weir, 2016***; ***Cabreros and Storey, 2019***; ***Chiu et al., 2022***), and they work as well as global 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***). Another approach closely related to PCA is nonmetric multidimensional scaling (***Zhu and Yu, 2009***). PCs are also proposed for modeling environment effects that are correlated to ancestry, for example, through geography (***Novembre et al., 2008***; ***Zhang and Pan, 2015***; ***Lin et al., 2021***). The strength of PCA is its simplicity, which as covariates can be readily included in more complex models, such as haplotype association (***Xu and Guan, 2014***) and polygenic models (***Qian et al., 2020***). However, PCA assumes that the underlying relatedness space is low dimensional (or low rank), so it can be well modeled with a small number of PCs, which may limit its applicability. PCA is known to be inadequate for family data (***Patterson et al., 2006***; ***Zhu and Yu, 2009***; ***Thornton and McPeek, 2010***; ***Price et al., 2010***), which is called ‘cryptic relatedness’ when it is unknown to the researchers, but no other troublesome cases have been confidently identified. Recent work has focused on developing more scalable versions of the PCA algorithm (***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 under relatedness is the LMM, which includes a random effect parameterized by the kinship matrix. Unlike PCA, LMM does not assume that relatedness is low-dimensional, and explicitly models families via the kinship matrix. Early LMMs used kinship matrices estimated from known pedigrees or using methods that captured recent relatedness only, and modeled population structure (ancestry) as fixed effects (***Yu et al., 2006***; ***Zhao et al., 2007***; ***Zhu and Yu, 2009***). Modern LMMs estimate kinship from genotypes using a non-parametric estimator, often referred to as a genetic relationship matrix, that captures the combined covariance due to family relatedness and ancestry (***Kang et al., 2008***; ***Astle and Balding, 2009***; ***Ochoa and Storey, 2021***). Like PCA, LMM has also been proposed for modeling environment correlated to genetics (***Vilhjálmsson and Nordborg, 2013***; ***Wang et al., 2022***). The classic LMM assumes a quantitative (continuous) complex trait, the focus of our work. Although case-control (binary) traits and their underlying ascertainment are theoretically a challenge (***Yang et al., 2014***), LMMs have been applied successfully to balanced case-control studies (***Astle and Balding, 2009***; ***Kang et al., 2010***) and simulations (***Price et al., 2010***; ***Wu et al., 2011***; ***Sul and Eskin, 2013***), and have been adapted for unbalanced case-control studies (***Zhou et al., 2018***). However, LMMs tend to be considerably slower than PCA and other models, so much effort has focused on improving their runtime and scalability (***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***).

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

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

	Sim. Genotypes			General				
Publication	**Type***	K†	F_{ST}‡	**Real** §	**Trait** ¶	**Power**	PCs(r)	**Best**
Zhao et al., 2007				✓	Q	✓	8	LMM
Zhu and Yu, 2009	I, A, F	3, 8	≤0.15	✓	Q	✓	1–22	LMM
Astle and Balding, 2009	I	3	0.10		CC	✓	10	Tie
Kang et al., 2010				✓	Both		2–100	LMM
Price et al., 2010	I, F	2	0.01		CC		1	Mixed
Wu et al., 2011	I, A	2–4	0.01		CC	✓	10	Mixed
Liu et al., 2011	S, A	2–3	R		Q	✓	10	Tie
Sul and Eskin, 2013	I	2	0.01		CC		1	Tie
Tucker et al., 2014	I	2	0.05	✓	Both	✓	5	Tie
Yang et al., 2014				✓	CC	✓	5	Tie
Song et al., 2015	S, A	2–3	R		Q		3	LMM
Loh et al., 2015				✓	Q	✓	10	LMM
Zhang and Pan, 2015				✓	Q	✓	20–100	LMM
Liu et al., 2016				✓	Q	✓	3–6	LMM
Sul et al., 2018				✓	Q		100	LMM
Loh et al., 2018				✓	Both	✓	20	LMM
Mbatchou et al., 2021				✓	Both		1	LMM
This work	A, T, F	10–243	≤0.25	✓	Q	✓	0–90	LMM

*Genotype simulation types. I: Independent subpopulations; S: subpopulations (with parameters drawn from real data); A: Admixture; T: Subpopulation Tree; F: Family.

†Model dimension (number of subpopulations or ancestries).

‡R: simulated parameters based on real data, F_{ST} not reported.

§Evaluations using unmodified real genotypes.

¶Q: quantitative; CC: case-control.

apparent redundancy of an LMM with PCs beneficial (***Price et al., 2010***; ***Tucker et al., 2014***; ***Zhang and Pan, 2015***), while others did not (***Liu et al., 2011***; ***Janss et al., 2012***), and the approach continues to be used (***Zeng et al., 2018***; ***Mbatchou et al., 2021***), although not always (***Matoba et al., 2020***). Recall that early LMMs used kinship to model family relatedness only, so population structure had to be modeled separately in those models, in practice as admixture fractions instead of PCs (***Yu et al., 2006***; ***Zhao et al., 2007***; ***Zhu and Yu, 2009***). The LMM with PCs (vs no PCs) is also believed to help better model loci that have experienced selection (***Price et al., 2010***; ***Vilhjálmsson and Nordborg, 2013***) and environment effects correlated with genetics (***Zhang and Pan, 2015***).

LMM and PCA are closely related models (***Astle and Balding, 2009***; ***Janss et al., 2012***; ***Hoffman and Dubé, 2013***; ***Zhang and Pan, 2015***), so similar performance is expected particularly under low-dimensional relatedness. Direct comparisons have yielded mixed results, with several studies finding superior performance for LMM, notably from papers promoting advances in LMMs, while many others report comparable performance (***Table 1***). No papers find that PCA outperforms LMM decisively, although PCA occasionally performs better in isolated and artificial cases or individual measures, often with unknown significance. Previous studies generally used either only simulated or only real genotypes, with only two studies using both. The simulated genotype studies, which tended to have low model dimensions and F_{ST}, were more likely to report ties or mixed results (6/8), whereas real genotypes tended to clearly favor LMMs (9/11). Similarly, 10/12 papers with quantitative traits favor LMMs, whereas 6/9 papers with case-control traits gave ties or mixed results—the only factor we do not explore in this work. Additionally, although all previous evaluations measured type I error (or proxies such as genomic inflation factors ***Devlin and Roeder, 1999*** or QQ plots), a large fraction (6/17) did not measure power (or proxies such as ROC curves), and only four used more than one number of PCs for PCA. Lastly, no consensus has emerged as to why LMM might outperform PCA or vice versa (***Price et al., 2010***; ***Sul and Eskin, 2013***; ***Price et al., 2013***; ***Hoffman and Dubé, 2013***), or which features of the real datasets are critical for the LMM advantage other than family relatedness, resulting in unclear guidance for using PCA. Hence, our work includes real and simulated genotypes with higher model dimensions and F_{ST} matching that of multiethnic human cohorts (***Ochoa and Storey, 2021***; ***Ochoa and Storey, 2019***), we vary the number of PCs, and measure robust proxies for type I error control and calibrated power.

In this work, we evaluate the PCA and LMM association models under various numbers of PCs, which are included in LMMs too. We use genotype simulations (admixture, family, and subpopulation tree models) and three real datasets: the 1000 Genomes Project (***Abecasis et al., 2010***; ***Abecasis 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***). We simulate quantitative traits from two models: fixed effect sizes (FES) construct coefficients inverse to allele frequency, which matches real data (***Park et al., 2011***; ***Zeng et al., 2018***; ***O'Connor et al., 2019***) and corresponds to high pleiotropy and strong balancing selection (***Simons et al., 2018***) and strong negative selection (***Zeng et al., 2018***; ***O'Connor et al., 2019***), which are appropriate assumptions for diseases; and random coefficients (RC), which are drawn independent of allele frequency, and corresponds to neutral traits (***Zeng et al., 2018***; ***Simons et al., 2018***). LMM without PCs consistently performs best in simulations without environment, and greatly outperforms PCA in the family simulation and in all real datasets. The tree simulations, which model subpopulations with the tree but exclude family structure, do not recapitulate the real data results, suggesting that family relatedness in real data is the reason for poor PCA performance. Lastly, removing up to 4th degree relatives in the real datasets recapitulates poor PCA performance, showing that the more numerous distant relatives explain the result, and suggesting that PCA is generally not an appropriate model for real data. We find that both LMM and PCA are able to model environment effects correlated with genetics, and LMM with PCs gains a small advantage in this setting only, but direct modeling of environment performs much better. All together, we find that LMMs without PCs are generally a preferable association model, and present novel simulation and evaluation approaches to measure the performance of these and other genetic association approaches.

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

Dataset	Type	Loci(m)	Ind. (n)	Subpops.[*] (K)	Causal loci[†] (m_1)	F_{ST}[‡]
Admix. Large sim.	Admix.	100 000	1000	10	100	0.1
Admix. Small sim.	Admix.	100 000	100	10	10	0.1
Admix. Family sim.	Admix.+Pedig.	100 000	1000	10	100	0.1
Human Origins	Real	190 394	2922	11–243	292	0.28
HGDP	Real	771 322	929	7–54	93	0.28
1000 Genomes	Real	1 111 266	2504	5–26	250	0.22
Human Origins sim.	Tree	190 394	2922	243	292	0.23
HGDP sim.	Tree	771 322	929	54	93	0.25
1000 Genomes sim.	Tree	1 111 266	2504	26	250	0.21

Ind.(n)

[*] For admixed family, ignores additional model dimension of 20 generation pedigree structure. For real datasets, lower range is continental subpopulations, upper range is number of fine-grained subpopulations.

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

[‡] Model parameter for simulations, estimated value on real datasets.

Results

Overview of evaluations

We use three real genotype datasets and simulated genotypes from six population structure scenarios to cover various features of interest (***Table 2***). We introduce them in sets of three, as they appear in the rest of our results. Population kinship matrices, which combine population and family relatedness, are estimated without bias using popkin (***Ochoa and Storey, 2021***; ***Figure 1***). The first set of three simulated genotypes are based on an admixture model with 10 ancestries (***Figure 1A***; ***Ochoa and Storey, 2021***; ***Gopalan et al., 2016***; ***Cabreros and Storey, 2019***). The 'large' version (1000 individuals) illustrates asymptotic performance, while the 'small' simulation (100 individuals) illustrates model overfitting. The 'family' simulation has admixed founders and draws a 20-generation random pedigree with assortative mating, resulting in a complex joint family and ancestry structure in the last generation (***Figure 1B***). The second set of three are the real human datasets representing global human diversity: Human Origins (***Figure 1D***), HGDP (***Figure 1G***), and 1000 Genomes (***Figure 1J***), which are enriched for small minor allele frequencies even after MAF <1% filter (***Figure 1C***). Last are subpopulation tree simulations (***Figure 1F, I, L***) fit to the kinship (***Figure 1E, H and K***) and MAF (***Figure 1C***) of each real human dataset, which by design do not have family structure.

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

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

Both PCA and LMM are evaluated in each replicate dataset including a number of PCs r between 0 and 90 as fixed covariates. In terms of p-value calibration, for PCA the best number of PCs r (minimizing mean $|\mathrm{SRMSD}_p|$ over replicates) is typically large across all datasets (***Table 3***), although much smaller r values often performed as well (shown in following sections). Most cases have a mean $|\mathrm{SRMSD}_p| < 0.01$, whose p-values are effectively calibrated. However, PCA is often miscalibrated on the family simulation and real datasets (***Table 3***). In contrast, for LMM, $r = 0$ (no PCs) is always best, and is always calibrated. Comparing LMM with $r = 0$ to PCA with its best r, LMM always has significantly smaller $|\mathrm{SRMSD}_p|$ than PCA or is statistically tied. For $\mathrm{AUC}_{\mathrm{PR}}$ and PCA, the best r is always smaller than the best r for $|\mathrm{SRMSD}_p|$, so there is often a tradeoff between calibrated p-values versus classification performance. For LMM, there is no tradeoff, as $r = 0$ often has the best mean $\mathrm{AUC}_{\mathrm{PR}}$, and otherwise is not significantly different from the best r. Lastly, LMM with $r = 0$ always has significantly greater or statistically tied $\mathrm{AUC}_{\mathrm{PR}}$ than PCA with its best r.

Evaluations in admixture simulations

Now we look more closely at results per dataset. The complete SRMSD_p and $\mathrm{AUC}_{\mathrm{PR}}$ distributions for the admixture simulations and FES traits are in ***Figure 3***. RC traits gave qualitatively similar results (***Figure 3—figure supplement 1***).

In the large admixture simulation, the SRMSD_p of PCA is largest when $r = 0$ (no PCs) and decreases rapidly to near zero at $r = 3$, where it stays for up to $r = 90$ (***Figure 3A***). Thus, PCA has calibrated p-values for $r \geq 3$, smaller than the theoretical optimum for this simulation of $r = K - 1 = 9$. In contrast,

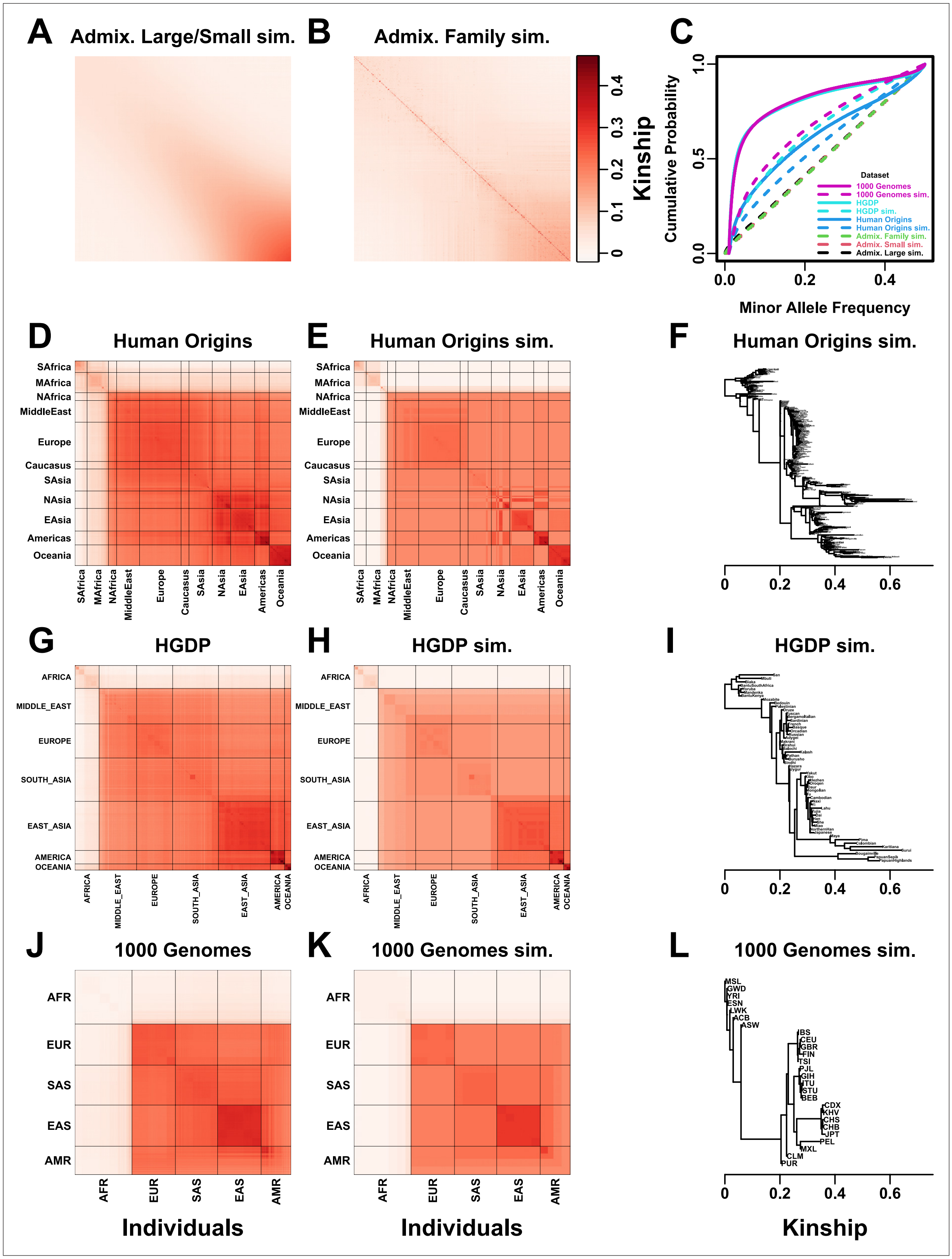

Figure 1. Population structures of simulated and real human genotype datasets. First two columns are population kinship matrices as heatmaps: individuals along x- and y-axis, kinship as color. Diagonal shows inbreeding values. (**A**) Admixture scenario for both Large and Small simulations. (**B**) Last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

Figure 1 continued on next page

Figure 1 continued

diversity of global populations. (**G**) Human Genome Diversity Panel (HGDP) is a WGS dataset from global native populations. (**J**) 1000 Genomes Project is a WGS dataset of global cosmopolitan populations. (**F, I, L**) Trees between subpopulations fit to real data. (**E, H, K**). Simulations from trees fit to the real data recapitulate subpopulation structure.

the SRMSD_p for LMM starts near zero for $r = 0$, but becomes negative as r increases (p-values are 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 $r > 3$, while the $\mathrm{AUC}_{\mathrm{PR}}$ distribution for LMM starts near its maximum at $r = 0$ and decreases with 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 performs nearly as well as LMM in this simulation.

The observed robustness to large r led us to consider smaller sample sizes. A model with large numbers of parameters r should overfit more as r approaches the sample size n. Rather than increase r beyond 90, 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

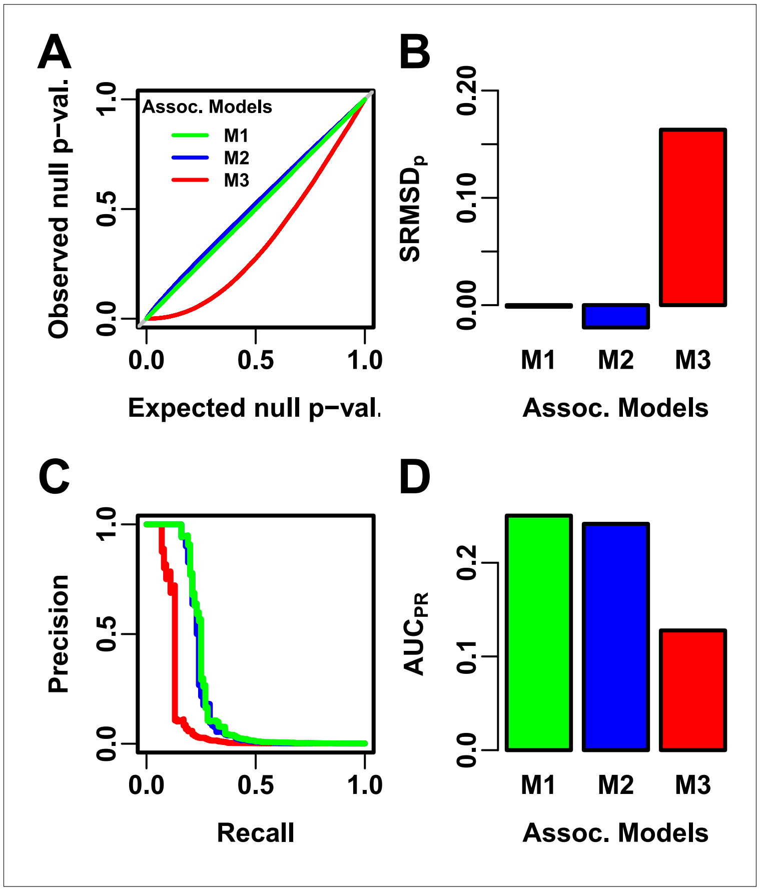

Figure 2. Illustration of evaluation measures. Three archetypal models illustrate our complementary measures: M1 is ideal, M2 overfits slightly, M3 is naive. (**A**) QQ plot of p-values of “null” (non-causal) loci. M1 has desired uniform p-values, M2/M3 are miscalibrated. (**B**)SRMSD_p (p-value Signed Root Mean Square Deviation) measures signed distance between observed and expected null p-values (closer to zero is better). (**C**) Precision and Recall (PR) measure causal locus classification performance (higher is better). (**D**) $\mathrm{AUC}_{\mathrm{PR}}$ (Area Under the PR Curve) reflects power (higher is better).

The online version of this article includes the following figure supplement(s) for figure 2:

Figure supplement 1. Comparison between SRMSD_p and inflation factor.

Figure supplement 2. Comparison between SRMSD_p and type I error rate.

Figure supplement 3. Comparison between $\mathrm{AUC}_{\mathrm{PR}}$ and calibrated power.

Table 3. Overview of PCA and LMM evaluations for high heritability simulations.

			LMM $r=0$ vs best r			PCA vs LMM $r=0$			
Dataset	**Metric**	**Trait**[*]	**Cal.**[†]	**Best r**[‡]	**P-value**[§]	**Best r**[‡]	**Cal.**[†]	**P-value**[§]	**Best model**[¶]
Admix. Large sim.	$\lvert\mathrm{SRMSD}_p\rvert$	FES	True	0	1	12	True	0.036	Tie
Admix. Small sim.	$\lvert\mathrm{SRMSD}_p\rvert$	FES	True	0	1	4	True	0.055	Tie
Admix. Family sim.	$\lvert\mathrm{SRMSD}_p\rvert$	FES	True	0	1	90	False	3.9e-10*	LMM
Human Origins	$\lvert\mathrm{SRMSD}_p\rvert$	FES	True	0	1	89	False	3.9e-10*	LMM
HGDP	$\lvert\mathrm{SRMSD}_p\rvert$	FES	True	0	1	87	True	4.4e-10*	LMM
1000 Genomes	$\lvert\mathrm{SRMSD}_p\rvert$	FES	True	0	1	90	False	3.9e-10*	LMM
Human Origins sim.	$\lvert\mathrm{SRMSD}_p\rvert$	FES	True	0	1	88	True	0.017	Tie
HGDP sim.	$\lvert\mathrm{SRMSD}_p\rvert$	FES	True	0	1	47	True	0.046	Tie
1000 Genomes sim.	$\lvert\mathrm{SRMSD}_p\rvert$	FES	True	0	1	78	True	9.6e-10*	LMM
Admix. Large sim.	$\lvert\mathrm{SRMSD}_p\rvert$	RC	True	0	1	26	True	0.11	Tie
Admix. Small sim.	$\lvert\mathrm{SRMSD}_p\rvert$	RC	True	0	1	4	True	0.00097	Tie
Admix. Family sim.	$\lvert\mathrm{SRMSD}_p\rvert$	RC	True	0	1	90	False	3.9e-10*	LMM
Human Origins	$\lvert\mathrm{SRMSD}_p\rvert$	RC	True	0	1	90	True	0.00065	Tie
HGDP	$\lvert\mathrm{SRMSD}_p\rvert$	RC	True	0	1	37	True	1.5e-05*	LMM
1000 Genomes	$\lvert\mathrm{SRMSD}_p\rvert$	RC	True	0	1	76	True	3.9e-10*	LMM
Human Origins sim.	$\lvert\mathrm{SRMSD}_p\rvert$	RC	True	0	1	85	True	0.14	Tie
HGDP sim.	$\lvert\mathrm{SRMSD}_p\rvert$	RC	True	0	1	44	True	8.8e-07*	LMM
1000 Genomes sim.	$\lvert\mathrm{SRMSD}_p\rvert$	RC	True	0	1	90	True	3.9e-10*	LMM
Admix. Large sim.	$\mathrm{AUC}_{\mathrm{PR}}$	FES		0	1	3		5.9e-06*	LMM
Admix. Small sim.	$\mathrm{AUC}_{\mathrm{PR}}$	FES		0	1	2		0.025	Tie
Admix. Family sim.	$\mathrm{AUC}_{\mathrm{PR}}$	FES		1	0.35	22		3.9e-10*	LMM
Human Origins	$\mathrm{AUC}_{\mathrm{PR}}$	FES		0	1	34		3.9e-10*	LMM
HGDP	$\mathrm{AUC}_{\mathrm{PR}}$	FES		1	0.33	16		4.4e-10*	LMM
1000 Genomes	$\mathrm{AUC}_{\mathrm{PR}}$	FES		1	0.11	8		3.9e-10*	LMM
Human Origins sim.	$\mathrm{AUC}_{\mathrm{PR}}$	FES		0	1	36		3.9e-10*	LMM
HGDP sim.	$\mathrm{AUC}_{\mathrm{PR}}$	FES		0	1	17		1.7e-05*	LMM
1000 Genomes sim.	$\mathrm{AUC}_{\mathrm{PR}}$	FES		0	1	10		5e-10*	LMM
Admix. Large sim.	$\mathrm{AUC}_{\mathrm{PR}}$	RC		0	1	3		1.4e-05*	LMM
Admix. Small sim.	$\mathrm{AUC}_{\mathrm{PR}}$	RC		0	1	1		0.095	Tie
Admix. Family sim.	$\mathrm{AUC}_{\mathrm{PR}}$	RC		0	1	34		3.9e-10*	LMM
Human Origins	$\mathrm{AUC}_{\mathrm{PR}}$	RC		3	0.4	36		9.6e-10*	LMM
HGDP	$\mathrm{AUC}_{\mathrm{PR}}$	RC		4	0.21	16		0.013	Tie
1000 Genomes	$\mathrm{AUC}_{\mathrm{PR}}$	RC		5	0.004	9		0.00043	Tie
Human Origins sim.	$\mathrm{AUC}_{\mathrm{PR}}$	RC		0	1	37		4.1e-10*	LMM

Table 3 continued on next page

Table 3 continued

			LMM $r=0$ vs best r			PCA vs LMM $r=0$			
Dataset	Metric	Trait[*]	Cal.[†]	Best r[‡]	P-value[§]	Best r[‡]	Cal.[†]	P-value[§]	Best model[¶]
HGDP sim.	$\mathrm{AUC}_{\mathrm{PR}}$	RC		3	0.087	17		0.0014	Tie
1000 Genomes sim.	$\mathrm{AUC}_{\mathrm{PR}}$	RC		3	0.37	10		8.5e-10*	LMM

[*] FES: Fixed Effect Sizes, RC: Random Coefficients.

[†] Calibrated: whether mean $|\mathrm{SRMSD}_p| < 0.01$ over 50 replicates.

[‡] Value of r (number of PCs) with minimum mean $|\mathrm{SRMSD}_p|$ or maximum mean $\mathrm{AUC}_{\mathrm{PR}}$.

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

[¶] Tie if no significant difference using Bonferroni threshold.

due to reducing n, we also reduce the number of causal loci (see Trait Simulation), which increases per-locus effect sizes. We found a large decrease in performance for both models as r increases, and best performance for $r = 1$ for PCA and $r = 0$ for LMM (***Figure 3B***). Remarkably, LMM attains much larger negative SRMSD_p values than in our other evaluations. LMM with $r = 0$ is significantly better than PCA ($r = 1$ to 4) in both measures (***Table 3***), but qualitatively the difference is negligible.

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

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 F_{ST}, numerous correlated subpopulations, and potential cryptic family relatedness.

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

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

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

Our results are qualitatively different for RC traits, which had smaller $\mathrm{AUC}_{\mathrm{PR}}$ gaps between LMM and PCA (***Figure 4—figure supplement 1***). Maximum $\mathrm{AUC}_{\mathrm{PR}}$ were smaller in RC compared to FES in Human Origins and 1000 Genomes, suggesting lower power for RC traits across association models. Nevertheless, LMM with $r = 0$ was significantly better than PCA for all measures in the real datasets and RC traits (***Table 3***).

Evaluations in subpopulation tree simulations fit to human data

To better understand which features of the real datasets lead to the large differences in performance between LMM and PCA, we carried out subpopulation tree simulations. Human subpopulations are related roughly by trees, which induce the strongest correlations, so we fit trees to each real dataset

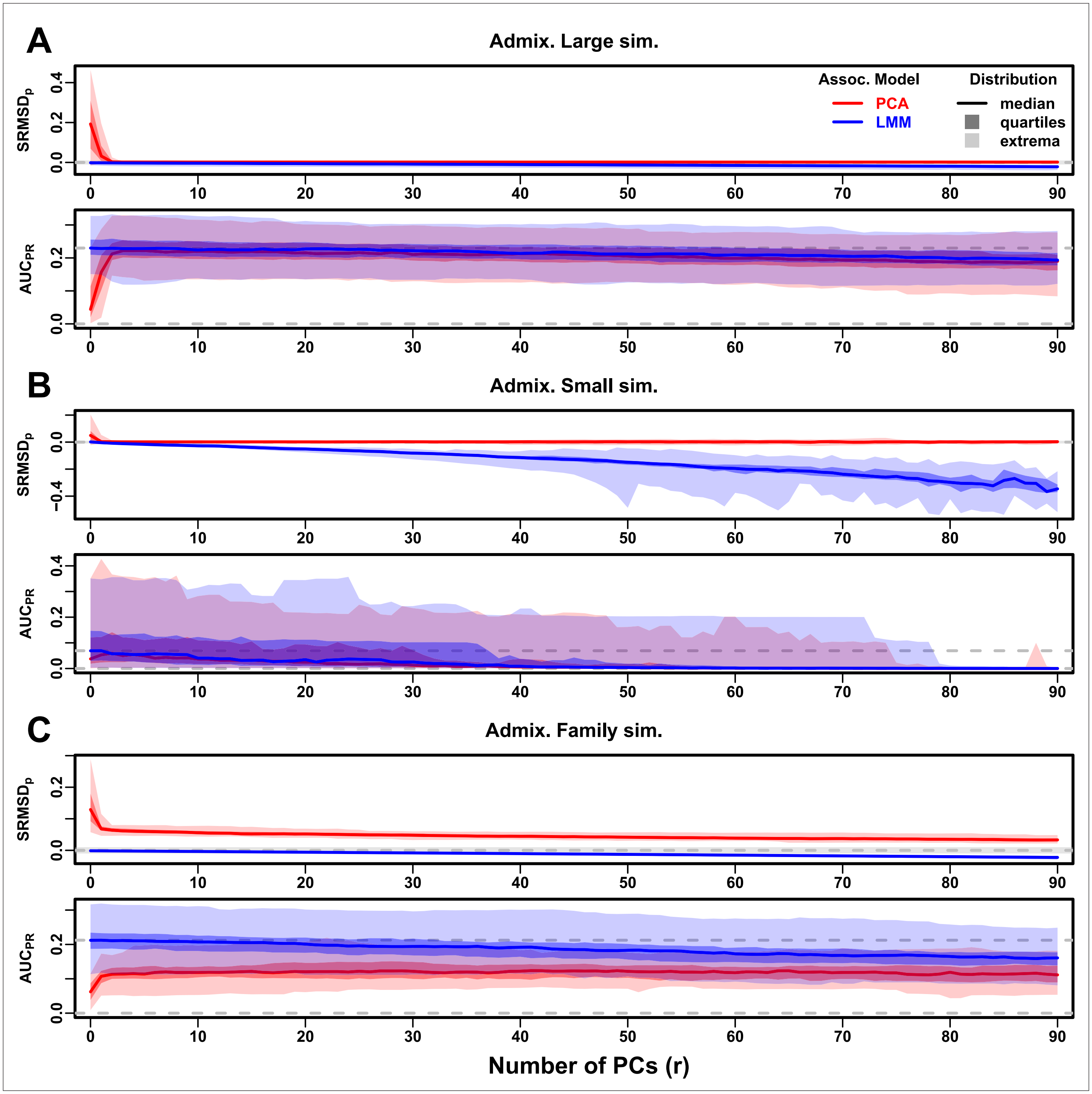

Figure 3. Evaluations in admixture simulations with FES traits, high heritability. PCA and LMM models have varying number of PCs ($r \in \{0, ..., 90\}$ on x-axis), with the distributions (y-axis) of SRMSD_p (top subpanel) and $\mathrm{AUC}_{\mathrm{PR}}$ (bottom subpanel) for 50 replicates. Best performance is zero SRMSD_p and large $\mathrm{AUC}_{\mathrm{PR}}$. Zero and maximum median $\mathrm{AUC}_{\mathrm{PR}}$ values are marked with horizontal gray dashed lines, and $|\mathrm{SRMSD}_p| < 0.01$ is marked with a light gray area. LMM performs best with $r = 0$, PCA with various r. (**A**) Large simulation ($n = 1,000$ individuals). (**B**) Small simulation ($n = 100$) shows overfitting for large r. (**C**) Family simulation ($n = 1,000$) has admixed founders and large numbers of close relatives from a realistic random 20-generation pedigree. PCA performs poorly compared to LMM: $\mathrm{SRMSD}_p > 0$ for all r and large $\mathrm{AUC}_{\mathrm{PR}}$ gap.

The online version of this article includes the following figure supplement(s) for figure 3:

Figure supplement 1. Evaluations in admixture simulations with RC traits, high heritability.

Figure supplement 2. Evaluations in admixture simulations with FES traits, low heritability.

Figure 3 continued on next page

Figure 3 continued

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

Figure supplement 4. Evaluations in admixture simulations with FES traits, environment.

Figure supplement 5. Evaluations in admixture simulations with RC traits, environment.

and tested if data simulated from these complex tree structures could recapitulate our previous results (***Figure 1***). These tree simulations also feature non-uniform ancestral allele frequency distributions, which recapitulated some of the skew for smaller minor allele frequencies of the real datasets (***Figure 1C***). The SRMSD_p and $\mathrm{AUC}_{\mathrm{PR}}$ distributions for these tree simulations (***Figure 5***) resembled our admixture simulation more than either the family simulation (***Figure 3***) or real data results (***Figure 4***). Both LMM with $r = 0$ and PCA (various r) achieve mean $|\mathrm{SRMSD}_p| < 0.01$ (***Table 3***). The $\mathrm{AUC}_{\mathrm{PR}}$ distributions of both LMM and PCA track closely as r is varied, although there is a small gap resulting in LMM ($r = 0$) besting PCA in all three simulations. The results are qualitatively similar for RC traits (***Figure 5—figure supplement 1***, ***Table 3***). Overall, these subpopulation tree simulations do not recapitulate the large LMM advantage over PCA observed on the real data.

Numerous distant relatives explain poor PCA performance in real data

In principle, PCA performance should be determined by the dimension of relatedness, or kinship matrix rank, since PCA is a low-dimensional model whereas LMM can model high-dimensional relatedness without overfitting. We used the Tracy-Widom test (***Patterson et al., 2006***) with $p < 0.01$ to estimate kinship matrix rank as the number of significant PCs (***Figure 6—figure supplement 1A***). The true rank of our simulations is slightly underestimated (***Table 2***), but we confirm that the family simulation has the greatest rank, and real datasets have greater estimates than their respective subpopulation tree simulations, which confirms our hypothesis to some extent. However, estimated ranks do not separate real datasets from tree simulations, as required to predict the observed PCA performance. Moreover, the HGDP and 1000 Genomes rank estimates are 45 and 61, respectively, yet PCA performed poorly for all $r \leq 90$ numbers of PCs (***Figure 4***). The top eigenvalue explained a proportion of variance proportional to F_{ST} (***Table 2***), but the rest of the top 10 eigenvalues show no clear differences between datasets, except the small simulation had larger variances explained per eigenvalue (expected since it has fewer eigenvalues; ***Figure 6—figure supplement 1***). Comparing cumulative variance explained versus rank fraction across all eigenvalues, all datasets increase from their starting point almost linearly until they reach 1, except the family simulation has much greater variance explained by mid-rank eigenvalues (***Figure 6—figure supplement 1***). We also calculated the number of PCs that are significantly associated with the trait, and observed similar results, namely that while the family simulation has more significant PCs than the non-family admixture simulations, the real datasets and their tree simulated counterparts have similar numbers of significant PCs (***Figure 6—figure supplement 2***). Overall, there is no separation between real datasets (where PCA performed poorly) and subpopulation tree simulations (where PCA performed relatively well) in terms of their eigenvalues or kinship matrix rank estimates.

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

To try to improve PCA performance, we followed the standard practice of removing 4th degree relatives, which reduced sample sizes between 5% and 10% (***Table 4***). Only $r = 0$ for LMM and $r = 20$

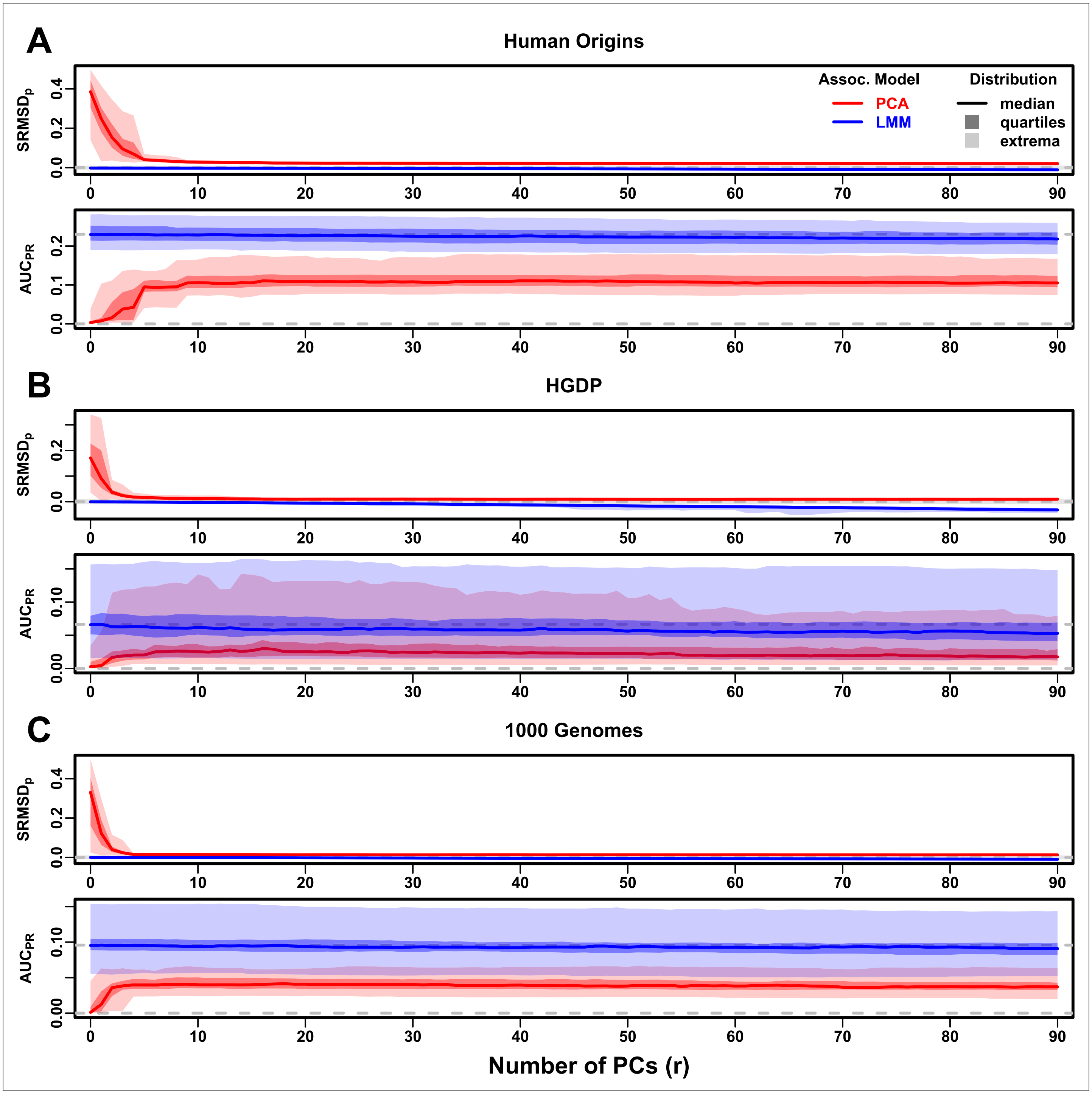

Figure 4. Evaluations in real human genotype datasets with FES traits, high heritability. Same setup as ***Figure 3***, see that for details. These datasets strongly favor LMM with no PCs over PCA, with distributions that most resemble the family simulation. (**A**) Human Origins. (**B**) Human Genome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

The online version of this article includes the following figure supplement(s) for figure 4:

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

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

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

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

Figure supplement 5. Evaluations in real human genotype datasets with RC traits, environment.

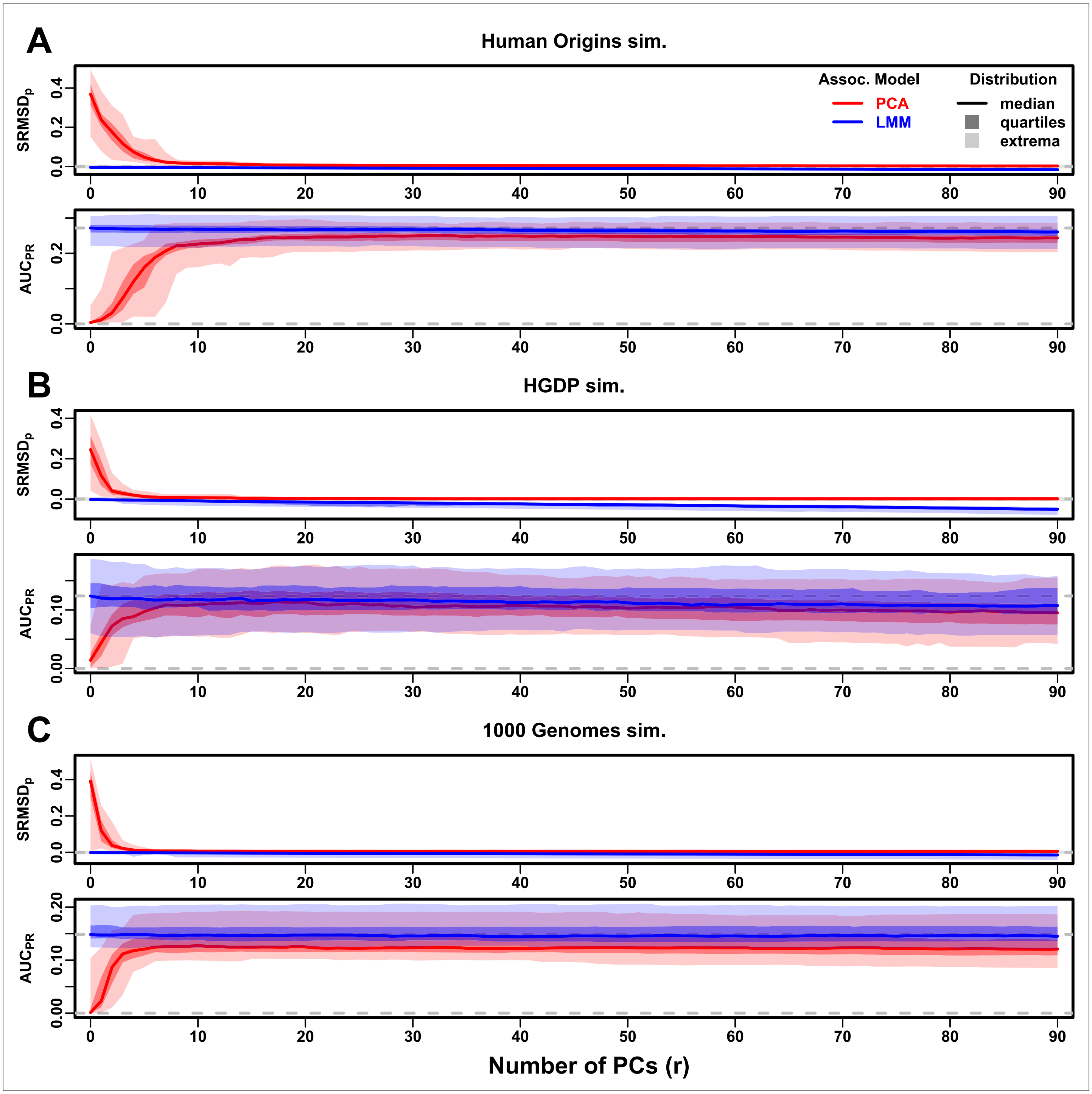

Figure 5. Evaluations in subpopulation tree simulations fit to human data with FES traits, high heritability. Same setup as ***Figure 3***, see that for details. These tree simulations, which exclude family structure by design, do not explain the large gaps in LMM-PCA performance observed in the real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

The online version of this article includes the following figure supplement(s) for figure 5:

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

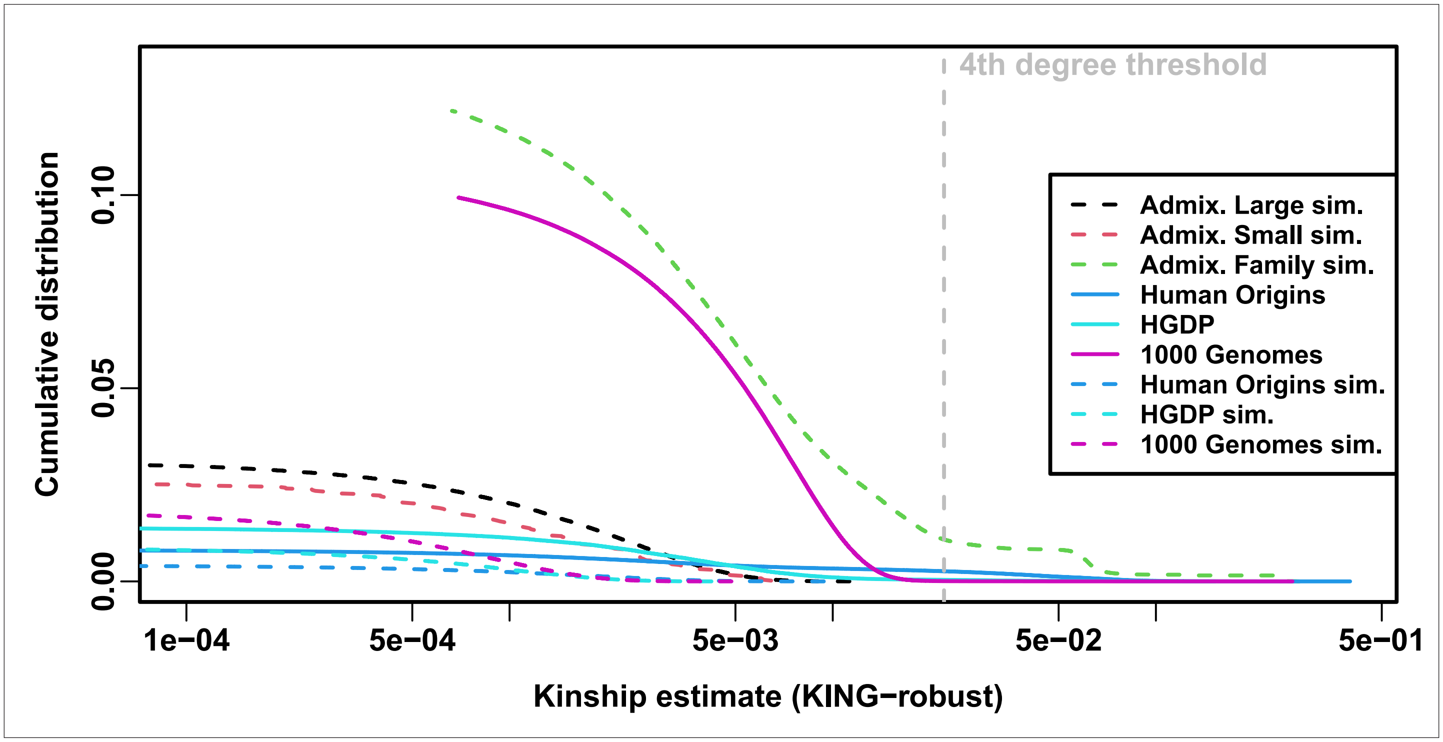

Figure 6. Local kinship distributions. Curves are complementary cumulative distribution of lower triangular kinship matrix (self kinship excluded) from KING-robust estimator. Note log x-axis; negative estimates are counted but not shown. Most values are below 4th degree relative threshold. Each real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

The online version of this article includes the following figure supplement(s) for figure 6:

Figure supplement 1. Estimated relatedness dimensions of datasets.

Figure supplement 2. Number of PCs significantly associated with traits.

for PCA were tested, as these performed well in our earlier evaluation, and only FES traits were tested because they previously displayed the large PCA-LMM performance gap. LMM significantly outperforms PCA in all these cases (Wilcoxon paired 1-tailed $p < 0.01$; ***Figure 7***). Notably, PCA still had miscalibrated p-values two of the three real datasets ($|\mathrm{SRMSD}_p| > 0.01$), the only marginally calibrated case being HGDP which is also the smallest of these datasets. Otherwise, $\mathrm{AUC}_{\mathrm{PR}}$ and SRMSD_p ranges were similar here as in our earlier evaluation. Therefore, the removal of 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 in the real datasets.

Low heritability and environment simulations

Our main evaluations were repeated with traits simulated under a lower heritability value of $h^2 = 0.3$. We reduced the number of causal loci in response to this change in heritability, to result in equal average effect size per locus compared to the previous high heritability evaluations (see Trait Simulation). Despite that, these low heritability evaluations measured lower $\mathrm{AUC}_{\mathrm{PR}}$ values than their high heritability counterparts (***Figure 3—figure supplement 2***, ***Figure 3—figure supplement 3***, ***Figure 4—figure supplement 2***, ***Figure 4—figure supplement 3***, ***Figure 7—figure supplement 1***).

Table 4. Dataset sizes after 4th degree relative filter.

Dataset	Loci (m)	Ind. (n)	Ind. removed (%)
Human Origins	189 722	2636	9.8
HGDP	758 009	847	8.8
1000 Genomes	1 097 415	2390	4.6

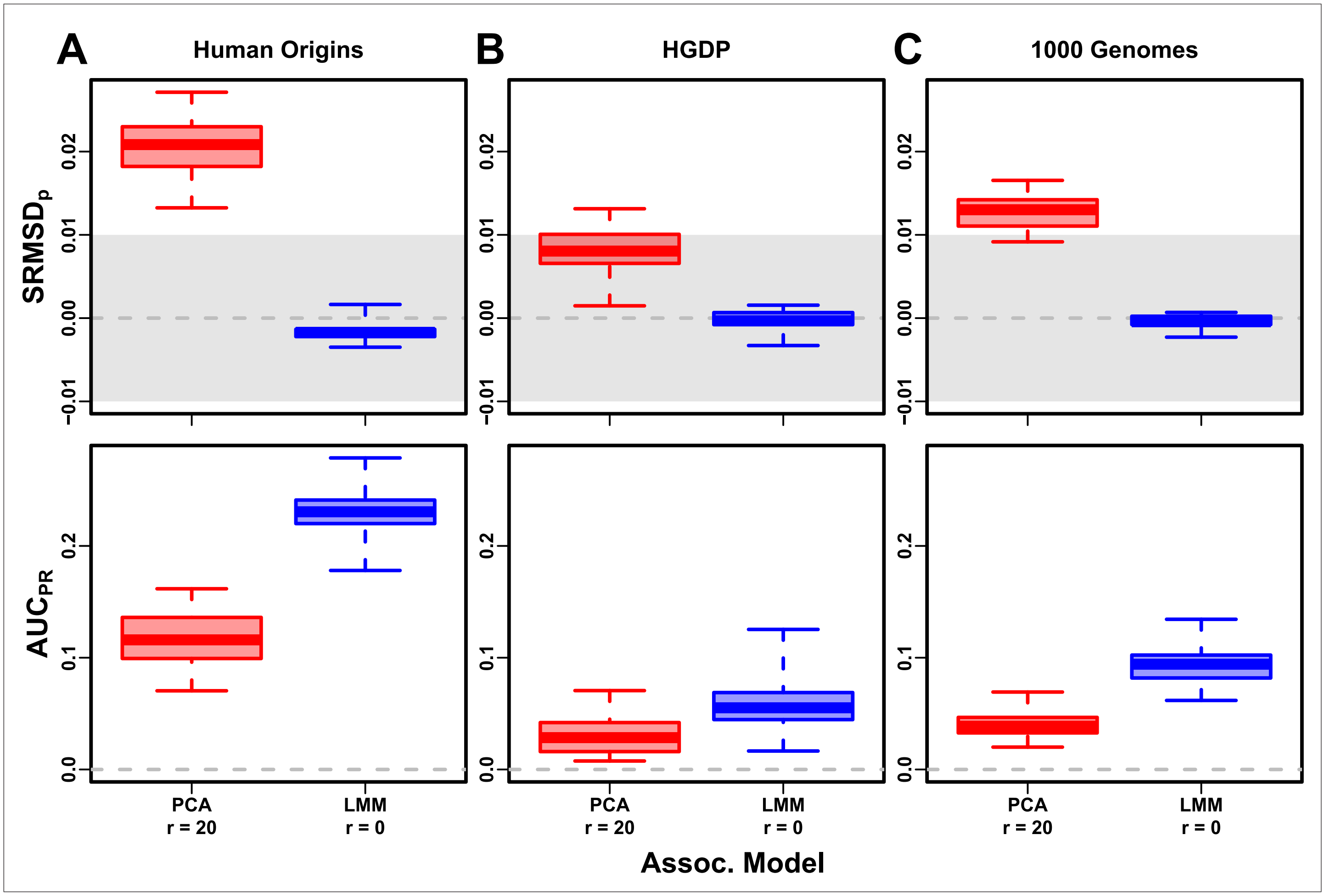

Figure 7. Evaluation in real datasets excluding 4th degree relatives, FES traits, high heritability. Each dataset is a column, rows are measures. Boxplot whiskers are extrema over 50 replicates. First row has $|\mathrm{SRMSD}_p| < 0.01$ band marked as gray area.

The online version of this article includes the following figure supplement(s) for figure 7:

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

The gap between LMM and PCA was reduced in these evaluations, but the main conclusion of the high heritability evaluation holds for low heritability as well, namely that LMM with $r = 0$ significantly outperforms or ties LMM with $r > 0$ and PCA in all cases (***Table 5***).

Lastly, we simulated traits with both low heritability and large environment effects determined by geography and subpopulation labels, so they are strongly correlated to the low-dimensional population structure. For that reason, PCs may be expected to perform better in this setting (in either PCA or LMM). However, we find that both PCA and LMM (even without PCs) increase their $\mathrm{AUC}_{\mathrm{PR}}$ values compared to the low-heritability evaluations (***Figure 8—figure supplement 1***; ***Figure 8*** also shows representative numbers of PCs, which performed optimally or nearly so in individual simulations shown in ***Figure 3—figure supplement 4***, ***Figure 3—figure supplement 5***, ***Figure 4—figure supplement 4***, ***Figure 4—figure supplement 5***). p-Value calibration is comparable with or without environment effects, for LMM for all r and for PCA once r is large enough (***Figure 8—figure supplement 1***). These simulations are the only where we occasionally observed for both metrics a significant, though small, advantage of LMM with PCs versus LMM without PCs (***Table 6***). Additionally, on RC traits only, PCA significantly outperforms LMM in the three real human datasets (***Table 6***), the only cases in all of our evaluations where this is observed. For comparison, we also evaluate an 'oracle' LMM without PCs but with the finest group labels, the same used to simulate environment, as fixed categorical covariates ('LMM lab.'), and see much larger $\mathrm{AUC}_{\mathrm{PR}}$ values than either LMM with PCs or PCA (***Figure 8***, ***Figure 3—figure supplement 4***, ***Figure 3—figure supplement 5***, ***Figure 4—figure***

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

			LMM $r = 0$ vs best r			PCA vs LMM $r = 0$			
Dataset	**Metric**	**Trait**[*]	**Cal.**[†]	**Best r**[‡]	**p-value**[§]	**Best r**[‡]	**Cal.**[†]	**p-value**[§]	**Best model**[¶]
Admix. Large sim.	$\lvert\mathrm{SRMSD}_p\rvert$	FES	True	0	1	62	True	0.00012*	LMM
Admix. Small sim.	$\lvert\mathrm{SRMSD}_p\rvert$	FES	True	0	1	3	True	0.27	Tie
Admix. Family sim.	$\lvert\mathrm{SRMSD}_p\rvert$	FES	True	0	1	90	False	3.9e-10*	LMM
Human Origins	$\lvert\mathrm{SRMSD}_p\rvert$	FES	True	0	1	81	True	3.9e-10*	LMM
HGDP	$\lvert\mathrm{SRMSD}_p\rvert$	FES	True	0	1	37	True	6.2e-09*	LMM
1000 Genomes	$\lvert\mathrm{SRMSD}_p\rvert$	FES	True	0	1	84	True	3.9e-10*	LMM
Admix. Large sim.	$\lvert\mathrm{SRMSD}_p\rvert$	RC	True	0	1	35	True	0.00094	Tie
Admix. Small sim.	$\lvert\mathrm{SRMSD}_p\rvert$	RC	True	0	1	3	True	0.087	Tie
Admix. Family sim.	$\lvert\mathrm{SRMSD}_p\rvert$	RC	True	0	1	90	False	4.1e-10*	LMM
Human Origins	$\lvert\mathrm{SRMSD}_p\rvert$	RC	True	0	1	75	True	0.00016*	LMM
HGDP	$\lvert\mathrm{SRMSD}_p\rvert$	RC	True	0	1	23	True	1.7e-05*	LMM
1000 Genomes	$\lvert\mathrm{SRMSD}_p\rvert$	RC	True	0	1	41	True	6.7e-10*	LMM
Admix. Large sim.	AUC_PR	FES		0	1	3		0.11	Tie
Admix. Small sim.	AUC_PR	FES		0	1	0		0.58	Tie
Admix. Family sim.	AUC_PR	FES		0	1	7		2.2e-06*	LMM
Human Origins	AUC_PR	FES		0	1	16		8e-10*	LMM
HGDP	AUC_PR	FES		11	0.68	6		0.0043	Tie
1000 Genomes	AUC_PR	FES		6	0.34	4		2.3e-07*	LMM
Admix. Large sim.	AUC_PR	RC		0	1	3		0.14	Tie
Admix. Small sim.	AUC_PR	RC		0	1	0		0.1	Tie
Admix. Family sim.	AUC_PR	RC		0	1	5		1.9e-06*	LMM
Human Origins	AUC_PR	RC		4	0.16	12		0.003	Tie
HGDP	AUC_PR	RC		2	0.14	5		0.14	Tie
1000 Genomes	AUC_PR	RC		0	1	4		0.078	Tie

[*] FES: Fixed Effect Sizes, RC: Random Coefficients.

[†] Calibrated: whether mean $\lvert\mathrm{SRMSD}_p\rvert < 0.01$ over 50 replicates.

[‡] Value of r (number of PCs) with minimum mean $\lvert\mathrm{SRMSD}_p\rvert$ or maximum mean AUC_PR.

[§] Wilcoxon paired 1-tailed test of distributions ($\lvert\mathrm{SRMSD}_p\rvert$ or AUC_PR) between models in header. Asterisk marks significant value using Bonferroni threshold ($p < \alpha/n_\mathrm{tests}$ with $\alpha = 0.01$ and $n_\mathrm{tests} = 48$ is the number of tests in this table).

[¶] Tie if no significant difference using Bonferroni threshold.

supplement 4, ***Figure 4—figure supplement 5***, ***Table 6***). However, LMM with labels is often more poorly calibrated than LMM or PCA without labels, which may be since these numerous labels are inappropriately modeled as fixed rather than random effects. Overall, we find that association studies with correlated environment and genetic effects remain a challenge for PCA and LMM, that addition of PCs to an LMM improves performance only marginally, and that if the environment effect is driven by geography or ethnicity then use of those labels greatly improves performance compared to using PCs.

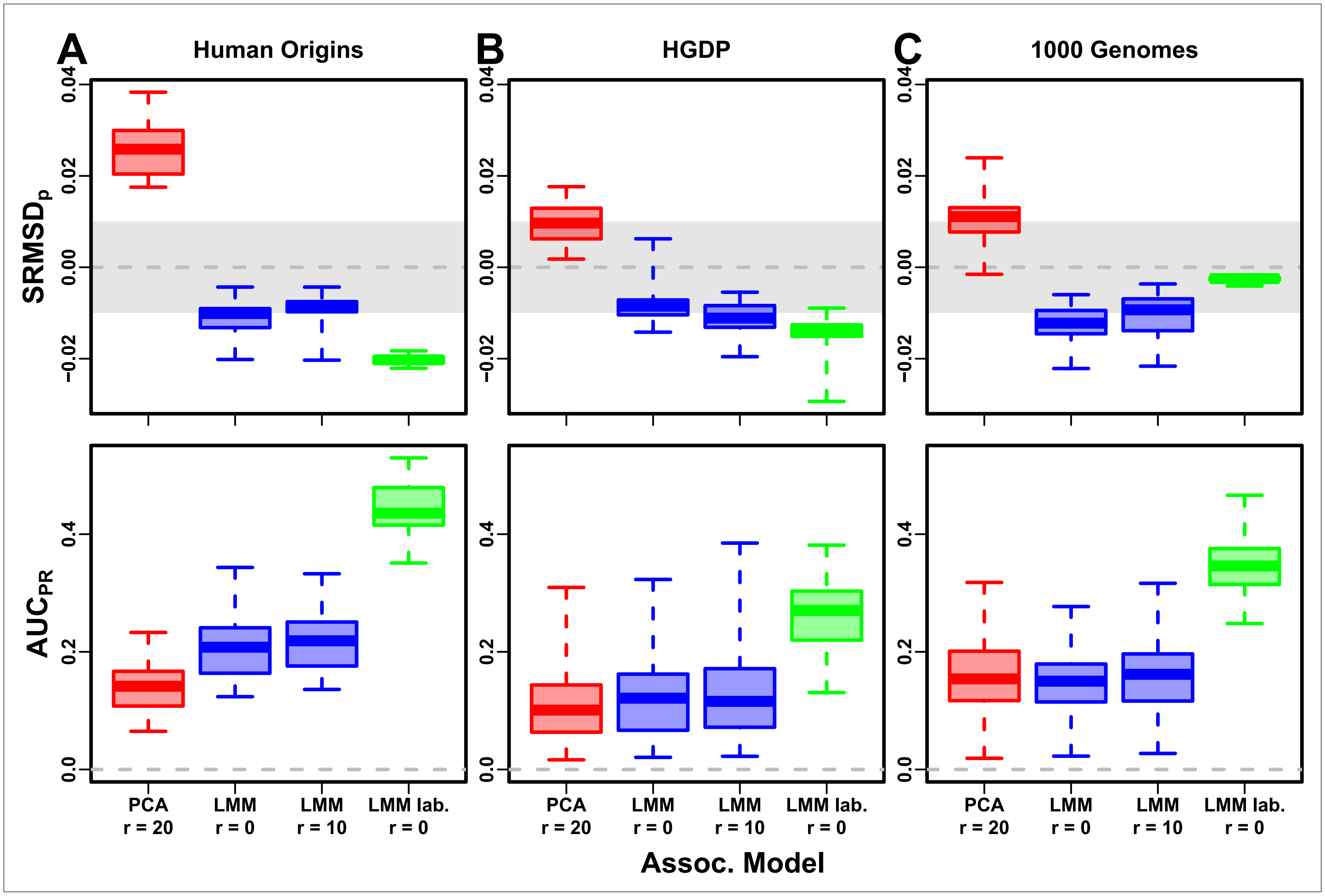

Figure 8. Evaluation in real datasets excluding 4th degree relatives, FES traits, environment. Traits simulated with environment effects, otherwise the same as ***Figure 7***. 'LMM lab.' includes as fixed effects true groups from which environment was simulated.

The online version of this article includes the following figure supplement(s) for figure 8:

Figure supplement 1. Comparison of performance in low heritability vs environment simulations.

Discussion

Our evaluations conclusively determined that LMM without PCs performs better than PCA (for any number of PCs) across all scenarios without environment effects, including all real and simulated genotypes and two trait simulation models. Although the addition of a few PCs to LMM does not greatly hurt its performance (except for small sample sizes), they generally did not improve it either (***Table 3***, ***Table 5***), which agrees with previous observations (***Liu et al., 2011***; ***Janss et al., 2012***) but contradicts others (***Zhao et al., 2007***; ***Price et al., 2010***). Our findings make sense since PCs are the eigenvectors of the same kinship matrix that parameterized random effects, so including both is redundant.

The presence of environment effects that are correlated to relatedness presents the only scenario where occasionally PCA and LMM with PCs outperform LMM without PCs (***Table 6***). It is commonly believed that PCs model such environment effects well (***Novembre et al., 2008***; ***Zhang and Pan, 2015***; ***Lin et al., 2021***). However, we observe that LMM without PCs models environment effects nearly as well as with PCs (***Figure 8***), consistent with previous findings (***Vilhjálmsson and Nordborg, 2013***; ***Wang et al., 2022***) and with environment inflating heritability estimates using LMM (***Heckerman et al., 2016***). Moreover, modeling the true environment groups as fixed categorical effects always substantially improved $\mathrm{AUC}_{\mathrm{PR}}$ compared to modeling them with PCs (***Figure 8***, ***Table 6***). Modeling numerous environment groups as fixed effects does result in deflated p-values (***Figure 8***, ***Table 6***), which we expect would be avoided by modeling them as random effects, a strategy we chose not to

Table 6. Overview of PCA and LMM evaluations for environment simulations.

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

*FES: Fixed Effect Sizes, RC: Random Coefficients.

†Calibrated: whether mean $|\text{SRMSD}_p| < 0.01$ over 50 replicates.

‡Value of r (number of PCs) with minimum mean $|\text{SRMSD}_p|$ or maximum mean AUC_PR.

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

¶Tie if no significant difference using Bonferroni threshold; in last column, pairwise ties are specified and "Tie" is three-way tie.

pursue here as it is both a circular evaluation (the true effects were drawn from that model) and out of scope. Overall, including PCs to model environment effects yields limited power gains if at all, even in an LMM, and is no replacement for more adequate modeling of environment whenever possible.

Previous studies found that PCA was better calibrated than LMM for unusually differentiated markers (***Price et al., 2010***; ***Wu et al., 2011***; ***Yang et al., 2014***), which as simulated were 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***). Our evaluations on real human data, which contain such loci in relevant proportions if they exist, do not replicate that result. Family relatedness strongly favors LMM, an advantage that probably outweighs this potential PCA benefit in real data.

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

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

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

Our tests also found PCA robust to large numbers of PCs, far beyond the optimal choice, agreeing with previous anecdotal observations (***Price et al., 2006***; ***Kang et al., 2010***), in contrast to using too few PCs for which there is a large performance penalty. The exception was the small sample size simulation, where only small numbers of PCs performed well. In contrast, LMM is simpler since there is no need to choose the number of PCs. However, an LMM with a large number of covariates may have conservative p-values, as observed for LMM with large numbers of PCs, which is a weakness of the score test used by the LMM we evaluated that may be overcome with other statistical tests. Simulations or post hoc evaluations remain crucial for ensuring that statistics are calibrated.

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

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

Another limitation of this work is ignoring rare variants, a necessity given our smaller sample sizes, where rare variant association is miscalibrated and underpowered. Using simulations mimicking the UK Biobank, recent work has found that rare variants can have a more pronounced structure than common variants, and that modeling this rare variant structure (with either PCA and LMM) may better model environment confounding, reduce inflation in association studies, and ameliorate stratification in polygenic risk scores (***Zaidi and Mathieson, 2020***). Better modeling rare variants and their structure is a key next step in association studies.

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

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

Materials and methods

The complex trait model and PCA and LMM approximations

Let $x_{ij} \in \{0, 1, 2\}$ be the genotype at the biallelic 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 of individual trait values. The additive linear model for a quantitative (continuous) trait is:

$$\mathbf{y} = \mathbf{1}\alpha + \mathbf{X}'\boldsymbol{\beta} + \mathbf{Z}'\boldsymbol{\eta} + \boldsymbol{\epsilon}, \tag{1}$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Simulations

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

Genotype simulation from the admixture model

The basic admixture model is as described previously (***Ochoa and Storey, 2021***) and is implemented in the R package bnpsd. Both Large and Family simulations have $n = 1,000$ individuals, while Small has $n = 100$. The number of loci is $m = 100,000$. Individuals are admixed from $K = 10$ intermediate subpopulations, or ancestries. Each subpopulation S_u ($u \in \{1, ..., K\}$) is 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 S_u arise from a random walk with spread σ on the 1D geography, and τ and σ are fit to give $F_{\mathrm{ST}} = 0.1$ and mean kinship $\bar{\theta}^T = 0.5F_{\mathrm{ST}}$ for the admixed individuals (***Ochoa and Storey, 2021***). Random ancestral allele frequencies p_i^T, subpopulation allele frequencies $p_i^{S_u}$, individual-specific allele frequencies π_{ij}, and genotypes x_{ij} are drawn from this hierarchical model:

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

where this Beta is the Balding-Nichols distribution (***Balding and Nichols, 1995***) with mean p_i^T and variance $p_i^T \left(1 - p_i^T\right) f_{S_u}^T$. 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 genotypes starting from p_i^T.

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

Genotype simulation from random admixed families

We simulated a pedigree with admixed founders, no close relative pairings, assortative mating based on a 1D geography (to preserve admixture structure), random family sizes, and arbitrary numbers of generations (20 here). This simulation is implemented in the R package simfam. Generations are drawn iteratively. Generation 1 has $n = 1000$ individuals from the above admixture simulation ordered by their 1D geography. Local kinship measures pedigree relatedness; in the first generation, everybody is locally unrelated and outbred. Individuals are randomly assigned sex. In the next generation, individuals are paired iteratively, removing random males from the pool of available males and pairing them with the nearest available female with local kinship $< 1/4^3$ (stay unpaired if there are no matches), until there are no more available males or females. Let $n = 1000$ be the desired population size, $n_m = 1$ the minimum number of children per family and n_f the number of families (paired parents) in the current generation, then the number of additional children (beyond the minimum) is drawn from $\mathrm{Poisson}(n/n_f - n_m)$. 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. Children draw alleles from their parents independently per locus. A new random pedigree is drawn for each replicate, as well as new founder genotypes from the admixture model.

Genotype simulation from a subpopulation tree model

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

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

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

Fitting subpopulation tree to real data

We developed new methods to fit trees to real data based on unbiased kinship estimates from popkin, implemented in bnpsd. A tree with given inbreeding coefficients $f_{S_w}^{P_w}$ for its edges (between subpopulation S_w and its parent P_w) gives rise to a coancestry matrix ϑ_{uv}^T for a subpopulation pair (S_u, S_v), and the goal is to recover these edge inbreeding coefficients from coancestry estimates. Coancestry values are total inbreeding coefficients of the MRCA population of each subpopulation pair. Therefore, we calculate $f_{S_w}^T$ for every S_w recursively from the root as follows. Nodes with parent $P_w = T$ are already as desired. Given $f_{P_w}^T$, the desired $f_{S_w}^T$ is calculated via the 'additive edge' δ_w (***Ochoa and Storey, 2021***):

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

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

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

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

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

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

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

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

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

To estimate tree edge lengths, first δ_w are estimated from $\hat{\vartheta}^T_{uv}$ and the topology using ***Equation 7*** and non-negative least squares linear regression (***Lawson and Hanson, 1974***) (implemented in nnls; ***Mullen, 2012***) to yield non-negative δ_w, and $f^{P_w}_{S_w}$ are calculated from δ_w by reversing ***Equation 5***. To account for small biases in coancestry estimation, an intercept term δ_0 is included ($I_0(u, v) = 1$ for all u, v), and when converting δ_w to $f^{P_w}_{S_w}$, δ_0 is treated as an additional edge to the root, but is ignored when drawing allele frequencies from the tree.

Trait simulation

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

$$\beta_i = \frac{1}{\sqrt{2v^T_i}}$$

for known p^T_i; otherwise v^T_i is replaced by the unbiased estimator (***Ochoa and Storey, 2021***) $\hat{v}^T_i = \hat{p}^T_i \left(1 - \hat{p}^T_i\right) / (1 - \bar{\varphi}^T)$, where $\bar{\varphi}^T$ is the mean kinship estimated with popkin. Each causal locus is multiplied by –1 with probability 0.5. Alternatively, under the *random coefficients* (RC) model, initial causal coefficients are drawn independently from $\beta_i \sim \text{Normal}(0, 1)$. For both models, the initial genetic variance is $\sigma^2_0 = \sum_{i \in C} 2v^T_i \beta^2_i$, replacing v^T_i with $\hat{v}^T_i$ for unknown p^T_i (so σ^2_0 is an unbiased estimate), so

Table 7. Variance parameters of trait simulations.

Trait variance type	h^2	σ^2_η	σ^2_ϵ
High heritability	0.8	0.0	0.2
Low heritability	0.3	0.0	0.7
Environment	0.3	0.5	0.2

we multiply every initial β_i by $\frac{h}{\sigma_0}$ to have the desired heritability. Lastly, for known p_i^T, the intercept coefficient is $\alpha = -\sum_{i \in C} 2p_i^T \beta_i$. When p_i^T are unknown, $\hat{p}_i^T$ should not replace p_i^T since that distorts the trait covariance (for the same reason the standard kinship estimator in ***Equation 5*** is biased), which is avoided with

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

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

We performed trait simulations with the following variance parameters (***Table 7***): *high heritability* used $h^2 = 0.8$ and no environment effects; *low heritability* used $h^2 = 0.3$ and no environment effects; lastly, *environment* used $h^2 = 0.3, \sigma^2_{\eta 1} = 0.3, \sigma^2_{\eta 2} = 0.2$ (total $\sigma^2_\eta = \sigma^2_{\eta 1} + \sigma^2_{\eta 2} = 0.5$). For real genotype datasets, the groups are the continental (environment 1) and fine-grained (environment 2) subpopulation labels given (see next subsection). For simulated genotypes, we created these labels by grouping by the index j (geographical coordinate) of each simulated individual, assigning group $g = \text{ceiling}(jk_i/n)$ where k_i is the number of groups in environment i, and we selected $k_1 = 5$ and $k_2 = 25$ to mimic the number of groups in each level of 1000 Genomes (***Table 2***).

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***), and analysis was uniquely enabled by the R packages BEDMatrix (***Grueneberg and de Los Campos, 2019***) and genio. Each dataset groups individuals in a two-level hierarchy: continental and fine-grained subpopulations. Final dataset sizes are in ***Table 2***.

We obtained the full (including non-public) Human Origins by contacting the authors and agreeing to their usage restrictions. The Pacific data (***Skoglund et al., 2016***) was obtained separately 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. Our analysis of both the whole-genome sequencing (WGS) version of HGDP (***Bergström et al., 2020***) and the high-coverage NYGC version of 1000 Genomes (***Fairley et al., 2020***) was restricted to autosomal biallelic SNP loci with filter “PASS”.

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

Evaluation of performance

All approaches are evaluated using two complementary metrics: SRMSD_p quantifies p-value uniformity, and $\mathrm{AUC_{PR}}$ measures causal locus classification performance and reflects power while ranking miscalibrated models fairly. These measures are more robust alternatives to previous measures from the literature (Appendix 2), and are implemented in simtrait.

P-values for continuous test statistics have a uniform distribution when the null hypothesis holds, a crucial assumption for type I error and FDR control (***Storey, 2003***; ***Storey and Tibshirani, 2003***). We use the Signed Root Mean Square Deviation (SRMSD_p) to measure the difference between the observed null p-value quantiles and the expected uniform quantiles:

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

where $m_0 = m - m_1$ is the number of null (non-causal) loci, here i indexes null loci only, $p_{(i)}$ is the i th ordered null p-value, $u_i = (i - 0.5)/m_0$ is its expectation, p_{median} is the median observed null p-value, $u_{\mathrm{median}} = \frac{1}{2}$ is its expectation, and sgn is the sign function (1 if $u_{\mathrm{median}} \geq p_{\mathrm{median}}$, –1 otherwise). Thus, $\mathrm{SRMSD}_p = 0$ corresponds to calibrated p-values, $\mathrm{SRMSD}_p > 0$ indicate anti-conservative p-values, and $\mathrm{SRMSD}_p < 0$ 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

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

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

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

$$\mathrm{Precision}(t) = \frac{\mathrm{TP}(t)}{\mathrm{TP}(t) + \mathrm{FP}(t)},$$
$$\mathrm{Recall}(t) = \frac{\mathrm{TP}(t)}{\mathrm{TP}(t) + \mathrm{FN}(t)}.$$

Precision and Recall trace a curve as t is varied, and the area under this curve is $\mathrm{AUC_{PR}}$. We use the R package PRROC to integrate the correct non-linear piecewise function when interpolating between points. A model obtains the maximum $\mathrm{AUC_{PR}} = 1$ if there is a t 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 causal loci: m_1/m.

Data and code availability

The data and code generated during this study are available on GitHub at https://github.com/OchoaLab/pca-assoc-paper (copy archived at ***Ochoa, 2023***). The public subset of Human Origins is available on the Reich Lab website at https://reich.hms.harvard.edu/datasets; non-public samples have to be requested from David Reich. The WGS version of HGDP was downloaded from the Wellcome Sanger Institute FTP site at ftp://ngs.sanger.ac.uk/production/hgdp/hgdp_wgs.20190516/. The high-coverage version of the 1000 Genomes Project was downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20190425_NYGC_GATK/.

Web resources

plink2, https://www.cog-genomics.org/plink/2.0/ ; GCTA, https://yanglab.westlake.edu.cn/software/gcta/ ; Eigensoft, https://github.com/DReichLab/EIG ; bnpsd, https://cran.r-project.org/package=

bnpsd ; simfam, https://cran.r-project.org/package=simfam ; simtrait, https://cran.r-project.org/package=simtrait ; genio, https://cran.r-project.org/package=genio ; popkin, https://cran.r-project.org/package=popkin ; ape, https://cran.r-project.org/package=ape ; nnls, https://cran.r-project.org/package=nnls ; PRROC, https://cran.r-project.org/package=PRROC ; BEDMatrix, https://cran.r-project.org/package=BEDMatrix.

Acknowledgements

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

Additional information

Competing interests

Yiqi Yao: is affiliated with BenHealth Consulting. The author has no financial interests to declare. The other author declares that no competing interests exist.

Funding

Funder	Grant reference number	Author
Whitehead Foundation		Alejandro Ochoa

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Yiqi Yao, Software, Formal analysis, Investigation, Visualization, Writing – original draft, Writing – review and editing; Alejandro Ochoa, Conceptualization, Resources, Data curation, Software, Formal analysis, Supervision, Funding acquisition, Validation, Investigation, Visualization, Methodology, Writing – original draft, Project administration, Writing – review and editing

Author ORCIDs

Alejandro Ochoa http://orcid.org/0000-0003-4928-3403

Decision letter and Author response

Decision letter https://doi.org/10.7554/eLife.79238.sa1
Author response https://doi.org/10.7554/eLife.79238.sa2

Additional files

Supplementary files

- MDAR checklist

Data availability

The current manuscript is a computational study, so no data have been generated for this manuscript. Code is available at https://github.com/OchoaLab/pca-assoc-paper (copy archived at ***Ochoa, 2023***).

The following previously published datasets were used:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Fairley S	2020	1000 Genomes Project, high-coverage version	ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20190425_NYGC_GATK/	International Genome Sample Resource, NYGC_GATK/
Bergstrom A	2020	Human Genome Diversity Panel, whole-genome sequencing version	ftp://ngs.sanger.ac.uk/production/hgdp/hgdp_wgs.20190516/	Wellcome Sanger Institute, wgs.20190516/
Lazaridis I	2016	Human Origins	https://reich.hms.harvard.edu/datasets	David Reich Lab, datasets

References

Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA, 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**:1061–1073. DOI: https://doi.org/10.1038/nature09534, PMID: 20981092

Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA, 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human Genomes. *Nature* **491**:56–65. DOI: https://doi.org/10.1038/nature11632, PMID: 23128226

Abraham G, Inouye M. 2014. Fast principal component analysis of large-scale genome-wide data. *PLOS ONE* **9**:e93766. DOI: https://doi.org/10.1371/journal.pone.0093766, PMID: 24718290

Abraham G, Qiu Y, Inouye M. 2017. Flashpca2: Principal component analysis of Biobank-scale genotype Datasets. *Bioinformatics* **33**:2776–2778. DOI: https://doi.org/10.1093/bioinformatics/btx299, PMID: 28475694

Agrawal A, Chiu AM, Le M, Halperin E, Sankararaman S. 2020. Scalable probabilistic PCA for large-scale genetic variation data. *PLOS Genetics* **16**:e1008773. DOI: https://doi.org/10.1371/journal.pgen.1008773, PMID: 32469896

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* **19**:1655–1664. DOI: https://doi.org/10.1101/gr.094052.109, PMID: 19648217

Al Khudhair A, Qiu S, Wyse M, Chowdhury S, Cheng X, Bekbolsynov D, Saha-Mandal A, Dutta R, Fedorova L, Fedorov A. 2015. Inference of distant genetic relations in humans using "1000 Genomes" *Genome Biology and Evolution* **7**:481–492. DOI: https://doi.org/10.1093/gbe/evv003, PMID: 25573959

Astle W, Balding DJ. 2009. Population structure and cryptic relatedness in genetic Association studies. *Statistical Science* **24**:451–471. DOI: https://doi.org/10.1214/09-STS307

Aulchenko YS, de Koning D-J, Haley C. 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**:577–585. DOI: https://doi.org/10.1534/genetics.107.075614, PMID: 17660554

Balding DJ, Nichols RA. 1995. A method for Quantifying differentiation between populations at multi-Allelic Loci and its implications for investigating identity and Paternity. *Genetica* **96**:3–12. DOI: https://doi.org/10.1007/BF01441146, PMID: 7607457

Bergström A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, Chen Y, Felkel S, Hallast P, Kamm J, Blanché H, Deleuze JF, Cann H, Mallick S, Reich D, Sandhu MS, Skoglund P, Scally A, Xue Y, Durbin R, et al. 2020. Insights into human genetic variation and population history from 929 diverse Genomes. *Science* **367**:eaay5012. DOI: https://doi.org/10.1126/science.aay5012, PMID: 32193295

Bouaziz M, Ambroise C, Guedj M. 2011. Accounting for population stratification in practice: A comparison of the main strategies dedicated to genome-wide Association studies. *PLOS ONE* **6**:e28845. DOI: https://doi.org/10.1371/journal.pone.0028845, PMID: 22216125

Cabreros I, Storey JD. 2019. A likelihood-free estimator of population structure bridging admixture models and principal components analysis. *Genetics* **212**:1009–1029. DOI: https://doi.org/10.1534/genetics.119.302159, PMID: 31028112

Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A, Chen Z, Chu J, Carcassi C, Contu L, Du R, Excoffier L, Ferrara GB, Friedlaender JS, Groot H, Gurwitz D, et al. 2002. A human genome diversity cell line panel. *Science* **296**:261–262. DOI: https://doi.org/10.1126/science.296.5566.261b, PMID: 11954565

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK: Rising to the challenge of larger and richer Datasets. *GigaScience* **4**:7. DOI: https://doi.org/10.1186/s13742-015-0047-8, PMID: 25722852

Chiu AM, Molloy EK, Tan Z, Talwalkar A, Sankararaman S. 2022. Inferring population structure in biobank-scale genomic data. *American Journal of Human Genetics* **109**:727–737. DOI: https://doi.org/10.1016/j.ajhg.2022.02.015, PMID: 35298920

Conomos MP, Laurie CA, Stilp AM, Gogarten SM, McHugh CP, Nelson SC, Sofer T, Fernández-Rhodes L, Justice AE, Graff M, Young KL, Seyerle AA, Avery CL, Taylor KD, Rotter JI, Talavera GA, Daviglus ML, Wassertheil-Smoller S, Schneiderman N, Heiss G, et al. 2016a. Genetic diversity and Association studies in US Hispanic/Latino populations: Applications in the Hispanic community health study/study of Latinos. *The American Journal of Human Genetics* **98**:165–184. DOI: https://doi.org/10.1016/j.ajhg.2015.12.001, PMID: 26748518

Conomos MP, Reiner AP, Weir BS, Thornton TA. 2016b. Model-free estimation of recent genetic relatedness. *The American Journal of Human Genetics* **98**:127–148. DOI: https://doi.org/10.1016/j.ajhg.2015.11.022, PMID: 26748516

Coram MA, Duan Q, Hoffmann TJ, Thornton T, Knowles JW, Johnson NA, Ochs-Balcom HM, Donlon TA, Martin LW, Eaton CB, Robinson JG, Risch NJ, Zhu X, Kooperberg C, Li Y, Reiner AP, Tang H. 2013. Genome-wide characterization of shared and distinct genetic components that influence blood lipid levels in ethnically diverse human populations. *American Journal of Human Genetics* **92**:904–916. DOI: https://doi.org/10.1016/j.ajhg.2013.04.025, PMID: 23726366

Devlin B, Roeder K. 1999. Genomic control for Association studies. *Biometrics* **55**:997–1004. DOI: https://doi.org/10.1111/j.0006-341x.1999.00997.x, PMID: 11315092

Fairley S, Lowy-Gallego E, Perry E, Flicek P. 2020. The International genome sample resource (IGSR) collection of open human Genomic variation resources. *Nucleic Acids Research* **48**:D941–D947. DOI: https://doi.org/10.1093/nar/gkz836, PMID: 31584097

Fedorova L, Qiu S, Dutta R, Fedorov A. 2016. Atlas of cryptic genetic relatedness among 1000 human Genomes. *Genome Biology and Evolution* **8**:777–790. DOI: https://doi.org/10.1093/gbe/evw034, PMID: 26907499

Galinsky KJ, Bhatia G, Loh PR, Georgiev S, Mukherjee S, Patterson NJ, Price AL. 2016. Fast principal-component analysis reveals CONVERGENT evolution of Adh1B in Europe and East Asia. *American Journal of Human Genetics* **98**:456–472. DOI: https://doi.org/10.1016/j.ajhg.2015.12.022, PMID: 26924531

Gazal S, Sahbatou M, Babron MC, Génin E, Leutenegger AL. 2015. High level of inbreeding in final phase of 1000 Genomes project. *Scientific Reports* **5**:17453. DOI: https://doi.org/10.1038/srep17453, PMID: 26625947

Gopalan P, Hao W, Blei DM, Storey JD. 2016. Scaling probabilistic models of genetic variation to millions of humans. *Nature Genetics* **48**:1587–1590. DOI: https://doi.org/10.1038/ng.3710, PMID: 27819665

Grau J, Grosse I, Keilwagen J. 2015. PRROC: Computing and Visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics* **31**:2595–2597. DOI: https://doi.org/10.1093/bioinformatics/btv153, PMID: 25810428

Grueneberg A, de Los Campos G. 2019. Bgdata - A suite of R packages for Genomic analysis with big data. G3: Genes, Genomes, Genetics **9**:1377–1383. DOI: https://doi.org/10.1534/g3.119.400018, PMID: 30894453

Hayashi K, Yuan KH, Liang L. 2018. On the bias in Eigenvalues of sample covariance matrix. Wiberg M, Culpepper S, Janssen R, González J, Molenaar D (Eds). *Quantitative Psychology Springer Proceedings in Mathematics & Statistics* Cham: Springer International Publishing. p. 221–233. DOI: https://doi.org/10.1007/978-3-319-77249-3_19

Heckerman D, Gurdasani D, Kadie C, Pomilla C, Carstensen T, Martin H, Ekoru K, Nsubuga RN, Ssenyomo G, Kamali A, Kaleebu P, Widmer C, Sandhu MS. 2016. Linear mixed model for Heritability estimation that explicitly addresses environmental variation. *PNAS* **113**:7377–7382. DOI: https://doi.org/10.1073/pnas.1510497113, PMID: 27382152

Henn BM, Hon L, Macpherson JM, Eriksson N, Saxonov S, Pe’er I, Mountain JL. 2012. Cryptic distant relatives are common in both isolated and cosmopolitan genetic samples. *PLOS ONE* **7**:e34267. DOI: https://doi.org/10.1371/journal.pone.0034267, PMID: 22509285

Hindorff LA, Bonham VL, Brody LC, Ginoza MEC, Hutter CM, Manolio TA, Green ED. 2018. Prioritizing diversity in human Genomics research. *Nature Reviews. Genetics* **19**:175–185. DOI: https://doi.org/10.1038/nrg.2017.89, PMID: 29151588

Hodonsky CJ, Jain D, Schick UM, Morrison JV, Brown L, McHugh CP, Schurmann C, Chen DD, Liu YM, Auer PL, Laurie CA, Taylor KD, Browning BL, Li Y, Papanicolaou G, Rotter JI, Kurita R, Nakamura Y, Browning SR, Loos RJF, et al. 2017. Genome-wide Association study of red blood cell traits in Hispanics/Latinos: The Hispanic community health study/study of Latinos. *PLOS Genetics* **13**:e1006760. DOI: https://doi.org/10.1371/journal.pgen.1006760, PMID: 28453575

Hoffman GE, Dubé M-P. 2013. Correcting for population structure and kinship using the linear mixed model: theory and extensions. *PLOS ONE* **8**:e75707. DOI: https://doi.org/10.1371/journal.pone.0075707

Hoffmann TJ, Choquet H, Yin J, Banda Y, Kvale MN, Glymour M, Schaefer C, Risch N, Jorgenson E. 2018. A large Multiethnic genome-wide Association study of adult body mass index identifies novel Loci. *Genetics* **210**:499–515. DOI: https://doi.org/10.1534/genetics.118.301479, PMID: 30108127

Hou K, Ding Y, Xu Z, Wu Y, Bhattacharya A, Mester R, Belbin GM, Buyske S, Conti DV, Darst BF, Fornage M, Gignoux C, Guo X, Haiman C, Kenny EE, Kim M, Kooperberg C, Lange L, Manichaikul A, North KE, et al. 2023a. Causal effects on complex traits are similar for common variants across segments of different Continental Ancestries within admixed individuals. *Nature Genetics* **55**:549–558. DOI: https://doi.org/10.1038/s41588-023-01338-6, PMID: 36941441

Hou Z, Ochoa A, Browning S. 2023b. Genetic Association models are robust to common population kinship estimation biases. *GENETICS* **224**:iyad030. DOI: https://doi.org/10.1093/genetics/iyad030, PMID: 36843304

Hu Y, Graff M, Haessler J, Buyske S, Bien SA, Tao R, Highland HM, Nishimura KK, Zubair N, Lu Y, Verbanck M, Hilliard AT, Klarin D, Damrauer SM, Ho Y-L, Wilson PWF, Chang K-M, Tsao PS, Cho K, O’Donnell CJ, et al. 2020. Minority-centric meta-analyses of blood lipid levels identify novel Loci in the population architecture using

Genomics and epidemiology (page) study. *PLOS Genetics* **16**:e1008684. DOI: https://doi.org/10.1371/journal.pgen.1008684, PMID: 32226016

Jacquard A. 1970. *Structures Génétiques Des Populations* Paris: Masson et Cie.

Janss L, de Los Campos G, Sheehan N, Sorensen D. 2012. Inferences from Genomic models in stratified populations. *Genetics* **192**:693–704. DOI: https://doi.org/10.1534/genetics.112.141143, PMID: 22813891

Jolliffe IT. 2002. *Principal Component Analysis* New York: Springer-Verlag.

Kamariza M, Crawford L, Jones D, Finucane H. 2021. Misuse of the term 'Trans-ethnic' in Genomics research. *Nature Genetics* **53**:1520–1521. DOI: https://doi.org/10.1038/s41588-021-00952-6, PMID: 34741159

Kang H.M, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E. 2008. Efficient control of population structure in model organism association mapping. *Genetics* **178**:1709–1723. DOI: https://doi.org/10.1534/genetics.107.080101

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S-Y, Freimer NB, Sabatti C, Eskin E. 2010. Variance component model to account for sample structure in genome-wide Association studies. *Nature Genetics* **42**:348–354. DOI: https://doi.org/10.1038/ng.548, PMID: 20208533

Lawson CL, Hanson RJ. 1974. *Solving Least Squares Problems* Englewood Cliffs: Prentice Hall.

Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, Sudmant PH, Schraiber JG, Castellano S, Lipson M, Berger B, Economou C, Bollongino R, Fu Q, Bos KI, Nordenfelt S, Li H, de Filippo C, Prüfer K, Sawyer S, et al. 2014. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**:409–413. DOI: https://doi.org/10.1038/nature13673

Lazaridis I, Nadel D, Rollefson G, Merrett DC, Rohland N, Mallick S, Fernandes D, Novak M, Gamarra B, Sirak K, Connell S, Stewardson K, Harney E, Fu Q, Gonzalez-Fortes G, Jones ER, Roodenberg SA, Lengyel G, Bocquentin F, Gasparian B, et al. 2016. Genomic insights into the origin of farming in the ancient near East. *Nature* **536**:419–424. DOI: https://doi.org/10.1038/nature19310, PMID: 27459054

Lee S, Epstein MP, Duncan R, Lin X. 2012. Sparse principal component analysis for identifying ancestry-informative markers in genome-wide Association studies. *Genetic Epidemiology* **36**:293–302. DOI: https://doi.org/10.1002/gepi.21621, PMID: 22508067

Lin M, Park DS, Zaitlen NA, Henn BM, Gignoux CR. 2021. Admixed populations improve power for variant discovery and Portability in genome-wide Association studies. *Frontiers in Genetics* **12**:673167. DOI: https://doi.org/10.3389/fgene.2021.673167, PMID: 34108994

Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. 2011. Fast linear mixed models for genome-wide Association studies. *Nature Methods* **8**:833–835. DOI: https://doi.org/10.1038/nmeth.1681, PMID: 21892150

Listgarten J, Lippert C, Kadie CM, Davidson RI, Eskin E, Heckerman D. 2012. Improved linear mixed models for genome-wide Association studies. *Nature Methods* **9**:525–526. DOI: https://doi.org/10.1038/nmeth.2037, PMID: 22669648

Liu N, Zhao H, Patki A, Limdi NA, Allison DB. 2011. Controlling population structure in human genetic association studies with samples of unrelated individuals. *Statistics and Its Interface* **4**:317–326. DOI: https://doi.org/10.4310/sii.2011.v4.n3.a6, PMID: 22308192

Liu X, Huang M, Fan B, Buckler ES, Zhang Z, Listgarten J. 2016. Iterative usage of fixed and random effect models for powerful and efficient genome-wide Association studies. *PLOS Genetics* **12**:e1005767. DOI: https://doi.org/10.1371/journal.pgen.1005767, PMID: 26828793

Loh PR, Tucker G, Bulik-Sullivan BK, Vilhjálmsson BJ, Finucane HK, Salem RM, Chasman DI, Ridker PM, Neale BM, Berger B, Patterson N, Price AL. 2015. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics* **47**:284–290. DOI: https://doi.org/10.1038/ng.3190

Loh PR, Kichaev G, Gazal S, Schoech AP, Price AL. 2018. Mixed-model association for biobank-scale datasets. *Nature Genetics* **50**:906–908. DOI: https://doi.org/10.1038/s41588-018-0144-6, PMID: 29892013

Mahajan A, Spracklen CN, Zhang W, Ng MCY, Petty LE, Kitajima H, Yu GZ, Rüeger S, Speidel L, Kim YJ, Horikoshi M, Mercader JM, Taliun D, Moon S, Kwak S-H, Robertson NR, Rayner NW, Loh M, Kim B-J, Chiou J, et al. 2022. Multi-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation. *Nature Genetics* **54**:560–572. DOI: https://doi.org/10.1038/s41588-022-01058-3, PMID: 35551307

Malécot G. 1948. *Mathématiques de l'hérédité* Paris: Masson et Cie.

Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. 2010. Robust relationship inference in genome-wide Association studies. *Bioinformatics* **26**:2867–2873. DOI: https://doi.org/10.1093/bioinformatics/btq559, PMID: 20926424

Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, Daly MJ, Bustamante CD, Kenny EE. 2017a. Human demographic history impacts genetic risk prediction across diverse populations. *American Journal of Human Genetics* **100**:635–649. DOI: https://doi.org/10.1016/j.ajhg.2017.03.004, PMID: 28366442

Martin AR, Lin M, Granka JM, Myrick JW, Liu X, Sockell A, Atkinson EG, Werely CJ, Möller M, Sandhu MS, Kingsley DM, Hoal EG, Liu X, Daly MJ, Feldman MW, Gignoux CR, Bustamante CD, Henn BM. 2017b. An unexpectedly complex architecture for skin Pigmentation in Africans. *Cell* **171**:1340–1353. DOI: https://doi.org/10.1016/j.cell.2017.11.015, PMID: 29195075

Matoba N, Akiyama M, Ishigaki K, Kanai M, Takahashi A, Momozawa Y, Ikegawa S, Ikeda M, Iwata N, Hirata M, Matsuda K, Murakami Y, Kubo M, Kamatani Y, Okada Y. 2020. GWAS of 165,084 Japanese individuals identified nine Loci associated with dietary habits. *Nature Human Behaviour* **4**:308–316. DOI: https://doi.org/10.1038/s41562-019-0805-1, PMID: 31959922

Mbatchou J, Barnard L, Backman J, Marcketta A, Kosmicki JA, Ziyatdinov A, Benner C, O'Dushlaine C, Barber M, Boutkov B, Habegger L, Ferreira M, Baras A, Reid J, Abecasis G, Maxwell E, Marchini J. 2021. Computationally efficient whole-genome regression for quantitative and binary traits. *Nature Genetics* **53**:1097–1103. DOI: https://doi.org/10.1038/s41588-021-00870-7, PMID: 34017140

McVean G. 2009. A Genealogical interpretation of principal components analysis. *PLOS Genetics* **5**:e1000686. DOI: https://doi.org/10.1371/journal.pgen.1000686, PMID: 19834557

Medina-Gomez C, Felix JF, Estrada K, Peters MJ, Herrera L, Kruithof CJ, Duijts L, Hofman A, van Duijn CM, Uitterlinden AG, Jaddoe VWV, Rivadeneira F. 2015. Challenges in conducting genome-wide Association studies in highly admixed multi-ethnic populations: The generation R study. *European Journal of Epidemiology* **30**:317–330. DOI: https://doi.org/10.1007/s10654-015-9998-4, PMID: 25762173

Mogil LS, Andaleon A, Badalamenti A, Dickinson SP, Guo X, Rotter JI, Johnson WC, Im HK, Liu Y, Wheeler HE. 2018. Genetic architecture of gene expression traits across diverse populations. *PLOS Genetics* **14**:e1007586. DOI: https://doi.org/10.1371/journal.pgen.1007586, PMID: 30096133

Mullen KM. 2012. Stokkum Ihmv, Nnls: The Lawson-Hanson algorithm for non-negative least squares (NNLS). The Comprehensive R Archive Network. https://CRAN.R-project.org/package=nnls

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, Stephens M, Bustamante CD. 2008. Genes mirror geography within Europe. *Nature* **456**:98–101. DOI: https://doi.org/10.1038/nature07331, PMID: 18758442

Ochoa A, Storey JD. 2019. New Kinship and FST Estimates Reveal Higher Levels of Differentiation in the Global Human Population. *bioRxiv*. DOI: https://doi.org/10.1101/653279

Ochoa A, Storey JD. 2021. Estimating FST and kinship for arbitrary population structures. *PLOS Genetics* **17**:e1009241. DOI: https://doi.org/10.1371/journal.pgen.1009241, PMID: 33465078

Ochoa A. 2023. Pca-Assoc-paper. swh:1:rev:8549eafe6c27583894640e6cd8639232ed15cade. Software Heritage. https://archive.softwareheritage.org/swh:1:dir:54f4600c823ac0f1c3b17eb03185aa49a8232d56;origin=https://github.com/OchoaLab/pca-assoc-paper;visit=swh:1:snp:fcc0d7bc50b88ce0b091fd4a89d811fb26f3ddd7;anchor=swh:1:rev:8549eafe6c27583894640e6cd8639232ed15cade

O'Connor LJ, Schoech AP, Hormozdiari F, Gazal S, Patterson N, Price AL. 2019. Extreme Polygenicity of complex traits is explained by negative selection. *American Journal of Human Genetics* **105**:456–476. DOI: https://doi.org/10.1016/j.ajhg.2019.07.003, PMID: 31402091

Paradis E, Schliep K. 2019. Ape 5.0: An environment for modern Phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**:526–528. DOI: https://doi.org/10.1093/bioinformatics/bty633, PMID: 30016406

Park JH, Wacholder S, Gail MH, Peters U, Jacobs KB, Chanock SJ, Chatterjee N. 2010. Estimation of effect size distribution from genome-wide Association studies and implications for future discoveries. *Nature Genetics* **42**:570–575. DOI: https://doi.org/10.1038/ng.610, PMID: 20562874

Park JH, Gail MH, Weinberg CR, Carroll RJ, Chung CC, Wang Z, Chanock SJ, Fraumeni JF, Chatterjee N. 2011. Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *PNAS* **108**:18026–18031. DOI: https://doi.org/10.1073/pnas.1114759108, PMID: 22003128

Patterson N, Price AL, Reich D. 2006. Population structure and Eigenanalysis. *PLOS Genetics* **2**:e190. DOI: https://doi.org/10.1371/journal.pgen.0020190, PMID: 17194218

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. 2012. Ancient admixture in human history. *Genetics* **192**:1065–1093. DOI: https://doi.org/10.1534/genetics.112.145037

Peterson RE, Kuchenbaecker K, Walters RK, Chen C-Y, Popejoy AB, Periyasamy S, Lam M, Iyegbe C, Strawbridge RJ, Brick L, Carey CE, Martin AR, Meyers JL, Su J, Chen J, Edwards AC, Kalungi A, Koen N, Majara L, Schwarz E, et al. 2019. Genome-wide Association studies in Ancestrally diverse populations: Opportunities, methods, pitfalls, and recommendations. *Cell* **179**:589–603. DOI: https://doi.org/10.1016/j.cell.2019.08.051, PMID: 31607513

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide Association studies. *Nature Genetics* **38**:904–909. DOI: https://doi.org/10.1038/ng1847, PMID: 16862161

Price AL, Zaitlen NA, Reich D, Patterson N. 2010. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics* **11**:459–463. DOI: https://doi.org/10.1038/nrg2813, PMID: 20548291

Price AL, Zaitlen NA, Reich D, Patterson N. 2013. Response to Sul and Eskin. *Nature Reviews. Genetics* **14**:300. DOI: https://doi.org/10.1038/nrg2813-c2, PMID: 23438870

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. 2000. Association mapping in structured populations. *American Journal of Human Genetics* **67**:170–181. DOI: https://doi.org/10.1086/302959, PMID: 10827107

Privé F, Luu K, Blum MGB, McGrath JJ, Vilhjálmsson BJ, Schwartz R. 2020. Efficient Toolkit implementing best practices for principal component analysis of population genetic data. *Bioinformatics* **36**:4449–4457. DOI: https://doi.org/10.1093/bioinformatics/btaa520, PMID: 32415959

Qian J, Tanigawa Y, Du W, Aguirre M, Chang C, Tibshirani R, Rivas MA, Hastie T. 2020. A fast and Scalable framework for large-scale and Ultrahigh-dimensional sparse regression with application to the UK Biobank. *PLOS Genetics* **16**:e1009141. DOI: https://doi.org/10.1371/journal.pgen.1009141, PMID: 33095761

Rakitsch B, Lippert C, Stegle O, Borgwardt K. 2013. A lasso multi-marker mixed model for Association mapping with population structure correction. *Bioinformatics* **29**:206–214. DOI: https://doi.org/10.1093/bioinformatics/bts669, PMID: 23175758

Roselli C, Chaffin MD, Weng LC, Aeschbacher S, Ahlberg G, Albert CM, Almgren P, Alonso A, Anderson CD, Aragam KG, Arking DE, Barnard J, Bartz TM, Benjamin EJ, Bihlmeyer NA, Bis JC, Bloom HL, Boerwinkle E, Bottinger EB, Brody JA, et al. 2018. Multi-ethnic genome-wide Association study for atrial fibrillation. *Nature Genetics* **50**:1225–1233. DOI: https://doi.org/10.1038/s41588-018-0133-9, PMID: 29892015

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW. 2002. Genetic structure of human populations. *Science* **298**:2381–2385. DOI: https://doi.org/10.1126/science.1078311, PMID: 12493913

Rosenberg NA, Huang L, Jewett EM, Szpiech ZA, Jankovic I, Boehnke M. 2010. Genome-wide Association studies in diverse populations. *Nature Reviews. Genetics* **11**:356–366. DOI: https://doi.org/10.1038/nrg2760, PMID: 20395969

Schlauch D, Fier H, Lange C. 2017. Identification of genetic Outliers due to sub-structure and cryptic relationships. *Bioinformatics* **33**:1972–1979. DOI: https://doi.org/10.1093/bioinformatics/btx109, PMID: 28334167

Shchur V, Nielsen R. 2018. On the number of siblings and p-th cousins in a large population sample. *Journal of Mathematical Biology* **77**:1279–1298. DOI: https://doi.org/10.1007/s00285-018-1252-8, PMID: 29876645

Simonin-Wilmer I, Orozco-Del-Pino P, Bishop DT, Iles MM, Robles-Espinoza CD. 2021. An overview of strategies for detecting genotype-phenotype associations across Ancestrally diverse populations. *Frontiers in Genetics* **12**:703901. DOI: https://doi.org/10.3389/fgene.2021.703901, PMID: 34804113

Simons YB, Bullaughey K, Hudson RR, Sella G. 2018. A population genetic interpretation of GWAS findings for human quantitative traits. *PLOS Biology* **16**:e2002985. DOI: https://doi.org/10.1371/journal.pbio.2002985, PMID: 29547617

Skoglund P, Posth C, Sirak K, Spriggs M, Valentin F, Bedford S, Clark GR, Reepmeyer C, Petchey F, Fernandes D, Fu Q, Harney E, Lipson M, Mallick S, Novak M, Rohland N, Stewardson K, Abdullah S, Cox MP, Friedlaender FR, et al. 2016. Genomic insights into the peopling of the southwest Pacific. *Nature* **538**:510–513. DOI: https://doi.org/10.1038/nature19844

Sokal RR, Michener CD. 1958. A statistical method for evaluating systematic relationships. *Univ Kansas, Sci Bull* **38**:1409–1438.

Song M, Hao W, Storey JD. 2015. Testing for genetic associations in arbitrarily structured populations. *Nature Genetics* **47**:550–554. DOI: https://doi.org/10.1038/ng.3244, PMID: 25822090

Speed D, Hemani G, Johnson MR, Balding DJ. 2012. Improved Heritability estimation from genome-wide SNPs. *American Journal of Human Genetics* **91**:1011–1021. DOI: https://doi.org/10.1016/j.ajhg.2012.10.010, PMID: 23217325

Storey JD. 2003. The positive false discovery rate: A Bayesian interpretation and the Q-value. *The Annals of Statistics* **31**:2013–2035. DOI: https://doi.org/10.1214/aos/1074290335

Storey JD, Tibshirani R. 2003. Statistical significance for Genomewide studies. *PNAS* **100**:9440–9445. DOI: https://doi.org/10.1073/pnas.1530509100, PMID: 12883005

Sul JH, Eskin E. 2013. Mixed models can correct for population structure for Genomic regions under selection. *Nature Reviews. Genetics* **14**:300. DOI: https://doi.org/10.1038/nrg2813-c1, PMID: 23438871

Sul JH, Martin LS, Eskin E. 2018. Population structure in genetic studies: Confounding factors and mixed models. *PLOS Genetics* **14**:e1007309. DOI: https://doi.org/10.1371/journal.pgen.1007309, PMID: 30589851

Sun G, Zhu C, Kramer MH, Yang SS, Song W, Piepho HP, Yu J. 2010. Variation explained in mixed-model Association mapping. *Heredity* **105**:333–340. DOI: https://doi.org/10.1038/hdy.2010.11, PMID: 20145669

Svishcheva GR, Axenovich TI, Belonogova NM, van Duijn CM, Aulchenko YS. 2012. Rapid variance components–based method for whole-genome Association analysis. *Nature Genetics* **44**:1166–1170. DOI: https://doi.org/10.1038/ng.2410, PMID: 22983301

Thornton T, McPeek MS. 2010. ROADTRIPS: Case-control Association testing with partially or completely unknown population and pedigree structure. *American Journal of Human Genetics* **86**:172–184. DOI: https://doi.org/10.1016/j.ajhg.2010.01.001, PMID: 20137780

Tucker G, Price AL, Berger B. 2014. Improving the power of GWAS and avoiding confounding from population stratification with PC-select. *Genetics* **197**:1045–1049. DOI: https://doi.org/10.1534/genetics.114.164285, PMID: 24788602

Vilhjálmsson BJ, Nordborg M. 2013. The nature of confounding in genome-wide Association studies. *Nature Reviews. Genetics* **14**:1–2. DOI: https://doi.org/10.1038/nrg3382, PMID: 23165185

Voight BF, Pritchard JK. 2005. Confounding from cryptic relatedness in case-control Association studies. *PLOS Genetics* **1**:e32. DOI: https://doi.org/10.1371/journal.pgen.0010032, PMID: 16151517

Wang H, Aragam B, Xing EP. 2022. Trade-offs of linear mixed models in genome-wide Association studies. *Journal of Computational Biology* **29**:233–242. DOI: https://doi.org/10.1089/cmb.2021.0157, PMID: 35230156

Wojcik GL, Graff M, Nishimura KK, Tao R, Haessler J, Gignoux CR, Highland HM, Patel YM, Sorokin EP, Avery CL, Belbin GM, Bien SA, Cheng I, Cullina S, Hodonsky CJ, Hu Y, Huckins LM, Jeff J, Justice AE, Kocarnik JM, et al. 2019. Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**:514–518. DOI: https://doi.org/10.1038/s41586-019-1310-4, PMID: 31217584

Wright S. 1949. The Genetical structure of populations. *Annals of Eugenics* **15**:323–354. DOI: https://doi.org/10.1111/j.1469-1809.1949.tb02451.x

Wu C, DeWan A, Hoh J, Wang Z. 2011. A comparison of Association methods correcting for population stratification in case-control studies. *Annals of Human Genetics* **75**:418–427. DOI: https://doi.org/10.1111/j.1469-1809.2010.00639.x, PMID: 21281271

Xu H, Guan Y. 2014. Detecting local haplotype sharing and haplotype Association. *Genetics* **197**:823–838. DOI: https://doi.org/10.1534/genetics.114.164814, PMID: 24812308

Yang J, Lee SH, Goddard ME, Visscher PM. 2011. GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics* **88**:76–82. DOI: https://doi.org/10.1016/j.ajhg.2010.11.011

Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. 2014. Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics* **46**:100–106. DOI: https://doi.org/10.1038/ng.2876

Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES. 2006. A unified mixed-model method for Association mapping that accounts for multiple levels of relatedness. *Nature Genetics* **38**:203–208. DOI: https://doi.org/10.1038/ng1702, PMID: 16380716

Zaidi AA, Mathieson I. 2020. Demographic history mediates the effect of stratification on Polygenic scores. *eLife* **9**:e61548. DOI: https://doi.org/10.7554/eLife.61548, PMID: 33200985

Zeng J, de Vlaming R, Wu Y, Robinson MR, Lloyd-Jones LR, Yengo L, Yap CX, Xue A, Sidorenko J, McRae AF, Powell JE, Montgomery GW, Metspalu A, Esko T, Gibson G, Wray NR, Visscher PM, Yang J. 2018. Signatures of negative selection in the genetic architecture of human complex traits. *Nature Genetics* **50**:746–753. DOI: https://doi.org/10.1038/s41588-018-0101-4, PMID: 29662166

Zhang S, Zhu X, Zhao H. 2003. On a Semiparametric test to detect associations between quantitative traits and candidate genes using unrelated individuals. *Genetic Epidemiology* **24**:44–56. DOI: https://doi.org/10.1002/gepi.10196, PMID: 12508255

Zhang Z, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordovas JM, Buckler ES. 2010. Mixed linear model approach adapted for genome-wide Association studies. *Nature Genetics* **42**:355–360. DOI: https://doi.org/10.1038/ng.546, PMID: 20208535

Zhang Y, Pan W. 2015. Principal component regression and linear mixed model in association analysis of structured samples: Competitors or complements. *Genetic Epidemiology* **39**:149–155. DOI: https://doi.org/10.1002/gepi.21879, PMID: 25536929

Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, Tang C, Toomajian C, Zheng H, Dean C, Marjoram P, Nordborg M. 2007. An Arabidopsis example of Association mapping in structured samples. *PLOS Genetics* **3**:e4. DOI: https://doi.org/10.1371/journal.pgen.0030004, PMID: 17238287

Zheng X, Weir BS. 2016. Eigenanalysis of SNP data with an identity by descent interpretation. *Theoretical Population Biology* **107**:65–76. DOI: https://doi.org/10.1016/j.tpb.2015.09.004, PMID: 26482676

Zhong Y, Perera MA, Gamazon ER. 2019. On using local ancestry to characterize the genetic architecture of human traits: Genetic regulation of gene expression in Multiethnic or admixed populations. *American Journal of Human Genetics* **104**:1097–1115. DOI: https://doi.org/10.1016/j.ajhg.2019.04.009, PMID: 31104770

Zhou X, Stephens M. 2012. Genome-Wide efficient mixed-model analysis for association studies. *Nature Genetics* **44**:821–824. DOI: https://doi.org/10.1038/ng.2310, PMID: 22706312

Zhou Q, Zhao L, Guan Y, Akey JM. 2016. Strong selection at MHC in Mexicans since admixture. *PLOS Genetics* **12**:e1005847. DOI: https://doi.org/10.1371/journal.pgen.1005847

Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, Wolford BN, LeFaive J, VandeHaar P, Gagliano SA, Gifford A, Bastarache LA, Wei WQ, Denny JC, Lin M, Hveem K, Kang HM, Abecasis GR, Willer CJ, Lee S. 2018. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics* **50**:1335–1341. DOI: https://doi.org/10.1038/s41588-018-0184-y, PMID: 30104761

Zhu C, Yu J. 2009. Nonmetric multidimensional Scaling corrects for population structure in association mapping with different sample types. *Genetics* **182**:875–888. DOI: https://doi.org/10.1534/genetics.108.098863, PMID: 19414565

Appendix 1

Fitting ancestral allele frequency distribution to real data

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

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

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

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

Since $V^T \leq \frac{1}{4}$ and $\bar{\varphi}^T \geq 0$, then $W^T \geq V^T$. Thus, the goal is to construct a new distribution with the original, lower variance of

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

We use the unbiased estimator $\hat{W}^T = \frac{1}{m}\sum_{i=1}^m \left(\hat{p}_i^T - \frac{1}{2}\right)^2$, while $\bar{\varphi}^T$ is calculated from the tree parameters: the subpopulation coancestry matrix (***Equation 7***), expanded from subpopulations to individuals, the diagonal converted to kinship (reversing ***Equation 5***), and the matrix averaged. However, since our model ignores the MAF filters imposed in our simulations, $\bar{\varphi}^T$ was adjusted. For Human Origins the true model $\bar{\varphi}^T$ of 0.143 was used. For 1000 Genomes and HGDP the true $\bar{\varphi}^T$ are 0.126 and 0.124, respectively, but 0.4 for both produced a better fit.

Lastly, we construct new allele frequencies,

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

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

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

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

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

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

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

Appendix 2

Comparisons between SRMSD_p, $\mathrm{AUC}_{\mathrm{PR}}$, and evaluation measures from the literature

2.1 The inflation factor λ

Test statistic inflation has been used to measure model calibration (***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***). To compare p-values from non-χ^2 tests (such as t-statistics), λ can be calculated from p-values using

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

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

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 $\mathrm{SRMSD}_p = 0$. However, non-uniform p-values with the expected median, such as from genomic control (***Devlin and Roeder, 1999***), result in $\lambda = 1$, but $\mathrm{SRMSD}_p \neq 0$ except for uniform p-values, a key flaw of λ that SRMSD_p overcomes. Inflated statistics (anti-conservative p-values) give $\lambda > 1$ and $\mathrm{SRMSD}_p > 0$. Deflated statistics (conservative p-values) give $\lambda < 1$ and $\mathrm{SRMSD}_p < 0$. Thus, $\lambda \neq 1$ always implies $\mathrm{SRMSD}_p \neq 0$ (where $\lambda - 1$ and SRMSD_p have the same sign), but not the other way around. Overall, λ depends only on the median p-value, while SRMSD_p uses the complete distribution. However, SRMSD_p requires knowing which loci are null, so unlike λ it is only applicable to simulated traits.

2.2 Empirical comparison of SRMSD_p and λ

There is a near one-to-one correspondence between λ and SRMSD_p in our data (***Figure 2—figure supplement 1***). PCA tended to be inflated ($\lambda > 1$ and $\mathrm{SRMSD}_p > 0$) whereas LMM tended to be deflated ($\lambda < 1$ and $\mathrm{SRMSD}_p < 0$), otherwise the data for both models fall on the same contiguous curve. We fit a sigmoidal function to this data,

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

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

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

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

2.3 Type I error rate

The type I error rate is the proportion of null p-values with $p \leq t$. Calibrated p-values have type I error rate near t, which may be evaluated with a binomial test. This measure may give different results for different t, for example be significantly miscalibrated only for large t (due to lack of power for smaller t), and it requires large simulations to estimate well as it depends on the tail of the distribution. In contrast, SRMSD_p uses the entire distribution so it is easier to estimate, $\mathrm{SRMSD}_p = 0$ guarantees calibrated type I error rates at all t, while large $|\mathrm{SRMSD}_p|$ indicates incorrect type I errors for a range of t. Empirically, we find the expected agreement and monotonic relationship between SRMSD_p and type I error rate (***Figure 2—figure supplement 2***).

2.4 Statistical power and comparison to $\mathrm{AUC}_{\mathrm{PR}}$

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

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

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

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

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

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

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

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

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

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

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

Empirically, we find the predicted positive correlation between $\mathrm{AUC}_{\mathrm{PR}}$ and calibrated power (***Figure 2—figure supplement 3***). The correlation is clear when considered separately per dataset, but the slope varies per dataset, which is expected because the proportion of alternative cases π_1 varies per dataset.