Estimation of inbreeding and kinship coefficients via latent identity-by-descent states (supplementary)

Yongtao Guan and Daniel Levy
National Heart, Lung, and Blood Institute

September 12, 2023

1 Optimal bin size.

We examined the deviation from the true kinship, separately for African, East Asian, and European samples. For each population, we simulated 124 pairs of first cousins, and plotted the mean deviation between inferred kinship and the true kinship for different bin sizes (from 0 to 0.09). The optimal bins curve are consistent between continental populations with the optimal bin size 0.04 (Figure S1).

2 Simulate kinship.

We chose ped-sim `https://github.com/williamslab/ped-sim` to simulate different degree of kinships (Caballero et al., 2019). Ped-sim requires phased haplotypes to mimic crossover events, 1000 genomes dataset have high quality phased haplotypes. We separately analyzed each populations and examined their individual inbreeding and pairwise kinship coefficients. Figure S2 suggests CHB, CHS, CEU, and YRI have no excessive inbreeding and kinship. We used CHB and CHS samples as founders to simulate non-admixed pedigrees, and CHB, CHS, CEU, and YRI samples as founders to simulate admixed pedigrees. After selecting samples, we selected bi-allelic common SNPs (minor allele frequencies greater than 0.05) and required neighboring SNPs to be at least 1000 bp apart. The command line we used was:

```
ped-sim -d first-cousin.def -m refined_mf.simmap -i \
     1000g.subset.vcf.gz  -o first-cousin --pois --nogz
```

We simulated pedigrees with different degree of relatedness including parent-child, full siblings, grandparents, first cousins, second cousins, and third cousins. Ped-sim chose founders by sampling without replacement, this limited the number of samples we can simulate in each run. For example, a pair of first cousins requires four founders. When we simulate with Han Chinese samples (CHB

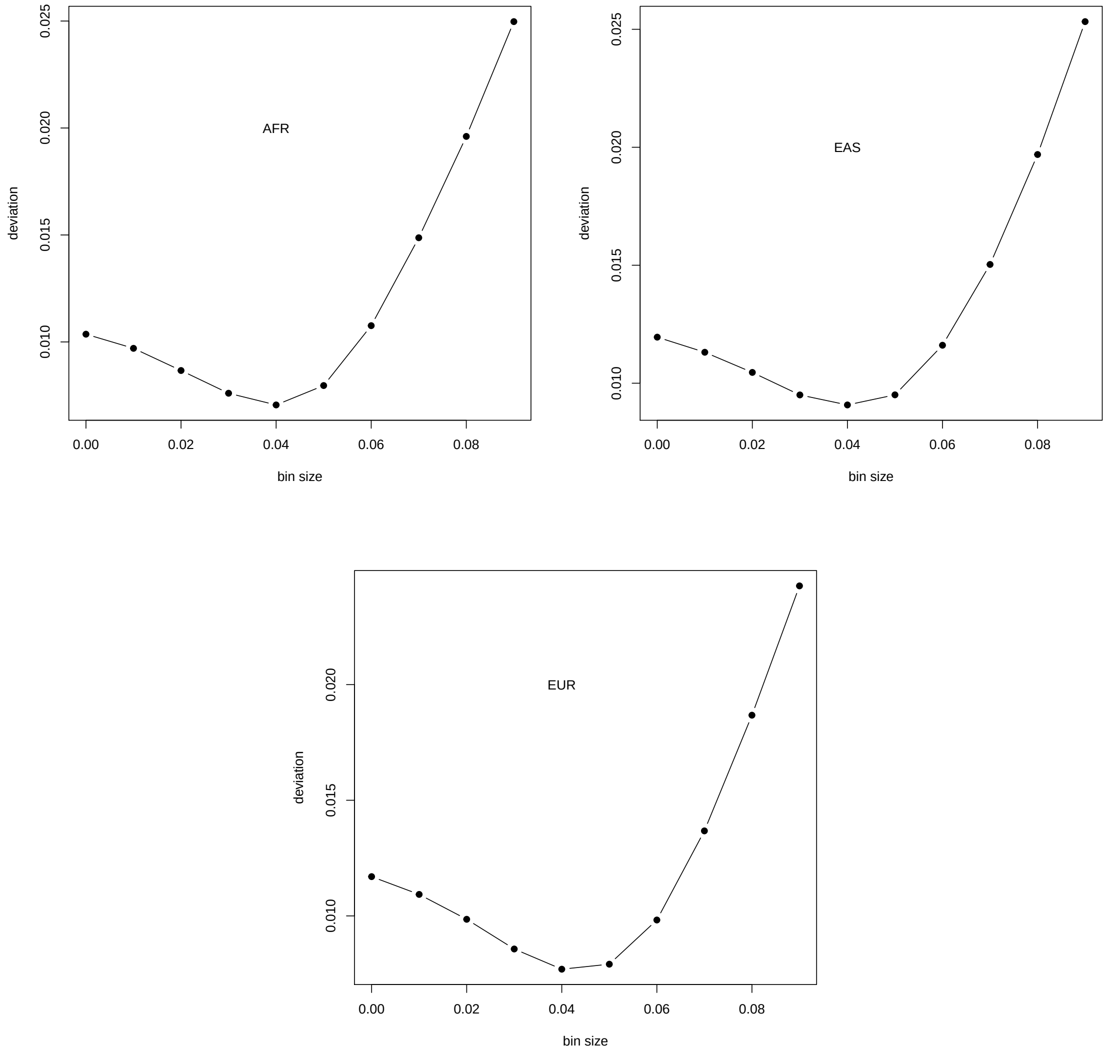


Figure S1: Optimal bin size. The x-axis is the bin size, and the y-axis is mean deviations between inferred kinship and simulated truth (over 124 pairs). The simulations were done using African, East Asian, and European samples separately for each panel.

and CHS), we have 208 samples (103 for CHB and 105 for CHS) as candidate founders, and this put a limit on how many pairs we can simulate in one run. For example, each pair of first cousin simulation requires 4 founders, and each pair of third cousin simulation requires 8 founders. We therefore run three sets of simulations for each pedigree. To determine the optimal bin sizes, we simulated 124 pairs of first cousins, each for one population, using 661 African samples, 504 East Asian samples, and 503 European samples.

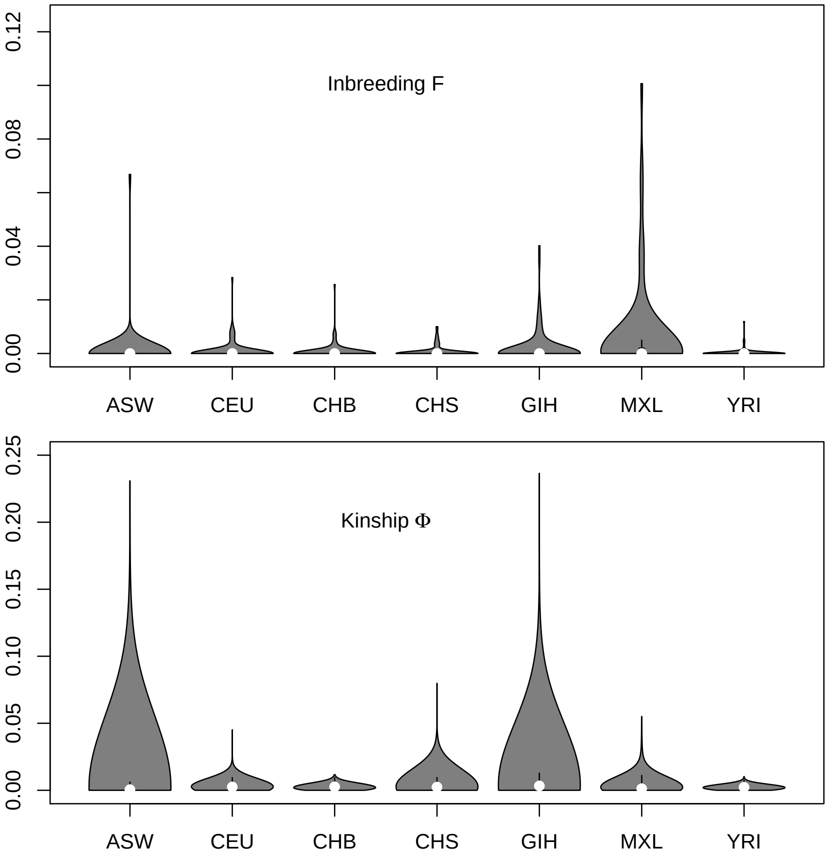


Figure S2: Violin plots for inbreeding (top panel) and kinship (bottom panel) coefficients of seven populations in 1000 genomes project, calculated with Kindred using sample allele frequencies.

3 Detailed comparison between different methods

We include a table detailing the deviations (mean and standard error) between inferred kinship and expected from pedigree, as a companion to Figure 1 in the main text (Table S1). We also included a table detailing the deviations (mean and standard error) between inferred kinship and true kinship (Table S2).

	scGRM	popkin	UKin	King	Kindred
Parent-child	0.199 ± 0.004	0.26 ± 0.003	0.249 ± 0.002	0.248 ± 0.002	0.248 ± 0.004
Full siblings	0.191 ± 0.018	0.259 ± 0.023	0.249 ± 0.023	0.248 ± 0.023	0.248 ± 0.023
Grandparent	0.101 ± 0.017	0.140 ± 0.020	0.126 ± 0.020	0.124 ± 0.021	0.124 ± 0.020
1st cousin	0.046 ± 0.011	0.075 ± 0.013	0.069 ± 0.013	0.060 ± 0.014	0.062 ± 0.013
2nd cousin	0.005 ± 0.006	0.030 ± 0.008	0.016 ± 0.008	0.013 ± 0.008	0.016 ± 0.008
3rd cousin	−0.008 ± 0.003	0.016 ± 0.005	0.005 ± 0.004	0.002 ± 0.005	0.004 ± 0.004
Negative percent	99.5%	0	41.5%	64.4%	0%

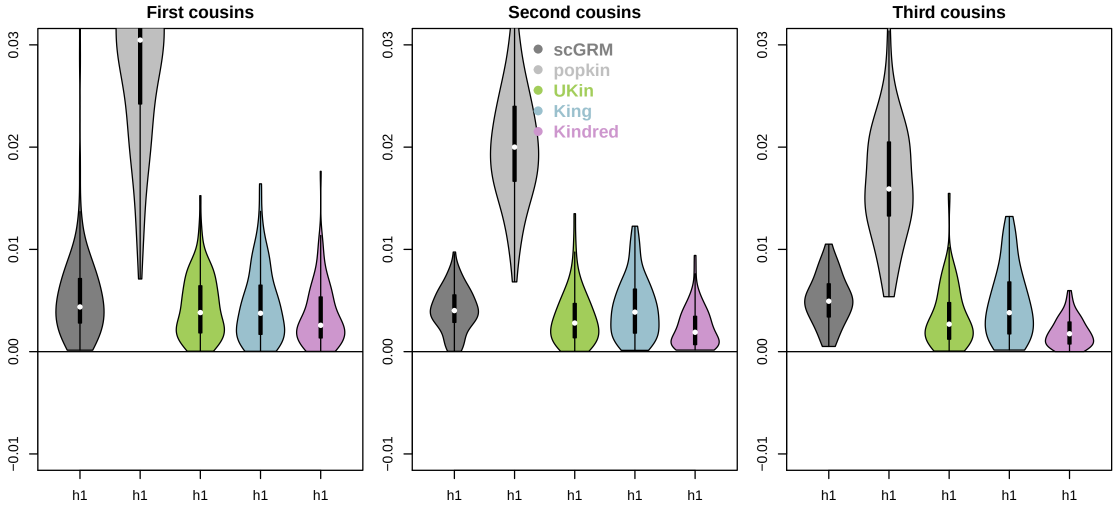
Table S1: Comparison of different methods of inferred kinship vs the expected kinship of the pedigree. The last row is the percent of the negative estimates for different methods in all simulations. The expected kinships for six simulated kinships are 0.250, 0.250, 0.125, 0.063, 0.0156, and 0.004.

	scGRM	popkin	UKin	King	Kindred
Parent-child	0.051 ± 0.004	0.014 ± 0.004	0.002 ± 0.001	0.003 ± 0.002	0.003 ± 0.003
Full siblings	0.059 ± 0.007	0.013 ± 0.004	0.005 ± 0.004	0.006 ± 0.004	0.005 ± 0.004
Grandparent	0.024 ± 0.005	0.017 ± 0.005	0.004 ± 0.003	0.005 ± 0.004	0.005 ± 0.004
1st cousin	0.017 ± 0.004	0.013 ± 0.004	0.004 ± 0.003	0.005 ± 0.004	0.004 ± 0.003
2nd cousin	0.011 ± 0.003	0.015 ± 0.004	0.003 ± 0.003	0.005 ± 0.003	0.003 ± 0.003
3rd cousin	0.012 ± 0.002	0.012 ± 0.004	0.003 ± 0.002	0.004 ± 0.003	0.002 ± 0.002

Table S2: Comparison of different methods of deviation between inferred kinship and the simulated truths. The results agreed well with Table S1.

4 Founders from different continental populations

In Figure 1 of the main text, the founders are Han Chinese samples (CHB+CHS) in the 1000G. Here as a supplement to Figure 1, we extended the simulation to diverse populations, and used founders from three continental populations (African, East Asian, and European). Figures S3 S4 S5 showed that Kindred consistently outperform other competing methods in all three continental populations. Here we plot the deviation between inferred and the truth (instead of expectation from the pedigree). Results of scGRM (dark gray) were obtained with PLINK, popkin (light gray) with its R package, UKin (green) with re-implemention in Kindred software, and King (blue) with its software. Results of Kindred is in plum. and Kindred (plum).



		First cousin	Second cousin	Third cousin
AFR	scGRM	0.005 ± 0.005	0.004 ± 0.002	0.005 ± 0.002
	popkin	0.030 ± 0.008	0.020 ± 0.006	0.016 ± 0.006
	UKin	0.004 ± 0.003	0.003 ± 0.003	0.003 ± 0.003
	King	0.004 ± 0.004	0.004 ± 0.003	0.005 ± 0.003
	Kindred	0.004 ± 0.003	0.002 ± 0.002	0.002 ± 0.001

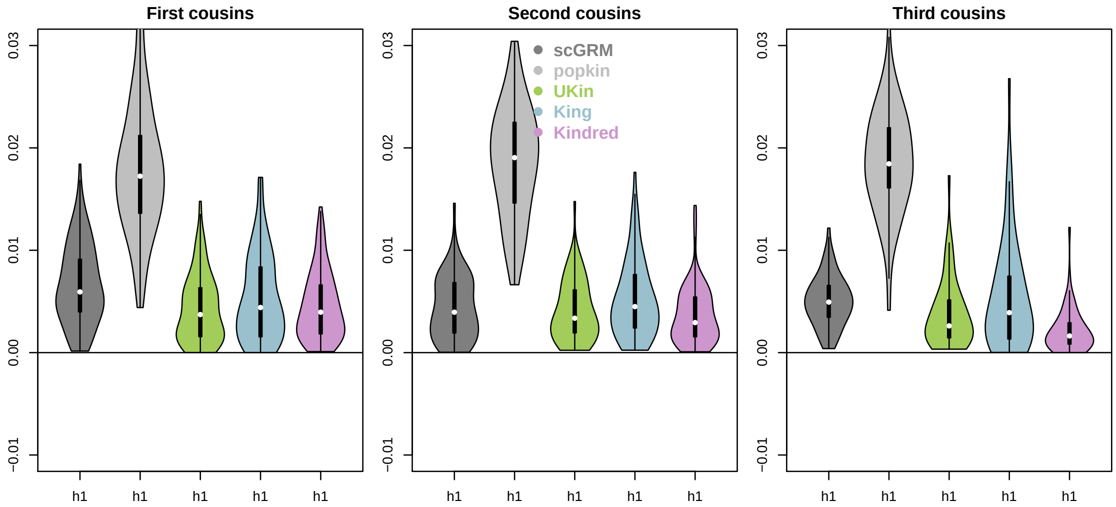
Figure S3: Comparison of kinship estimates simulated from African samples from 1000 Genomes project. Plots show deviation between inferred and the truth. Table shows mean ± sd for each violin plot.

5 SNPs of small population divergence.

We prepared an annotation file that contains a FST tag for each SNP (defined below). To calculate FST tag, we first obtained all bi-allelic SNPs with minimum count of minor alleles 50 (among 2504 samples). For each SNP we counted AN-AC for different populations. Below is an example to process SNPs on chr22 of CHB population.

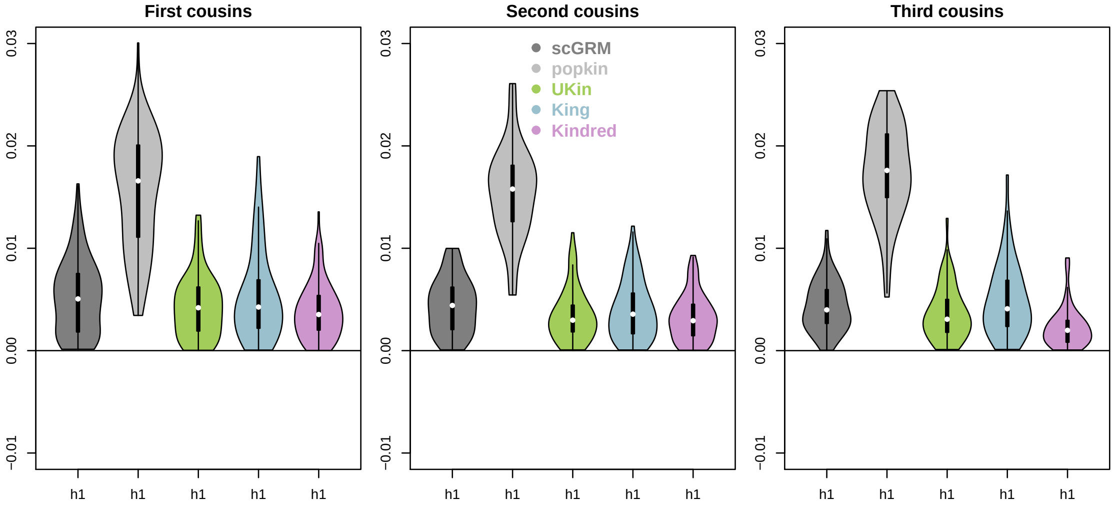
```
bcftools view -m2 -M2 -v snps -c 50:minor ALL.chr22.phase3.genotypes.vcf.gz | \
    bcftools view -S samples.CHB.txt | bcftools annotate --remove INFO |\
    bcftools plugin fill-AN-AC | \
    bcftools query -f "%CHROM %POS %REF %ALT %INFO/AC %INFO/AN\n" > chb.chr22.an-ac
```

After we obtained for each chromosome the AN-AC files for populations of CEU, YRI, and CHB, We did a chisq test for each SNP to obtain a p-values, with null hypothesis as equal allele frequency



		First cousin	Second cousin	Third cousin
EAS	scGRM	0.007 ± 0.004	0.004 ± 0.003	0.005 ± 0.002
	popkin	0.018 ± 0.006	0.019 ± 0.005	0.019 ± 0.005
	UKin	0.004 ± 0.003	0.004 ± 0.003	0.004 ± 0.003
	King	0.005 ± 0.004	0.005 ± 0.004	0.006 ± 0.005
	Kindred	0.005 ± 0.003	0.004 ± 0.003	0.002 ± 0.002

Figure S4: Comparison of kinship estimates simulated from East Asian samples from 1000 Genomes project. Plots show deviation between inferred and the truth. Table shows mean ± sd for each violin plot.



		First cousin	Second cousin	Third cousin
EUR	scGRM	0.005 ± 0.004	0.004 ± 0.003	0.004 ± 0.002
	popkin	0.016 ± 0.006	0.015 ± 0.004	0.017 ± 0.005
	UKin	0.004 ± 0.003	0.003 ± 0.003	0.004 ± 0.003
	King	0.005 ± 0.004	0.004 ± 0.003	0.005 ± 0.003
	Kindred	0.004 ± 0.003	0.003 ± 0.002	0.002 ± 0.002

Figure S5: Comparison of kinship estimates simulated from European samples from 1000 Genomes project. Plots show deviation between inferred and the truth. Table shows mean ± sd for each violin plot.

across three populations. Lastly, we write a vcf file with allele frequencies and an additional tag INFO/FST = -10log10(p-value). SNPs whose $FST < 3$ appears to be working well for admixed samples, and there are about 1.2 million such bi-allelic SNPs in 1000 genomes.

6 Simulate phenotypes with a prescribed heritability.

We randomly chose m causal SNPs to obtain a matrix X of $n \times m$ where n is the sample size, simulated m effect sizes by sampling from standard normal distribution, independently for each SNP to obtain a m-vector β, and obtain $y = X\beta$. We then draw an n-vector of independent and identically distributed (with standard normal distribution) ϵ, and renormalized to have variance 1. For a given heritability h, we computed $v = h/(1-h)$ and renormalize y to y' that has variance v. The $z = y' + \epsilon$ is the phenotype with heritability h. In this paper, we used $h = 0.30$ and $m = 100$ and 10000.

7 Comparisons between with and without inbreeding

By setting $\Delta_j = 0$ for $j = 1, \cdots, 6$, we can estimates kinship without modeling inbreeding. Below are the comparison using the same datasets that produced Figure 1 in the main text. Without modeling inbreeding, kinship estimates were biased upward (Figure S6). Note in the simulation, the inbreeding of the founders are small, so that the observed biases are small.

8 Clustering on Chr17

The plot shows that the clusters on PC2 vs PC3 plot is independent to population structure, and in each population there appear to be three distinct clusters on PC2 vs PC3 plot.

9 Bayesian PVE estimates are reliable

We simulated two sets 100 phenotypes of nominal heritability 0.30. One set with 100 causal SNPs, the other with 10000 causal SNPs. The simulation was done using the genotypes from the height dataset. We obtained both GCTA estimates and our own Bayes MAP estimates and compared them in Figure S8. Both plots showed consistency between GCTA results and Bayesian MAP estimates. So that the difference observed in the resampling study was not due to different method, but different kinship estimates.

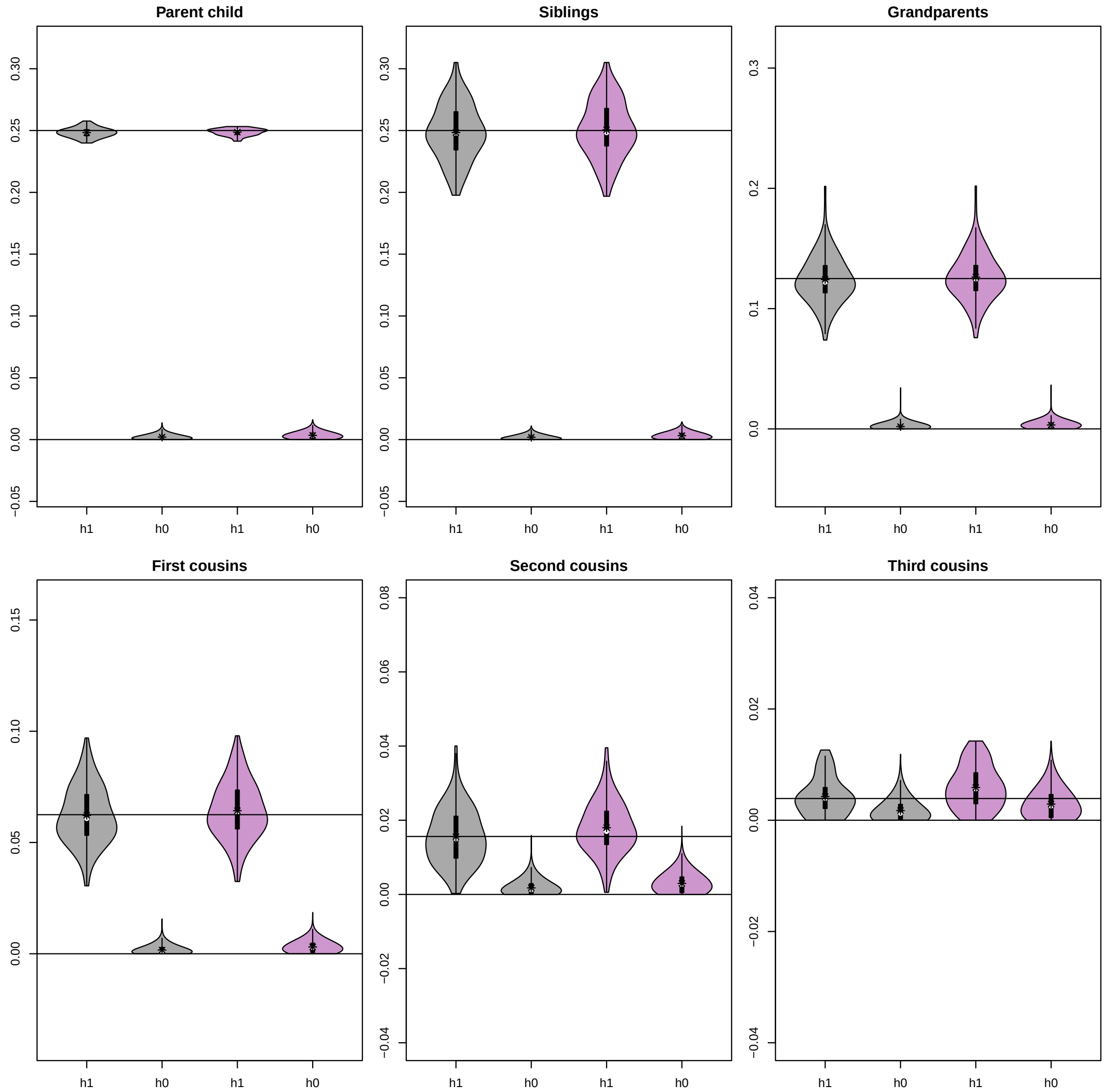


Figure S6: Not modeling inbreeding bias kinship estimates. Gray colors are results from using all nine latent IBD states (with inbreeding); Plum colors are results from using the last three latent IBD states (without inbreeding). Due to the scale of the plot on y-axis, the biases are more visible between more distant relatedness.

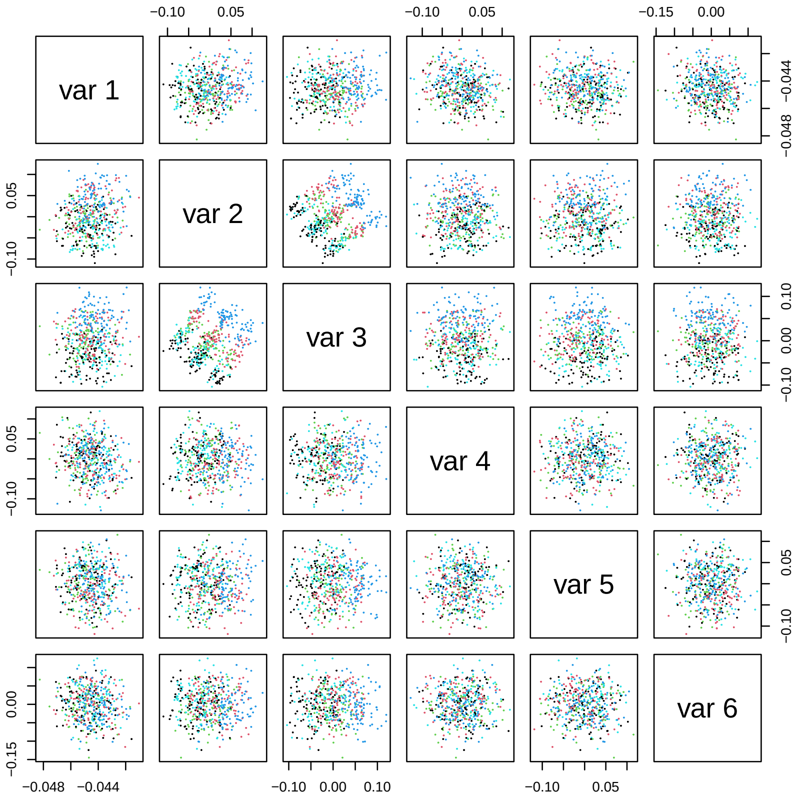


Figure S7: Pairwise PC plots for East Asians on Chromosome 17. Black: CDX (Chinese Dai in Xishuangbanna, China); Red: CHB (Han Chinese in Beijing, China): Green: CHS (Han Chinese South); Blue: JPT (Japanese in Tokyo); Cyan: KHV (Kinh in Ho Chi Minh City, Vietnam).

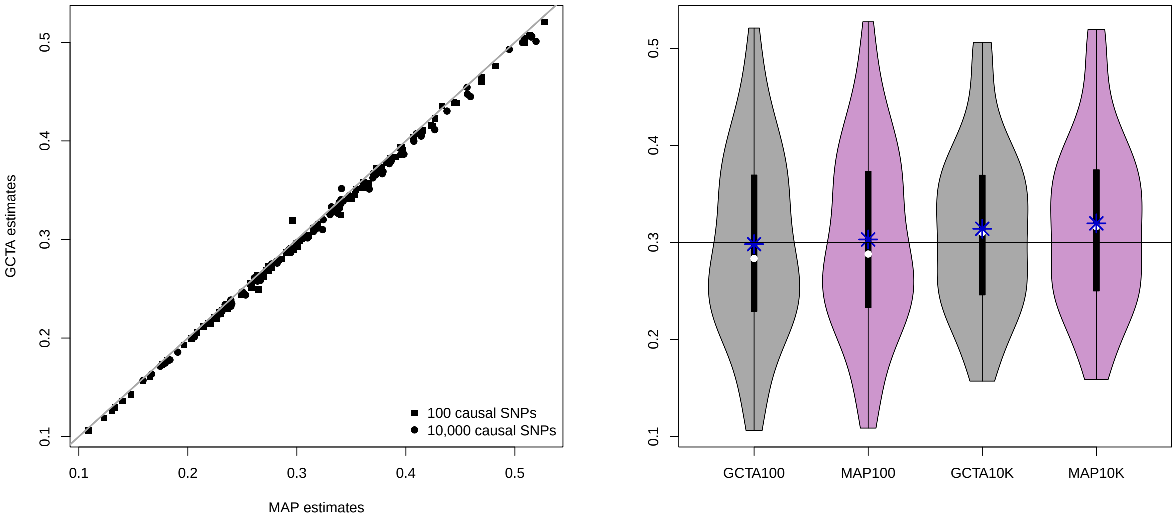


Figure S8: Simulations study demonstrate Bayesian MAP estimates are consistent with GCTA estimates. The left panel compared MAP estimates and GCTA estimates for each simulated phenotypes. The right panels compared the overall distribution between MAP estimates and GCTA estimates.

10 Kindred run time

The run time is wall time and the experiments were conducted on a machine with 256G RAM, 112 Core, AMD EPYC 7B13 CPU at 3049.998 MHz, running Linux Ubuntu 22.04. The Kindred version is 0.81.

11 Latent state model to explain scGRM.

Let X and Y denote allele counts of genotypes $G1$ and $G2$ in Table 1 of main text, p is allele frequency of A allele, and consider

$$\frac{(X-Y)^2}{2p(1-p)} = \frac{(X-2p)^2}{2p(1-p)} + \frac{(Y-2p)^2}{2p(1-p)} - 2\frac{(X-2p)(Y-2p)}{2p(1-p)}, \tag{1}$$

where $Q = \frac{(X-2p)(Y-2p)}{2p(1-p)}$ is the quantity calculated in scGRM.

Let's reproduce Table 1 with an extra column $D = (G1 - G2)^2$ (Figure S4). Then the expectation of the left hand side $E(LHS) = \frac{1}{2pq}\left(\Sigma_1^t D \cdot \Delta_1 + \cdots + \Sigma_9^t D \cdot \Delta_9\right)$. Direct compute each term to get $E(LHS) = 4\Delta_2 + \Delta_3 + 3\Delta_4 + \Delta_5 + 3\Delta_6 + \Delta_8 + 2\Delta_9$. The genotype distribution of $G1$ is $P(X = 2) = p^2 + pqF_1$, $P(X = 0) = q^2 + pqF_1$, and $P(X = 1) = 2pq(1-F_1)$, and the expectation of $(X-2p)^2$ is $(2-2p)^2(p^2+pqF_1)+$

		Number of threads		
Num Samples	Num SNPs	16	32	64
1439	0.5M	1.3	0.7	0.6
	1M	2.3	1.2	0.8
	2M	4.5	2.3	1.4
2878	0.5M	5.3	2.7	1.7
	1M	9.3	4.7	2.8
	2M	17.6	8.9	5.2
5757	0.5M	20.6	10.8	6.4
	1M	36.9	18.8	11.1
	2M	71.0	36.0	20.6

Table S3: Wall time (minutes) for Kindred model fitting for different number of samples, number of SNPs (in million), and number of threads used for computation. Note the complexity of Kindred is quadratic in number of samples, linear in number of SNPs, and inverse linear to number of threads used.

Σ_1	Σ_2	Σ_3	Σ_4	Σ_5	Σ_6	Σ_7	Σ_8	Σ_9	G1	G2	D
p	p^2	p^2	p^3	p^2	p^3	p^2	p^3	p^4	AA	AA	0
0	0	pq	$2p^2q$	0	0	0	p^2q	$2p^3q$	AA	AB	1
0	pq	0	pq^2	0	p^2q	0	0	p^2q^2	AA	BB	4
0	0	0	0	pq	$2p^2q$	0	p^2q	$2p^3q$	AB	AA	1
0	0	0	0	0	0	$2pq$	pq	$4p^2q^2$	AB	AB	0
0	0	0	0	pq	$2pq^2$	0	pq^2	$2pq^3$	AB	BB	1
0	pq	0	p^2q	0	pq^2	0	0	p^2q^2	BB	AA	4
0	0	pq	$2pq^2$	0	0	0	pq^2	$2pq^3$	BB	AB	1
q	q^2	q^2	q^3	q^2	q^3	q^2	q^3	q^4	BB	BB	0

Table S4: Reproduction of Table 1 in main text with an extra column $D = (G1 - G2)^2$.

$(1-2p)^2 2pq(1-F_1)+(0-2p)^2(q^2+pqF_1)$ and direct computation to get $2pq(1+F_1)$, therefore the expectation of the first term on the right hand side is $(1+F_1)$. The calculation on $G2$ gets the expectation of the second term on the right hand side as $(1+F_2)$. Thus the expectation of RHS is $E(RHS) = 1 + F_1 + 1 + F_2 - 2Q$.

12 R code to evaluate Bayes factor for a linear mixed model.

```
library(MASS)
logBF = function(y,W, Q,D,lam){
   #W contains a column of 1 and has full rank.
   #D is a vector of eigenvalues;
    n=length(y)
    y=cbind(y);
    X = cbind(Q, W)
    yty = sum(y*y);
    ytW = t(y) %*% W;
    wtw = t(W) %*% W;
    var0=ytW %*% ginv(wtw) %*% t(ytW)
    #wtw have low dimension.
    ytX = t(y) %*% X;
    lD=lam*D;
    mF=diag(lD/(1+lD));
    QtW = t(Q) %*% W;    #QtW;
    M = ginv(wtw - t(QtW) %*% mF %*% QtW);
    fQw = mF%*%QtW;
    a11=mF+ fQw %*% M %*% t(fQw);
    a12=-fQw %*% M;
    a21=-M %*% t(fQw);
    a22=M;
    aa=rbind(cbind(a11,a12), cbind(a21,a22));
    res1= 0.5*log10(det(wtw))-0.5*sum(log10(1+lD))+0.5*log10(det(M));
    f1=log10(yty - ytX %*% aa %*% t(ytX));
    f2=log10(yty - var0)
    return(res1 - (n/2)*(f1-f2));
}
```

References

Caballero, M., D. N. Seidman, Y. Qiao, J. Sannerud, T. D. Dyer, D. M. Lehman, J. E. Curran, R. Duggirala, J. Blangero, S. Carmi, and A. L. Williams (2019, 12). Crossover interference and sex-specific genetic maps shape identical by descent sharing in close relatives. *PLOS Genetics 15*(12), 1–29.