Estimation of inbreeding and kinship coefficients via latent identity-by-descent states [1]

Yongtao Guan

Framingham Heart Study
National Heart, Lung, and Blood Institute

present at
Department of Biostatistics, Boston University

April 21, 2023

[1] This is joint work with Dan Levy.

Kinship and inbreeding coefficients

- Kinship (denoted by ϕ) is the probability that two alleles sampled each from two individuals are identical by descent (IBD).
- Kinship between two individuals is the inbreeding coefficient (denoted by F) of their (hypothetical) children.
- Between one and oneself (or between monozygotic twins) $\phi = (1 + F)/2$.
- Inbreeding can be treated as a derived concept of kinship.

Identity by descent (IBD)

The definitions of inbreeding and kinship hinge on IBD, while IBD is defined relative to a reference population, where different alleles in that reference population are considered *not* IBD [2].

[2] Wang, J. (2016). Theor. Popul. Biol. 107, 4–13

Existing methods to infer kinship

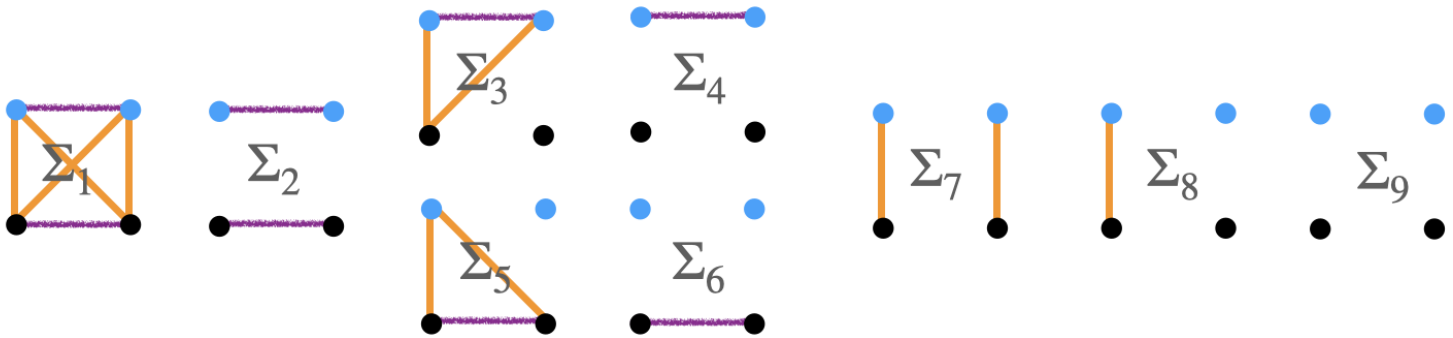
- Sample correlation based genetic relatedness matrix (scGRM).
- UKin, aims to correct bias observed in scGRM methods [3].
- King, based on counts of a subset of joint genotypes [4]

All three produced substantial amount of negative estimates for kinship, difficult to interpret.

[3] Jiang, W., et. al, and H. Zhao (2022). BMC Bioinformatics 23(1), 525

[4] Manichaikul, A, and et. al, and W.-M. Chen (2010). Bioinformatics 26(22), 2867–2873.

Jacquard IBD states



Kinship can be computed from the loading probabilities, Δ_j for j-th latent IBD state Σ_j, as follows [5]

$$\phi = \Delta_1 + \frac{1}{2}(\Delta_3 + \Delta_5 + \Delta_7) + \frac{1}{4}\Delta_8. \tag{1}$$

Inbreeding coefficients can also be computed:

$$\begin{aligned} F_1 &= \Delta_1 + \Delta_2 + \Delta_3 + \Delta_4, \\ F_2 &= \Delta_1 + \Delta_2 + \Delta_5 + \Delta_6. \end{aligned} \tag{2}$$

[5] Jacquard, A. (1972). Biometrics 28(4), 1101–1114.

Jacquard IBD states

Each latent IBD states emit joint genotypes at a probability distribution that is a function of allele frequency p. [6]

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

where $q = 1 - p$. Our aim is to infer Δ_j, the loading probabilities of Σ_j.

[6] Thompson, E. A. (2013). Genetics 194(2), 301–326.

Fit the model

We consider SNPs with allele frequency p so that they share the same Σ matrix. Denote $\hat{\theta}$ estimates of fractions of joint genotypes (AA AA, AA AB, ... etc).

$$\arg\min_{\Delta} \; ||\mathbf{S}_p\Delta - \hat{\theta}_p||_2 \tag{3a}$$

$$\text{s.t} \quad \Delta_j \geq 0 \;\text{ for all } j, \;\text{ and } \sum \Delta_j = 1 \tag{3b}$$

where $\mathbf{S}_p = (\Sigma_1, \ldots, \Sigma_9)$, $\Delta = (\Delta_1, \ldots, \Delta_9)$ is the vector of loading probabilities.

Fit the model

For the i-th SNP with allele frequency p_i, we can compute $\mathbf{S}_{p_i}$ and we observe $\hat{\theta} = e_i$, where e_i has a single entry 1 and the rest 8 entries 0.

$$\arg\min_{\Delta} \ ||\mathbf{S}\Delta - \hat{\theta}||_2 \tag{4}$$

with constraints (3b). The above system generalizes (3) to SNPs of arbitrary allele frequencies. In practice, we find binning SNPs and re-estimate allele frequencies in each bin improves performance.

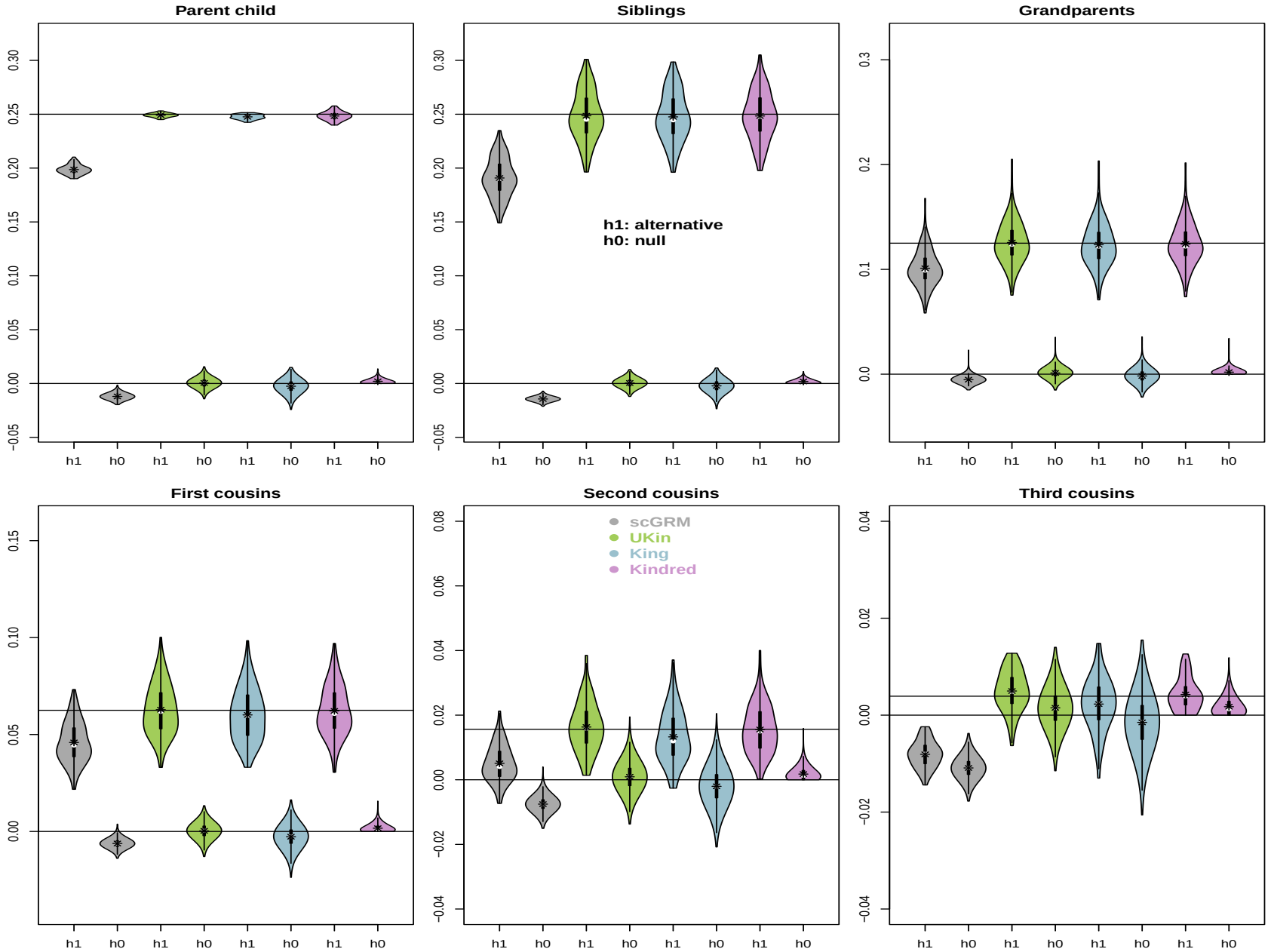
Fit the model: invariant

It can be verified that there are two linear dependence in $\mathbf{S}_p$. One is $\Sigma_2 + 2\Sigma_8 = \Sigma_4 + \Sigma_6 + \Sigma_7$ and the other is $pq(\Sigma_1 + \Sigma_2 - 2\Sigma_3 - 2\Sigma_5 + 2\Sigma_7) = \Sigma_7 - 2\Sigma_8 + \Sigma_9$. Therefore, the solution to the system $\mathbf{S}\Delta = \hat{\theta}$ is not unique. Let $\mathbf{S}^+$ be Moore-Penrose inverse of $\mathbf{S}$, then $\Delta = \mathbf{S}^+\hat{\theta} + (I - \mathbf{S}^+\mathbf{S})v$ for any vector v. Denote $C = (I - \mathbf{S}^+\mathbf{S})v$, it can be verified that

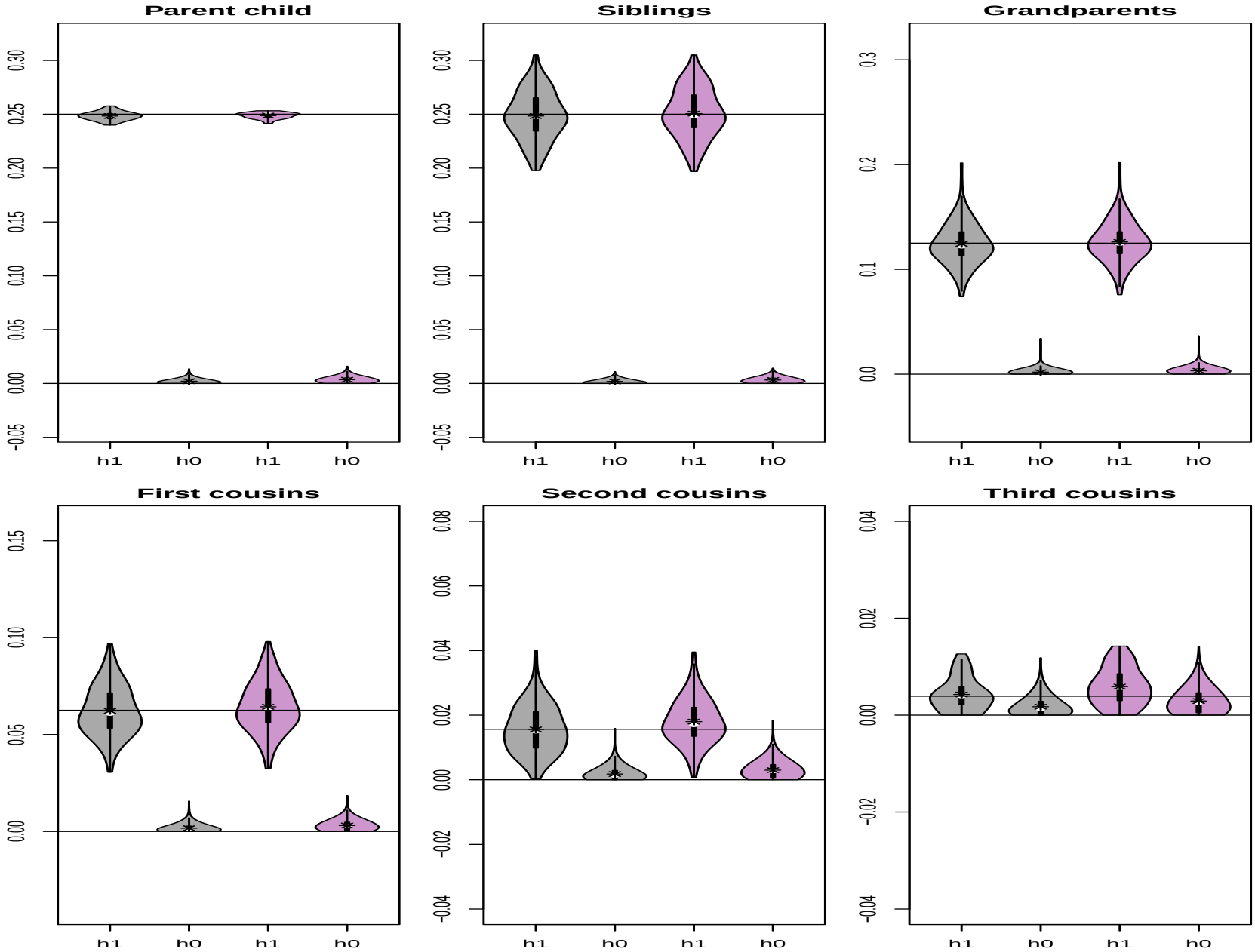
$$
\begin{aligned}
C_1 = C_3 = C_5 = C_9 &= 0 \\
C_2 &= \frac{1}{8}v_2 - \frac{1}{8}v_4 - \frac{1}{8}v_6 - \frac{1}{8}v_7 + \frac{1}{4}v_8 \\
C_4 = C_6 = C_7 &= -\frac{1}{8}v_2 + \frac{1}{8}v_4 + \frac{1}{8}v_6 + \frac{1}{8}v_7 - \frac{1}{4}v_8 \\
C_8 &= \frac{1}{4}v_2 - \frac{1}{4}v_4 - \frac{1}{4}v_6 - \frac{1}{4}v_7 + \frac{1}{2}v_8.
\end{aligned}
\tag{5}
$$

1) $\Delta_1, \Delta_3, \Delta_5$, and Δ_9 are not affected by v and these components have unique solutions. 2) $C_2 + C_4 = 0$, $C_2 + C_6 = 0$ and $C_7 + \frac{1}{2}C_8 = 0$, which means, $\Delta_2 + \Delta_4$, $\Delta_2 + \Delta_6$, and $\Delta_7 + \frac{1}{2}\Delta_8$ are invariant. 3) Consequently ϕ in Equation (1) and F_1 and F_2 in Equation (2) are unique.

Simulated non-admixed samples



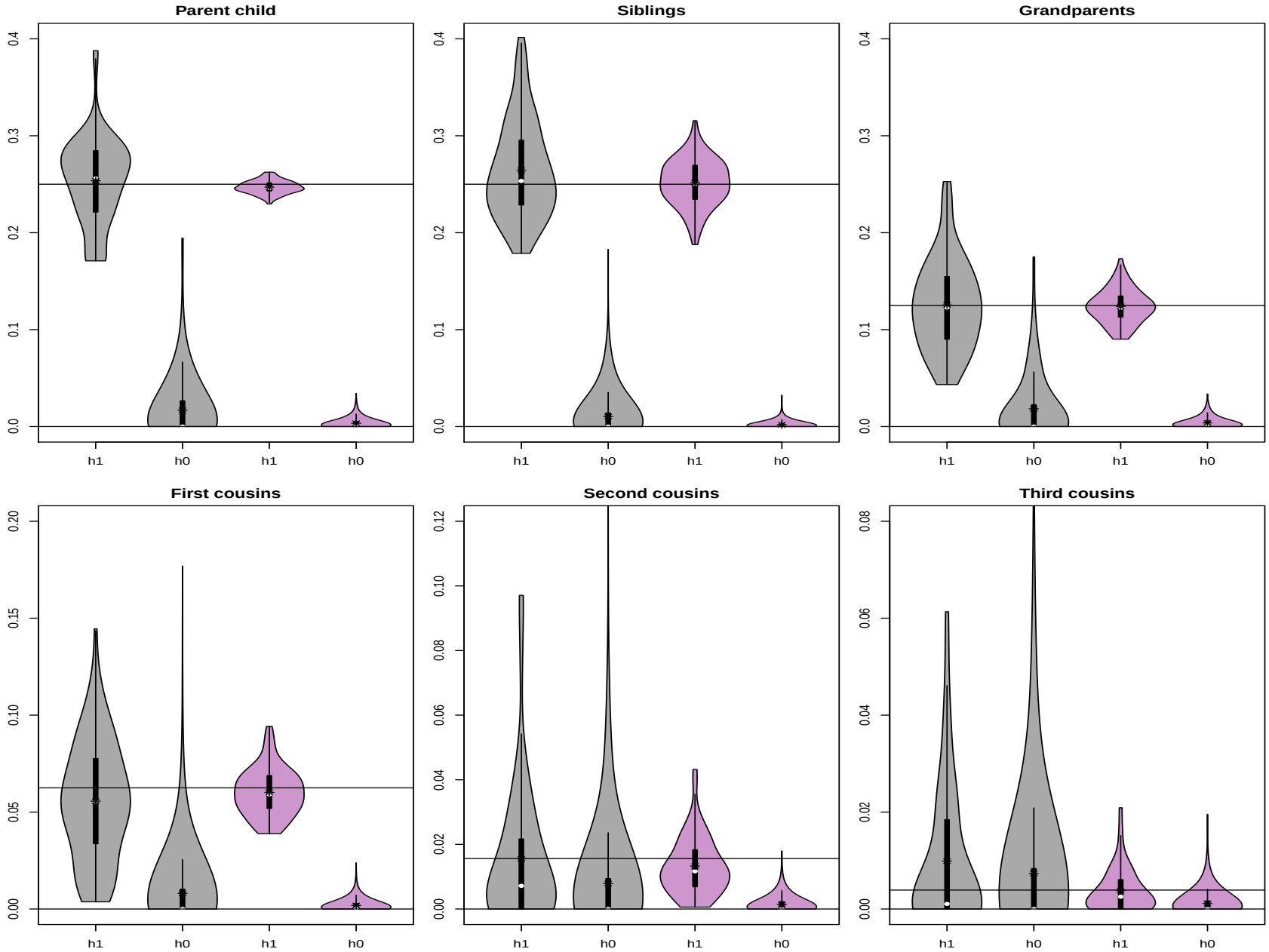
Comparisons between with and without inbreeding



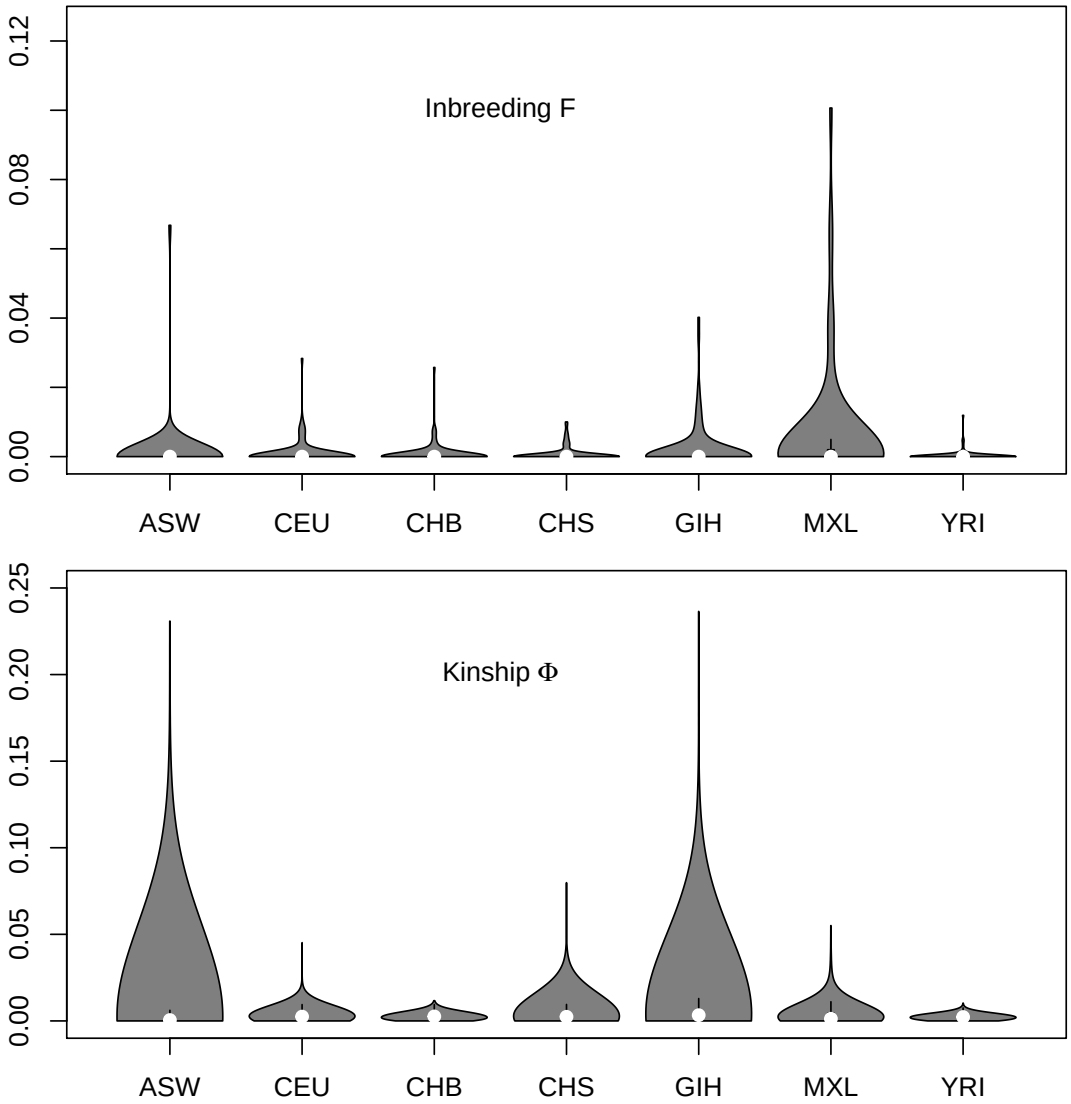
Kinship between admixed samples

- The reference population for IBD for admixed samples has to be the ancestral population predates continental population divergence.
- This ancestral population can be partially mimicked by selecting a set of SNPs whose allele frequencies are similar across different continental populations.
- Among 12 million bi-allelic SNPs with minimum 50 minor allele counts (out of total 2504 diplotypes) in 1000 genomes project, there are 1.2 million such SNPs.
- We also randomly selected common bi-allelic SNPs of 1.2 million, and used these to compute kinship for comparison.

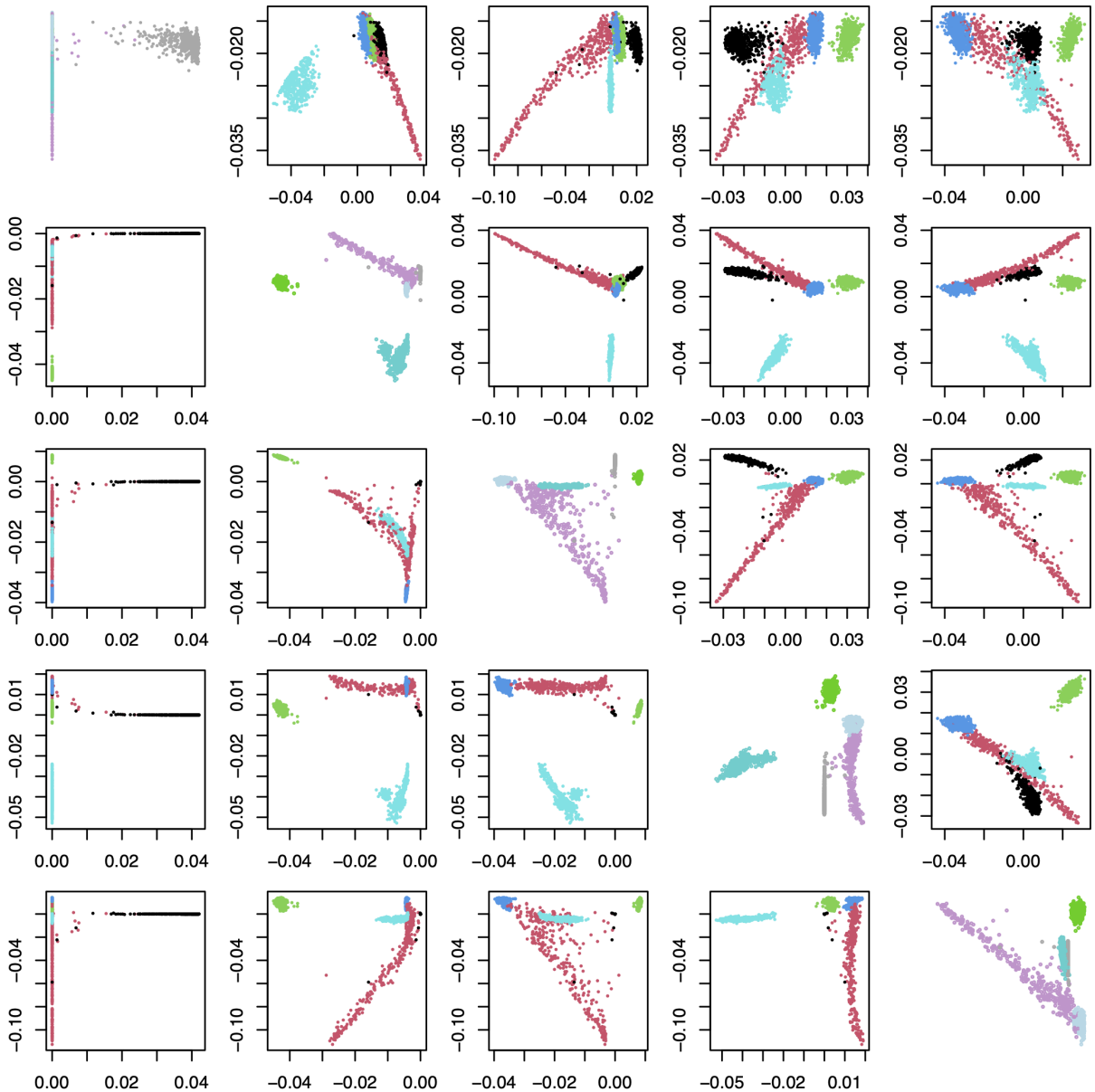
Simulated admixed samples



1000 genomes samples

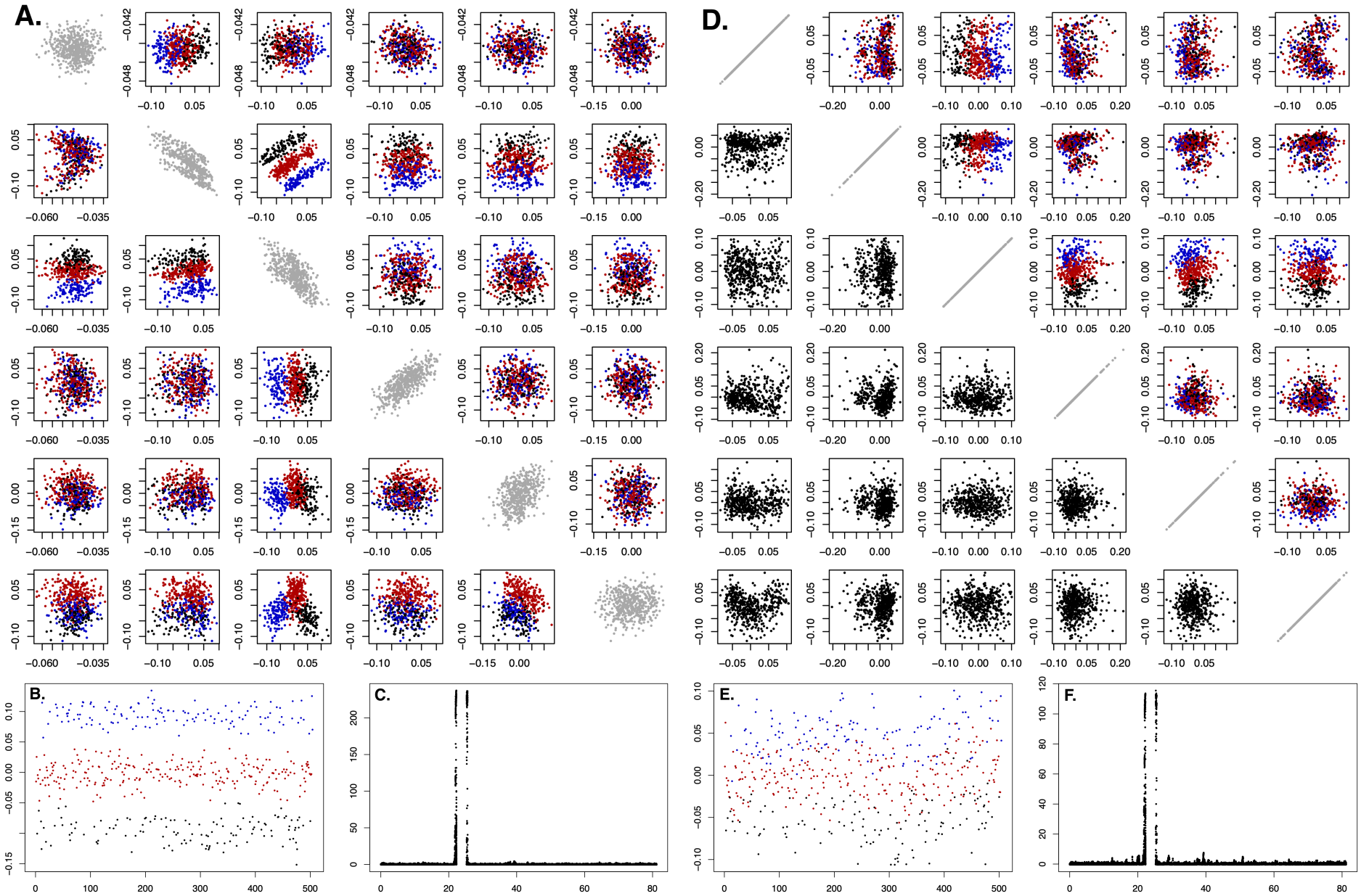


random common vs small population divergence SNPs



Africans (in black), Americans (in red), East Asians (in green), Europeans (in blue), and South asians (in cyan).

Chr17: Asian vs African freq. (Kindred), and scGRM.



Genomic control

λ	None	scGRM	UKin	King	Kindred
BD	1.115	0.988	0.993	1.024	1.014
CAD	1.078	0.985	0.993	1.019	1.014
CD	1.079	0.999	1.003	1.015	1.019
HT	1.063	0.997	1.004	1.016	1.014
RA	1.053	0.974	0.980	1.000	0.989
T1D	1.067	0.940	0.950	0.966	0.972
T2D [7]	1.076	0.986	0.995	1.024	1.008
Height [8]	1.029	0.992	0.990	0.993	0.995
Mean	1.070	0.983	0.989	1.007	1.003
SSD	0.024	0.019	0.017	0.020	0.016

[7] Wellcome Trust Case Control Consortium (2007). Nature 447, 661–678.

[8] Yang, J., et. al. and P. M. Visscher (2010). Nat Genet 42 (7), 565–569.

Heritability of height

	scGRM	UKin	King	Kindred
GCTA	0.449 ± 0.084	0.436 ± 0.078	0.376 ± 0.072	0.474 ± 0.085
Gemma[9]	0.446 ± 0.084	0.426 ± 0.079	0.393 ± 0.071	0.473 ± 0.085

We tried to do down-sampling study, but GCTA produced nonsensical results.

[9] Zhou, X. and M. Stephens (2012). Nat Genet 44(7), 821–824

MAP estimates for heritability

Consider a linear model

$$\begin{aligned} M0 : \mathbf{y} &= \mathbf{W}\alpha + \epsilon \\ \alpha &\sim MVN(0, \tau^{-1} V_W) \\ \epsilon &\sim MVN_n(0, \tau^{-1} I_n). \end{aligned} \tag{6}$$

Adding a random effect $\mathbf{u}$ we have a new model

$$\begin{aligned} M1 : \mathbf{y} &= \mathbf{W}\alpha + \mathbf{Z}\mathbf{u} + \epsilon \\ \mathbf{u} &\sim MVN_m(0, \tau^{-1}\eta K) \end{aligned} \tag{7}$$

Let Q and D be eigen decomposition such that $ZKZ^t = QDQ^t$, we have

$$\begin{aligned} M2 : \mathbf{y} &= \mathbf{W}\alpha + Q\gamma + \epsilon \\ \gamma &\sim MVN_n(0, \tau^{-1} V_Q) \end{aligned} \tag{8}$$

where $V_Q = \eta D$. To see this,

$$E(Z\mathbf{u}\mathbf{u}^t Z^t) = \eta ZKZ^t = \eta QDQ^t = E(Q\gamma\gamma^t Q^t).$$

MAP estimates for heritability

Denote $X = (Q, W)$ and $V = \begin{pmatrix} \eta D & 0 \\ 0 & V_W \end{pmatrix}$, using normal inverse gamma prior to get

$$BF(\eta) = \frac{\det(W^t W)^{1/2}}{\det(\mathbf{X}^t\mathbf{X} + V^{-1})^{1/2} \det(V_Q)^{1/2}} \left(\frac{\mathbf{y}^t\mathbf{y} - \mathbf{y}^t\mathbf{X}(\mathbf{X}^t\mathbf{X} + V^{-1})^{-1}\mathbf{X}^t\mathbf{y}}{\mathbf{y}^t\mathbf{y} - \mathbf{y}^t W(W^t W)^{-1} W^t\mathbf{y}} \right)^{-n/2} \tag{9}$$

Bayes factor can be evaluated efficiently for different η.

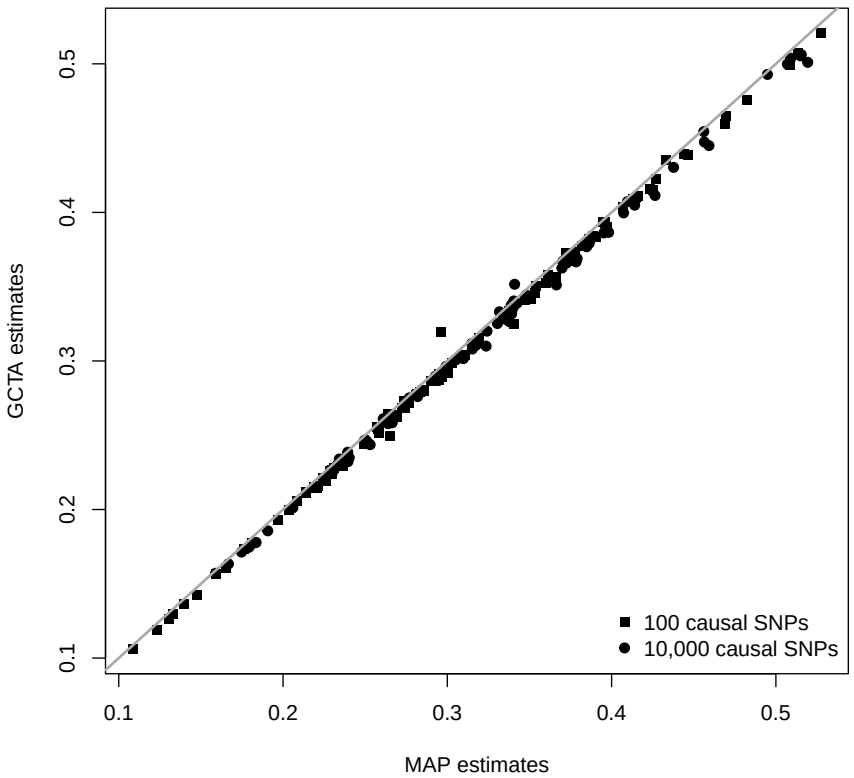
$$\det(X^t X + V^{-1}) = \det(I_n + \frac{1}{\eta D}) \det(W^t W - W^t Q (I_n + \frac{1}{\eta D})^{-1} Q^t W);$$

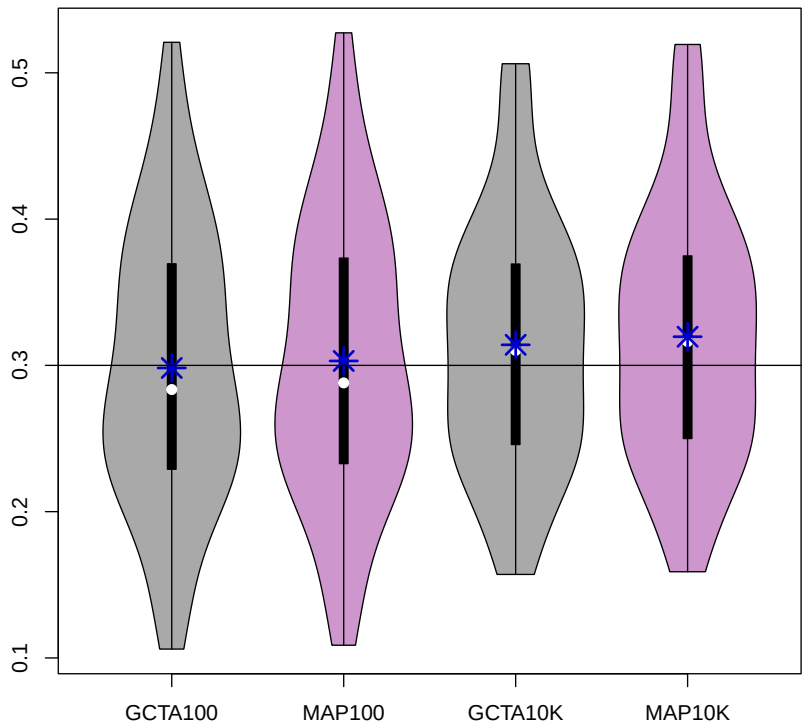
Denote $F = (I_n + \frac{1}{\eta D})^{-1}$, $M = (W^t W - W^t Q F Q^t W)^{-1}$ to get

$$(X^t X + V^{-1})^{-1} = \begin{pmatrix} F + FQ^t WMW^t QF & -FQ^t WM \\ -MW^t QF & M \end{pmatrix}.$$

We can use Nelder-Mead method to optimize $BF(\eta)$ over η.

Simulations heritability 30%





Heritability of height: down sampling study

	scGRM	UKin	King	Kindred
90%	0.455 $\pm$ 0.045	0.445 $\pm$ 0.038	0.410 $\pm$0.032	0.486 $\pm$ 0.041
70%	0.448 $\pm$ 0.086	0.470 $\pm$ 0.086	0.406 $\pm$0.078	0.500 $\pm$ 0.088
50%	0.477 $\pm$ 0.151	0.486 $\pm$ 0.111	0.395 $\pm$0.126	0.533 $\pm$ 0.160

Summary

- Kindred use non-negative least square to compute loadings of latent IBD states to infer kinship and inbreeding coefficients.
- Kindred allows one to specify reference populations (through specifying the set of allele frequencies).
- By choosing a set of small population divergence SNPs, kindred is also effective for admixed samples.
- Kindred is effective to control for population structure and relatedness in GWAS.
- Kindred produces slightly large but statistically significant estimates on heritability.
- Software is available at `www.haplotype.org`.