Estimate inbreeding and kinship coefficients via latent identity-by-descent states

Yongtao Guan (grant.guan@nih.gov) and Daniel Levy (levyd@nhlbi.nih.gov)

Framingham Heart Study & National Heart, Lung, and Blood Institute

1 Introduction

1.1 Motivtation

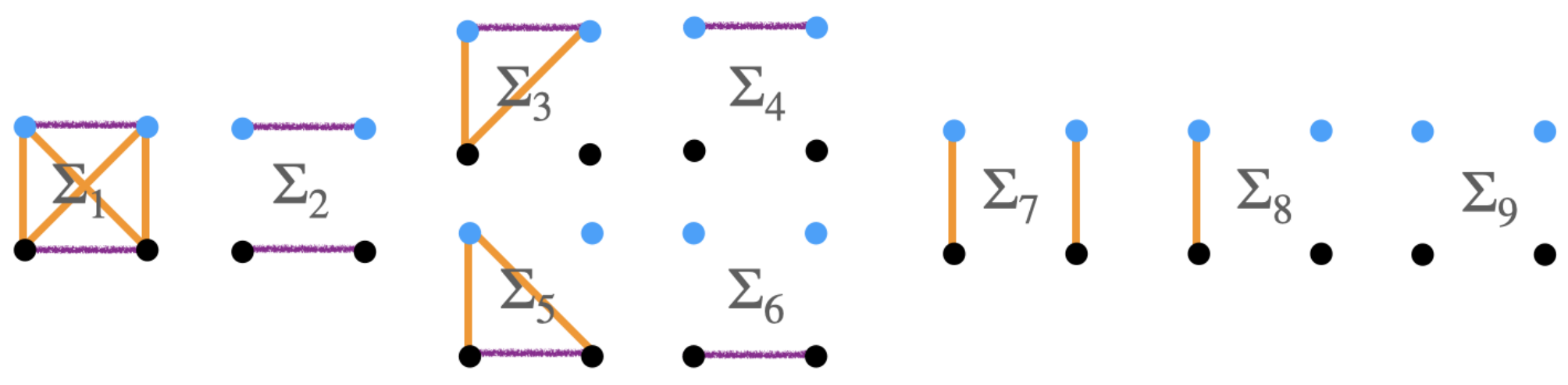
- Estimating the individual inbreeding coefficient and pairwise kinship is an important problem in human disease mapping, forensics, animal and plant breeding, conservation and evolutionary biology.
- Existing methods such as scGRM, UKin, and King are either biased, or assume no inbreeding when estimating kinships.
- Large proportion of estimates are negative, difficult to interpret.

1.2 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.
- 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 Methods

2.1 Jacquard IBD states



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

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

Inbreeding coefficients can also be computed:

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

2.2 Latent states emit joint genotypes

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

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.

2.3 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).

$$\text{argmin}_\Delta \ \ ||\mathbf{S}_p\Delta - \hat{\theta}_p||_2 \tag{3a}$$

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

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

$$\text{argmin}_\Delta \ \ ||\mathbf{S}\Delta - \hat{\theta}||_2 \tag{4}$$

with constraints $(3b)$. We may also fit (4) without constraint.

2.4 Invariant properties

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

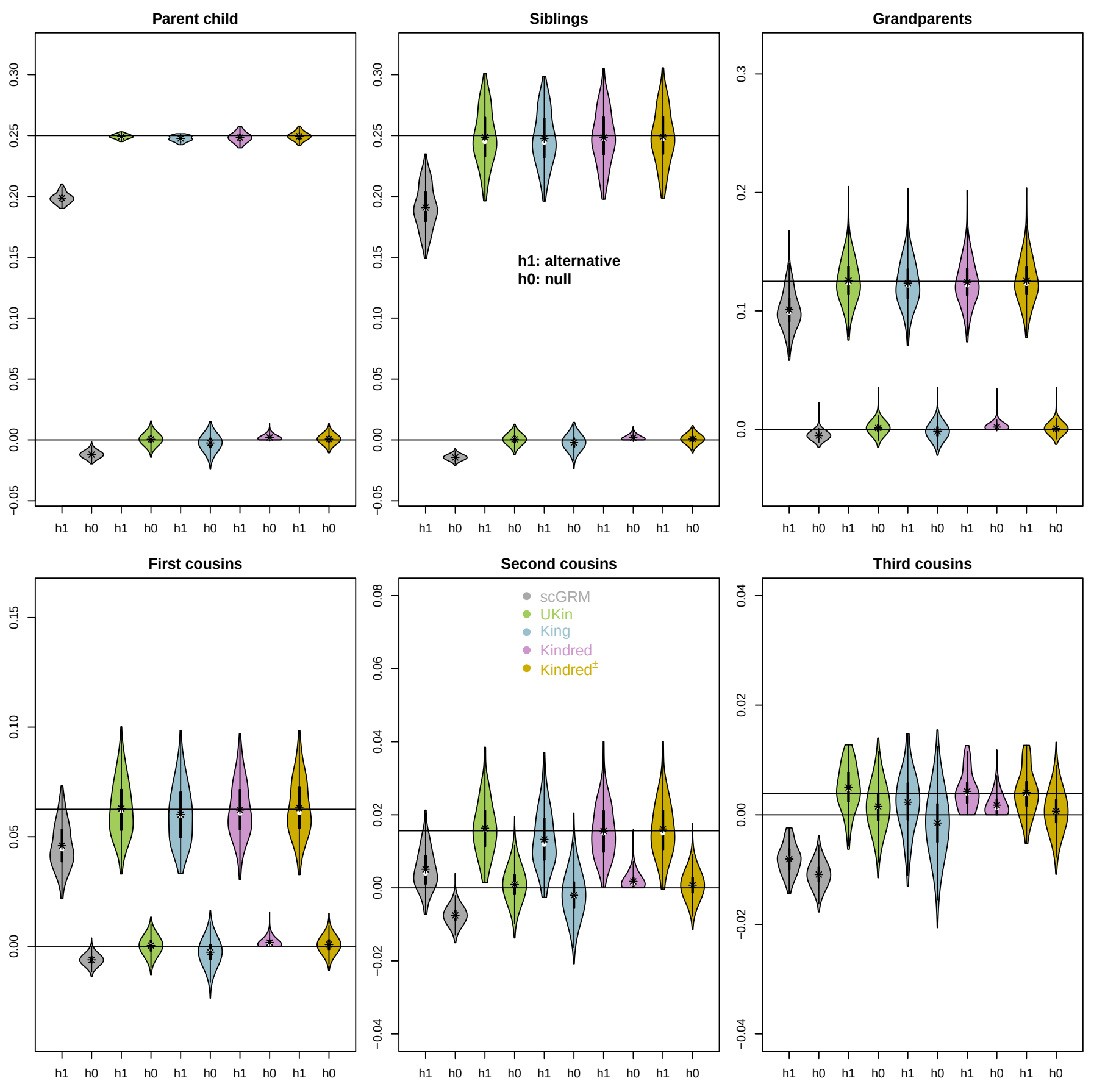
$$\begin{aligned} &C_1 = C_3 = C_5 = C_9 = 0 \\ &C_2 = \frac{1}{8}v_2 - \frac{1}{8}v_4 - \frac{1}{8}v_6 - \frac{1}{8}v_7 + \frac{1}{4}v_8 \\ C_4 = C_6 = \,&C_7 = -\frac{1}{8}v_2 + \frac{1}{8}v_4 + \frac{1}{8}v_6 + \frac{1}{8}v_7 - \frac{1}{4}v_8 \\ &C_8 = \frac{1}{4}v_2 - \frac{1}{4}v_4 - \frac{1}{4}v_6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

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.

3 Results

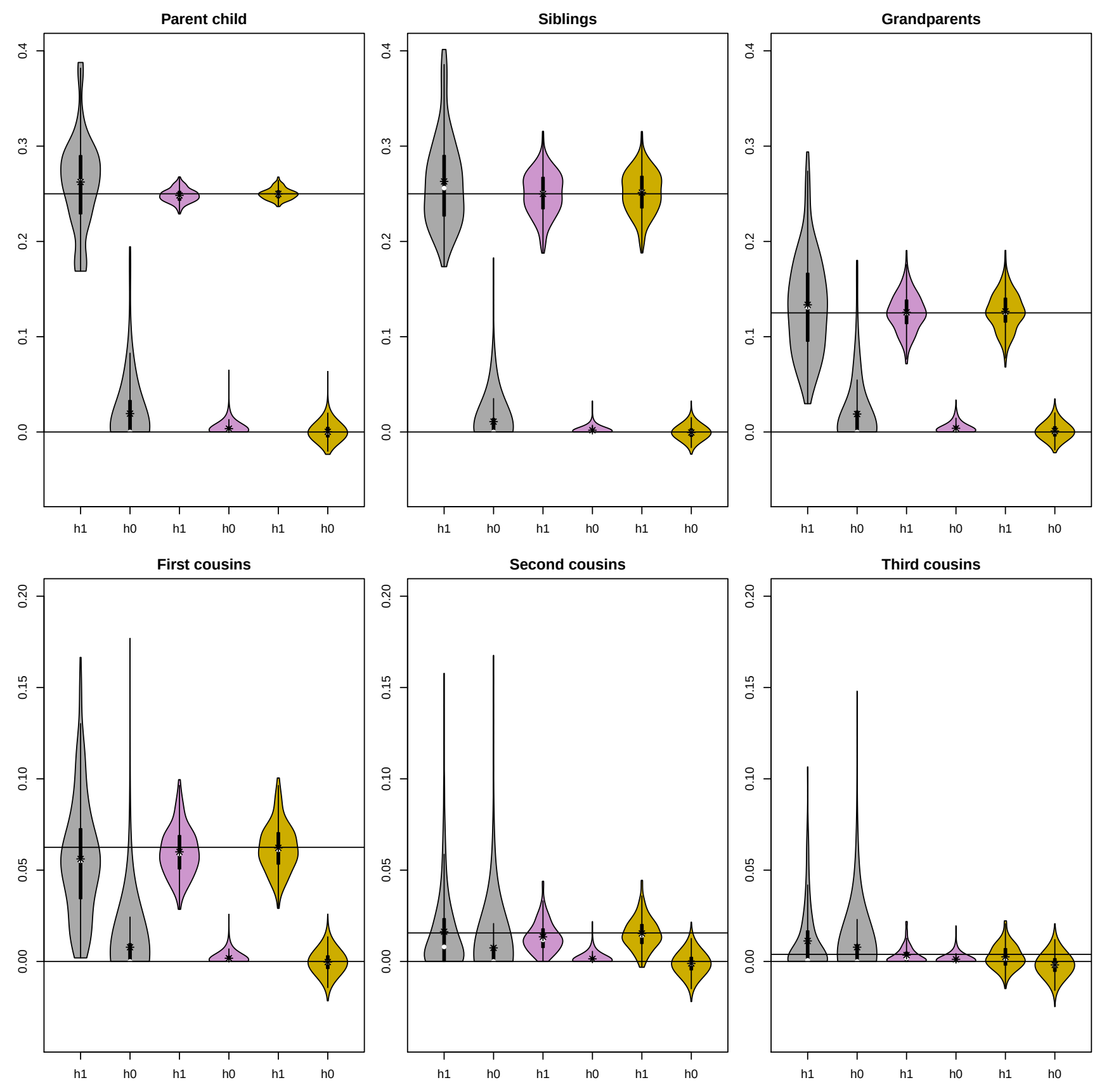
Software Kindred is available at www.haplotype.org.

3.1 Non-admixed samples

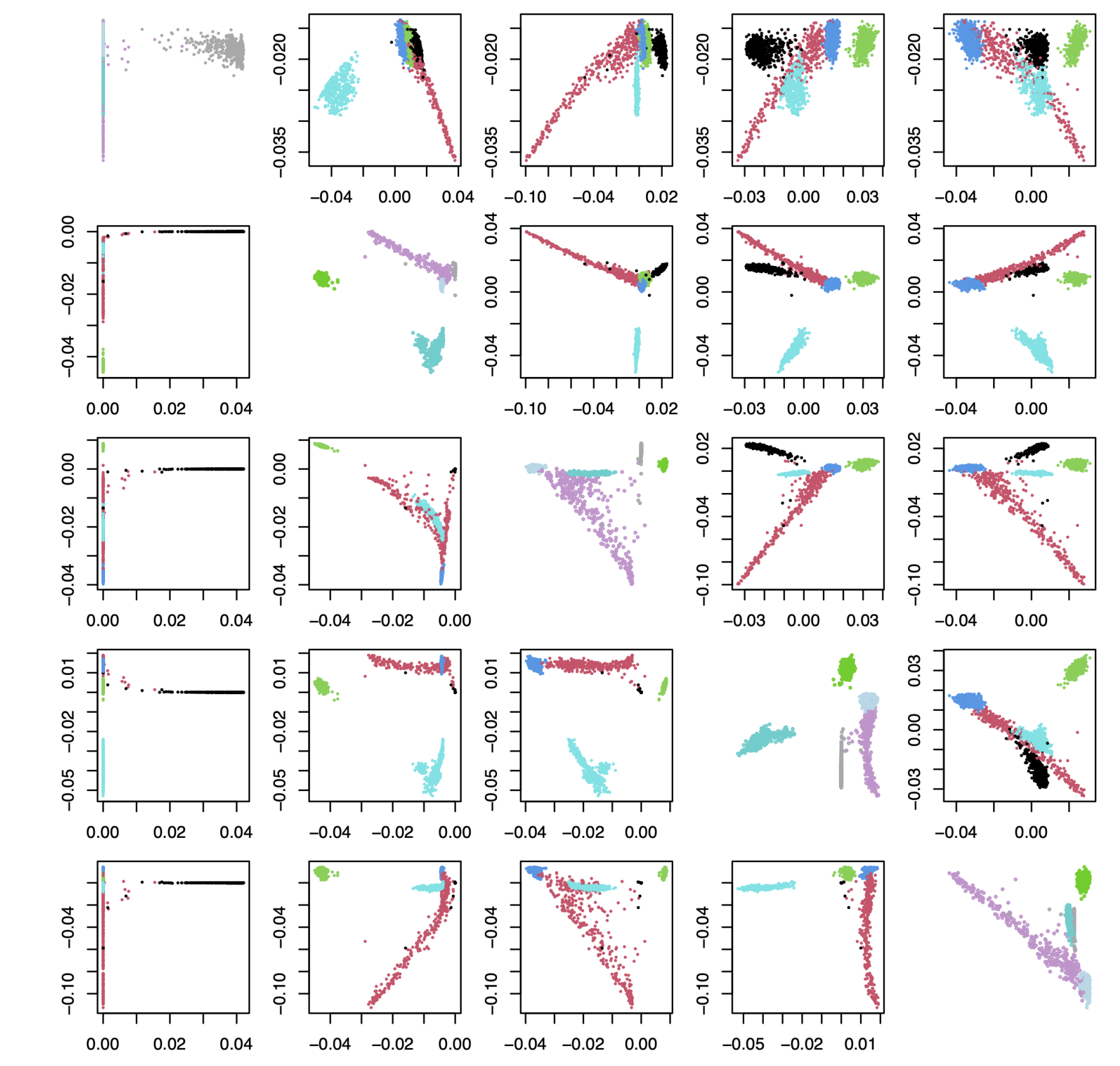


3.2 Admixed samples

- If the continental population were taken as the homogenous reference population for IBD for non-admixed samples, then for 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.

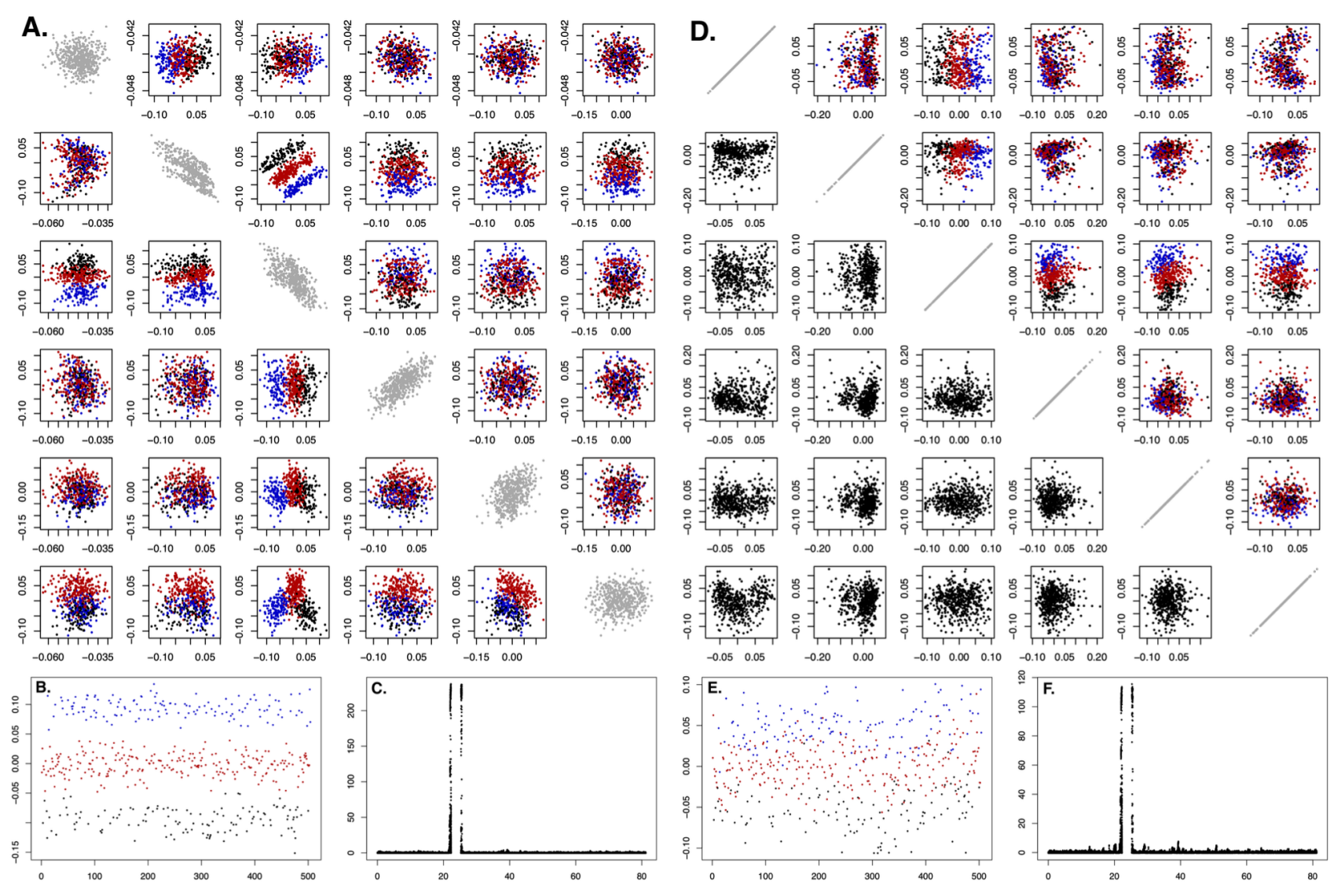


3.3 1000 genomes populations



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

3.4 CenHap on Chr17: Asian vs African



3.5 Genomic control

λ	None	scGRM	UKin	King	Kindred	Kindred$^\pm$
lpa	1.377	0.985	0.995	0.984	1.017	1.008
pon1	1.350	0.978	0.984	0.975	1.001	0.990
MPO	1.287	0.997	0.996	1.002	1.007	1.001
resistin	1.278	0.997	0.998	1.001	1.011	1.002
srage	1.262	0.997	1.000	1.004	0.999	1.004
cd56	1.260	0.988	0.993	0.993	1.003	0.999
cntn1	1.207	0.994	0.998	0.998	1.002	1.003
CD5L	1.190	1.001	1.003	1.009	1.003	1.006
Mean	0.276	0.008	0.005	0.008	0.005	0.004
SSD	0.064	0.007	0.005	0.008	0.006	0.003

3.6 Heritability of height

	scGRM	UKin	King	Kindred	Kindred$^\pm$
GCTA	0.45	0.44	0.38	0.47	0.41
Gemma	0.45	0.43	0.39	0.47	0.41
90%	0.46	0.45	0.41	0.49	0.41
70%	0.45	0.47	0.41	0.50	0.41
50%	0.48	0.49	0.40	0.53	0.44

4 Summary

- Kindred non-negative estimates for kinship and inbreeding coefficients.
- Kindred allows one to specify reference populations.
- Kindred works for admixed samples.
- Kindred is effective to control for relatedness in GWAS.
- Slightly larger but statistically significant estimates of heritability.
- Can be extended to infer gene kinship via hidden Markov models.
- Software is available at www.haplotype.org.

References

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

[Thompson] Thompson, E. A. (2013, 06). *Genetics 194*(2), 301–326.

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

[GCTA] Yang, J., S. H. Lee, M. E. Goddard, and P. M. Visscher (2011, 2022/12/25). *The American Journal of Human Genetics 88*(1), 76–82.

[King] Manichaikul, A., J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale, and W.-M. Chen (2010, 10). *Bioinformatics 26*(22), 2867–2873.

[UKin] Jiang, W., X. Zhang, S. Li, S. Song, and H. Zhao (2022). *BMC Bioinformatics 23*(1), 525.