Bioinformatics, 2024, pp. 1–11

doi: tbd
Advance Access Publication Date: Day Month Year
Paper

PAPER

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

Yongtao Guan[1,2,*] and Daniel Levy[1,2]

[1], Framingham Heart Study, 73 Mt. Wayte, Framingham, 01702, MA, USA and [2]Cardiovascular Epidemiology and Genomics, National Heart, Lung, and Blood Institute, 31 Center Drive, Bethesda, 20892, DC, USA
[*]Corresponding author. grant.guan@nih.gov
FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Motivation: Estimating the individual inbreeding coefficient and pairwise kinship is an important problem in human genetics (e.g., in disease mapping) and in animal and plant genetics (e.g., inbreeding design). Existing methods such as sample correlation-based genetic relationship matrix, KING, and UKin are either biased, or not able to estimate inbreeding coefficients, or produce a large proportion of negative estimates that are difficult to interpret. This limitation of existing methods is partly due to failure to explicitly model inbreeding. Since all humans are inbred to various degrees by virtue of shared ancestries, it is prudent to account for inbreeding when inferring kinship between individuals.

Results: We present "Kindred", an approach that estimates inbreeding and kinship by modeling latent identity-by-descent states that accounts for all possible allele sharing – including inbreeding – between two individuals. Kindred used non-negative least squares method to fit the model, which not only increases computation efficiency compared to the maximum likelihood method, but also guarantees non-negativity of the kinship estimates. Through simulation, we demonstrate the high accuracy and non-negativity of kinship estimates by Kindred. By selecting a subset of SNPs that are similar in allele frequencies across different continental populations, Kindred can accurately estimate kinship between admixed samples. In addition, we demonstrate that the realized kinship matrix estimated by Kindred is effective in reducing genomic control values via linear mixed model in genome-wide association studies. Finally, we demonstrate that Kindred produces sensible heritability estimates on an Australian height dataset.

Availability and implementation: Kindred is implemented in C with multithreading. It takes vcf file or stream as input and works seamlessly with bcftools. Kindred is freely available at `https://github.com/haplotype/kindred`.

Introduction

Kinship (denoted by ϕ) between two individuals is the probability that two alleles sampled at a random locus each from one individual are identical by descent (IBD). The inbreeding coefficient (denoted by F) of an individual is the probability that two parental alleles sampled at a random locus in the genome are IBD. Thus inbreeding coefficient of an individual is the kinship between his or her parents. In addition, between one and oneself (or between monozygotic twins) $\phi = (1 + F)/2$ (Wright, 1922). Therefore, inbreeding can be treated as a derived concept of kinship, and a statistical model designed to estimate kinship automatically applies to estimating the inbreeding coefficient as well. Kinship estimated from genotype data is called realized kinship. Due to the stochastic nature of recombination and gamete segregation, realized kinship may have significant variation from pedigree estimates (Visscher et al., 2006). Estimating the individual inbreeding coefficient and pairwise kinship is an important problem in human disease mapping, forensics, animal and plant breeding, and conservation and evolutionary biology (Wang, 2016).

In an effort to study gene IBD, Jacquard (1972) documented nine IBD states between any two individuals within a pedigree. These IBD states are partially observable within a pedigree, and the mean probabilities of each state can be computed purely based on the pedigree. These IBD states, however, are completely latent between two individuals not linked by a known pedigree. Thompson (2013) described connections between the Jacquard IBD states and Ewen's sampling partition in coalescence, and provided the joint distribution of genotypes between two individuals conditioning on latent gene IBD states. These formed the basis for our strategy to estimate kinship by inferring latent IBD states between two individuals via joint genotype distributions.

Existing methods of kinship estimation, such as KING and the recent UKin approach (Manichaikul et al., 2010; Jiang et al., 2022) ignore the possible inbreeding of each sample. Inbreeding can affect the abundance of homozygous markers in the genome; failure to account for inbreeding results not only in the inability to estimate inbreeding, but also biases kinship estimates (more in the Discussion). The sample correlation-based genetic relatedness matrix (scGRM) has been widely used as the kinship matrix (Price et al., 2006; Yang et al., 2011; Wang et al., 2017). But it has been shown that scGRM estimates are biased downward (Weir and Goudet, 2017; Ochoa and Storey, 2021; Jiang et al., 2022), and this bias is due to scGRM using sample allele frequency in computation, which are evidently biased compare to the allele frequencies of the reference population. Several methods have been developed to correct for the bias. In particular, Weir and Goudet (2017) used least related samples to recalibrate the kinship estimates; Jiang et al. (2022) quantified the bias under an assumption of no inbreeding and correct the bias. But both approaches still produce negative estimates for unrelated samples. Ochoa and Storey (2021) achieved non-negativity by subtracting the smallest estimates; their method also ignored inbreeding by only considering locally outbred individuals.

We present Kindred, which estimates kinship and inbreeding coefficients by inferring the loading probabilities (or weights) of latent IBD states. A similar approach was implemented by Milligan (2003) using maximum likelihood, but its computation was expensive and deemed not practical for modern GWAS datasets (Zheng et al., 2012). Kindred, on the other hand, used nonnegative least squares method to fit the model. This is not only computationally more efficient, but also enforces non-negativity of kinship estimates. Through simulations, we demonstrate its high accuracy and non-negativity in kinship estimates. In addition, by selecting a subset of SNPs that are similar in allele frequencies across different populations, we demonstrate that Kindred accurately estimates kinship between admixed samples. A major application of the kinship matrix in genome-wide association studies is the variance-covariance matrix for random effect in linear mixed models, and we demonstrate that realized kinship matrix estimated by Kindred is effective in reducing genomic control values. Another application of the kinship matrix is to estimate heritability through variance component models, and we demonstrate that the realized kinship matrix estimated by Kindred produced a higher, statistically significant, heritability estimate than other methods.

Materials and Methods

Latent IBD states model

At an arbitrary marker, there are 15 detailed IBD states between four alleles of two individuals. If the parental origin of an allele is not of interest, these 15 states can be reduced to nine condensed IBD states (Jacquard, 1972; Thompson, 2013). Here we work with the nine condensed identity states detailed in Table 1. Conditional on each state, the distribution of joint genotypes for a bi-allelic SNP is given in Table 1. This distribution is a bi-allelic special case of what is presented for the four allelic case (Thompson, 2013). Kinship can be computed from the loading probabilities (or weights), Δ_j for j-th latent IBD state Σ_j, as follows (Jacquard, 1972):

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

With reference to the diagram in Table 1, all latent states that have IBD between two samples (yellow segments) contribute to the kinship calculation in Equation 1. The fractional coefficients multiplying each loading probabilities in Equation 1 come from their respective numbers of equivalent and mutually exclusive IBD states. (Combining equivalent and mutually exclusive states is how 15 states were condensed to nine.) Inbreeding coefficients can also be computed from the loading probabilities of latent IBD states (Jacquard, 1972):

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

With Table 1, we can model the observed genotype as emanating from the mixture of the nine latent IBD states to estimate the loading probabilities (Δ's), and from which to estimate kinship using Equation (1). We start by considering a subset of SNPs with allele frequency p so that they share the same matrix in Table 1. Denote $n_{p,1}$ the count of AA AA, $n_{p,2}$ the count of AA AB, ..., and $n_{p,9}$ the count of BB BB (Table 1), the multinomial likelihood is

$$(n_{p,1}, \ldots, n_{p,9}) \sim \text{Multinomial}(\theta_{p,1}, \ldots, \theta_{p,9}),$$

and the maximum likelihood estimate for $\theta_{p,1}, \ldots, \theta_{p,9}$ are $\hat{\theta}_{p,1} = n_{p,1}/\sum n_{p,j}, \ldots, \hat{\theta}_{p,9} = n_{p,9}/\sum n_{p,j}$. We can then model joint genotypes distribution $\hat{\theta}_p$ as a linear combination of Σ_j's to seek the least squares fit of a constrained system

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

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

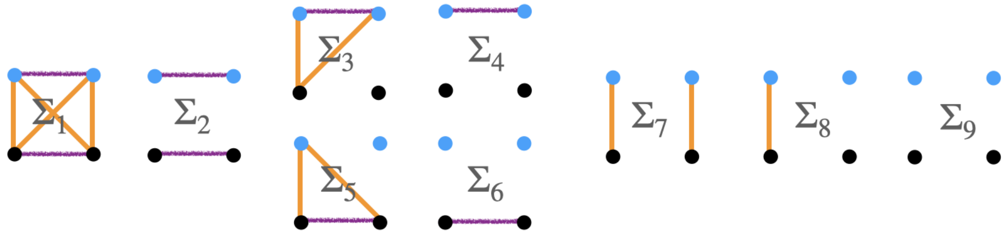
where $\mathbf{S}_p = (\Sigma_1, \ldots, \Sigma_9)$ is a 9×9 matrix detailed in Table 1, $\Delta = (\Delta_1, \ldots, \Delta_9)$ is the vector of loading probabilities, and $|| \cdot ||_2$ denotes the L_2 norm, which is the square root of sum of squares of all components. From the estimates of the loading probabilities $\hat{\Delta}$, we can compute the kinship estimate $\hat{\phi}$ using Equation (1). In essence we treat $\hat{\theta}$ as observed and seek a non-negative least squares fit to obtain $\hat{\Delta}$.

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 equals 1 (depending on the joint genotypes) and the rest 8 entries 0. We append $\mathbf{S}_{p_i}$'s together to obtain the design matrix $\mathbf{S}$ and append e_i's together to obtain a vector $\hat{\theta}$. For total m SNPs, $\mathbf{S}$ is an $9m \times 9$ matrix, and $\hat{\theta}$ is an $9m$ vector, and we seek the least squares fit of a constrained system

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

with constraints (3b). The above system generalizes (3a) to SNPs of arbitrary allele frequencies.

The least square approach has an implicit assumption of independence between units of observations. When the independence assumption was violated, such as dependence among SNPs in our application, the least square approach is still statistically valid in light of composite likelihood (Varin et al., 2011). The model can be fit via non-negative least squares method (Lawson and Hanson, 1974; Bro and De Jong, 1997). The non-negative least squares method respects the constraints (3b), which guarantee non-negative estimates; while ordinary least squares method does not. In practice, we find binning SNPs and re-estimate allele frequencies in each bin improves performance (Supplementary).



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

Table 1. The diagram above the table shows nine states of IBD sharing between two individuals. These are colored re-rendering of Table 2 in (Jacquard, 1972). Two alleles of individual 1 are colored in blue and two alleles of individual 2 in black. Two alleles IBD within an individual are connected via a purple segment. Two alleles IBD between individuals are connected by an orange segment. Σ_7, Σ_8, and Σ_9 are two non-inbreed individuals share two, one, and zero alleles IBD respectively. The table details joint genotype distribution for each latent IBD states, reproduced from Table 3 of (Thompson, 2013). Each of the nine columns labelled by Σ_j corresponds to a latent state in the diagram above. The joint genotypes are listed in columns G1 and G2 where the order matters. p denotes the frequency of A allele and q frequency of B allele, and $p + q = 1$. Note each column labelled by Σ_j sums to 1. When G1 and G2 are the genotypes of the same individual, only Σ_1 and Σ_7 are relevant. When inbreedings are ignored, only last three Σ columns are relevant.

Results

Kinship between non-admixed samples

We simulated related samples with two Han Chinese populations (CHB and CHS) from 1000 Genomes project (Auton et al., 2015) as founders. The simulated kinships have a wide range, including parent offspring, full sibling, first-degree, second-degree, and third-degree cousins. We selected 1.7 million bi-allelic common SNPs to infer kinship (more details in Supplementary). Figure 1 compared performance of different methods (numerical comparisons can be found in Supplementary Table S1 and S2), and we make following observations: 1) The classical method scGRM had noticeable bias in all simulations; 2) popkin (Ochoa and Storey, 2021) appeared to over-corrected and biased upwards, while UKin worked well in correcting the bias; 3) KING performed similarly well with UKin, with UKin performed only slightly better, presumably because KING was benefited from the large number of SNPs we used here compared to the simulations in (Jiang et al., 2022); 4) Kindred and UKin performed similarly well in mean estimates under the alternatives, with Kindred produced slightly better results (mean of the inferred kinships are closest to the theoretical value) in simulations of more distant relationship (cousins); 5) Kindred had the smallest variations under the null, which were the majority of the kinship estimates; 6) Kindred and popkin were the only two methods that produced non-negative estimates on all kinship estimates.

Table 1 was obtained using Han Chinese founders from 1000 Genomes project. To investigate whether Kindred performs similarly well in other founding populations, we conducted additional simulations using diverse founders. Supplementary Figure S3, S4, and S5 compared the deviation between inferred and the truth (instead of the expectation from the pedigree used in Figure 1) for African, East Asian, and European founders respectively. Kindred outcompeted other methods, particularly for harder problems such as kinship between third cousins.

Kinship between admixed samples.

Inferring kinship between admixed samples is a difficult problem. The difficulty lies in estimating appropriate allele frequencies. Thornton et al. (2012) solved this problem by using individual-specific allele frequencies, which are admixture proportion-weighted ancestral population frequencies. Here we propose to use a subset of SNPs that have similar allele frequencies between continental populations. 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 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, which we call SNPs of small population divergence (SPD). Among 12 million bi-allelic SNPs with minimum 50 minor allele counts (out of total 2504 diplotypes) in the 1000 Genomes project, there are 1.2 million SPD SNPs (details in Supplementary), and we used these SNPs to compute kinship for simulated admixed samples. We also randomly selected common bi-allelic SNPs of 1.2 million, and used these to compute kinship for comparison. We simulated related admixed samples in the same manner as simulating non-admixed samples, the only difference was to choose founders from multiple continents. We chose CEU, YRI, CHB, and CHS as founders to simulate related admixed samples, for these populations show a small extent of inbreeding and a low level of pairwise kinships. Figure 2 demonstrates that our

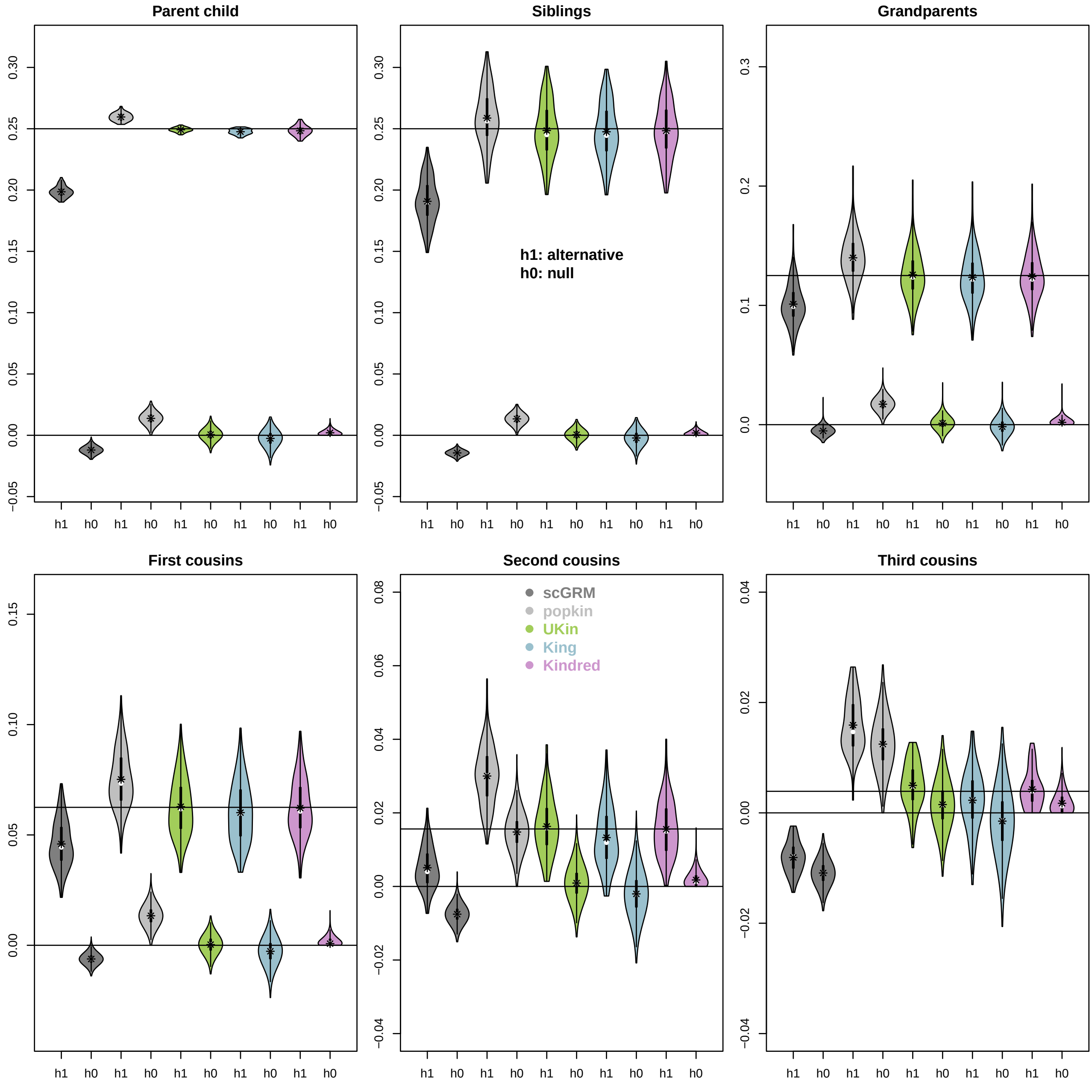


Fig. 1. Comparison of kinship estimates by different methods. Results of scGRM (dark gray) were obtained with PLINK. Results of popkin (light gray) were obtained with its R package. Results of UKin (green) were obtained from reimplementation in software Kindred (to take advantage of its multi-threading capacity). Results of King (blue) were obtained from its software King. Results of Kindread (plum) were obtained from its software Kindred. Expected kinship based on pedigree and zero values were marked by horizontal lines. For each method we showed two violin plots: under the alternative (h1) and under the null (h0). The mean was marked by $\star$. Supplementary Table S1 contains numeric comparisons with mean $\pm$ sd under h1, and the percent of negative estimates under h0.

strategy of using SPD SNPs worked well for admixed samples. While randomly selected SNPs produced kinship estimates with much larger variation and the results were perhaps only useful for the first degree relatedness, the selected SPD SNPs produced kinship estimates that were comparable to estimates for non-admixed samples.

Genomic control

One important application of kinship estimates is to control for population stratification and (cryptic) relatedness in genome-wide association studies, either via controlling for top principal components (Price et al., 2006), or incorporate kinship into test statistics (Thornton and McPeek, 2010), or via linear mixed model (Kang et al., 2010; Zhou and Stephens, 2012; Chen et al., 2016). The cohorts in the Framingham Heart Study (FHS), funded by the National Heart, Lung, and

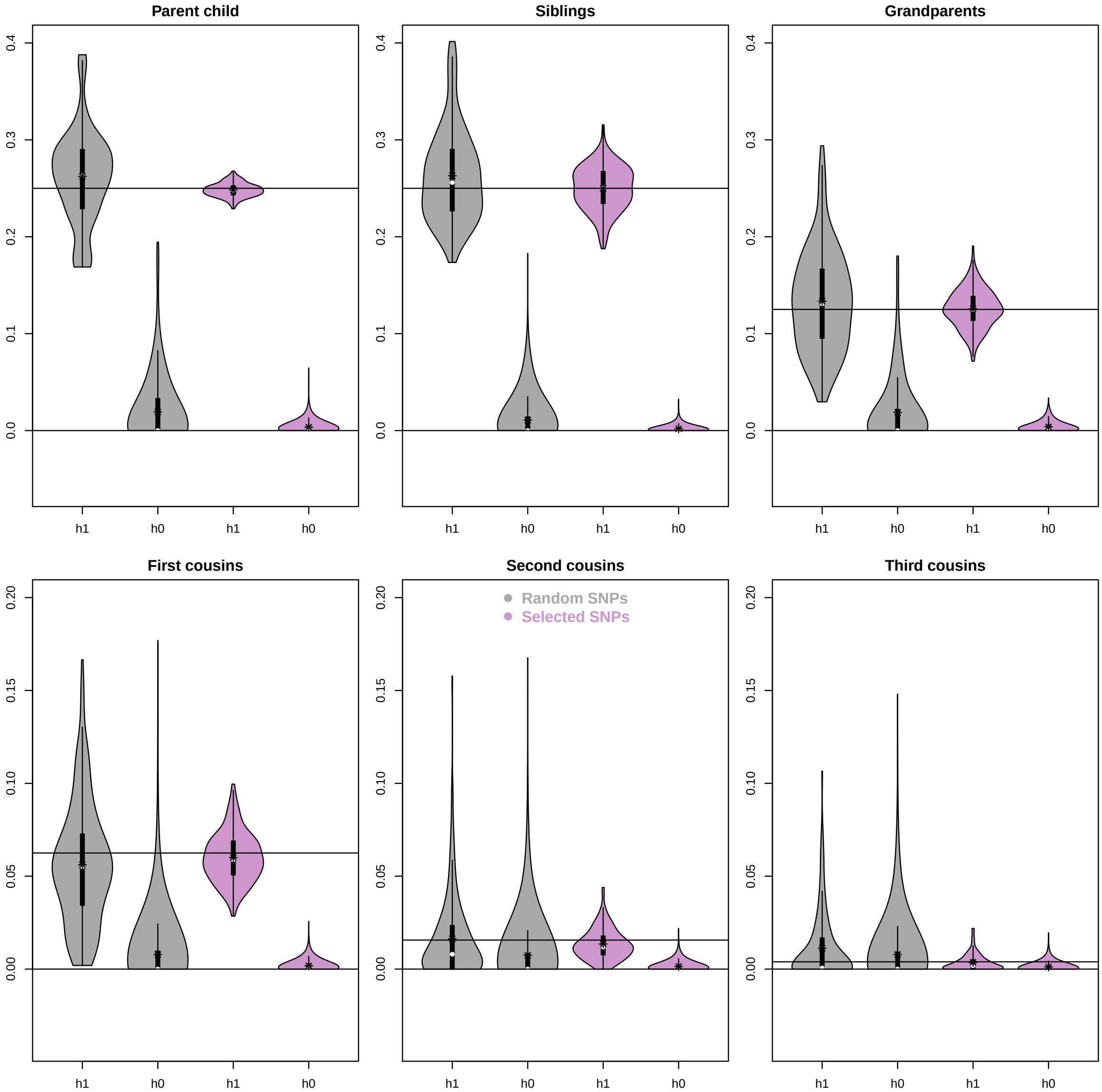


Fig. 2. Kindred is effective for admixed samples using SNPs of small population divergence (color plum). Kinship estimated using randomly selected common SNPs (colored in gray) tend to be biased and with large variation. In each panel, theoretical values and zero values were marked by horizontal lines.

Blood Institute (NHLBI), consists many independent three generational pedigrees, nuclear families, trios, duos, and singletons (Kannel et al., 1979). This is a setting in which linear mixed model is effective in controlling for relatedness in the samples. We analyzed 5757 samples with whole genome sequencing data through NHLBI's TOPMed program (Taliun et al., 2021) and who also have protein immunoassays obtained through the NHLBI's Systems Approach to Biomarker Research in Cardiovascular Disease (SABRe CVD) Initiative (Ho et al., 2018). Here we analyzed eight proteins that show substantial inflated genomic control values λ (Devlin and Roeder, 1999).

In Table 2 we compared different kinship estimates in their ability to reduce inflation of test statistics due to the relatedness. All methods are effective in reducing λ to near 1, suggesting biases in kinship estimates were well tolerated in the linear mixed model. If measured by deviation from 1, scGRM and King are tied to be the second best, and popkin, UKin and Kindred are tied to be the best. Noticeably, Kindred has the largest mean of λ among all methods and is the only one whose mean of λ slightly larger than 1.

Heritability of height

Using the height data from (Yang et al., 2010), we estimated phenotype variation explained (PVE) by different kinship matrices, where PVE estimates by GCTA were confirmed by GEMMA. The Kindred kinship matrix produced the highest PVE estimates (Table 3). To investigate whether the higher PVE estimates was due to chance or not, we resort to down-sampling. But for the down sampled datasets GCTA produced untenable results due to small sample sizes. With the understanding that Bayesian method performs better with

λ	None	scGRM	popkin	UKin	King	Kindred
lpa	1.377	0.985	0.994	0.995	0.984	1.017
pon1	1.350	0.978	0.985	0.984	0.975	1.001
MPO	1.287	0.997	0.994	0.996	1.002	1.007
resistin	1.278	0.997	0.998	0.998	1.001	1.011
srage	1.262	0.997	1.000	1.000	1.004	0.999
cd56	1.260	0.988	0.994	0.993	0.993	1.003
cntn1	1.207	0.994	0.999	0.998	0.998	1.002
CD5L	1.190	1.001	1.003	1.003	1.009	1.003
Deviation	0.276	0.008	0.005	0.005	0.008	0.005
Mean	1.276	0.992	0.995	0.995	0.996	1.005

Table 2. Comparison of kinship estimates by their effects on genomic control λ values . The single SNP test (to derive genomic control values) using linear mixed model was done with GEMMA. The deviation for each column are calculated from mean of $(|\lambda - 1|)$.

		scGRM	popkin	UKin	King	Kindred
GCTA	μ	0.449	0.450	0.436	0.376	0.474
	σ	0.084	0.077	0.078	0.072	0.085
GEMMA	μ	0.446	0.445	0.426	0.393	0.473
	σ	0.084	0.077	0.079	0.071	0.085
Resample 90%	μ	0.455	0.454	0.445	0.410	0.486
	σ	0.045	0.033	0.038	0.032	0.041
Resample 70%	μ	0.448	0.473	0.470	0.406	0.500
	σ	0.086	0.081	0.086	0.078	0.088
Resample 50%	μ	0.477	0.452	0.486	0.395	0.533
	σ	0.151	0.137	0.111	0.126	0.160

Table 3. Mean (μ) and standard deviation (σ) of heritability estimates of height data and its resampling. The top part of the table was obtained from a single dataset of 3925 samples, where PVE were estimated by GCTA and GEMMA respectively for different kinship estimates; the bottom part of the table was obtained from the resampling study, where 90%, 70% and 50% of the 3925 samples were sampled without replacement in each trial. For each percentage, 100 resampling trials were run and PVEs were estimated using our Bayesian MAP estimates, and mean and standard deviation were computed from the 100 estimates.

small sample sizes, we reframed the linear mixed model as Bayesian linear regression with a specific prior. Within this framework, we can estimate PVE by seeking MAP (maximum a posterior) estimate of a hyperparameter η, which determines the (relative) size of the random effect (details in Appendix). Our down-sampling study showed that Kindred indeed outperformed scGRM (and others) in estimating PVE of the height data, and the difference over 100 down-sampling was statistically significant (for 90% resampling rank test $P = 4 \times 10^{-6}$; for 70% resampling rank test $P = 4 \times 10^{-5}$; and for 50% resampling rank test $P = 0.02$). In addition to down-sampling, we masked phenotypes of randomly selected subset of samples, and used kinship to predict those masked phenotypes, Kindred performed best in these genomic prediction cross validation experiments (Supplementary Figure S6). It remains to be determined whether this gain in PVE can be translated to power gain in genetic association studies, particularly in multi-omics phenotypes such as proteomics and gene expression assay.

Discussion

We developed a latent IBD state model to infer kinship and inbreeding coefficients. Our method explicitly models inbreeding when inferring kinships between samples. Kindred makes use of the non-negative least squares method (Lawson and Hanson, 1974) for model fitting, so that our kinship estimates are non-negative, which overcomes difficulties of producing negative estimates that other methods have. In order to use least squares method to fit the latent state model, a unit observation of joint genotypes at a SNP was expanded as a frequency vector, albeit trivial, with one entry as 1 (depending on joint genotypes) and other entries as 0. Compared to a multinomial maximum likelihood approach, the least squares approach we chose here is efficient in computation (Supplementary Tables S3 and S4), particularly with the readily available modern non-negative least squares method (Bro and De Jong, 1997). The system is rank deficient, but it can be shown that the quantities of interest, namely kinship and inbreeding coefficients, are invariant (Supplementary).

The latent state model can be used as a theoretical framework to analyze other methods such as scGRM. Let X and Y denote allele counts of genotypes $G1$ and $G2$ in Table 1, p is allele frequency of A allele, and consider

$$\frac{(X-Y)^2}{2p(1-p)} = \frac{(X-2p)^2}{2p(1-p)} + \frac{(Y-2p)^2}{2p(1-p)} - 2\frac{(X-2p)(Y-2p)}{2p(1-p)}, \tag{5}$$

where $Q = \frac{(X-2p)(Y-2p)}{2p(1-p)}$ is the quantity calculated in scGRM. The expection of the left hand side in Equation (5) can be directly computed in light of Table 1, and we get $E(LHS) = 4\Delta_2 + \Delta_3 + 3\Delta_4 + \Delta_5 + 3\Delta_6 + \Delta_8 + 2\Delta_9$. Identifying the expectation of the right hand side is $1 + F_1 + 1 + F_2 - 2Q$ and plugging F_1 and F_2 in Equations (2) to get $E(RHS) = 2 + 2\Delta_1 + 2\Delta_2 + \Delta_3 + \Delta_4 + \Delta_5 + \Delta_6 - 2Q$ (more details in Supplementary). Equating the expectations of LHS and RHS and making use the identity $\sum_{j=1}^{9} \Delta_i = 1$, to get the expectation of Q as $E(Q) = 2\Delta_1 + (\Delta_3 + \Delta_5 + \Delta_7) + \frac{1}{2}\Delta_8 = 2\phi$, as defined in Equation (1).

A natural question is if scGRM produces unbiased estimates of kinship according to the latent state model framework, then why does it have such noticeable bias, particularly negative estimates for unrelated or distant related samples? One contributing factor is that the sample allele frequencies were biased from the reference population. Ochoa and Storey (2021) attempted to correct the bias by recalibrating based on the smallest estimates, but our simulations suggested popkin tends to over-correct. On the other hand, Jiang et al. (2022) provided an effective way (UKin) to correct the bias, which dramatically reduced negative estimates for kinship between unrelated samples. Kindred produced non-negative kinship estimates even with biased allele frequencies, owning to the non-negative least squares method used in model fitting. From above analysis we know $Q = (1 + F_1 + 1 + F_2 - LHS)/2$, and not accounting for inbreeding effectively sets $F_1 = F_2 = 0$, which reduces Q and thus contributes to excessive negative entries in scGRM. From the perspective of latent state model, however, not modeling inbreeding is effectively setting loading probabilities of the first six latent states as zero, and those excessive allele sharing caused by inbreeding are more likely to be assigned to latent states Σ_7 and Σ_8 than Σ_9 (Figure 1), resulting in slight upward bias (the amount of bias were limited by the kinship between founders). Supplementary Figure S7 demonstrated this intuition via simulations.

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 (Wang, 2016; Goudet et al., 2018). The defining feature of a reference population is its marginal allele frequencies. (Arguably, the best feature of a reference population is its haplotype frequencies, but that is difficult to obtain and not easy to work with.) Both scGRM and UKin used allele frequencies estimated from samples, which implicitly assumed samples were taken from reference populations and allele frequencies estimates were unbiased. Both were strong assumptions. It has been pointed out that one has to take into account of pedigree or cryptic relatedness to obtain unbiased estimates of allele frequencies (Sara McPeek et al., 2004). So an iterative procedure to refine estimates of allele frequencies and kinship over and over might help to alleviate bias and negativeness observed in these methods. Kindred allows users to specify allele frequencies through a tag in VCF files, which effectively allows users to specify a reference population.

Although mis-specifying a reference population will bias the kinship (and inbreeding coefficient) estimates, there in fact has nothing wrong with such a practice if the results were interpreted accordingly. For example, when using Africans as reference population, kinship estimates between East Asians were elevated compared to using East Asian as reference population. Since African alleles are more diverse than East Asian alleles, it is only reasonable that East Asians are modestly "inbred" with reference to the African alleles. In fact, interesting patterns can emerge in East Asian samples through the lens of African allele frequencies. Figure 3 showed such an example when computing kinship using SNPs on chr17. Panel A upper triangular contain PC plots from Kindred kinship estimates using African allele frequencies as reference. Panel A lower triangular using East Asian allele frequencies as reference. Three distinct clusters (between PC2 and PC3) formed in the Panel A upper triangular; clusters in the Panel A lower triangle (between PC6 and PC3) are not distinct. There is no visible clustering in PCs from scGRM (Panel D lower triangle), although the coloring (Panel D upper triangle) suggested that samples in the same cluster in Panel A upper triangle tend to aggregate in PC3 from scGRM. In other words, using African alleles as reference to examine kinships between East Asian samples makes the discovery possible. The clustering was not caused by population structure in East Asians (Supplementary Figure S8). To investigate genetics variants that associated with the clustering, we derived a phenotype that cleanly separate samples into three clusters as shown in Figure 3B (details in Supplementary), and performed single SNP test using BIMBAM (Servin and Stephens, 2007). Interestingly, strong associations clustered around centromere region (Figure 3C). We used PC3 of scGRM as a derived phenotype (Figure 3E) and observed similar association pattern but with much reduced magnitude in $\log_{10}$ Bayes factors (Figure 3F). A recent study suggested that haplotype spanning centromere regions may be introgressed from archaic DNA (Langley et al., 2019), which renders a plausible explanation to why the clustering in East Asians is more distinct through the lens of African alleles.

Using a set of SNPs with small population divergence (SPD) to mimic allele frequencies of the ancestral population, Kindred can estimate kinship for admixed samples with high accuracy. (We chose not to use the traditional F_{st} to the measure the population divergence out of the consideration that F_{st} emphasize pairwise difference, while we wanted similar allele frequencies in at least three continental populations.) We recommend SPD SNPs to be used when analyzing samples with more than one continental origin. Arguably, when jointly analyzing diverse populations (in addition to analyzing admixed samples), using the SNPs of small population divergence is more sensible than all alternatives. We inferred realized kinship matrices for all 2504 samples in the 1000 Genomes project, using either 1.2 million SPD SNPs or 1.2 million randomly selected common SNPs. Figure 4 compared their top five PCs. PCs from SPD SNPs still cleanly separates continental populations, which suggests that although each individual SNPs are non-informative to ancestry, combined they are distinctly informative to ancestry. For PCs from randomly selected common SNPs, a striking feature is that PC1 was dominated by Africans samples, which have very little variation in other four leading PCs. This feature appears to be unique to Kindred.

Appendices

Bayes factor for linear mixed model

Consider a linear model

$$\begin{aligned} M1 : \mathbf{y} &= \mathbf{W}\alpha + \mathbf{e} \\ \alpha &\sim MVN(0, \tau^{-1} V_W) \\ \mathbf{e} &\sim MVN_n(0, \tau^{-1} I_n) \end{aligned} \tag{6}$$

where $\mathbf{W}$ is a $n \times w$ representing w nuisance covariates. Adding a random effect $\mathbf{u}$ we have a new model

$$\begin{aligned} M2Z : \mathbf{y} &= \mathbf{W}\alpha + \mathbf{Z}\mathbf{u} + \mathbf{e} \\ \mathbf{u} &\sim MVN_m(0, \tau^{-1} \eta K) \end{aligned} \tag{7}$$

where Z is $n \times m$ matrix presenting the loading, and K is $m \times m$ covariance matrix. Let Q and D be eigen decomposition such that $ZKZ^t = QDQ^t$, where $D = diag(d_1, \ldots, d_n)$ with $d_1 \geq d_2 \geq \cdots \geq d_n$ and $QQ^t = I$. Equations 7 can be rewritten

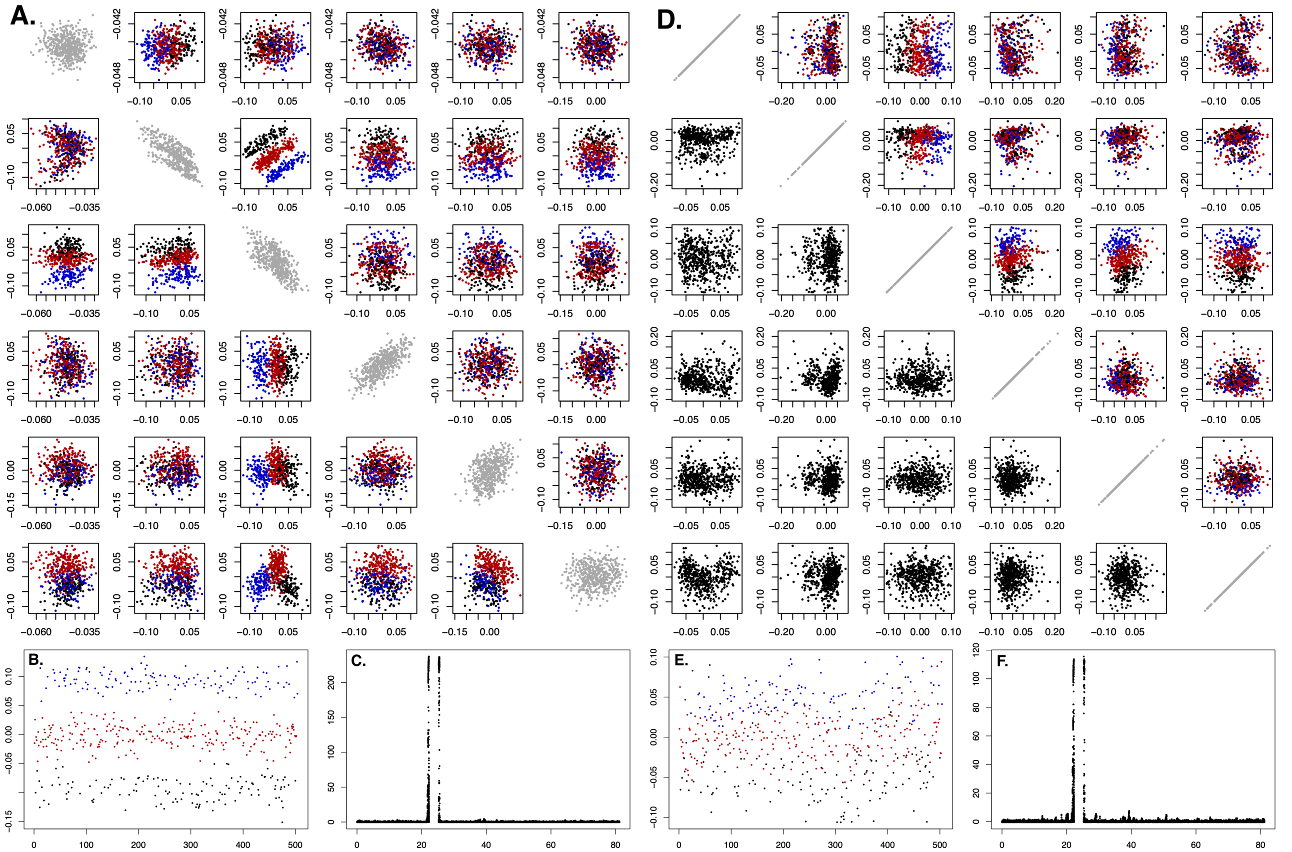


Fig. 3. East Asian samples clustering pattern on Chr17. Individuals were assigned colors based on PC2 and PC3 clustering in Panel A, and the same color assignments were used in Panels A,B,D,E. Panel A: pairwise plot of top six Kindred PCs. The upper triangular plots were PCs of kinship matrix inferred using African allele frequencies as reference. The lower triangular plots, East Asian allele frequencies as reference. (The diagonals are plots of j-th PC from one vs j-th PC from the other.) Panel B: phenotypes derived from PC2 and PC3 in upper triangular plots in Panel A. Panel C: Manhattan plot of single SNP ($\log_{10}$) Bayes factors between common biallelic SNPs on Chr17 and derived phenotypes in Panel B. Panel D: pairwise plots of top six scGRM PCs. Without coloring, samples form no distinct clusters (lower triangle plots). With coloring (upper triangle), three groups of samples aggregate. Panel E: derived phenotype from PC3 in Panel D. Panel F: Manhattan plot of single SNP ($\log_{10}$) Bayes factors between common biallelic SNPs on Chr17 and derived phenotypes in Panel E. Note the y-axis range in Panel F is half of that in Panel C.

as

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

where $V_Q = \eta D$. To see this, $E(Z\mathbf{u}\mathbf{u}^t Z^t) = \eta Z K Z^t = \eta Q D Q^t = E(Q\gamma\gamma^t Q^t)$.

Since models M1 and M2 are nested, it is understood that the distribution assumption and prior specification used in a simpler model carry over to the more complex model. By specifying a Gamma prior on τ we have normal-inverse-gamma on a linear model (c.f. Servin and Stephens, Zhou and Guan).

$$\tau \sim \Gamma(\kappa_1/2, \kappa_2/2) \tag{9}$$

It's clear that from a Bayesian perspective, a linear mixed model is just a linear model with a specific prior. For $M1$ after integrating out α and τ and letting $\kappa_1, \kappa_2 \to 0$, we have

$$\begin{aligned} p(y|\eta) = &\frac{(2\pi)^{-n/2}\Gamma(n/2)}{\det(W^tW + V_W^{-1})^{1/2}\det(V_W)^{1/2}} \\ &\cdot \left(\frac{\mathbf{y}^t\mathbf{y} - \mathbf{y}^t W(W^tW + V_W^{-1})^{-1}W^t\mathbf{y}}{2}\right)^{-n/2} \end{aligned} \tag{10}$$

Treating $M1$ as null and $M2$ as alternative, we compute BF_{21} in a closed form with the above prior specification. Denote $X = (Q, W)$ and $V = \begin{pmatrix} \eta D & 0 \\ 0 & V_W \end{pmatrix}$, we have

$$\begin{aligned} BF(\eta) = &\frac{\det(W^tW + V_W^{-1})^{1/2}}{\det(\mathbf{X}^t\mathbf{X} + V^{-1})^{1/2}\det(V_Q)^{1/2}} \\ &\cdot \left(\frac{\mathbf{y}^t\mathbf{y} - \mathbf{y}^t\mathbf{X}(\mathbf{X}^t\mathbf{X} + V^{-1})^{-1}\mathbf{X}^t\mathbf{y}}{\mathbf{y}^t\mathbf{y} - \mathbf{y}^t W(W^tW + V_W^{-1})^{-1}W^t\mathbf{y}}\right)^{-n/2}. \end{aligned} \tag{11}$$

Let $V_W \to \infty$, we have $V^{-1} = \begin{pmatrix} \eta^{-1}D^{-1} & 0 \\ 0 & 0 \end{pmatrix}$ and

$$\begin{aligned} BF(\eta) = &\frac{\det(W^tW)^{1/2}}{\det(\mathbf{X}^t\mathbf{X} + V^{-1})^{1/2}\det(V_Q)^{1/2}} \\ &\cdot \left(\frac{\mathbf{y}^t\mathbf{y} - \mathbf{y}^t\mathbf{X}(\mathbf{X}^t\mathbf{X} + V^{-1})^{-1}\mathbf{X}^t\mathbf{y}}{\mathbf{y}^t\mathbf{y} - \mathbf{y}^t W(W^tW)^{-1}W^t\mathbf{y}}\right)^{-n/2} \end{aligned} \tag{12}$$

Bayesian estimates of PVE.

Bayes factor (12) can be evaluated efficiently for different η. As $X^tX + V^{-1} = \begin{pmatrix} I_n + \frac{1}{\eta D} & Q^tW \\ W^tQ & W^tW \end{pmatrix}$, we compute its determinant

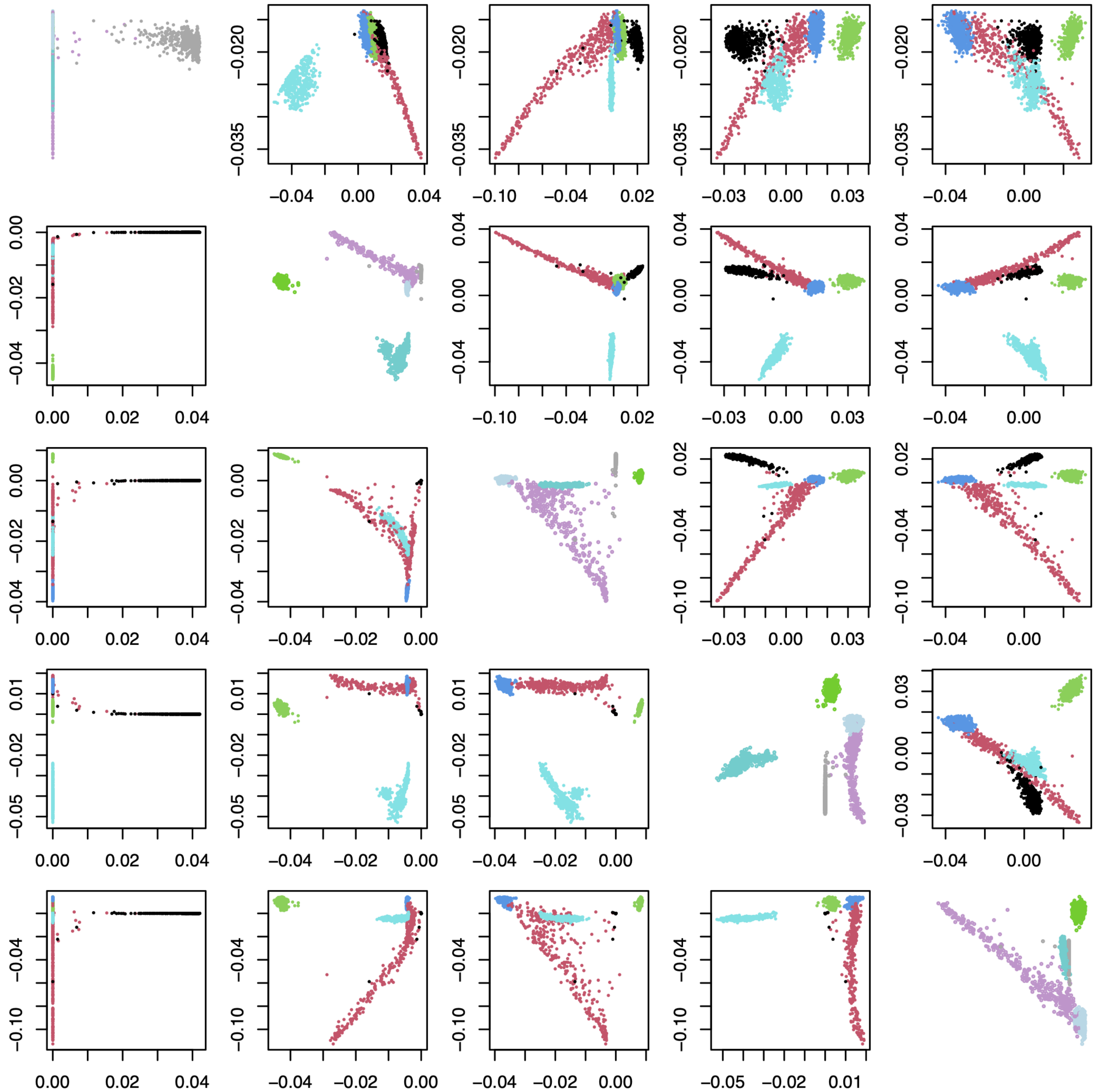


Fig. 4. PC plots of 1000 genomes samples. The upper triangle are pairwise plots of top five PCs inferred with small population divergence (SPD) SNPs. The lower triangle, randomly selected common (RSC) SNPs. The diagonals are plots of j-th PC from SPD SNPs vs j-th PC from RSC SNPs, with $j = 1, 2, 3, 4, 5$. Five continental samples are Africans (in black), Americans (in red), East Asians (in green), Europeans (in blue), and South asians (in cyan). On diagonal their colors are in gray, plum, light green, light blue, and cyan.

using the identity

$$\det\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \det(A)\det(D - CA^{-1}B)$$

to get

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

and its inverse using the identities

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1}+A^{-1}B(D-CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D-CA^{-1}B)^{-1} \\ -(D-CA^{-1}B)^{-1}CA^{-1} & (D-CA^{-1}B)^{-1} \end{pmatrix}.$$

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

$$(X^tX + V^{-1})^{-1} = \begin{pmatrix} F+FQ^tWMW^tQF & -FQ^tWM \\ -MW^tQF & M \end{pmatrix}.$$

Note these computations only involve inexpensive matrix multiplying vector, without expensive matrix multiplication and matrix inversion. The only expensive calculation is the eigendecomposition to obtain Q and D, which only needs to be done once. A R script to compute Bayes factor (Equation 12) is available in the Supplementary.

With efficient evaluation of BF at hand, we can find $\hat{\eta}$ that maximize $BF(\eta)$ by Nelder-Meade algorithm, and this is a MAP (maximum a posterior) estimator. With $\hat{\eta}$ we can compute $PVE = \hat{\eta}/(1+\hat{\eta})$. In Supplementary Figure S9, we demonstrate two properties of the PVE via simulations: 1) Bayesian estimates recovered the true PVE; and 2) Bayesian estimates are consistent with those GCTA estimates.

Supplementary information

Supplementary data are available at *Bioinformatics* online.

Acknowledgements

We thank Nick Martin and the Queensland Institute of Medical Research for making the Australia height data available to us. We thank editors and reviewers whose comments greatly improved the presentation of the paper.

Conflicts of interests

The authors declare no competing interests.

Funding

This research is supported by the Division of Intramural Research of the National Heart, Lung, and Blood Institute, Bethesda, MD (D.L. Principal Investigator).

Author contributions

Y.G. conceived the study, developed the methodology, implemented the computational software and performed experiments, analyzed results, and wrote the manuscript. D.L. supervised the work, provided access to data and computation, edited and approved the final manuscript.

Data and code availability

This study analyzed existing datasets from public domain and simulated datasets. Data from the 1000 Genomes project can be found at `https://www.internationalgenome.org/`, and datasets from Framingham Heart Study can be obtained via dbGaP. The software for Kindred and its source code are freely available at `https://github.com/haplotype/kindred`.

References

Auton, A., G. R. Abecasis, D. M. Altshuler, R. M. Durbin, D. R. Bentley, A. Chakravarti, A. G. Clark, and et al. (2015). A global reference for human genetic variation. *Nature 526*(7571), 68–74.

Bro, R. and S. De Jong (1997). A fast non-negativity-constrained least squares algorithm. *Journal of Chemometrics 11*(5), 393–401.

Chen, H., C. Wang, M. P. Conomos, A. M. Stilp, Z. Li, T. Sofer, A. A. Szpiro, W. Chen, J. M. Brehm, J. C. Celedón, S. Redline, G. J. Papanicolaou, T. A. Thornton, C. C. Laurie, K. Rice, and X. Lin (2016). Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *The American Journal of Human Genetics 98*(4), 653–666.

Devlin, B. and K. Roeder (1999). Genomic control for association studies. *Biometrics 55*(4), 997–1004.

Goudet, J., T. Kay, and B. S. Weir (2018, October). How to estimate kinship. *Mol. Ecol. 27*(20), 4121–4135.

Ho, J. E., A. Lyass, P. Courchesne, G. Chen, C. Liu, X. Yin, S.-J. Hwang, J. M. Massaro, M. G. Larson, and D. Levy (2018, July). Protein biomarkers of cardiovascular disease and mortality in the community. *J. Am. Heart Assoc. 7*(14).

Jacquard, A. (1972). Genetic information given by a relative. *Biometrics 28*(4), 1101–1114.

Jiang, W., X. Zhang, S. Li, S. Song, and H. Zhao (2022). An unbiased kinship estimation method for genetic data analysis. *BMC Bioinformatics 23*(1), 525.

Kang, H. M., J. H. Sul, Service, Susan K, N. A. Zaitlen, S.-Y. Kong, N. B. Freimer, C. Sabatti, and E. Eskin (2010, March). Variance component model to account for sample structure in genome-wide association studies. *Nat Genet 42*(4), 348–354.

Kannel, W. B., M. Feinleib, P. M. McNamara, R. J. Garrison, and W. P. Castelli (1979, September). An investigation of coronary heart disease in families. the framingham offspring study. *Am. J. Epidemiol. 110*(3), 281–290.

Langley, S. A., K. H. Miga, G. H. Karpen, and C. H. Langley (2019, June). Haplotypes spanning centromeric regions reveal persistence of large blocks of archaic DNA. *Elife 8*.

Lawson, C. and R. Hanson (1974). Prentice Hall, Englewood Cliffs, NJ.

Manichaikul, A., J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale, and W.-M. Chen (2010, 10). Robust relationship inference in genome-wide association studies. *Bioinformatics 26*(22), 2867–2873.

Milligan, B. G. (2003, 03). Maximum-Likelihood Estimation of Relatedness. *Genetics 163*(3), 1153–1167.

Ochoa, A. and J. D. Storey (2021, 01). Estimating fst and kinship for arbitrary population structures. *PLOS Genetics 17*(1), 1–36.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics 38*(8), 904–909.

Sara McPeek, M., X. Wu, and C. Ober (2004). Best linear unbiased allele-frequency estimation in complex pedigrees. *Biometrics 60*(2), 359–367.

Servin, B. and M. Stephens (2007, 07). Imputation-based analysis of association studies: Candidate regions and quantitative traits. *PLOS Genetics 3*(7), 1–13.

Taliun, D., D. N. Harris, M. D. Kessler, J. Carlson, Z. A. Szpiech, R. Torres, and et al. (2021, February). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed program. *Nature 590*(7845), 290–299.

Thompson, E. A. (2013, 06). Identity by Descent: Variation in Meiosis, Across Genomes, and in Populations. *Genetics 194*(2), 301–326.

Thornton, T. and M. S. McPeek (2010, February). ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure. *Am J Hum Genet 86*(2), 172–184.

Thornton, T., H. Tang, T. J. Hoffmann, H. M. Ochs-Balcom, B. J. Caan, and N. Risch (2012, July). Estimating kinship in admixed populations. *Am. J. Hum. Genet. 91*(1), 122–138.

Varin, C., N. Reid, and D. Firth (2011). An overview of composite likelihood methods. *Statistica Sinica 21*(1), 5–42.

Visscher, P. M., S. E. Medland, M. A. R. Ferreira, K. I. Morley, G. Zhu, B. K. Cornes, G. W. Montgomery, and N. G. Martin (2006, 03). Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLOS Genetics 2*(3), 1–10.

Wang, B., S. Sverdlov, and E. Thompson (2017, January). Efficient estimation of realized kinship from single nucleotide polymorphism genotypes. *Genetics 205*(3), 1063–1078.

Wang, J. (2016, February). Pedigrees or markers: Which are better in estimating relatedness and inbreeding coefficient? *Theor. Popul. Biol. 107*, 4–13.

Weir, B. S. and J. Goudet (2017, August). A unified characterization of population structure and relatedness. *Genetics 206*(4), 2085–2103.

Wright, S. (1922). Coefficients of inbreeding and relationship. *Am Nat 56*, 330–338.

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher (2010, June). Common SNPs explain a large proportion of the heritability for human height. *Nat Genet 42*(7), 565–569.

Yang, J., S. H. Lee, M. E. Goddard, and P. M. Visscher (2011, 2022/12/25). Gcta: A tool for genome-wide complex trait analysis. *The American Journal of Human Genetics 88*(1), 76–82.

Zheng, X., D. Levine, J. Shen, S. M. Gogarten, C. Laurie, and B. S. Weir (2012, 10). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics 28*(24), 3326–3328.

Zhou, X. and M. Stephens (2012, June). Genome-wide efficient mixed-model analysis for association studies. *Nat Genet 44*(7), 821–824.