Correcting statistical bias in correlation-based kinship estimators

Wei Jiang[1]❂, Xiangyu Zhang[2]❂, Siting Li[3], Shuang Song[4,5], Hongyu Zhao[1]*

1 Department of Biostatistics, School of Public Health, Yale University, New Haven, Connecticut, United States of America
2 Department of Statistics, University of Science and Technology of China, Hefei, Anhui, China
3 Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College, Hanover, New Hampshire, United States of America
4 Center for Statistical Science, Tsinghua University, Beijing, China
5 Department of Industrial Engineering, Tsinghua University, Beijing, China

❂These authors contributed equally to this work.
* E-mail: hongyu.zhao@yale.edu

Abstract

Accurate estimate of relatedness is important for genetic data analyses, such as association mapping and heritability estimation based on data collected from genome-wide association studies. Inaccurate relatedness estimates may lead to spurious associations and biased heritability estimations. Individual-level genotype data are often used to estimate kinship coefficient between individuals. The commonly used sample correlation-based genomic relationship matrix (scGRM) method estimates kinship coefficient by calculating the average sample correlation coefficient among all single nucleotide polymorphisms (SNPs), where the observed allele frequencies are used to calculate both the expectations and variances of genotypes. Although this method is widely used, a substantial proportion of estimated kinship coefficients are negative, which are difficult to interpret. In this paper, through mathematical derivation, we show that there indeed exists bias in the estimated kinship coefficient using the scGRM method when the observed allele frequencies are regarded as true frequencies. This leads to negative bias for the average estimate of kinship among all individuals, which explains the estimated negative kinship coefficients. Based on this observation, we propose an unbiased estimation method, UKin, which can reduce the bias. We justify our improved method with rigorous mathematical proof. We have conducted simulations as well as two real data analyses to demonstrate that both bias and root mean square error in kinship coefficient estimation can be reduced by using UKin. Further simulations indicate that the power in association mapping can also be improved by using our unbiased kinship estimates to adjust for cryptic relatedness.

Author summary

Inference of relatedness plays an important role in genetic data analysis. Many methods have been proposed to estimate kinship coefficients, including the commonly used genomic relationship matrix method. However, a substantial proportion of the kinship coefficients estimated by this method are negative, which is difficult to interpret. In this paper, through mathematical derivation, we show that there indeed exists a negative bias in this approach. To correct for this bias, we propose a new kinship coefficient estimation method, UKin, which is unbiased without requiring extra genetic information nor added computational complexity. The better performance of UKin in reducing bias and root mean squared error is demonstrated through theory, simulations and analysis of data from the young-onset breast cancer and familial intracranial aneurysm studies.

Introduction

Accurate estimation of relatedness among individuals is important in genetic data analysis. For example, in both population-based and family-based genome-wide association studies (GWAS) with uncertain relationships among study subjects, it is critical to appropriately account for cryptic relatedness because incorrect estimates can decrease power and inflate false positive rates of association tests [1–3]. Several methods have been proposed to adjust for relatedness in GWAS, such as introducing a genomic relationship matrix (GRM) as an augment into

well-developed linear mixed model (LMM) [4–7]. It has been demonstrated that proper consideration of genetic relatedness can also benefit heritability estimation based on GWAS data in the presence of pedigree structures [8, 9].

In order to adjust for cryptic relatedness in genetic studies like association mapping and heritability estimation, individual-level genotype data are often used to estimate pairwise kinship coefficients. The sample correlation-based genomic relationship matrix (scGRM) method estimates kinship coefficient by calculating the average sample correlation coefficient among all genetic variants, in which the observed allele frequencies are used for the calculation of both expectation and variance of genotypes [10–12]. We note that most association mapping and heritability estimation packages use this method as their default setting for calculating GRM, such as GCTA, GEMMA and FaSTLMM [6, 8, 13]. Although this method is widely used, researchers have noted that a substantial proportion of the estimated kinship coefficients are negative. As kinship coefficient is defined to be a positive number (see in Materials and Methods), it is difficult to interpret these negative estimates [14–16].

In this paper, through mathematical derivation, we first show that there indeed exists bias in the estimated kinship coefficients using the scGRM method. The bias exists because the observed allele frequencies are regarded as true frequencies. We also prove analytically that the bias essentially results in a negative average for all estimates, which explains the large proportion of negative values. Based on this observation, we propose an improved kinship estimation method, UKin, which can remove bias. We provide a mathematical proof for the unbiasedness of the UKin estimator. Simulations and real data analyses also demonstrate that both bias and root mean square error (RMSE) can be reduced by replacing the scGRM method with our UKin method. For real data analyses, we apply our method to two studies, young-onset breast cancer (BC) and familial intracranial aneurysm (FIA), which have pedigree information to evaluate our results. Finally, as an application of our method in association mapping, we conduct a simulation study to show the power of detecting genetic associations can be improved by correcting cryptic relatedness using our unbiased kinship estimates.

The paper is organized as follows. In the Results section, we evaluate the performance of UKin through two simulations and two real data sets in BC and FIA to validate our theoretical derivation and demonstrate the effectiveness of UKin estimator in reducing bias and RMSE. In the Materials and Methods section, we present the theoretical details which show the scGRM method is biased, propose our UKin estimation method and give the correctness proof, as well as its connection with the scGRM estimator. Lastly, we conclude with a simulation study in the Discussion section to demonstrate that UKin method can further improve the power of association mapping. Technical details such as mathematical derivations are provided in S1 Appendix.

Results

1. Simulation experiments

An illustrative example

We start our discussion with a simple but extreme example. In this experiment, we assumed that there were 500 full siblings from the same family. Although unlikely to exist in reality, this

example serves as a good illustration of our theoretical derivation. As every two individuals selected from the same family were full siblings, the true value of their kinship coefficient should be 0.25 (see in Table 5). However, following Property 3 in the Materials and Methods section, their average kinship coefficient estimated by scGRM, denoted by $\bar{\bar{\phi}}$, should have the expectation:

$$E\bar{\bar{\phi}} = \frac{n-1}{4}(\bar{\rho}-1)/(\frac{n(n-1)}{2}) = \frac{\bar{\rho}-1}{2n} = \frac{0.5-1}{2 \times 500} = -5 \times 10^{-4},$$

where n is the sample size and $\bar{\rho}$ is the average of their true genetic correlation coefficients. Property 1 together with Table 5 in the Materials and Methods section suggest that $\bar{\rho} = 0.5$ for full siblings.

This result shows the unexpected phenomenon that although all individuals in our simulated samples are full siblings to each other, the average of the estimated kinship coefficients has a negative value. To illustrate Property 3 in practice, we simulated 200 unrelated families each consisting of 500 full siblings with the method provided by the package CorBin [17]. Each individual was genotyped at 10,000 single nucleotide polymorphisms (SNPs). Following the scGRM method and the UKin method proposed in Materials and Methods, we estimated pairwise kinship coefficients and calculated their mean values, respectively. The histograms of these estimated average kinship coefficients are shown in Fig 1. From this plot, we could see the distribution of average kinship estimated by the scGRM method centered around -5×10^{-4}, which is consistent with our expectation from the analytical results. By contrast, the UKin approach performed better in dealing with this extreme situation, with the average estimates centered at 0.25, the true value of pairwise kinship coefficient for full-sibling pairs. Besides, from Fig 1 we could observe that the two distributions have similar shapes, which could be explained by Equation 5 in Materials and Methods which suggests that unbiased estimator of correlation coefficient $\tilde{\rho}_{ii'}$ could be expressed as a linear combination of the scGRM estimators $\hat{\rho}_{ii'}$. Considering there were 500 full siblings from the same family, we calculated the average on both sides of Equation 5 among all the simulated individual pairs, which is $\bar{\tilde{\rho}} = 500\bar{\hat{\rho}} + 1$, where $\bar{\tilde{\rho}}$ and $\bar{\hat{\rho}}$ represent the average of correlation coefficients between full siblings from the same family, estimated by the UKin method and the scGRM method respectively, i.e.

$$\bar{\tilde{\rho}} = \frac{\sum\limits_{i=1}^{n}\sum\limits_{i'=i+1}^{n}\tilde{\rho}_{ii'}}{n(n-1)/2} \qquad \bar{\hat{\rho}} = \frac{\sum\limits_{i=1}^{n}\sum\limits_{i'=i+1}^{n}\hat{\rho}_{ii'}}{n(n-1)/2}.$$

As there was a linear relationship between kinship coefficient and correlation coefficient (see Property 1 in Materials and Methods), the distributions of the average kinship coefficients estimated by the two methods should have the same shape.

A more general simulation

To evaluate the performance of the UKin method in kinship coefficient estimation and to compare it with the scGRM estimate in a more general situation, we performed the following simulations in which population homogeneity was assumed. To include different kinds of relationships in our experiment, we simulated 6,000 people including 1,000 pairs with kinship coefficient 0.125, 1,000 pairs with coefficient 0.25, and 500 pairs with coefficient 0.5. For
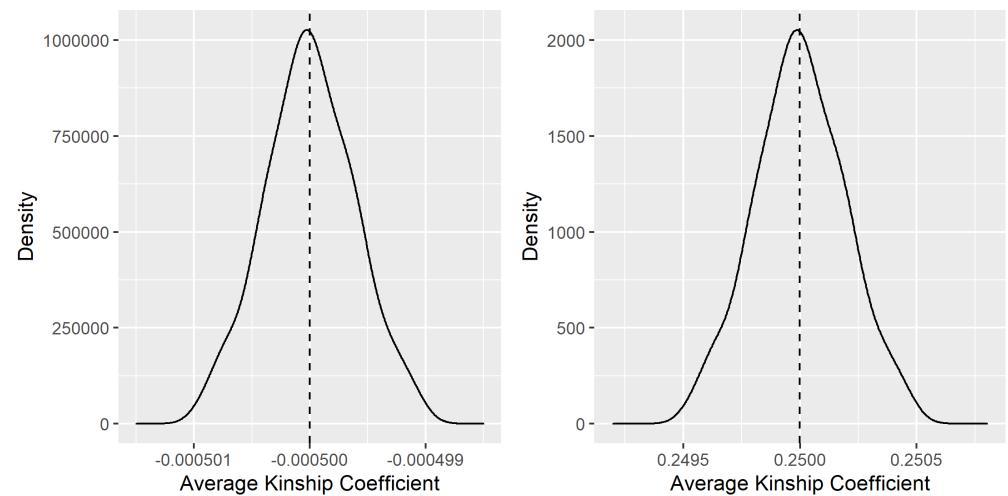
Fig 1. Distribution of average kinship coefficients estimated by the scGRM (left) and UKin (right) methods in this extreme example. Two hundred unrelated families each consisting of 500 full siblings were simulated, with each sibling genotyped at 10,000 SNPs. The averages of kinship coefficients among all individual pairs from the same family were calculated and the distribution of these averages is displayed. The true value of kinship coefficient between full siblings is 0.25. The vertical dashed line in each plot corresponds to the mean value of these averages estimated by the corresponding method.

simplicity, different relative pairs were set to be unrelated. In addition, we also included 1,000 people who had no relationship with other individuals. For each subject, genotype data were generated for 10,000 random and independent SNPs. The minor allele frequencies (MAFs) of genotyped variants were drawn uniformly from $[0.05, 0.5]$.

With the UKin and scGRM estimators, we estimated kinship coefficients between all simulated individual pairs and divided those coefficients into four groups according to their true relationships. Fig 2 shows the distribution of the estimated kinship coefficients in each group respectively. As shown in this plot and summarized in Table 1, for groups with true kinship coefficient 0.25 and 0.5, our UKin method achieved lower RMSE than the scGRM method in estimating kinship coefficients, while the opposite was true for the independent pairs. For the group consisting of pairs having kinship coefficients 0.125, the two methods had similar performance.

Although Fig 2 clearly demonstrates the RMSE for the two methods, it is difficult to compare their biases from the plots. More detailed comparisons are shown in Table 1. As shown in the second column of this table, UKin always performed better than scGRM when we compared the mean values of estimated kinship coefficients, as the results of the UKin method were closer to true values for all four groups. Besides, results in the third column of Table 1 show that UKin could reduce RMSE for close relatives, which is consistent with the conclusion we get from Fig 2. Furthermore, UKin shows a downward trend of RMSE with increasing true kinship coefficients, while scGRM is completely on the opposite. It is also notable that when we consider individual pairs with kinship coefficient 0.5, i.e. Monozygotic twins (M-Z twins), both bias and RMSE are extremely close to zero if we utilize UKin to estimate.
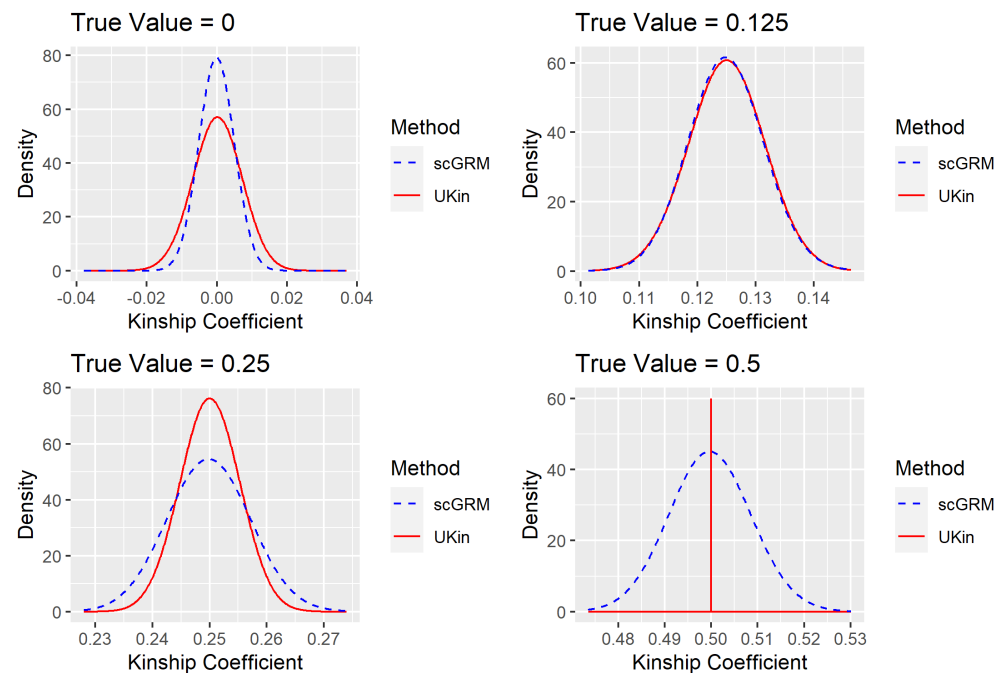
2. Real data analyses

Fig 2. Distributions of kinship coefficients estimated by the UKin method and the scGRM method in our simulation study including 6,000 individuals with different relationships. The four plots correspond to the four groups divided by the true value of estimated kinship coefficients.

Table 1. Comparison of UKin and scGRM in biases and RMSEs

True Value	Bias from True Value$(\times 10^{-5})$		Root Mean Square Error$(\times 10^{-3})$	
	(UKin)	(scGRM)	(UKin)	(scGRM)
0.000	**3.943**	-11.81	6.915	**5.000**
0.125	**1.543**	-7.486	5.724	**5.668**
0.250	**3.414**	-13.80	**4.537**	6.329
0.500	**0.000**	-22.82	**0.000**	7.431

The Young-Onset Breast Cancer Study

To demonstrate our unbiased method could get more accurate results in estimating kinship coefficients, we applied the UKin method to real data from a family-based study of genes and environment in young-onset BC (*dbGaP Study Accession: phs000678.v1.p1*). This study recruited families from the US and Puerto Rico with a daughter who was recently diagnosed with breast cancer and another unaffected daughter. For each family, only the diseased daughter and her unaffected full sister were genotyped for analysis. As for data quality control, we removed individuals with more than 10% missing genotypes as well as SNPs with a missing genotype rate greater than 5% or a minor allele frequency less than 5%. After further removing individuals with missing phenotypes, we got 1,983 subjects (1,458 cases and 525 controls) with 614,310 variants in total. The processed data included 511 pairs of full sisters, with one affected by breast cancer. We assumed individuals from different families were unrelated, then the true

101
102
103
104
105
106
107
108
109
110
111
112

values of all estimated kinship coefficients should be either 0.25 (511 full sister pairs) or 0 (all the other individual pairs). 113 114

We first applied the scGRM method to estimate the kinship coefficients, which had poor performance. For the 511 within-family pairs (full sister pairs), only 473 pairs were estimated to have kinship coefficients between $2^{-5/2}$ and $2^{-3/2}$, which means 7.4% of full sisters were incorrectly inferred to be other kinds of relative pairs. Estimation of $1,964,642 = \binom{1,983}{2} - 511$ between-family pairs with the scGRM method also performed poorly, as 0.45% unrelated pairs between families were misspecified as 1st-degree relative pairs (such as sibling pairs), 2nd-degree relative pairs (such as half-sibs, avuncular pairs and grandparent-grandchild pairs) or 3rd-degree relative pairs (such as first cousins). In contrast, our UKin method had more accurate estimates with 501 out of 511 (98.0%) full sisters pairs correctly estimated and only 177 unrelated pairs (less than 0.01%) were misspecified as 3rd-degree relative pairs (Table 2 and Table 3). 115 116 117 118 119 120 121 122 123 124 125

Table 2. Distribution of estimated kinship coefficients of 511 full siblings in the two sister data studying young-onset breast cancer

Relationship	unrelated	3rd-degree relative pairs	2nd-degree relative pairs	**1st-degree relative pairs**	M-Z twins
Inference criteria	$< \frac{1}{2^{9/2}}$	$\left(\frac{1}{2^{9/2}}, \frac{1}{2^{7/2}}\right)$	$\left(\frac{1}{2^{7/2}}, \frac{1}{2^{5/2}}\right)$	$\left(\frac{1}{2^{5/2}}, \frac{1}{2^{3/2}}\right)$	$> \frac{1}{2^{3/2}}$
scGRM	0	0	6	**473**	32
UKin	0	0	10	**501**	0
True	0	0	0	**511**	0

Table 3. Distribution of estimated kinship coefficients of 1,964,642 unrelated individual pairs in the two sister data studying young-onset breast cancer

Relationship	**unrelated**	3rd-degree relative pairs	2nd-degree relative pairs	1st-degree relative pairs	M-Z twins
Inference criteria	$< \frac{1}{2^{9/2}}$	$\left(\frac{1}{2^{9/2}}, \frac{1}{2^{7/2}}\right)$	$\left(\frac{1}{2^{7/2}}, \frac{1}{2^{5/2}}\right)$	$\left(\frac{1}{2^{5/2}}, \frac{1}{2^{3/2}}\right)$	$> \frac{1}{2^{3/2}}$
scGRM	**1,955,711**	2,129	2,081	4,721	0
UKin	**1,964,465**	177	0	0	0
True	**1,964,642**	0	0	0	0

The histograms of kinship coefficients estimated by scGRM and UKin for all pairs (including both full sister pairs and unrelated pairs) in the BC study are given in Fig 3. To make the comparison more clearly, we only took individual pairs with estimated kinship coefficients between 2^{-5} and $2^{-1.5}$ into consideration. It is obvious that the histogram corresponding to scGRM contains more pairs with estimated kinship coefficients larger than $2^{-7/2}$. From previous analysis, we know most of them are misspecified unrelated pairs. In contrast, our approach is much less likely to make such mistakes. Besides, the UKin histogram shows a peak centered close to 0.25 and has a distinct separation from estimated kinship coefficients near zero. However, scGRM does not work well in this aspect because the distribution of non-zero kinship coefficients is centered around 0.23 and has an obvious distribution overlap with unrelated pairs, 126 127 128 129 130 131 132 133 134 135

which suggests that UKin performs better in separating relatives from unrelated pairs. Furthermore, if we only consider the 511 full sister pairs, the true kinship coefficient should be 0.25. However, the average and mean square error (MSE) of the kinship coefficients estimated by scGRM were 0.263 and 4.03×10^{-3}, respectively. In contrast, the corresponding UKin results were 0.248 and 6.72×10^{-4}, respectively. To visualize the difference between UKin and scGRM, we also draw the scatter plot of the estimated kinship coefficients for the 511 full sister pairs between the two methods (Fig 4). The scatter plot demonstrates that while the distribution of UKin estimates is more concentrated at its true value, scGRM tends to overestimate the kinship coefficients for many full sister pairs. These results clearly show the better performance of UKin than scGRM.
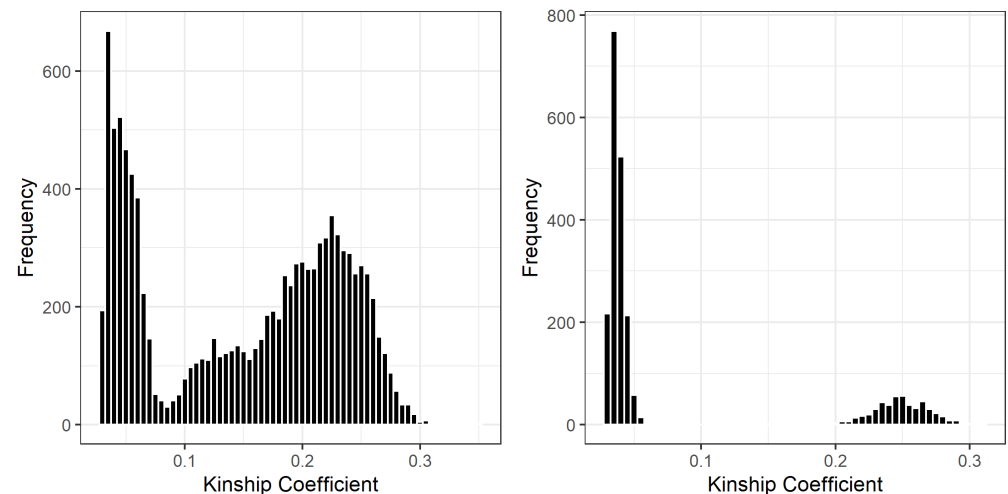

Fig 3. Comparison of distributions of kinship coefficients estimated by scGRM (left) and UKin (right) in breast cancer study. This study genotyped 1,983 individuals at 614,310 variants. Pairwise relationships in this dataset included 511 full sister pairs from irrelevant families and unrelated pairs. In this figure, we only considered estimated kinship coefficients between 2^{-5} and $2^{-1.5}$. Class interval of the histogram for each method is set to be 0.005.

The Familial Intracranial Aneurysm Linkage Study

To further investigate the effectiveness of the UKin method in kinship coefficient estimation, we applied UKin to infer pedigree structure using genotype data from the FIA linkage study (*dbGaP Study Accession: phs000293.v1.p1*). This study recruited 400 families with multiple individuals who have an intracranial aneurysm (IA) through 23 (25) referral centers throughout North America, Australia, and New Zealand that represent 35 (40) recruitment sites. After a standardized procedure of quality control and discarding subjects with missing phenotype, we obtained 990 individuals from 371 families and each of them was genotyped at 5,505 SNPs. In this FIA dataset, the confirmed relationships include 137 1st-degree relative pairs (including 19 full siblings and 118 parent-child pairs).

We compared the performance of UKin and scGRM in identifying these 1st-degree relative pairs and estimating their kinship coefficients. UKin was able to correctly recognize all the 137 1st-degree pairs (with estimated kinship coefficients between $2^{-2.5}$ and $2^{-1.5}$), while scGRM misspecified one parent-child pair as monozygotic twins, with an estimated kinship coefficient
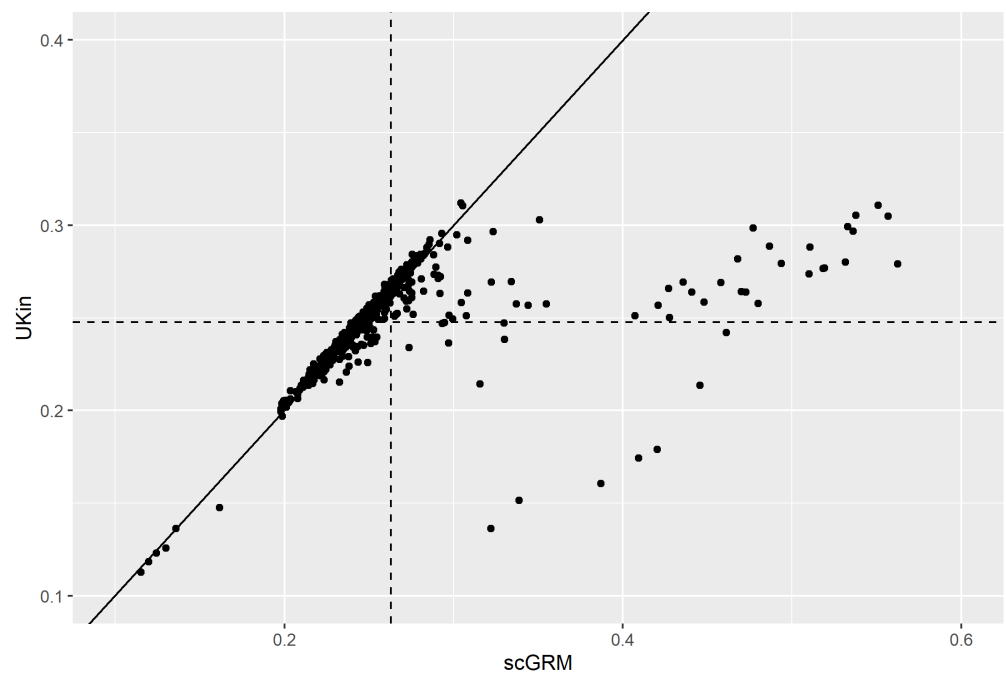
Fig 4. The scatter plot of the estimated kinship coefficients between the UKin and scGRM methods. For this plot we only consider the 511 full sister pairs in the BC data set. The oblique solid line stands for the equation $y = x$, while the vertical and horizontal dashed lines correspond to the mean values of scGRM and UKin estimates, respectively.

of 0.442. The histograms of the kinship coefficients of these 137 individual pairs estimated by   160
the two methods (Fig 5) indicate that unbiased estimations are more concentrated, taking values   161
between 0.21 and 0.3. However, the distribution of scGRM estimations is more dispersed,   162
including a distinct outlier. This fact is more clearly shown in the scatter plot including all the   163
137 1st-degree pairs in the IA data set (Fig 6). We further calculated the bias from the true value   164
(0.25) and RMSE of the estimated coefficients for each estimator. As summarized in Table 4, the   165
estimation bias of UKin was 1/6 of the bias estimated by scGRM, while the RMSE of UKin was   166
half of scGRM. We also note that scGRM misspecified 15 parent-child pairs or unrelated pairs   167
as MZ twins, while UKin only made five such mistakes which were all included in the   168
misspecified pairs of scGRM. These results demonstrate that our UKin method achieves more   169
accurate outcomes in relationship inference and kinship estimation, even when the number of   170
genotyped SNPs is small.   171

Table 4. Bias and RMSE of estimated kinship coefficients for the 137 pairs of 1st-degree relatives in the FIA study

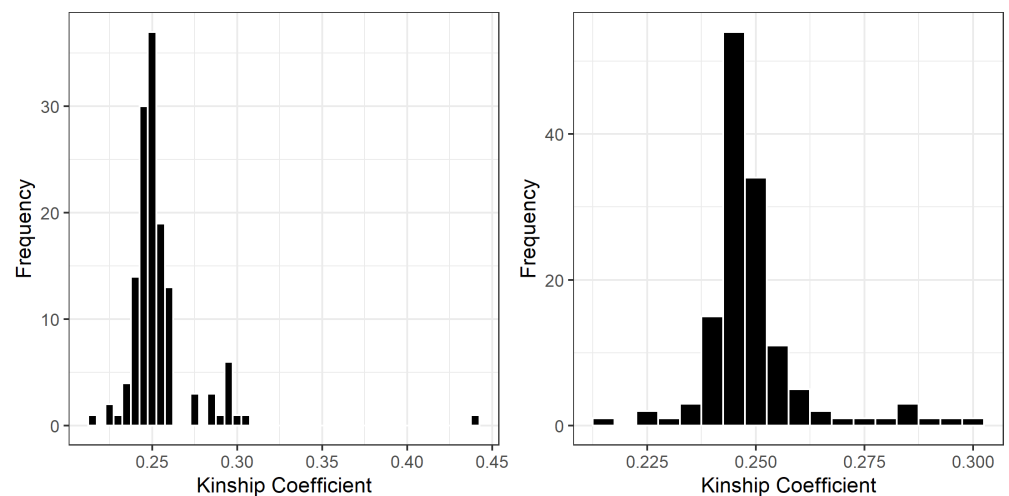| Estimation Method	Bias($\times 10^{-3}$)	RMSE($\times 10^{-2}$)
UKin	0.667	1.175
scGRM	4.045	2.244

Fig 5. Distributions of estimated kinship coefficients of 1st-degree relatives in the FIA study with scGRM (left) and UKin (right). Among all the 137 1st-degree relative pairs in this dataset, there are 19 full siblings and 118 parent-child pairs. Class interval of the histogram for each method is set to be 0.005.

Discussion

Among the many kinship coefficient estimation methods, the most commonly applied estimator uses dense SNP genotypes and allele frequencies in the samples to calculate average pairwise correlation coefficients among SNPs. Although this method is intuitive and easy to calculate, we prove that it is actually biased because it treats the observed allele frequencies as true frequencies. Through rigorous derivation, we showed that pairwise kinship coefficients estimated by scGRM add up to be a negative value, which explains the phenomenon that a substantial proportion of kinship coefficient estimates are negative.

When conducting large scale estimates of kinship coefficients, the existing bias in scGRM can lead to incorrect inference of relationships, and this problem can be extremely severe if the subjects in the dataset are highly related. Our method, UKin, solved this issue by incorporating genetic information from the whole population to adjust for the bias in the estimated kinship coefficient between every single pair. This unbiased estimator can be expressed as a polynomial of scGRM estimators, and leveraging only information of dense genotypes from the population. As demonstrated by our simulations and applications to the BC and FIA family data, UKin performed better in reducing both estimation bias and RMSEs. For the two sister study, the results suggest that while scGRM could lead to severe spurious inference of relative pairs, UKin rarely made such mistakes. Even when the number of genotyped SNPs was limited for the FIA study, UKin could reduce statistical bias and RMSE while avoiding spurious relationship inference.

In our theoretical derivations and simulation studies, we made assumptions like linkage equilibrium (LE) and absence of inbreeding, that is, genotypes at different markers are independent. During our derivation, we used the same weights for all SNPs, and our simulated datasets were also generated under this assumption. Although there is linkage disequilibrium (LD) in reality, empirical results from the analyses of the BC and FIA family data show that the
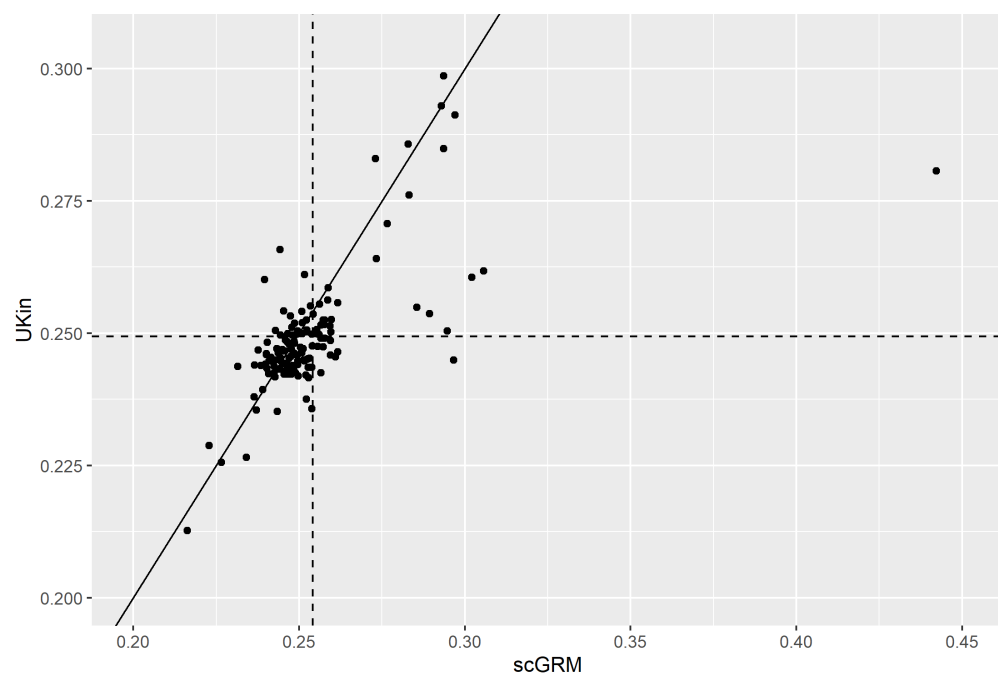
Fig 6. The scatter plot of the estimated kinship coefficients between the UKin and scGRM methods in the FIA study. Only the 137 1st-degree pairs are shown. The oblique solid line stands for the equation $y = x$, while the vertical and horizontal dashed lines correspond to the mean values of scGRM and UKin estimates, respectively.

bias and RMSEs can also be reduced greatly with the application of UKin to real data. To consider the problem of LD in practice, we can give different weights based on LD to these SNPs. Following the approach of Wang (2017) [15], these LD weights $\mathbf{w} = (w_1, w_2, ..., w_m)^T$ can be calculated by solving the following minimization problem:

$$\min_{\mathbf{w}}[\mathbf{w}^T \mathbf{R} \mathbf{w} - \mathbf{w}^T 1] \quad : \quad w_l \geq 0, \forall l.$$

where $\mathbf{R} = [\rho_{lk}^2]$ is the matrix of squared LD correlations. Theoretically, this result can be directly applied to UKin by assigning the correlation coefficient at each SNP marker its corresponding weight, which might make our approach adapt to LD situation.

Another assumption throughout our study is a homogeneous population so that the allele frequencies can be calculated once and applied to all subjects. Some methods have been proposed to estimate kinship coefficients in admixed populations, where the assumption of population homogeneity is untenable [11, 18, 19]. However, as most of these methods are based on the scGRM method, they are also likely to be biased estimators, too. How to extend our UKin method to deal with admixed populations is a topic for future studies.

More accurate kinship estimation will improve the performance of different genetic analyses such as association mapping. In recent years, GWAS have seen great success in identifying genetic loci contributing to complex human traits [20, 21]. By studying a genome-wide data set of genetic variants in different individuals, GWAS looks for SNPs correlated with traits in the samples. Accurate specification of familial relationships is expected to bring more powerful

association results in GWAS with unknown (or unrecognized) family structure. We have investigated whether association mapping can be improved by applying UKin to account for cryptic relatedness.

We conducted a simulation study to compare the performance of UKin with scGRM in GWAS. In our experiments, we simulated 4,000 samples including 2,000 cases and 2,000 controls. We included subjects with various pairwise kinship coefficients in both cases and controls. More specifically, we simulated 250 1st-degree relative pairs, 250 2nd-degree relative pairs, 250 3rd-degree relative pairs, and 500 unrelated subjects. The total number of SNPs genotyped for each individual was set to be 10,000 and the MAFs of non-risk SNPs were drawn uniformly from $[0.05, 0.5]$. The proportion of risk SNPs was set at 0.05 or 0.1. For these risk SNPs, a variable following the Gaussian distribution $\mathcal{N}(0, 0.05^2)$ was added to the previous uniform distribution to obtain their MAFs in cases. We set those MAFs below 0.05 or greater than 0.95 to be 0.05 and 0.95, respectively.

We applied GEMMA [13], which was developed to implement the genome-wide mixed model association algorithm for a standard linear mixed model for association analysis. In our simulations, we performed likelihood ratio tests in a univariate LMM for marker association mappings with a single phenotype. PLINK binary file format was [22] adopted as input files containing phenotypes and genetic information. A standardized relatedness matrix file estimated by either scGRM or UKin was included to appropriately account for relatedness among subjects.

We applied GEMMA to analyze the simulated GWAS dataset and selected all SNPs with P-value below the threshold 5×10^{-4}. Statistical power and type I error rate were calculated to evaluate the performance of marker association tests when the relatedness matrix used in LMMs was estimated by scGRM and UKin, respectively. The results suggest that the type I error rate was appropriately controlled at a low level (less that 5×10^{-4}) for both methods. We compared the power of association mapping which suggests that for the two risk SNP proportions considered (i.e. 0.05 and 0.1), the power of identifying risk variants was always improved after we replaced scGRM with UKin in estimating pairwise kinship coefficients. For example, when the proportion of risk SNPs was set at 0.05, the power was improved from 0.154 to 0.167 by adopting the UKin method. This simulation demonstrates that the application of UKin can improve statistical power while controlling the type I error rate in GWAS. However, further simulations and real data experiments are required to evaluate the advantages of UKin over the scGRM comprehensively, which is the subject of future research.

Materials and Methods

Alleles are said to be identical by descent (IBD) if they are inherited from a common ancestor. To describe the average amount of IBD sharing at the genome level, we often adopt the concept of kinship coefficient [12]. For two individuals indexed by a and b, their kinship coefficient, ϕ_{ab}, is defined as the probability that two alleles sampled at random from two individuals at the same autosomal locus are IBD. Let k_{0ab}, k_{1ab}, k_{2ab} denote the probability that individuals a and b share zero, one and two alleles IBD, respectively. The definition of kinship coefficient indicates that ϕ_{ab} can be expressed as a function of those IBD-sharing probabilities, to be more explicit, $\phi_{ab} = k_{1ab}/4 + k_{2ab}/2$. Table 5 lists values of kinship coefficients, their corresponding

IBD-sharing probabilities and the inference criteria of ϕ_{ab} derived using powers of 2 [18] for various relative pairs under the assumption of no inbreeding.

Table 5. Kinship coefficients for different relative pairs

Relationship	ϕ_{ab}	$(k_{0ab}, k_{1ab}, k_{2ab})$	Inference criteria
Monozygotic twins	0.5	$(0,0,1)$	$> \frac{1}{2^{3/2}}$
Parent-offspring	0.25	$(0,1,0)$	$(\frac{1}{2^{5/2}}, \frac{1}{2^{3/2}})$
Full sibs	0.25	$(0.25, 0.5, 0.25)$	$(\frac{1}{2^{5/2}}, \frac{1}{2^{3/2}})$
Half sibs	0.125	$(0.5, 0.5, 0)$	$(\frac{1}{2^{7/2}}, \frac{1}{2^{5/2}})$
Uncle-niece	0.125	$(0.5, 0.5, 0)$	$(\frac{1}{2^{7/2}}, \frac{1}{2^{5/2}})$
First cousin	0.0625	$(0.75, 0.25, 0)$	$(\frac{1}{2^{9/2}}, \frac{1}{2^{7/2}})$
Unrelated	0	$(1,0,0)$	$< \frac{1}{2^{9/2}}$

Suppose we have genotype data of n individuals, for each person we consider his/her genotypes at m SNP markers respectively. For $1 \leq i \leq n, 1 \leq j \leq m$, let X_{ij} be the number of reference alleles (with label A) for individual i at SNP marker j. Thus X_{ij} takes values 0, 1, or 2 according to whether individual i has, respectively, 0,1, or 2 copies of allele A at marker j.

To simplify the illustration, we denote μ_j and σ_j^2 as the expectation and variance of X_{ij}, respectively. In other words, $E(X_{ij}) = \mu_j, Var(X_{ij}) = \sigma_j^2$. We assume the population variance for each marker is already known throughout our derivation. In practice, we can use sample variance, an unbiased estimator of population variance, as a substitute. Now we consider a pair of individuals i and i'. We use $\rho_{ii',j}$ to denote the correlation coefficient between X_{ij} and $X_{i'j}$. Besides, we let $\bar{\rho}_j$ be the average of $\rho_{ii',j}$ among all the individual pairs, i.e.

$$\bar{\rho}_j = \frac{\sum\limits_{i=1}^{n} \sum\limits_{i'=i+1}^{n} \rho_{ii',j}}{n(n-1)/2}.$$

If we further assume all individuals are sampled from a homogeneous population, we can derive the following relationship among those correlations:

Property 1. Assume all individuals are sampled from a homogeneous population, then for $1 \leq i, i' \leq n, 1 \leq j \leq m$, we have

$$i. \rho_{ii',j} = \rho_{ii'}, \bar{\rho}_j = \bar{\rho}.$$
$$ii. \rho_{ii'} = 2\phi_{ii'}.$$

This property has also been mentioned in other articles, for example, see [11]. A proof of this property is given in S1 Appendix. Now we summarize the conclusions of this property as follows:

Result i. implies that the correlation between X_{ij} and $X_{i'j}$ is irrelevant to which SNP we choose and depends only on the pair of individuals we select. Result ii. provides the quantitative relation between the kinship coefficient and the correlation of genotypes, which indicates that the estimation of kinship coefficient $\phi_{ii'}$ is equivalent to estimating the correlation coefficient of

genotypes between individual i and i' ($\rho_{ii'}$).

Estimating kinship coefficient by calculating the average sample pairwise correlation among all genetic variants has been taken by many methods. Following this principle, a natural estimator of $\rho_{ii'}$ is

$$\hat{\rho}_{ii'} = \frac{1}{m} \sum_{j=1}^{m} \frac{(X_{ij} - \bar{X}_j)(X_{i'j} - \bar{X}_j)}{\sigma_j^2}. \tag{1}$$

where $\bar{X}_j = \frac{1}{n} \sum_{i=1}^{n} X_{ij}$ is the average counts of reference alleles (with label A) at SNP j in the whole population. We call $\hat{\phi}_{ii'} = \frac{1}{2}\hat{\rho}_{ii'}$ the scGRM estimator.

However, as we are going to demonstrate, $\hat{\rho}_{ii'}$ is actually a biased estimator of $\rho_{ii'}$. To illustrate this, we need the following property:

Property 2. For $1 \le i, i' \le n, 1 \le j \le m$, the estimated correlation coefficient between X_{ij} and $X_{i'j}$ has a systematic bias from $\rho_{ii'}$. More specifically, we have

$$E\left[\frac{(X_{ij} - \bar{X}_j)(X_{i'j} - \bar{X}_j)}{\sigma_j^2}\right] = \rho_{ii'} - \frac{1}{n} \sum_{\substack{a=1 \\ a \neq i}}^{n} \rho_{ia} - \frac{1}{n} \sum_{\substack{a=1 \\ a \neq i'}}^{n} \rho_{ai'} - \frac{1}{n} + \frac{n-1}{n}\bar{\rho}. \tag{2}$$

The proof is given in S1 Appendix.

Equation (2) also reveals that the expected value of $\frac{1}{\sigma_j^2}(X_{ij} - \bar{X}_j)(X_{i'j} - \bar{X}_j)$ is not related to which SNP we select. Now we consider the expectation of estimator (1), it comes to the conclusion that

$$
\begin{aligned}
E\hat{\rho}_{ii'} &= E\left[\frac{1}{m} \sum_{j=1}^{m} \frac{(X_{ij} - \bar{X}_j)(X_{i'j} - \bar{X}_j)}{\sigma_j^2}\right] \\
&= \frac{1}{m} \sum_{j=1}^{m} E\left[\frac{(X_{ij} - \bar{X}_j)(X_{i'j} - \bar{X}_j)}{\sigma_j^2}\right] \\
&= \rho_{ii'} - \frac{1}{n} \sum_{\substack{a=1 \\ a \neq i}}^{n} \rho_{ia} - \frac{1}{n} \sum_{\substack{a=1 \\ a \neq i'}}^{n} \rho_{ai'} - \frac{1}{n} + \frac{n-1}{n}\bar{\rho}.
\end{aligned}
$$

If $\hat{\rho}_{ii'}$ is an unbiased estimator of $\rho_{ii'}$, then we should have $E\hat{\rho}_{ii'} = \rho_{ii'}$. However, the result we derive is obviously contradictory to it. The existence of bias means a systematic error when we estimate kinship coefficient via the scGRM method mentioned above. To make this fact clearer, we sum the expectation of $\frac{1}{\sigma_j^2}(X_{ij} - \bar{X}_j)(X_{i'j} - \bar{X}_j)$ up over all the individual pairs in the population, which leads to the following property:

Property 3. For every SNP marker j, where $1 \le j \le m$, we have

$$E\left[\sum_{i=1}^{n} \sum_{i'=i+1}^{n} \frac{(X_{ij} - \bar{X}_j)(X_{i'j} - \bar{X}_j)}{\sigma_j^2}\right] = \frac{n-1}{2}(\bar{\rho} - 1).$$

The proof is given in S1 Appendix.

Recall that $E\hat{\rho}_{ii'} = E\frac{1}{\sigma_j^2}(X_{ij} - \bar{X}_j)(X_{i'j} - \bar{X}_j)$, thus Property 3 also suggests

$$\sum_{i=1}^{n}\sum_{i'=i+1}^{n} E\hat{\rho}_{ii'} = \frac{n-1}{2}(\bar{\rho} - 1).$$

From Property 1 we know $\bar{\rho}$ is the theoretical mean value of correlations between pair-wise individuals, therefore it must take the value between 0 and 1. This fact together with Property 3 reveals that the mean value of estimator $\hat{\rho}_{ii'}$ is negative on average, which explains the empirical observation that a substantial proportion of estimated kinship coefficients are negative.

This bias problem makes $\hat{\phi}_{ii'}$ less desirable as an estimator of kinship between individuals i and i'. We can design an improved kinship estimation method which can eliminate the bias for each pair of individuals based on the scGRM estimator $\hat{\phi}_{ii'}$. The improved estimation method, UKin, which stands for the unbiased kinship estimator, solves the bias problem without adding much computational complexity. To understand how this method guarantees the unbiasedness, we need the following property:

Property 4. For every SNP marker $j, 1 \leq j \leq m$, and every pair of individuals i and i', $1 \leq i$, $i' \leq n$, we have

$$E\left[\frac{(X_{ij} - \bar{X}_j)(X_{i'j} - \bar{X}_j)}{\sigma_j^2} + \frac{1}{2}\sum_{\substack{k=1\\k\neq i}}^{n}\frac{(X_{ij} - \bar{X}_j)(X_{kj} - \bar{X}_j)}{\sigma_j^2}\right.$$
$$\left. + \frac{1}{2}\sum_{\substack{l=1\\l\neq i'}}^{n}\frac{(X_{lj} - \bar{X}_j)(X_{i'j} - \bar{X}_j)}{\sigma_j^2} + 1\right] = \rho_{ii'}. \tag{3}$$

The proof is given in S1 Appendix.

For ease of presentation, we set

$$u_{ii'}^j = 1 + \frac{1}{2}\sum_{\substack{k=1\\k\neq i}}^{n}\frac{(X_{ij} - \bar{X}_j)(X_{kj} - \bar{X}_j)}{\sigma_j^2} + \frac{1}{2}\sum_{\substack{l=1\\l\neq i'}}^{n}\frac{(X_{lj} - \bar{X}_j)(X_{i'j} - \bar{X}_j)}{\sigma_j^2} + \frac{(X_{ij} - \bar{X}_j)(X_{i'j} - \bar{X}_j)}{\sigma_j^2}.$$

Using (3), we also conclude that the expectation of $u_{ii'}^j$ does not depend on which SNP we select. Based on this fact, a reasonable estimator of $\rho_{ii'}$ is

$$\tilde{\rho}_{ii'} = \frac{1}{m}\sum_{j=1}^{m} u_{ii'}^j. \tag{4}$$

As Property 4 shows $Eu_{ii'}^j = \rho_{ii'}$ holds for every $1 \leq j \leq m$, the expectation of $\tilde{\rho}_{ii'}$ is still $\rho_{ii'}$. In other words, $\tilde{\rho}_{ii'}$ is an unbiased estimator of $\rho_{ii'}$, thus $\tilde{\phi}_{ii'} = \frac{1}{2}\tilde{\rho}_{ii'}$ is an unbiased kinship estimator. Besides, as we can observe from the expression of $Eu_{ii'}^j$, $\tilde{\rho}_{ii'}$ is the sum of a group of scGRM estimators $\hat{\rho}_{ii'}$ and a few correction terms, which means the UKin estimator relies on the same information we need for calculating the scGRM estimator $\hat{\phi}_{ii'}$. Thus the implementation of the UKin method doesn't require extra data.

It is worth noting that there exists some relationship between the scGRM and UKin estimator.

Substituting the expression of $u_{ii'}^{j}$ into (4), we get

$$\tilde{\rho}_{ii'} = \hat{\rho}_{ii'} + \frac{1}{2}\sum_{\substack{k=1 \\ k \neq i}}^{n} \hat{\rho}_{ik} + \frac{1}{2}\sum_{\substack{l=1 \\ l \neq i'}}^{n} \hat{\rho}_{li'} + 1. \tag{5}$$

Equation (5) indicates that the UKin estimator $\tilde{\phi}_{ii'}$ is a linear combination of some scGRM estimators $\hat{\phi}_{ii'}$ and constants. Thus $\tilde{\phi}_{ii'}$ and $\hat{\phi}_{ii'}$ are based on the same genetic information. Besides, this conclusion also shows that the UKin method won't bring a significant increase in computational complexity than the scGRM method.

Throughout our above analysis, we make assumptions of no inbreeding, LE and population homogeneity. In the Discussion we have analyzed these assumptions in detail.

Supporting information

S1 Appendix. Mathematical derivations of the properties in the Materials and Methods section.

Acknowledgments

References

1. Newman DL, Abney M, McPeek MS, Ober C, Cox NJ. The importance of genealogy in determining genetic associations with complex traits. Am J Hum Genet. 2001;69(5):1146–1148.

2. Day-Williams AG, Blangero J, Dyer TD, Lange K, Sobel EM. Linkage analysis without defined pedigrees. Genet Epidemiol. 2011;35(5):360–370.

3. Choi Y, Wijsman EM, Weir BS. Case-control association testing in the presence of unknown relationships. Genet Epidemiol. 2009;33(8):668–678.

4. Korte A, Vilhjálmsson BJ, Segura V, Platt A, Long Q, Nordborg M. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. Nat Genet. 2012;44(9):1066–1071.

5. Hoffman GE. Correcting for population structure and kinship using the linear mixed model: theory and extensions. PloS One. 2013;8(10):e75707.

6. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. FaST linear mixed models for genome-wide association studies. Nat Meth. 2011;8(10):833–835.

7. Chung D, Yang C, Li C, Gelernter J, Zhao H. GPA: a statistical approach to prioritizing GWAS results by integrating pleiotropy and annotation. PLoS Genet. 2014;10(11):e1004787.

8. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88(1):76–82.

9. Xia C, Amador C, Huffman J, Trochet H, Campbell A, Porteous D, et al. Pedigree-and SNP-associated genetics and recent environment are the major contributors to anthropometric and cardiometabolic trait variation. PLoS Genet. 2016;12(2):e1005804.

10. Hayes BJ, Visscher PM, Goddard ME. Increased accuracy of artificial selection by using the realized relationship matrix. Genetics Research. 2009;91(1):47–60.

11. Thornton T, Tang H, Hoffmann TJ, Ochs-Balcom HM, Caan BJ, Risch N. Estimating kinship in admixed populations. Am J Hum Genet. 2012;91(1):122–138.

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

13. Zhou X, Stephens M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nat Meth. 2014;11(4):407–409.

14. Wang J. Marker-based estimates of relatedness and inbreeding coefficients: an assessment of current methods. J Evol Biol. 2014;27(3):518–530.

15. Wang J. Estimating pairwise relatedness in a small sample of individuals. Heredity. 2017;119(5):302–313.

16. Powell JE, Visscher PM, Goddard ME. Reconciling the analysis of IBD and IBS in complex trait studies. Nat Rev Genet. 2010;11(11):800–805.

17. Jiang W, Song S, Hou L, Zhao H. A set of efficient methods to generate high-dimensional binary data with specified correlation structures. The American Statistician. 2020; p. 1–13.

18. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010;26(22):2867–2873.

19. Dou J, Sun B, Sim X, Hughes JD, Reilly DF, Tai ES, et al. Estimation of kinship coefficient in structured and admixed populations using sparse sequencing data. PLoS Genet. 2017;13(9):e1007021.

20. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet. 2012;90(1):7–24.

21. Yang J, Ferreira T, Morris AP, Medland SE, Madden PA, Heath AC, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet. 2012;44(4):369–375.

22. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–575.

S1 Appendix

Proof of Property 1.

For the j-th single nucleotide polymorphism (SNP) $(1 \leq j \leq m)$, let f_j be the frequency of the reference allele (with label A) at that SNP. Consider a pair of individuals i and i' whose kinship coefficient is denoted by $\phi_{ii'}$, we derive the covariance of X_{ij} and $X_{i'j}$ from two different aspects. Recall that we denote $\rho_{ii',j}$ to be the correlation between X_{ij} and $X_{i'j}$, thus we have

$$Cov(X_{ij}, X_{i'j}) = \rho_{ii',j}\sigma_j^2. \tag{S.1}$$

On the other hand, X_{ij} can be treated as the sum of two independent Bernoulli random variables. That is, $X_{ij} = B_{ij(1)} + B_{ij(2)}$. For $k = 1, 2$,

$$B_{ij(k)} = \begin{cases} 0 & \text{if the } k \text{ th allele for } i \text{ at SNP } j \text{ is } a \\ 1 & \text{if the } k \text{ th allele for } i \text{ at SNP } j \text{ is } A \end{cases}.$$

With this expression of X_{ij}, we have

$$\begin{aligned} & Cov(X_{ij}, X_{i'j}) \\ &= Cov(B_{ij(1)} + B_{ij(2)}, B_{i'j(1)} + B_{i'j(2)}) \\ &= \sum_{k=1}^{2} \sum_{k'=1}^{2} Cov(B_{ij(k)}, B_{i'j(k')}) \\ &= \sum_{k=1}^{2} \sum_{k'=1}^{2} \{E[B_{ij(k)}B_{i'j(k')}] - E[B_{ij(k)}]E[B_{i'j(k')}]\}. \end{aligned} \tag{S.2}$$

As we denote f_j to be the probability that a random allele chosen from the j-th SNP is A. Notice that $B_{ij(k)}B_{i'j(k')} = 1$ only when the two alleles selected from i and i' at this marker are both with label A, under this circumstance, these two reference alleles are either identical by descent (IBD) or not. For simplicity, let $A_{ij}(k)$ represent the k-th alleles from individual i at SNP j, if we assume IBD genes have the same allelic types and non-IBD genes have independent allelic types, we obtain

$$\begin{aligned} & \sum_{k=1}^{2} \sum_{k'=1}^{2} E[B_{ij(k)}B_{i'j(k')}] \\ &= \sum_{k=1}^{2} \sum_{k'=1}^{2} [P(A_{ij}(k) \text{ and } A_{i'j}(k') \text{ are IBD})f_j + (1 - P(A_{ij}(k) \text{ and } A_{i'j}(k') \text{ are IBD}))f_j^2]. \end{aligned}$$

Consider the definitions of f_j and $\phi_{ii'}$, we obtain

$$E[B_{ij(k)}]E[B_{i'j(k')}] = f_j^2.$$

$$\phi_{ii'} = \frac{1}{4} \sum_{k=1}^{2} \sum_{k'=1}^{2} P(A_{ij}(k) \text{ and } A_{i'j}(k') \text{ are IBD}).$$

Substituting them into (S.2), we get

$$Cov(X_{ij}, X_{i'j}) = 4\phi_{ii'} f_j (1 - f_j). \tag{S.3}$$

As X_{ij} is the sum of two *i.i.d.* Bernoulli random variables whose probability of success is f_j, we can derive that $\sigma_j^2 = 2f_j(1 - f_j)$. Together with (S.1) and (S.3), we have

$$\rho_{ii',j} = 2\phi_{ii'}. \tag{S.4}$$

Equation (S.4) also reveals that the value of correlation $\rho_{ii',j}$ doesn't depend on which SNP is selected, thus we get

$$\rho_{ii',1} = \rho_{ii',2} = \cdots = \rho_{ii',m} = \rho_{ii'}.$$
$$\bar{\rho}_1 = \bar{\rho}_2 = \cdots = \bar{\rho}_m = \bar{\rho}.$$

Proof of Property 2.

To demonstrate this property, we need a few preparations:

i. Consider the result $Cov(X_{ij}, X_{i'j}) = \rho_{ii'} \sigma_j^2$, we have

$$
\begin{aligned}
& EX_{ij}X_{i'j} \\
&= Cov(X_{ij}, X_{i'j}) + EX_{ij}EX_{i'j} \\
&= \rho_{ii'} \sigma_j^2 + \mu_j^2.
\end{aligned} \tag{S.5}
$$

Directly applying this result yields

$$(1) EX_{ij}^2 = \sigma_j^2 + \mu_j^2.$$
$$(2) E(\sum_{i=1}^{n} X_{ij}^2) = n(\sigma_j^2 + \mu_j^2).$$
$$(3) E(\sum_{i=1}^{n} \sum_{i'<i} X_{ij}X_{i'j}) = \sum_{i=1}^{n} \sum_{i'<i} (\rho_{ii'} \sigma_j^2 + \mu_j^2)$$
$$= \frac{n(n-1)}{2} (\bar{\rho} \sigma_j^2 + \mu_j^2).$$

ii. Based on (1)-(3) stated above, we have

$$
\begin{aligned}
E\bar{X}_j^2 &= E(\frac{1}{n} \sum_{i=1}^{n} X_{ij})^2 \\
&= \frac{1}{n^2} E(\sum_{i=1}^{n} X_{ij}^2 + 2\sum_{i=1}^{n} \sum_{i'<i} X_{ij}X_{i'j}) \\
&= \frac{n(\sigma_j^2 + \mu_j^2) + n(n-1)(\bar{\rho} \sigma_j^2 + \mu_j^2)}{n^2} \\
&= \frac{(n-1)\bar{\rho} + 1}{n} \sigma_j^2 + \mu_j^2.
\end{aligned} \tag{S.6}
$$

With these preparations, we now work on the demonstration of Property 2.

Directly expand the expression on the left side of (S.2), we have

$$
E\left[\frac{(X_{ij} - \bar{X}_j)(X_{i'j} - \bar{X}_j)}{\sigma_j^2}\right]
$$

$$
= \frac{1}{\sigma_j^2}[EX_{ij}X_{i'j} - EX_{ij}\bar{X}_j - EX_{i'j}\bar{X}_j + E(\bar{X}_j^2)].
$$

Substituting (S.5) and (S.6) into this expansion, we have

$$
E\left[\frac{(X_{ij} - \bar{X}_j)(X_{i'j} - \bar{X}_j)}{\sigma_j^2}\right]
$$

$$
= \frac{1}{\sigma_j^2}[\rho_{ii'}\sigma_j^2 + \mu_j^2 - \frac{1}{n}(\sum_{\substack{a=1\\a\neq i}}^{n} EX_{ij}X_{aj} + EX_{ij}^2)
$$

$$
- \frac{1}{n}(\sum_{\substack{a=1\\a\neq i'}}^{n} EX_{i'j}X_{aj} + EX_{i'j}^2) + \frac{(n-1)\bar{\rho}+1}{n}\sigma_j^2 + \mu_j^2]
$$

$$
= \frac{1}{\sigma_j^2}[\rho_{ii'}\sigma_j^2 + \mu_j^2 - \frac{1}{n}(\sum_{\substack{a=1\\a\neq i}}^{n} \rho_{ia}\sigma_j^2 + \sigma_j^2 + n\mu_j^2)
$$

$$
- \frac{1}{n}(\sum_{\substack{a=1\\a\neq i'}}^{n} \rho_{ai'}\sigma_j^2 + \sigma_j^2 + n\mu_j^2) + \frac{(n-1)\bar{\rho}+1}{n}\sigma_j^2 + \mu_j^2]
$$

$$
= \frac{1}{\sigma_j^2}[\rho_{ii'}\sigma_j^2 - \frac{1}{n}\sum_{\substack{a=1\\a\neq i}}^{n} \rho_{ia}\sigma_j^2 - \frac{1}{n}\sigma_j^2 - \frac{1}{n}\sum_{\substack{a=1\\a\neq i'}}^{n} \rho_{ai'}\sigma_j^2 - \frac{1}{n}\sigma_j^2 + \frac{(n-1)\bar{\rho}+1}{n}\sigma_j^2]
$$

$$
= \rho_{ii'} - \frac{1}{n}\sum_{\substack{a=1\\a\neq i}}^{n} \rho_{ia} - \frac{1}{n}\sum_{\substack{a=1\\a\neq i'}}^{n} \rho_{ai'} + \frac{(n-1)}{n}\bar{\rho} - \frac{1}{n}. \tag{S.7}
$$

Thus we derive the conclusion in Property 2.

Proof of Property 3.

For ease of calculation, we make a complement to the value range of index i':

$$E\left[\sum_{i=1}^{n}\sum_{i'=i+1}^{n}\frac{(X_{ij}-\bar{X}_j)(X_{i'j}-\bar{X}_j)}{\sigma_j^2}\right]$$

$$= \frac{1}{2}E\left[\sum_{i=1}^{n}\sum_{i'\neq i}\frac{(X_{ij}-\bar{X}_j)(X_{i'j}-\bar{X}_j)}{\sigma_j^2}\right]$$

$$= \frac{1}{2}\sum_{i=1}^{n}\sum_{i'\neq i}E\left[\frac{(X_{ij}-\bar{X}_j)(X_{i'j}-\bar{X}_j)}{\sigma_j^2}\right]. \tag{S.8}$$

Equation(S.8) together with the conclusion (S.7) in the proof of Property 2 yields

$$E\left[\sum_{i=1}^{n}\sum_{i'=i+1}^{n}\frac{(X_{ij}-\bar{X}_j)(X_{i'j}-\bar{X}_j)}{\sigma_j^2}\right]$$

$$= \frac{1}{2}\sum_{i=1}^{n}\sum_{i'\neq i}\left[\rho_{ii'}-\frac{1}{n}\sum_{\substack{a=1\\a\neq i}}^{n}\rho_{ia}-\frac{1}{n}\sum_{\substack{a=1\\a\neq i'}}^{n}\rho_{ai'}+\frac{(n-1)\bar{\rho}}{n}-\frac{1}{n}\right]$$

$$= \frac{n(n-1)}{2}\bar{\rho}-\frac{1}{2n}\sum_{i=1}^{n}\sum_{i'\neq i}\sum_{\substack{a=1\\a\neq i}}^{n}\rho_{ia}-\frac{1}{2n}\sum_{i=1}^{n}\sum_{i'\neq i}\sum_{\substack{a=1\\a\neq i'}}^{n}\rho_{ai'}+\frac{n(n-1)^2}{2n}\bar{\rho}-\frac{n(n-1)}{2n}$$

$$= \frac{(2n-1)(n-1)}{2}\bar{\rho}-\frac{1}{2n}\sum_{i=1}^{n}\sum_{i'\neq i}\sum_{\substack{a=1\\a\neq i}}^{n}\rho_{ia}-\frac{1}{2n}\sum_{i=1}^{n}\sum_{i'\neq i}\sum_{\substack{a=1\\a\neq i'}}^{n}\rho_{ai'}-\frac{n(n-1)}{2n}.$$

We observe that $\displaystyle\sum_{\substack{a=1\\a\neq i}}^{n}\rho_{ia}$ is irrelevant to i', therefore

$$\frac{1}{2n}\sum_{i=1}^{n}\sum_{i'\neq i}\sum_{\substack{a=1\\a\neq i}}^{n}\rho_{ia}$$

$$= \frac{n-1}{2n}\sum_{i=1}^{n}\sum_{\substack{a=1\\a\neq i}}^{n}\rho_{ia}$$

$$= \frac{n-1}{2n}n(n-1)\bar{\rho} = \frac{(n-1)^2}{2}\bar{\rho}.$$

Besides, if we change the sequence of summation, we have

$$\frac{1}{2n}\sum_{i=1}^{n}\sum_{i'\neq i}\sum_{\substack{a=1\\a\neq i'}}^{n}\rho_{ai'}$$

$$=\frac{1}{2n}\sum_{i'=1}^{n}\sum_{i\neq i'}\sum_{\substack{a=1\\a\neq i'}}^{n}\rho_{ai'}$$

$$=\frac{n-1}{2n}n(n-1)\bar{\rho}$$

$$=\frac{(n-1)^2}{2}\bar{\rho}.$$

Substituting them into the expansion, we get

$$E\left[\sum_{i=1}^{n}\sum_{i'=i+1}^{n}\frac{(X_{ij}-\bar{X}_j)(X_{i'j}-\bar{X}_j)}{\sigma_j^2}\right]$$

$$=\frac{(2n-1)(n-1)}{2}\bar{\rho}-\frac{(n-1)^2}{2}\bar{\rho}-\frac{(n-1)^2}{2}\bar{\rho}-\frac{n(n-1)}{2n}$$

$$=(\frac{(2n-1)(n-1)}{2}-(n-1)^2)\bar{\rho}-\frac{n-1}{2}$$

$$=\frac{n-1}{2}\bar{\rho}-\frac{n-1}{2}$$

$$=\frac{n-1}{2}(\bar{\rho}-1).$$

Thus we finish the proof of Property 3.

Proof of Property 4.

At the start, we focus on a part of the expression on the left side of (3):

$$E\left[\sum_{\substack{i'=1\\i'\neq i}}^{n}\frac{(X_{ij}-\bar{X}_j)(X_{i'j}-\bar{X}_j)}{\sigma_j^2}\right]$$

$$=\sum_{\substack{i'=1\\i'\neq i}}^{n}\left[\rho_{ii'}-\frac{1}{n}\sum_{\substack{a=1\\a\neq i}}^{n}\rho_{ia}-\frac{1}{n}\sum_{\substack{a=1\\a\neq i'}}^{n}\rho_{ai'}+\frac{(n-1)}{n}\bar{\rho}-\frac{1}{n}\right]$$

$$=\sum_{\substack{i'=1\\i'\neq i}}^{n}\rho_{ii'}-\frac{n-1}{n}\sum_{\substack{i'=1\\i'\neq i}}^{n}\rho_{ii'}-\frac{1}{n}\sum_{\substack{i'=1\\i'\neq i}}^{n}\sum_{\substack{a=1\\a\neq i'}}^{n}\rho_{ai'}+\frac{(n-1)^2}{n}\bar{\rho}-\frac{n-1}{n}$$

$$=\frac{1}{n}\sum_{\substack{i'=1\\i'\neq i}}^{n}\rho_{ii'}-\frac{1}{n}\left(\sum_{i'=1}^{n}\sum_{\substack{a=1\\a\neq i'}}^{n}\rho_{ai'}-\sum_{\substack{a=1\\a\neq i}}^{n}\rho_{ai}\right)+\frac{(n-1)^2}{n}\bar{\rho}-\frac{n-1}{n}$$

$$=\frac{2}{n}\sum_{\substack{i'=1\\i'\neq i}}^{n}\rho_{ii'}-\frac{1}{n}n(n-1)\bar{\rho}+\frac{(n-1)^2}{n}\bar{\rho}-\frac{n-1}{n}$$

$$=\frac{2}{n}\sum_{\substack{i'=1\\i'\neq i}}^{n}\rho_{ii'}-\frac{n-1}{n}\bar{\rho}-\frac{n-1}{n}. \tag{S.9}$$

Substituting (S.7), together with (S.9), into the whole expansion, we get

$$E\left[\frac{(X_{ij}-\bar{X}_j)(X_{i'j}-\bar{X}_j)}{\sigma_j^2}+\frac{1}{2}\sum_{\substack{k=1\\k\neq i}}^{n}\frac{(X_{ij}-\bar{X}_j)(X_{kj}-\bar{X}_j)}{\sigma_j^2}\right.$$

$$\left.+\frac{1}{2}\sum_{\substack{l=1\\l\neq i'}}^{n}\frac{(X_{lj}-\bar{X}_j)(X_{i'j}-\bar{X}_j)}{\sigma_j^2}+1\right]$$

$$=\rho_{ii'}-\frac{1}{n}\sum_{\substack{a=1\\a\neq i}}^{n}\rho_{ia}-\frac{1}{n}\sum_{\substack{a=1\\a\neq i'}}^{n}\rho_{ai'}+\frac{(n-1)}{n}\bar{\rho}-\frac{1}{n}$$

$$+\frac{1}{2}\left(\frac{2}{n}\sum_{\substack{k=1\\k\neq i}}^{n}\rho_{ik}-\frac{n-1}{n}\bar{\rho}-\frac{n-1}{n}\right)$$

$$+\frac{1}{2}\left(\frac{2}{n}\sum_{\substack{l=1\\l\neq i'}}^{n}\rho_{li'}-\frac{n-1}{n}\bar{\rho}-\frac{n-1}{n}\right)+1$$

$$=\rho_{ii'}-\frac{1}{n}-\frac{n-1}{n}+1$$

$$=\rho_{ii'}.$$

Here we have proved the conclusion in Property 4.