Received: 29 November 2021 | Accepted: 10 June 2022

DOI: 10.1111/2041-210X.13963

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

RESEARCH ARTICLE

A joint likelihood estimator of relatedness and allele frequencies from a small sample of individuals

Jinliang Wang

Institute of Zoology, Zoological Society of London, London, UK

Correspondence
Jinliang Wang
Email: jinliang.wang@ioz.ac.uk

Handling Editor: Michael Morrissey

Abstract

1. As a key parameter in population genetics, **relatedness** has found wide applications in molecular ecology, evolutionary biology, conservation, forensics and in studies of human inheritable diseases. It is defined as the probability that two individuals share an allele due to recent common ancestry. Many estimators have been developed to estimate relatedness from genotype data. However, they are invariably biased when a sample is small or contains a high proportion of close relatives, because allele frequencies required for inferring relatedness are poorly estimated in both cases under the impracticable and yet indispensable assumption of a large sample of unrelated genotypes.
2. In this study, I develop a likelihood method to estimate relatedness and allele frequencies jointly from a sample of multilocus genotypes. I propose an expectation maximization (EM) algorithm to update allele frequencies and the nine condensed identical by descent (IBD) coefficients ($\Delta_i, i = 1, 2, \ldots, 9$) of each pair of sampled individuals iteratively till convergence. Relatedness between and inbreeding coefficients of individuals is then calculated from the estimated nine IBD coefficients. The EM algorithm is also implemented in the reduced non-inbreeding model ($\Delta_i \equiv 0, i = 1, 2, \ldots, 6$) to estimate three condensed IBD coefficients ($\Delta_i, i = 7, 8, 9$) and relatedness.
3. Using simulated and empirical data, I show that the new method is much less biased and more accurate than previous methods, providing almost unbiased relatedness and inbreeding estimates, when the sampled individuals are few or/and contain many close relatives. The EM algorithm for the likelihood estimator is fast enough to handle a sample with thousands of individuals and millions of markers, thanks to the parallelization using openMP and MPI. The method is implemented in a software package, EMIBD9, that runs on all major computer platforms.
4. This study shows allele frequencies and relatedness, although highly correlated and difficult to disentangle from each other when the only information available is a sample of multilocus genotypes, can be estimated jointly from genotype data of diallelic and multiallelic markers in a likelihood framework. The new method

and software are especially useful for analysing small samples (such as ancient samples from museums, or samples from endangered species) and samples with a strong genetic structure.

KEYWORDS
allele frequency, identical by descent, inbreeding, markers, relatedness

1 | INTRODUCTION

Two individuals are genetically related when they share recent common ancestors (Wright, 1922). The extent of relatedness is determined by the number of common ancestors and their distances (in generations) to the two individuals, as calculated by a path analysis of pedigree data (Wright, 1922). It can also be measured by the proportion of alleles shared between individuals that are identical by descent (IBD, Malécot, 1948), which are replicas of the same ancestral alleles (Lynch & Ritland, 1999). More precisely and frequently, it is quantified by coancestry coefficient or kinship coefficient (Harris, 1964; Jacquard, 1972), the probability that two alleles, one taken at random from each individual, are IBD (Ritland, 1996). According to this definition, outbred individuals have relatedness of 0.25 when they are parent–offspring or full sibs, 0.125 when they are half sibs and 0.0625 when they are first cousins (FC) (Jacquard, 1972).

With the rapid development of various genetic markers, many relatedness estimators have been proposed to estimate relatedness from genotype data (Wang, 2017). Invariably, these estimators assume known allele frequencies which define the reference population structure against which relatedness is measured. This is true no matter the estimators are based on likelihood or allele frequency moments, and no matter they are developed for use in a homogeneous population (e.g. Lynch & Ritland, 1999; Ritland, 1996; Wang, 2002) or a heterogeneous population with subpopulation structure and with admixture (e.g. Conomos et al., 2016; Manichaikul et al., 2010; Moltke & Albrechtsen, 2014; Thornton et al., 2012). The reality is that, however, allele frequencies are rarely known. Frequently, they must be deduced from the sample of individuals whose relatedness is being estimated.

Estimating both allele frequencies and relatedness from the same sample of individuals could lead to highly biased estimates of relatedness, because allele frequencies and relatedness are correlated in determining the genotype data (Wang, 2017). First, sampled individuals must be assumed unrelated and noninbred such that simple allele counting can be used for estimating allele frequencies. The assumption is frequently violated, resulting in poorly estimated allele frequencies and thus underestimates of relatedness between close relatives and overestimates of relatedness between unrelated or loosely related individuals. For example, many (n) full siblings (FS) from a single parent pair might be included in a sample of N ($>n$) individuals for relatedness analysis. This is possible when early-life stage individuals (e.g. eggs or juveniles) are sampled from a high fecund species (e.g. fish and frogs). If the parental genotypes at a locus are AA and BB, then the n FS will show the same genotype AB. With simple allele counting without recognizing and accounting for the sibship structure, the frequencies of A, B or both will be overestimated from the sample. The same is true with other parental genotypes. The overestimated frequencies of parental alleles of the full-sib family lead inevitably to underestimated relatedness of these FS (<0.25) and overestimated relatedness of other sampled individuals. The problem cannot be alleviated using many genomic markers and deteriorates with an increasing proportion of relatives included in the sample.

Second, even in the ideal (for estimating allele frequencies) case of all sampled individuals being outbred and unrelated, the estimates of powers and products of allele frequencies are still biased (Weir, 1996), although estimates of allele frequencies are expected to be unbiased, when sample size is small. For example, the frequency of allele A is estimated as $\hat{p} = m/(2N)$ when a sample of N diploid genotypes contains m copies of allele A. Estimator $\hat{p}$ is unbiased irrespective of sample size N, as its expectation is $E\left[\hat{p}\right] = p$ where p is the population allele frequency. However, $E\left[\hat{p}^k\right] \neq p^k$ when $k > 1$ (Weir, 1996) and $E\left[\hat{p}\hat{q}\right] \neq pq$ (where $\hat{q}$ and q are estimated and parametric frequency of another allele). The problem of biased estimates of allele frequency powers and products deteriorates with a decreasing sample size. Unfortunately, all relatedness estimators use powers and products of allele frequencies and can yield poor estimates of relatedness when applied to a small sample (Wang, 2017). The problem becomes severe with a decreasing sample size.

Wang (2017) tackled the small sample size problem by modifying the estimator of Lynch (1988) and Li et al. (1993) and the estimator of Wang (2002). He showed that, using the unbiased estimators of allele frequency powers and products, the two estimators become much less biased than the original estimators and other estimators which do not correct for small sample sizes (Wang, 2017). However, there have been no attempts made to estimate allele frequencies by accounting for the relatedness structure of sampled individuals. As a result, when relatives are included in a sample, both allele frequency and relatedness estimates will be inevitably biased. Ironically, inclusion of related individuals in a sample is exactly why we are interested in knowing the relatedness structure of the sample.

In this study, I propose a likelihood method for estimating both allele frequencies and relatedness jointly and iteratively from the same sample of multilocus genotypes, accounting for small sample sizes and allowing for the presence of inbreeding. From genotype data of markers with each having more than three alleles, the estimator infers the nine condensed IBD coefficients (Harris, 1964; Jacquard, 1972) reliably for each pair of sampled individuals, which are then used to calculate relatedness and inbreeding coefficients. From genotype data of markers with each having two or three alleles, the estimator infers accurately summary IBD statistics such as relatedness and inbreeding coefficients, although it is unable to infer each of the nine condensed IBD coefficients reliably (Csűrös, 2014). I develop an expectation maximization (EM) algorithm to maximize the likelihood function for estimates of IBD coefficients and allele frequencies. Using simulated and empirical data, I compare the accuracy of relatedness obtained from the new estimator, estimators accounting for small sample size only (Wang, 2017) and estimators accounting for neither small sample sizes nor the presence of relatives in a sample.

2 | MATERIALS AND METHODS

2.1 | IBD and relatedness

Two individuals are genetically related because they have common ancestors (e.g. sibs) or one is the ancestor of the other (e.g. parent–offspring) in their recent genealogy. This means that, at each locus, related individuals tend to share genes IBD, which are the replicas of the same gene from a common ancestor (Malécot, 1948). Barring mutations which are rare in the small time-scale of recent ancestry, genes IBD are always identical in state (IIS), showing an identical allele or an identical sequence of DNA. As a result, related individuals (or relatives) have similar genotypes. The IIS patterns (or modes) between the genotypes of two individuals, X and Y, observed at some marker loci can thus be used to infer the underlying IBD patterns and thus relatedness between X and Y, given the population marker allele frequencies which act as the reference in distinguishing probabilistically IIS due to IBD and due to chance. These inferences reflect the realized IBD and relatedness at the particular marker loci (Wang, 2016). When the markers are regarded as a random sample from the genome, the inferences would on average signify the realized genomic relatedness due to shared genealogy (Wang, 2016).

There are 15 mutually exclusive and exhaustive IBD modes (Jacquard, 1972) among the two genes from diploid individual X and the two genes from diploid individual Y. When paternal and maternal genes are not distinguished, these 15 modes reduce to 9 condensed IBD modes (Harris, 1964; Jacquard, 1972; Wang, 2011a), D_i (i = 1, 2, ..., 9), as depicted in Figure 1. The relatedness between X and Y is fully described by D_i. Unfortunately, D_i is not observable directly and cannot be ascertained in general even when X and Y have known pedigrees and known genotypes (Wang, 2007; Weir et al., 2006). However, the probability of D_i, Δ_i, called condensed IBD coefficients, can be deduced from the genotypes (or IIS modes, S_i) of X and Y, given the allele frequencies of the population from which X and Y come. The genetic structure of X and Y can then be fully quantified by the vector $\mathbf{\Delta} = \{\Delta_1, \Delta_2, \Delta_3, \Delta_4, \Delta_5, \Delta_6, \Delta_7, \Delta_8, \Delta_9\}$. Apparently as probabilities, $0 \le \Delta_i \le 1$ and $\sum_{i=1}^{9} \Delta_i \equiv 1$. For full sibs (FS), half sibs (HS), parent–offspring (PO), unrelated (UR) and full sibs whose parents are full sibs (FSFS), $\mathbf{\Delta} = \left\{0,0,0,0,0,0,\frac{1}{4},\frac{1}{2},\frac{1}{4}\right\}$, $\left\{0,0,0,0,0,0,0,\frac{1}{2},\frac{1}{2}\right\}$, {0,0,0,0,0,0,0,1,0}, {0,0,0,0,0,0,0,0,1} and $\left\{\frac{1}{16},\frac{1}{32},\frac{1}{8},\frac{1}{32},\frac{1}{8},\frac{1}{32},\frac{7}{32},\frac{5}{16},\frac{1}{16}\right\}$, respectively (Jacquard, 1972). In the absence of inbreeding (i.e. both X and Y are outbred, such as FS, HS, PO, UR), only three IBD coefficients (Δ_7, Δ_8 and Δ_9) are necessary, and the remaining six coefficients are always zero ($\Delta_i \equiv 0$ for $i = 1, 2, ..., 6$).

In many applications of conservation biology, evolutionary biology and quantitative genetics, only some simple summary statistics of $\mathbf{\Delta}$ are necessary. These include inbreeding (F) and coancestry (Θ) coefficients. Given $\mathbf{\Delta}$ between X and Y, the F of and Θ between X and Y are calculated by

$$\begin{aligned} F_X &= \Delta_1 + \Delta_2 + \Delta_3 + \Delta_4, \\ F_Y &= \Delta_1 + \Delta_2 + \Delta_5 + \Delta_6, \\ \theta_{XY} = \theta_{YX} &= \Delta_1 + \frac{1}{2}(\Delta_3 + \Delta_5 + \Delta_7) + \frac{1}{4}\Delta_8 \end{aligned} \tag{1}$$

Thus, F_X is the probability that the two homologous genes at a locus of X are IBD. It is equivalent to the coancestry coefficient between the parents of X. F_Y is similarly defined and interpreted. θ_{XY} is the probability that two homologous genes, one taken at random from X and one from Y, are IBD. It quantifies the degree of relatedness between X and Y, from 0 (unrelated, sharing no genes IBD, when $\Delta_1 = \Delta_3 = \Delta_5 = \Delta_7 = \Delta_8 = 0$) to 1 (completely inbred and related, when $\Delta_1 = 1$ and $\Delta_i = 0$ for i = 2, 3, ..., 9). When X and Y are FS, HS,

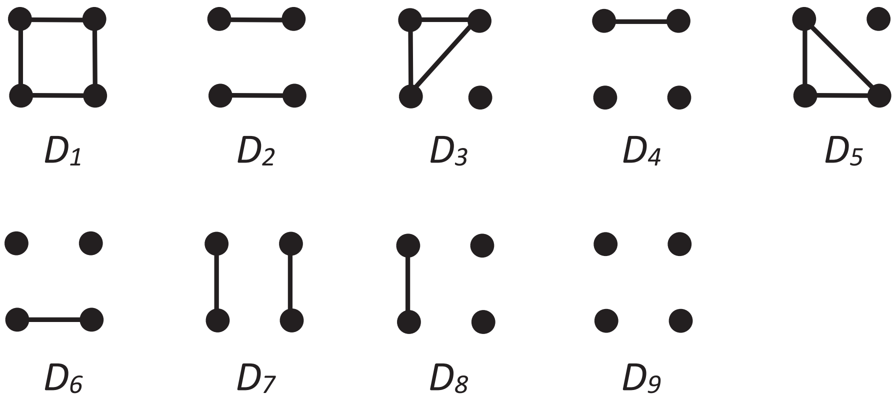

FIGURE 1 Nine condensed identical by descent (IBD) modes. Each mode is represented by two pairs of dots, with the top pair representing the two genes in individual X and the bottom pair representing the two genes in individual Y. Genes connected by lines are IBD. Genes of X or Y are not ordered.

PO, FC, UR and FSFS, θ_{XY} =0.25, 0.125, 0.25, 0.0625, 0 and 0.375, respectively. For non-inbred individuals X and Y (i.e. $F_X = F_Y = 0$), the maximal value of θ_{XY} is 0.5 when X and Y are identical twins or when $X \equiv Y$ (i.e. self coancestry, $\theta_{XX} = \theta_{YY} = 0.5$). When inbreeding is present, the maximal value of θ_{XY} is 1 when completely inbred ($F_X = F_Y = 1$) individuals X and Y are identical twins or when $X \equiv Y$.

2.2 | Probability of a pair of genotypes

Suppose a locus in a population has k codominant alleles $\{A_i\}$ of frequencies $\mathbf{p} = \{p_i\}$, with index i = 1, 2, ..., k. An individual W (=X or Y) from the population would have one of the $k(k+1)/2$ possible (ordered) genotypes, $\mathbf{G}_W = \{A_iA_j\}$ for $j \geq i = 1, 2, ..., k$. The frequency of $\mathbf{G}_W = \{A_iA_j\}$ is $\mathbf{Q}_W = p_i^2 + F_W p_i(1 - p_i)$ when $i = j$ (i.e. homozygote) and $\mathbf{Q}_W = 2p_ip_j(1 - F_W)$ when $i \neq j$ (i.e. heterozygote), where F_W is the inbreeding coefficient of W. The frequency of the joint genotypes $\mathbf{G}_X$ and $\mathbf{G}_Y$, $\mathbf{Q}_{XY}$, is not equal to the product of $\mathbf{Q}_X$ and $\mathbf{Q}_Y$ in general, except in the special case that X and Y are unrelated (i.e. $\theta_{XY} = 0$ which implies $\Delta_1 = \Delta_3 = \Delta_5 = \Delta_7 = \Delta_8 = 0$). Relatedness between X and Y (i.e. $\theta_{XY} > 0$) will cause them to share similar genotypes, with $\mathbf{Q}_{XY} > \mathbf{Q}_X\mathbf{Q}_Y$ when $\mathbf{G}_X$ is similar to $\mathbf{G}_Y$ and $\mathbf{Q}_{XY} < \mathbf{Q}_X\mathbf{Q}_Y$ when otherwise. The joint genotypes (IIS modes) of X and Y observed at marker loci can thus be used to infer $\mathbf{\Delta}$, which can then be used to calculate θ_{XY} between and F_X and F_Y of X and Y.

$\mathbf{Q}_{XY}$ given parameters $\mathbf{\Delta}$ and $\mathbf{p}$ can be calculated from Harris' (1964) genotypic array of pairs of individuals, or more conveniently from Table 1 (Anderson & Weir, 2007; Milligan, 2003; Wang, 2011a) by considering the nine IIS modes, $\mathbf{S} = \{S_1, S_2, ..., S_9\}$, separately. Note that while a locus with four or more alleles has nine IIS modes, a locus with 2 and 3 alleles has a set of 5 and 8 possible IIS modes, $\mathbf{S} = \{S_1, S_2, S_3, S_5, S_7\}$ and $\mathbf{S} = \{S_1, S_2, S_3, S_4, S_5, S_6, S_7, S_8\}$, respectively. This means that not all of the nine Δ coefficients are estimable from a marker with fewer than four alleles. As shown by Csűrös (2014), 'biallelic genotypes, however, do not convey enough information about the generic IBD structure, since different identity coefficients can generate the same joint genotype distribution'. Fortunately, several summary statistics of $\mathbf{\Delta} = \{\Delta_1, \Delta_2, \ldots, \Delta_9\}$, including Θ and F in (1), can still be estimated from diallelic marker data (Csűrös, 2014) as verified by my simulations below. I will discuss the issue more in the Discussion part.

2.3 | Likelihood function for Δ

Given parameter $\mathbf{\Delta}$ describing the (probabilistic) genetic structure of individuals X and Y, and the known parameter $\mathbf{p}_l$ describing the baseline genetic structure at locus l of the population from which X and Y are drawn, the probability of the observation $\mathbf{S}_l$ of X and Y is equal to the likelihood of $\mathbf{\Delta}$,

$$\mathcal{L}(\mathbf{\Delta}|\mathbf{S}_l, \mathbf{p}_l) = Prb(\mathbf{S}_l|\mathbf{\Delta}, \mathbf{p}_l) = \sum_{j=1}^{9} Prb(\mathbf{S}_l|\Delta_j, \mathbf{p}_l)\Delta_j. \qquad (2)$$

The observation S_{lj} in $\mathbf{S}_l = \{S_{l1}, S_{l2}, \ldots, S_{l9}\}$ is 1 when the genotypes of X and Y at locus l are observed to be in IIS mode j and is 0 when otherwise. Equation (2) is conveniently calculated using Table 1. For L loci in linkage and identity equilibria, the multilocus likelihood is simply the product of single locus likelihood calculated by Equation (2).

Maximizing the likelihood function with respect to $\mathbf{\Delta}$ leads to the maximum likelihood estimates (MLEs) of $\mathbf{\Delta}$. This estimation procedure can be carried out for each pair of sampled individuals independently for their estimates of $\mathbf{\Delta}$, from which more interesting summary quantities such as inbreeding (F) and coancestry (Θ) coefficients can be calculated.

In the absence of inbreeding (i.e. $F_X = F_Y = 0$), Equation (2) is greatly simplified because six out of the nine $\mathbf{\Delta}$ coefficients are zero ($\Delta_j \equiv 0$ for j = 1, 2,..., 6) and only three $\mathbf{\Delta}$ coefficients ($\Delta_7, \Delta_8, \Delta_9$) need to be estimated. The coancestry between X and Y is simply calculated as $\theta_{XY} = \Delta_7 + \frac{1}{4}\Delta_8$. It is worth noting that, in this special case of no inbreeding, even diallelic markers afford the estimation of each of the three $\mathbf{\Delta}$ coefficients (Δ_7, Δ_8 and Δ_9), because more IIS modes (5) than IBD modes (3) exist.

2.4 | Likelihood function for both p and Δ

Like previous likelihood or moment methods (Wang, 2017), the likelihood approach described above assumes that population allele frequencies, $\mathbf{p}$, are known, and IBD coefficients $\mathbf{\Delta}$ are the sole parameters to be estimated from the genotype data and allele frequency data. The definition and calculation of $\mathbf{\Delta}$ from either pedigree or genotype data require a reference population in which all homologous genes within and between individuals are assumed non-IBD or in which all individuals are assumed non-inbred and unrelated. For a marker-based analysis, it is the population whose allele frequencies are used as $\mathbf{p}$ that acts as the reference (Wang, 2016, 2017). The meaning and thus value of $\mathbf{\Delta}$ between two individuals are expected to vary with a change in the reference. This change could be simply a substitution of the population (e.g. from a global to a continental population) whose allele frequencies are used as $\mathbf{p}$ in marker-based analysis.

In a practical marker-based relatedness analysis, $\mathbf{p}$ is rarely known. Frequently, the only data available is a sample of multilocus genotypes. In such a situation, three possible approaches are adopted in relatedness and inbreeding coefficient estimation. The first is to estimate sample allele frequencies by simple allele counting under the assumption of non-inbred and unrelated individuals, and then to use them as $\mathbf{p}$ for inferring the subset of $\mathbf{\Delta} = \{\Delta_7, \Delta_8, \Delta_9\}$ by assuming the absence of inbreeding (e.g. Lynch & Ritland, 1999; Ritland, 1996; Wang, 2002) or the full set of $\mathbf{\Delta} = \{\Delta_1, \Delta_2, \ldots, \Delta_9\}$ (Wang, 2007). This approach effectively uses the sample as reference and tends to estimate $\theta'_{XY} = \frac{\theta_{XY} - \overline{\theta}_0}{1 - \overline{\theta}_0}$ as the relatedness between individuals X and Y, where $\overline{\theta}_0$ is the probability of a pair of genes taken at random from the sample being IBD (Wang, 2014). For a sample of N individuals, we have $\overline{\theta}_0 = \frac{4}{N(2N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \theta_{ij} + \frac{1}{N(2N-1)} \sum_{i=1}^{N} F_i$.

TABLE 1 Probability of identity-in-state modes S_l given identity-by-descent modes D_l

IIS mode	Allelic state	IBD modes								
		D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9
S_1	A_iA_i,A_iA_i	p_i	p_i^2	p_i^2	p_i^3	p_i^2	p_i^3	p_i^2	p_i^3	p_i^4
S_2	A_iA_i,A_jA_j	0	p_ip_j	0	$p_ip_j^2$	0	$p_i^2p_j$	0	0	$p_i^2p_j^2$
S_3	A_iA_i,A_iA_j	0	0	p_ip_j	$2p_i^2p_j$	0	0	0	$p_i^2p_j$	$2p_i^3p_j$
S_4	A_iA_i,A_jA_k	0	0	0	$2p_ip_jp_k$	0	0	0	0	$2p_i^2p_jp_k$
S_5	A_iA_j,A_iA_i	0	0	0	0	p_ip_j	$2p_i^2p_j$	0	$p_i^2p_j$	$2p_i^3p_j$
S_6	A_jA_k,A_iA_i	0	0	0	0	0	$2p_ip_jp_k$	0	0	$2p_i^2p_jp_k$
S_7	A_iA_j,A_iA_j	0	0	0	0	0	0	$2p_ip_j$	$p_ip_j(p_i+p_j)$	$4p_i^2p_j^2$
S_8	A_iA_j,A_iA_k	0	0	0	0	0	0	0	$p_ip_jp_k$	$4p_i^2p_jp_k$
S_9	A_iA_j,A_kA_l	0	0	0	0	0	0	0	0	$4p_ip_jp_kp_l$

Note: Alleles with different subscripts are distinct.

Abbreviations: IBD, identical by descent; IIS, identical in state.

When the sample is large and has no or weak IBD structure such that $\overline{\theta}_0$ is close to zero, θ'_{XY} is a good estimate of θ_{XY}, which is of the primary interest as defined in (1). Otherwise, θ'_{XY} tends to underestimate θ_{XY} when $\theta_{XY} > \overline{\theta}_0$ and to overestimate θ_{XY} when $\theta_{XY} < \overline{\theta}_0$, with the estimation bias increasing with an increasing difference between θ_{XY} and $\overline{\theta}_0$, and with an increasingly strong genetic structure of the sample measured by $\overline{\theta}_0$.

The second is to use the sample explicitly as the reference in estimating θ (Goudet et al., 2018; Weir & Goudet, 2017) and F (Zhang et al., 2022). Their formulation of the estimators avoids estimating and using **p** explicitly. However, like the first option, they are estimating $\theta''_{XY} = \frac{\theta_{XY}-\overline{\theta}_1}{1-\overline{\theta}_1}$ as the relatedness between individuals X and Y and $F''_X = \frac{F_X-\overline{\theta}_1}{1-\overline{\theta}_1}$ as the inbreeding coefficient of individual X, where $\overline{\theta}_1$ is the average coancestry of the N sampled individuals, $\overline{\theta}_1 = \frac{2}{N(N-1)}\sum_{i=1}^{N}\sum_{j=i+1}^{N}\theta_{ij}$. Except for a very small N, $\overline{\theta}_1$ is close to $\overline{\theta}_0$ and thus θ''_{XY} is close to θ'_{XY} in expectation. When a sample contains five FS sharing the same pair of parents and five individuals unrelated among themselves and with the siblings, for example, $\overline{\theta}_0$ and $\overline{\theta}_1$ are expected to be 0.0526 and 0.0555, respectively, θ'_{XY} and θ''_{XY} are expected to be 0.2084 and 0.2059 for each of the 10 full-sib pairs and are expected to be −0.0555 and −0.0588 for each of the other pairs (which are UR) in the 10 individual sample.

The third is to estimate **p** and IBD jointly from the sample. Depending on the depth of the sample's underlying pedigree, this approach can vary enormously in complexity in statistical methodology and computation. The simplest is a zero-generation pedigree where all individuals are unrelated to each other but are potentially inbred. In such a situation, Bayesian (Ayres & Balding, 1998; Vogl et al., 2002) and likelihood (Hall et al., 2012) methods were developed to estimate **p** and F jointly. For a sample known to contain simple one-generation (siblings) and two-generation (siblings as well as their parents) related individuals, the pedigree and **p** can be inferred jointly by a likelihood approach (Wang, 2004; Wang & Santure, 2009). For a general and more complex pedigree, however, it becomes tremendously difficult to reconstruct (Cannings et al., 1978; Cowell, 2009; Elston & Stewart, 1971) from marker data with known **p**, let alone to estimate pedigree and **p** jointly. Unfortunately, we have to deal with a general pedigree in IBD estimation because in practice a sample can have any pedigree structure.

Herein, I will refine the likelihood function and propose algorithms to estimate **p** and $\boldsymbol{\Delta}$ jointly from a sample of genotypes. Although $\boldsymbol{\Delta}$ is of the primary interest, **p** and its higher-order moments must be estimated accurately in an iterative process to account for both the small size and the genetic structure of a sample.

When allele frequencies are estimated from the sample being analysed for relatedness, the likelihood functions summarized in Table 1 must be modified to reduce biases due to small sample size. Suppose c_1 copies of allele A_1 are observed in the N sampled genotypes at a locus. Assuming the absence of genetic structure, $\widehat{p}_1 = c_1/(2N)$ provides the best unbiased (i.e. $E(\widehat{p}_1) = p_1$) estimate of p_1. However, $(\widehat{p}_1)^m$ is a poor estimate of p_1^m for $m > 1$ (Weir, 1996). It overestimates p_1^m in expectation, with the overestimation increasing with a decreasing sample size. It can be shown (Appendix 1) that the unbiased estimators of allele frequency powers and products are

$$\begin{aligned}
\widehat{p}_i^{\,m} &= \prod_{n=0}^{m-1}\frac{c_i-n}{2N-n}, \quad (m=1,2,3,4)\\
\widehat{p}_i^{\,m}\widehat{p}_j^{\,k} &= \frac{\left(\prod_{n=0}^{m-1}(c_i-n)\right)\left(\prod_{n=0}^{k-1}(c_j-n)\right)}{\prod_{n=0}^{m+k-1}(2N-n)}, \quad (m=1,2,3;k=1,2)\\
\widehat{p}_i^{\,m}\widehat{p}_j\widehat{p}_k &= \frac{c_jc_k}{2N(2N-1)}\prod_{n=0}^{m-1}\frac{c_i-n}{(2N-2-n)}, \quad (m=1,2)\\
\widehat{p}_i\widehat{p}_j\widehat{p}_k\widehat{p}_l &= \frac{c_ic_jc_kc_l}{2N(2N-1)(2N-2)(2N-3)}
\end{aligned} \tag{3}$$

Likelihood function (2) as detailed in Table 1 is applicable to a small sample when the parametric values of allele frequency powers and products are replaced by the unbiased estimators (3). However, the approach accounts for small sample size only and can still lead to biased IBD estimates when the sample contains related or/and inbred individuals. Ideally, both the small size and the relatedness/inbreeding of a sample should be accounted for in inferring **p**, which is then used in inferring $\boldsymbol{\Delta}$. This requires that both **p** and $\boldsymbol{\Delta}$ are estimated jointly, iteratively refining **p** by accounting for $\boldsymbol{\Delta}$ and estimating $\boldsymbol{\Delta}$ using refined **p**.

To estimate **p** and $\boldsymbol{\Delta}$ jointly, $\boldsymbol{\Delta}$ can no longer be estimated independently for each pair of individuals, as formulated in (2) and all previous likelihood and moment methods. Suppose N individuals are sampled and genotyped at L loci in linkage and identity equilibria, and the genotypes of individuals X and Y are observed to have IIS mode $S_{XYl} \in \mathbf{S}$ at locus l (=1, 2, ..., L). Assuming the probabilities of pairs of genotypes are independent, the likelihood function becomes

$$\begin{aligned}\mathcal{L}(\boldsymbol{\Delta},\mathbf{p}|\mathbf{S}) &= \prod_{X=1}^{N}\prod_{Y=1}^{N}\prod_{l=1}^{L} Prb\left(S_{XYl}|\boldsymbol{\Delta},\mathbf{p}\right)\\ &= \prod_{X=1}^{N}\prod_{Y=1}^{N}\prod_{l=1}^{L}\sum_{j=1}^{9} Prb\left(S_{XYl}|\Delta_{XYj},\mathbf{p}_l\right)\Delta_{XYj},\end{aligned} \qquad (4)$$

where $\boldsymbol{\Delta}$ includes $9N^2$ IBD coefficients (i.e. 9 for each of N^2 pairs of individuals, including self-pairs), and $\mathbf{p} = \left\{\mathbf{p}_1,\mathbf{p}_2, \dots ,\mathbf{p}_L\right\}$ has $\sum_{l=1}^{L} n_l$ allele frequencies where n_l is the number of alleles at locus l. Strictly, Equation (4) is not a proper likelihood function, because its component marginal likelihoods are not truly independent. The probabilities of pairs of genotypes are not independent because one individual is involved in N pairs and because the sample IBD structure can involve more than two related individuals (e.g. three half siblings (HS) sharing the same parent). More appropriately, Equation (4) is a composite likelihood function (Varin et al., 2011). Similar composite likelihood method was used to reconstruct simple one-generation and two-generation pedigrees, yielding results only slightly less accurate than a likelihood method (Wang, 2012).

For the non-inbreeding model, Equation (4) is greatly simplified because only $\frac{3}{2}N(N-1)$ of the $9N^2$ IBD coefficients need to be estimated. For each pair of individuals X and Y (X ≠ Y), only $\left\{\Delta_7,\Delta_8,\Delta_9\right\}$ need to be estimated and $\Delta_j \equiv 0$ for $j = 1, 2, ..., 6$. There are $\frac{1}{2}N(N-1)$ such (ordered) pairs of individuals, because Δ for X with Y is the same as Δ for Y with X in the absence of inbreeding and thus they are not distinguished. For each of the N self-pairs (e.g. X with X), we have $\left\{\Delta_7,\Delta_8,\Delta_9\right\} \equiv \{1, 0, 0\}$ and $\Delta_j \equiv 0$ for $j = 1, 2, ..., 6$.

2.5 | EM algorithm

For the simple case of a non-inbred population (i.e. $\Delta_j \equiv 0$ for j = 1, 2, ..., 6), several studies (e.g. Anderson & Weir, 2007; Milligan, 2003) used simplex algorithms and Choi et al. (2009) used the EM algorithm (Dempster et al., 1977; Wu, 1983) in searching for the MLE of $\{\Delta_7, \Delta_8, \Delta_9\}$. For the general case of an inbred population, Wang (2007, 2011a) used Powell's quadratically convergent method (Press et al., 1996) for estimating the entire set of nine IBD coefficients, $\boldsymbol{\Delta} = \{\Delta_j\}$ for j = 1, 2, ..., 9. In this study, I will develop an EM algorithm for MLE of $\boldsymbol{\Delta} = \{\Delta_1, \Delta_2, \dots , \Delta_9\}$ from likelihood functions (2) and (4). The same algorithm also applies to the much reduced non-inbreeding model (by setting $\Delta_j \equiv 0$ for j = 1, 2,..., 6).

The variables to be estimated in function (4), $9N^2$ IBD coefficients and $\sum_{l=1}^{L} n_l$ allele frequencies, are inter-dependent and numerous even in the simple case of a small sample of individuals (N) and loci (L). As a result, it is difficult to estimate them jointly from the genotype data. I develop an EM algorithm (Dempster et al., 1977) to solve (4) for MLEs of both **p** and $\boldsymbol{\Delta}$. The algorithm is similar to that of Hall et al. (2012) for estimating **p** and F, assuming unrelated individuals. However, it is necessarily more complicated as the assumed genetic structure (the presence of both relatedness and inbreeding) is more complicated. Details of the algorithm are described in Appendix 2.

2.6 | Comparison with other estimators

The new relatedness estimators (2) and (4) described above and implemented for the full model (with inbreeding) and the reduced model (without inbreeding) are compared with previous estimators that account for small sample sizes only (Wang, 2017) and those that account for neither small sizes nor relatedness structures of a sample. These previous estimators assume the absence of inbreeding. As numerous estimators are available that accommodate neither small size nor relatedness of a sample, I choose the popular Lynch and Ritland (1999) estimator, denoted as $\hat{r}_{LR}$ hereafter, as an example. In fact, in the case of a small sample size which is the focus of this study, all estimators in this category behave similarly, yielding highly biased estimates (Wang, 2017). Two estimators in the first category that accommodates small sample size only were developed by Wang (2017) and were shown to be unbiased regardless of N when all N sampled individuals are unrelated (Wang, 2017). One estimator is based on that of Lynch (1988) improved by Li et al. (1993), and the other is based on that of Wang (2002). Both estimators are modified using unbiased estimators of allele frequency powers as shown in Equation (3). The two estimators behave similarly in comparison with other estimators when sample sizes are small (Wang, 2017). However, the one based on Wang (2002) affords estimation of both Δ_7 and Δ_8, as well as relatedness (=$\frac{1}{2}\Delta_7 + \frac{1}{4}\Delta_8$). In the present study, I will consider this estimator, denoted as $\hat{r}_W$ hereafter, as an example.

It is worth noting that although I focus on comparing the accuracy of different estimators for relatedness, the new estimators offer much more. They give estimates of nine condensed IBD coefficients, $\{\Delta_1, \Delta_2, \dots , \Delta_9\}$, for each pair of individuals. Each IBD coefficient is reliably estimated when a marker has more than three alleles. For markers with each having two or three alleles, they can still yield estimates of $\{\Delta_1, \Delta_2, ..., \Delta_9\}$, although individually Δ_j (j = 1, 2, ..., 9) might be poorly estimated. However, thanks to the negative correlations among estimates of $\{\Delta_1, \Delta_2, \dots , \Delta_9\}$, these estimates can still be combined to yield reliably various summary statistics, including the most frequently used relatedness and inbreeding coefficients. For relatedness estimation, the new estimators allow for inbreeding, a property that only a couple of relatedness estimators (e.g. Loiselle et al., 1995; Ritland, 1996) possess. Alternatively, diallelic or triallelic markers can be used in the reduced non-inbreeding models (2) and (4) to estimate $\{\Delta_7, \Delta_8, \Delta_9\}$ just like previous estimators, but with improvements in accommodating the small sizes and relatedness of sampled individuals.

Hereafter, the new likelihood estimator (4) jointly estimating IBD and allele frequencies is denoted as $\hat{r}_{EM1}$, and the new likelihood estimator (2) estimating IBD only is denoted as $\hat{r}_{EM2}$. The corresponding estimators in the reduced non-inbreeding models are denoted by $\hat{r}_{EM1(3)}$ and $\hat{r}_{EM2(3)}$, respectively (the subscript '3' means only 3 coefficients, Δ_7, Δ_8 and Δ_9, are estimated). The four estimators are computed by the EM algorithm described in Appendix 2. They are implemented in a new computer program, EMIBD9, that runs on Windows, linux and Mac platforms. The program can be downloaded from webpage https://www.zsl.org/science/software/EMIBD9 and Zenodo at https://doi.org/10.5281/zenodo.6672390. The simulated and empirical datasets described below are analysed by EMIBD9 for estimators $\hat{r}_{EM1}$, $\hat{r}_{EM2}$, $\hat{r}_{EM1(3)}$ and $\hat{r}_{EM2(3)}$, and by COANCESTRY (Wang, 2011b) for estimators $\hat{r}_{LR}$ and $\hat{r}_{W}$.

2.7 | Simulations and accuracy assessment

2.7.1 | Simulations

Simulated data are generated and analysed to investigate the new estimators' statistical behaviour in estimating $\{\Delta_1, \Delta_2, \ldots, \Delta_9\}$ and the summary statistics of relatedness and inbreeding coefficients, and to understand the relative accuracy of the new estimators in comparison with previous estimators when samples are small or/and contain a substantial proportion of close relatives.

In simulation 1, I investigated the behaviour of estimator $\hat{r}_{EM1}$ for the inbreeding model in estimating $\{\Delta_1, \Delta_2, \ldots, \Delta_9\}$ and of estimator $\hat{r}_{EM1(3)}$ for the non-inbreeding model in estimating $\{\Delta_7, \Delta_8, \Delta_9\}$ when each marker locus has 2, 3 or 4 alleles. I considered a sample consisting of 498 non-inbred and unrelated individuals and one pair of related individuals, X and Y. X and Y are full sibs from a pair of full-sib parents (FSFS, $\boldsymbol{\Delta} = \{\frac{1}{16}, \frac{1}{32}, \frac{1}{8}, \frac{1}{32}, \frac{1}{8}, \frac{1}{32}, \frac{7}{32}, \frac{5}{16}, \frac{1}{16}\}$) or are outbred full sibs (FS, $\boldsymbol{\Delta} = \{0,0,0,0,0,0,\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\}$). For both FS and FSFS types of samples, I calculated relatedness from $\boldsymbol{\Delta}$ estimates. For FSFS samples, I also calculated several additional summary statistics from $\boldsymbol{\Delta}$ estimates. These include inbreeding coefficient as calculated by (1), the probability that at least one pair of alleles among three randomly selected ones from two individuals are IBD,

$$\theta_3 = \Delta_1 + \Delta_2 + \Delta_3 + \Delta_5 + \Delta_7 + (\Delta_4 + \Delta_6 + \Delta_8)/2, \quad (5)$$

the probability that three randomly chosen alleles from two individuals are IBD,

$$\theta_{3:3} = \Delta_1 + (\Delta_3 + \Delta_5)/2, \quad (6)$$

and the difference in probabilities of some inbred IBD modes,

$$\theta_6 = (\Delta_4 - \Delta_6)/2, \quad (7)$$

$$\theta_5 = (\Delta_3 - \Delta_5)/2, \quad (8)$$

as defined by Csűrös (2014). Correlation coefficients among Δ_j ($j = 1, 2, \ldots, 9$) estimates from $\hat{r}_{EM1}$ were also calculated.

In simulation 2, I considered a small sample of individuals of the same familial relationship. Specifically, I simulated N individuals ($N = 2, 3, 4, 5, 6, 7, 8, 10, 12, 14, 16$) of familial relationship FS, HS or FC. FS share both parents. HS share the same father but have different mothers. FC share the same paternal grandparents. All parents or grandparents are non-inbred and unrelated such that FS, HS and FC have expected relatedness of $\theta = 0.25, 0.125, 0.065$, respectively. The simulated data were analysed by estimators $\hat{r}_{EM1}$, $\hat{r}_{EM2}$, $\hat{r}_{W}$ and $\hat{r}_{LR}$ to compare their accuracy and to investigate the effect of jointly estimating allele frequencies and relatedness.

In simulation 3, I simulated a small sample of individuals of mixed familial relationships. Specifically, I considered four types of relatedness structure of a sample of individuals, reflecting different types and proportions of close relatives and varying sample sizes. A sample contains two individuals related as FSFS, FS or FC, or unrelated (UR). Additionally, it also contains a varying number n of unrelated individuals, with $n = 0, 1, 2, \ldots, 10$. The simulated data were analysed comparatively by estimators $\hat{r}_{EM1}$, $\hat{r}_{EM2}$, $\hat{r}_{W}$ and $\hat{r}_{LR}$ for relatedness. For the mixed samples containing an FSFS dyad and unrelated individuals (i.e. $n > 0$), the inbreeding coefficients of each of the $n + 2$ individuals were also estimated by $\hat{r}_{EM1}$, $\hat{r}_{EM2}$, the estimator of Li and Horvitz (1953) (denoted by F_{LH}) and the estimator derived by Wang (2011c, equation [9]) following the approach of Lynch and Ritland (1999) (denoted by F_{LR}). Both F_{LH} and F_{LR} are implemented in software COANCESTRY (Wang, 2011b).

For all sets of simulations, each sampled individual was genotyped at L loci, with each locus having a fixed number of k codominant alleles in a uniform frequency distribution in the population. In simulation 1, L = 10,000 and k = 2, 3 and 4. In simulations 2 and 3, L = 100,000 and k = 2. The marker loci were assumed to be in Hardy–Weinberg and linkage equilibria. For each parameter combination defined by the type of relatedness structure and values of N, L and k, a number of 1000 (when L is large) or 10,000 (when L is small) replicate datasets were simulated and comparatively analysed by different estimators. For each dataset, all sampled individuals were used to calculate allele frequencies and relatedness.

2.7.2 | Accuracy assessment

A relatedness estimator was assessed by its bias, precision measured by standard deviation (*SD*), and accuracy measured by root mean squared errors (RMSE),

$$Bias = \frac{1}{RM}\sum_{i=1}^{R}\sum_{j=1}^{M}\left(r - \hat{r}_{ij}\right), \quad (9)$$

$$SD = \left(\frac{1}{RM}\sum_{i=1}^{R}\sum_{j=1}^{M}\left(\bar{r} - \hat{r}_{ij}\right)^2\right)^{0.5},$$

$$RMSE = \left(\frac{1}{RM}\sum_{i=1}^{R}\sum_{j=1}^{M}\left(r - \hat{r}_{ij}\right)^2\right)^{0.5}, \quad (10)$$

where R is the number of replicates (=1,000 or 10,000) and M is the number of pairs of focal relationship in a replicate dataset (e.g. $M = 6$ in a dataset containing 4 full sibs where the focal relationship is FS), $\bar{r} = \frac{1}{RM}\sum_{i=1}^{R}\sum_{j=1}^{M}\hat{r}_{ij}$ is the mean estimate, and r and $\hat{r}_{ij}$ are the parameter value and estimate for the jth focal pair in the ith replicate. The quality of an estimator for inferring other parameters such as inbreeding coefficients and Δ_j ($j = 1, 2, ..., 9$) is evaluated similarly to relatedness.

2.8 | Empirical data analysis

2.8.1 | CEPH data

The genotype data of 65 CEPH reference families (https://cephb.fr/familles_CEPH.php) from different populations were analysed by $\hat{r}_{EM1}$, $\hat{r}_{EM2}$, $\hat{r}_{LR}$ and $\hat{r}_{W}$. Each family has 2 parents with 5 or more full sib children and between 0 and 4 grandparents. I choose both parents and the first two full-sib children from each family for the study of their relatedness. Therefore, the 65 datasets were analysed independently, each consisting of four persons in a nuclear family. Among the 6 dyads in each of the 65 datasets, 1 is unrelated (between the parents), 4 are parent–offspring and 1 is full sibs. Although genomic data are available, they are highly sparse with many missing data in these families. For each dataset, I choose loci at which at least three of the four persons have genotype data and show two or more alleles. Because of the sparse nature and the small sample size (only four persons), the number of usable loci is small and is highly variable, from 11 to 9,614 loci with an average of 2,342 and a *SD* of 2,121 among the 65 datasets. Also because of the small sample size (4 persons) and high relatedness (a single nuclear family), many selected loci have only two alleles although they are microsatellites and have many more alleles in the original large sample of individuals.

2.8.2 | Ant data

Hammond et al. (2001) sampled individuals from an ant species, *Leptothorax acervorum*, and genotyped them at six microsatellite loci to study the mating system. A number of 9, 7, 47, 45, 45, 45, 45, 45, 44 and 45 diploid workers were sampled from 10 colonies, constituting a sample of 377 individuals. For this species, we know that each colony is headed by a single diploid queen mated with a single haploid male. Therefore, workers from the same colony are full sibs and workers from different colonies are non-sibs. Full sibs are expected to have $\Delta_7 = \Delta_8 = 0.5$, $\Delta_9 = 0$ and $\theta = 0.375$ in the absence of any background relatedness in this haplo-diploid species. Since the sample size of 377 individuals is large, we might expect that it makes little difference whether to update allele frequencies or not in estimating IBD coefficients and relatedness. However, the few colonies and the large colony sizes of the sample mean a strong genetic structure of the sample, and thus allele frequencies could be poorly estimated by assuming a non-inbred and unrelated sample. Therefore, I made a comparative analysis of the sample using estimators $\hat{r}_{EM1}$ and $\hat{r}_{EM2}$ with and without updating allele frequencies.

2.8.3 | Human Genome Diversity Panel data

Li et al. (2008) studied the worldwide human population structure using 938 individuals sampled from 51 populations of the Human Genome Diversity Panel (HGDP). The data were later expanded to include genotypes of 1,043 individuals, each genotyped at 644,258 single nucleotide polymorphisms (SNPs), available from http://www.cephb.fr/en/hgdp_panel.php#basedonnees. In this study, I analyse the expanded data by estimators $\hat{r}_{EM1}$ and $\hat{r}_{EM2}$ for IBD coefficients and relatedness.

3 | RESULTS

3.1 | Simulation 1: Marker polymorphisms

For inbred and related dyads FSFS (Table 2), Δ_1, Δ_2,..., Δ_9 are all well estimated when $n_l = 4$ but some of them are poorly estimated (with a large bias and a large *SD*) when $n_l = 3$ or $n_l = 2$. The estimates are particularly poor for Δ_3, Δ_5, Δ_7, Δ_8 and Δ_9 when $n_l = 2$, and for Δ_3, Δ_5, Δ_8 and Δ_9 when $n_l = 3$. However, the estimable summary statistics (θ, F_X, F_Y, θ_3, $\theta_{3:3}$, θ_6 and θ_5) as identified by Csűrös (2014) are all very well estimated no matter the markers have 2, 3 or 4 alleles. At $n_l = 2$ or $n_l = 3$, estimates of these summary statistics are little biased, but have slightly higher *SD*s and thus slightly higher RMSEs than those at $n_l = 4$. This is expected as at the same number of $L = 10{,}000$ loci and the same allele frequency distribution, markers with more alleles ($n_l = 4$) are more informative about IBD coefficients than markers with fewer alleles ($n_l = 2$ or 3). These summary statistics should be better estimated using more diallelic markers than markers with 4 or more alleles.

For non-inbred FS dyads (Table 3), all nine IBD coefficients, Δ_1, Δ_2, ..., Δ_9, are well estimated by $\hat{r}_{EM1}$, regardless of the number of alleles (2, 3, 4) per locus. The six Δ coefficients involving IBD within individuals, Δ_1, Δ_2,..., Δ_6, are accurately estimated to be close to the simulated (expected) value of zero when n_l is 2, 3 or 4. When constrained to non-inbreeding (i.e. $\Delta_j \equiv 0$ for $j = 1, 2, ..., 6$), estimator $\hat{r}_{EM1(3)}$ yields estimates of Δ_7, Δ_8 and Δ_9 qualitatively very similar to those from inbred estimator $\hat{r}_{EM1}$ without the constraint. For diallelic markers ($n_l = 2$), $\hat{r}_{EM1(3)}$ should be able to estimate Δ_7, Δ_8 and Δ_9 reliably because there are 5 possible IIS modes against 3 IBD modes. Yet, the estimates are almost identical to those from $\hat{r}_{EM1}$, indicating that $\hat{r}_{EM1}$ affords accurate estimates of Δ_1, Δ_2,..., Δ_9 in the absence of inbreeding even when diallelic markers are used.

A complicated correlation structure was found among $\boldsymbol{\Delta}$ estimates from estimator $\hat{r}_{EM1}$ using markers with $n_l = 2$, $n_l = 3$ and $n_l = 4$ (Supplementary Appendix 1) alleles. Consistently across relationships FS and FSFS and across numbers of alleles per locus, highly negative correlations were found between $\hat{\Delta}_7$ and $\hat{\Delta}_8$ and between

TABLE 2 Estimates of Δ and some summary statistics for FSFS dyads

		$n_l = 2$			$n_l = 3$			$n_l = 4$		
Parameter	Truth	Mean	*SD*	RMSE	Mean	*SD*	RMSE	Mean	*SD*	RMSE
Δ_1	0.06250	0.048	0.010	0.018	0.050	0.009	0.015	0.060	0.007	0.007
Δ_2	0.03125	0.017	0.005	0.015	0.039	0.005	0.009	0.033	0.004	0.004
Δ_3	0.12500	0.173	0.014	0.050	0.145	0.011	0.023	0.128	0.008	0.008
Δ_4	0.03125	0.015	0.007	0.018	0.014	0.006	0.018	0.027	0.006	0.007
Δ_5	0.12500	0.173	0.012	0.050	0.146	0.011	0.024	0.129	0.007	0.008
Δ_6	0.03125	0.014	0.007	0.018	0.013	0.006	0.020	0.028	0.006	0.006
Δ_7	0.21875	0.280	0.011	0.062	0.234	0.011	0.019	0.223	0.008	0.008
Δ_8	0.31250	0.063	0.020	0.251	0.242	0.025	0.075	0.301	0.017	0.020
Δ_9	0.06250	0.217	0.021	0.156	0.117	0.015	0.056	0.072	0.011	0.015
$\boldsymbol{\theta}$	**0.37500**	**0.376**	**0.007**	**0.007**	**0.373**	**0.005**	**0.005**	**0.375**	**0.004**	**0.004**
$\boldsymbol{F_X}$	**0.25000**	**0.253**	**0.011**	**0.012**	**0.248**	**0.009**	**0.009**	**0.248**	**0.008**	**0.008**
$\boldsymbol{F_Y}$	**0.25000**	**0.253**	**0.012**	**0.012**	**0.248**	**0.008**	**0.008**	**0.249**	**0.008**	**0.008**
$\boldsymbol{\theta_3}$	**0.71875**	**0.737**	**0.014**	**0.019**	**0.749**	**0.009**	**0.009**	**0.750**	**0.006**	**0.006**
$\boldsymbol{\theta_{3:3}}$	**0.18750**	**0.221**	**0.009**	**0.034**	**0.196**	**0.007**	**0.011**	**0.188**	**0.006**	**0.006**
$\boldsymbol{\theta_4}$	**0.00000**	**0.000**	**0.005**	**0.005**	**0.001**	**0.004**	**0.004**	**0.001**	**0.003**	**0.003**
$\boldsymbol{\theta_5}$	**0.00000**	**0.000**	**0.008**	**0.008**	**−0.001**	**0.006**	**0.006**	**0.000**	**0.005**	**0.005**

Notes: Each sampled individual is genotyped at 10,000 loci, and each locus has an equal number of either $n_l = 2$, $n_l = 3$ or $n_l = 4$ alleles with frequencies in a uniform distribution. Estimates of $\Delta = \{\Delta_1, \Delta_2, \ldots, \Delta_9\}$ are obtained from $\hat{r}_{EM1}$, and the summary statistics are calculated from Δ estimates.

Abbreviations: FSFS, full sibs whose parents are full sibs; RMSE, root mean squared error.

TABLE 3 Estimates of Δ and θ for FS dyads

			$n_l = 2$			$n_l = 3$			$n_l = 4$		
Estimator	**Δ**	Truth	Mean	*SD*	RMSE	Mean	*SD*	RMSE	Mean	*SD*	RMSE
$\hat{r}_{EM1}$	Δ_1	0.000	0.002	0.001	0.002	0.002	0.001	0.002	0.001	0.001	0.002
	Δ_2	0.000	0.002	0.001	0.002	0.002	0.001	0.002	0.001	0.001	0.001
	Δ_3	0.000	0.008	0.005	0.009	0.003	0.002	0.004	0.003	0.002	0.004
	Δ_4	0.000	0.003	0.002	0.004	0.004	0.003	0.004	0.004	0.002	0.004
	Δ_5	0.000	0.008	0.005	0.009	0.003	0.002	0.004	0.003	0.002	0.004
	Δ_6	0.000	0.003	0.002	0.004	0.004	0.001	0.002	0.004	0.002	0.004
	Δ_7	0.250	0.260	0.013	0.016	0.258	0.010	0.012	0.259	0.008	0.012
	Δ_8	0.500	0.489	0.026	0.028	0.502	0.017	0.017	0.503	0.014	0.014
	Δ_9	0.250	0.225	0.018	0.031	0.224	0.014	0.030	0.222	0.012	0.030
	θ	0.250	0.262	0.005	0.013	0.259	0.004	0.010	0.260	0.004	0.011
$\hat{r}_{EM1(3)}$	Δ_7	0.250	0.252	0.013	0.013	0.260	0.009	0.014	0.263	0.007	0.015
	Δ_8	0.500	0.499	0.023	0.023	0.502	0.016	0.016	0.505	0.012	0.013
	Δ_9	0.250	0.249	0.017	0.017	0.239	0.012	0.016	0.232	0.010	0.020
	θ	0.250	0.251	0.005	0.005	0.255	0.004	0.006	0.258	0.003	0.008

Note: Each sampled individual is genotyped at 10,000 loci, and each locus has an equal number of either $n_l = 2$, $n_l = 3$ or $n_l = 4$ alleles with frequencies of a uniform distribution.

Abbreviations: FS, full sibling; RMSE, root mean squared error.

$\widehat{\Delta}_8$ and $\widehat{\Delta}_9$, with their correlation coefficients varying from −0.48 to −0.81. Consistently negative correlations were found between $\widehat{\Delta}_3$ and $\widehat{\Delta}_8$ and between $\widehat{\Delta}_5$ and $\widehat{\Delta}_8$, with correlation coefficients varying from −0.16 to −0.69.

3.2 | Simulation 2: Identical familial relationships

The estimator inferring **p** and **Δ** jointly, $\hat{r}_{EM1}$, is the least biased for all relationships simulated (Figure 2), yielding an average $\hat{\theta}$ close to the expected value, which is 0.25, 0.125 and 0.0625 for FS, HS and FC, respectively. It consistently overestimates θ slightly when it is small (HS and FC), and underestimates θ slightly when it is large (FS). Estimator $\hat{r}_W$ always underestimates θ, with the extent of underestimation decreasing with a decreasing true value of θ. $\hat{r}_{EM2}$ is always 0 (for HS and FC), or close to zero (for FS), while $\hat{r}_{LR}$ is always negative with values widely ranged from −0.5 to −0.05. In terms of overall accuracy measured by RMSE, $\hat{r}_{EM1}$ is the best, followed by $\hat{r}_W$, and $\hat{r}_{EM2}$ and $\hat{r}_{LR}$ are the least accurate estimators. The differences between $\hat{r}_{EM1}$ and $\hat{r}_W$ tend to decrease with a decreasing true θ value. This is expected as the main problem with $\hat{r}_W$ is its biasness caused by poor allele frequency estimates due to the inclusion of relatives in a sample, which tends to decrease to zero as θ decreases to zero.

With fewer markers than those in Figure 2, the accuracy advantages of $\hat{r}_{EM1}$ over other estimators largely remain (Supplementary Appendix 2). Only when the number of loci is small (i.e. $L < 2,000$) does $\hat{r}_{EM1}$ performs similarly to $\hat{r}_W$.

3.3 | Simulation 3: Mixed familial relationships

For relationship involving inbreeding (FSFS), $\hat{r}_{EM1}$ is the least biased and the most accurate estimator, regardless of the proportion

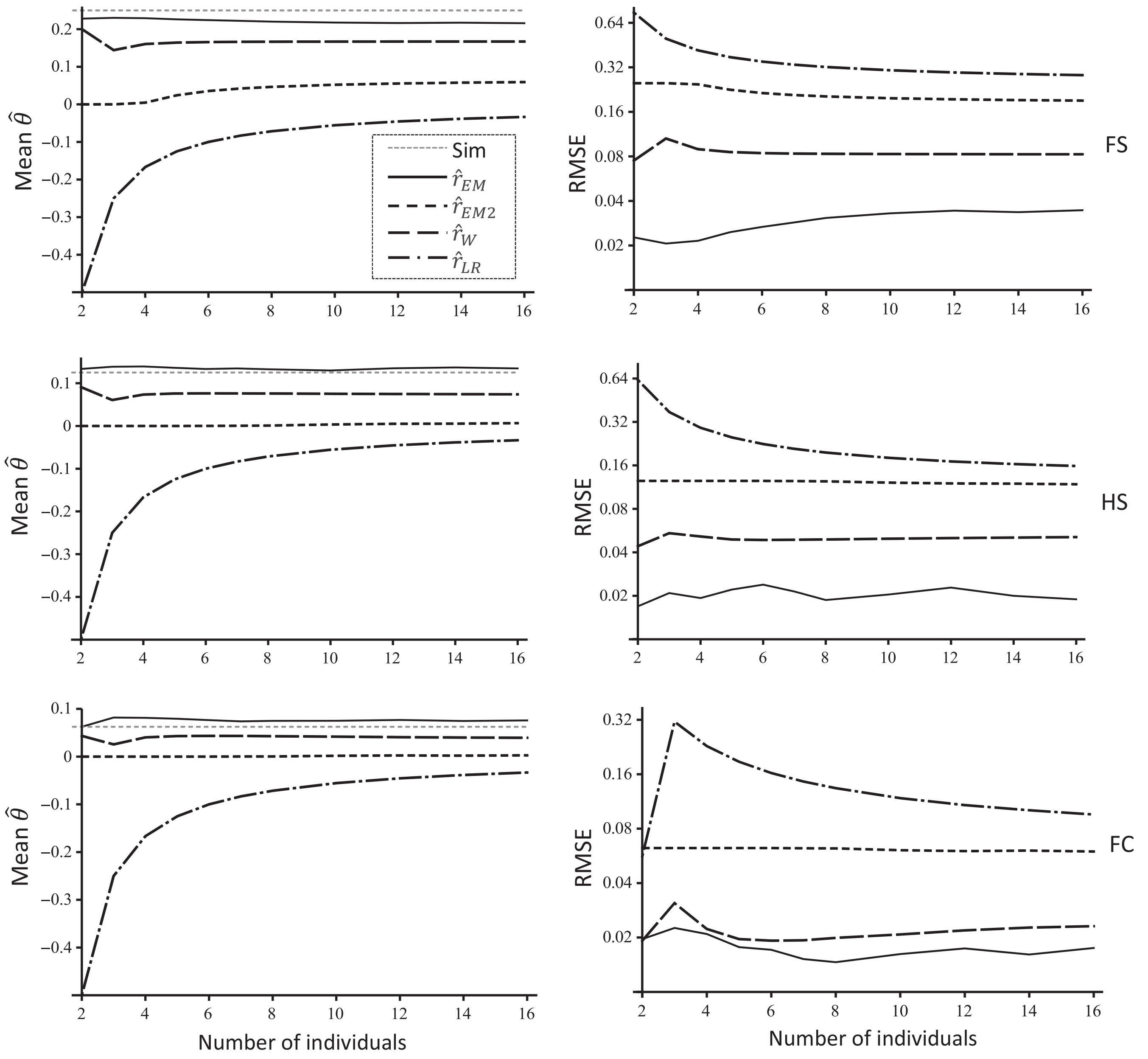

FIGURE 2 Mean and root mean squared error (RMSE) of $\hat{\theta}$ estimated by $\hat{r}_{EM1}$, $\hat{r}_{EM2}$, $\hat{r}_{LR}$ and $\hat{r}_W$ for relationships full sibling (FS), half sibling (HS) and first cousin (FC) as a function of the number of individuals (x axis) in a sample. All individuals in a sample are of the same relationship, FS, HS or FC with an expected θ value of 0.25, 0.125 or 0.0625 shown by the horizontal dotted lines. A number of $L = 100,000$ single nucleotide polymorphisms (SNPs) with allele frequencies in a uniform distribution are genotyped for each individual.

of close relatives (FSFS) included in a sample (Figure 3). As expected, $\hat{r}_{EM2}$ approaches $\hat{r}_{EM1}$ with an increasing proportion of unrelated individuals in a sample. $\hat{r}_{LR}$ and $\hat{r}_W$ assume non-inbreeding, and thus always underestimate the relatedness for FSFS dyads substantially.

Without inbreeding, $\hat{r}_{EM1}$ outperforms the other estimators substantially only when a sample contains a high proportion of close relatives (FS). Otherwise, it has a performance almost indistinguishable from that of $\hat{r}_W$. Although both $\hat{r}_{EM2}$ and $\hat{r}_W$ account for the small size only (not the genetic structure) of a sample, $\hat{r}_{EM2}$ underestimates the relatedness for relatives more than $\hat{r}_W$, perhaps because it attempts to estimate 9 (Δ_1, Δ_2,..., Δ_9) while $\hat{r}_W$ estimates 3 (Δ_7, Δ_8, Δ_9) IBD coefficients per dyad.

Comparing inbreeding coefficient estimates of a mixed sample containing two FSFS individuals and N-2 (N = 3, 4, ..., 10) unrelated individuals from diallelic marker data (Supplementary Appendix 4), it is clear that $\hat{r}_{EM1}$ is much less biased and much more accurate than the other three estimators for both non-inbred individuals (expected $F = 0$) and the FSFS inbred individuals (expected $F = 0.25$). By accounting for both the small size and the IBD structure of a sample, $\hat{r}_{EM1}$ yields highly accurate F estimates even when a sample has only three individuals (2 are FSFS). In contrast, the other three estimators are increasingly biased with a decreasing sample size N. The two moment estimators yield negative F estimates for non-inbred individuals, and for FSFS individuals as well when N is small (i.e. $N < 6$). $\hat{r}_{EM2}$ is better than the two moment estimators for both inbred and non-inbred individuals and regardless of sample size, because it accounts for small sample sizes.

3.4 | Analysis of the CEPH data

For UR, $\hat{r}_{EM1}$ and $\hat{r}_{EM2}$ yield perfect estimates of $\hat{\Delta}_7$, $\hat{\Delta}_8$, and θ (Figure 4). This is because the expected values, in the absence of background relatedness, are $\hat{\Delta}_7 = \hat{\Delta}_8 = \theta = 0$, and likelihood estimators are lower bounded by 0. In contrast, moment estimators $\hat{r}_{LR}$ and $\hat{r}_W$ give consistently negative estimates of Δ_8 and Δ_7, respectively. As a result, they yield negative estimates of θ.

For PO, $\hat{r}_{EM1}$ and $\hat{r}_{EM2}$ give 0 or close to 0 estimates of Δ_7 which has an expected value of 0. However, while $\hat{\Delta}_8$ estimates from $\hat{r}_{EM1}$ are close to the expected value of 1, $\hat{\Delta}_8$ estimates from $\hat{r}_{EM2}$ are scattered around 0. Therefore, $\hat{\theta}$ is well estimated by $\hat{r}_{EM1}$ with an average of 0.24, but is much underestimated by $\hat{r}_{EM2}$ with an average of 0.002. Estimator $\hat{r}_W$ gives negative estimates of Δ_7 and positive estimates of Δ_8. Both $\hat{\Delta}_7$ and $\hat{\Delta}_8$ are highly scattered, especially $\hat{\Delta}_8$. Because $\hat{\Delta}_7$ and $\hat{\Delta}_8$ are negatively correlated, $\hat{r}_W$ still yields good estimates of θ, which are only slightly lower on average than the expected value of 0.25. $\hat{\Delta}_7$ and $\hat{\Delta}_8$ from $\hat{r}_{LR}$ are both less scattered than those from $\hat{r}_W$. However, $\hat{\Delta}_7$ and $\hat{\Delta}_8$ are highly negatively correlated, and both tend to be negative. As a result, $\hat{\theta}$ is also negative, with an average of −0.1.

For FS, $\hat{r}_{EM1}$ underestimates both $\hat{\Delta}_7$ and $\hat{\Delta}_8$ slightly and thus $\hat{\theta}$ slightly. The average $\hat{\theta}$ is 0.16. If allele frequencies are not updated, however, $\hat{\Delta}_7$, $\hat{\Delta}_8$ and $\hat{\theta}$ are all severely underestimated by $\hat{r}_{EM2}$. $\hat{r}_W$ yields positive estimates of $\hat{\Delta}_7$, $\hat{\Delta}_8$ and $\hat{\theta}$, with an average $\hat{\theta}$ of 0.16. However, $\hat{\Delta}_8$ is consistently overestimated and $\hat{\Delta}_7$ is consistently underestimated. $\hat{r}_{LR}$ gives positive $\hat{\Delta}_7$ but highly negative $\hat{\Delta}_8$ and $\hat{\theta}$, with an average $\hat{\theta}$ of −0.25.

$\hat{r}_{EM1}$ also yields estimates of Δ_1, Δ_2,..., Δ_9 as listed in Table 4. All of the six Δ coefficients involving within individual IBD, Δ_1, Δ_2,..., Δ_6, are estimated to be close to zero for all of the three relationships (UR, PO and FS).

The best estimates are obtained obviously from estimator $\hat{r}_{EM1}$. It yields the least biased $\hat{\Delta}_7$, $\hat{\Delta}_8$ and $\hat{\theta}$ across the three relationships. For a small sample of four highly related individuals as is the case of this dataset, updating allele frequencies by accounting for the inferred relatedness structure of the sample and accommodating the small sample size by using unbiased estimates of allele frequency powers and products prove important for accurate estimates of relatedness.

3.5 | Analysis of the ant data

Both $\hat{r}_{EM1}$ and $\hat{r}_{EM2}$ reveal the 10-colony structure of the sampled 377 individuals (Figure 5) using genotype data from just 6 microsatellites (with a high rate of missing data). However, by estimating allele frequencies and IBD coefficients jointly, estimator $\hat{r}_{EM1}$ also reveals inter colony genetic structures which are absent or quite vague when allele frequencies are not updated by estimator $\hat{r}_{EM2}$. For example, colonies 1 and 3 (counting from bottom up, or from left to right), colonies 2 and 4, and colonies 3 and 4 are related according to estimator $\hat{r}_{EM1}$. However, this is not true according to estimator $\hat{r}_{EM2}$. This inter-colony structure is expected from the ant reproduction nature of budding.

Estimates of Δ_1, Δ_2, ..., Δ_9 were obtained from estimator $\hat{r}_{EM1}$ for each dyad in the sample of 377 ant workers. Similar to $\hat{\theta}$ shown in Figure 5, $\hat{\Delta}_7$, $\hat{\Delta}_8$ and $\hat{\Delta}_9$ display clearly the 10-colony structure (Supplementary Appendix 3). Estimates of each of the remaining IBD coefficients, Δ_j for j = 1, 2, ..., 6, do not show this colony structure, but still have vaguely the colony block structure. Colony 10 (on the top right corner of the heatmaps) is much more inbred than other colonies as it has higher Δ_1, Δ_3 and Δ_5 estimates (Supplementary Appendix 3).

3.6 | Analysis of the HGDP data

By updating allele frequencies jointly with IBD coefficients, $\hat{r}_{EM1}$ reveals the population structure measured by pairwise relatedness (Figure 6) much more clearly than $\hat{r}_{EM2}$, a structure in broad agreement with that inferred by a model-based admixture analysis by PopCluster (Wang, 2022a, 2022b) assuming $K = 5$ populations.

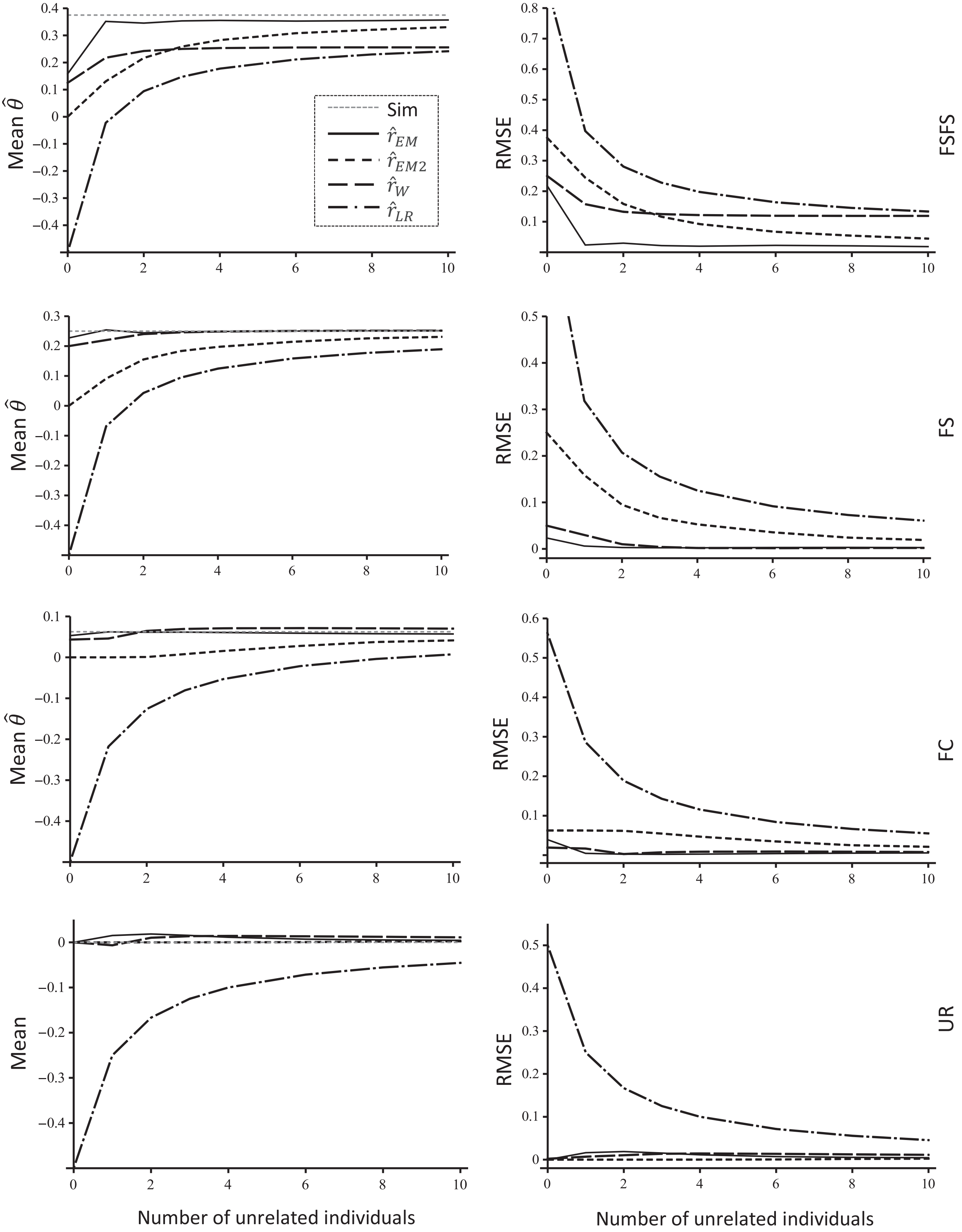
Mean $\hat{\theta}$
RMSE
FSFS
FS
FC
UR
Mean
Sim
$\hat{r}_{EM}$
$\hat{r}_{EM2}$
$\hat{r}_{W}$
$\hat{r}_{LR}$
0.4
0.3
0.2
0.1
0
−0.1
−0.2
−0.3
−0.4
0.8
0.7
0.6
0.5
0
2
4
6
8
10
Number of unrelated individuals
Number of unrelated individuals

FIGURE 3 Mean and root mean squared error (RMSE) of $\hat{\theta}$ estimated by $\hat{r}_{EM1}$, $\hat{r}_{EM2}$, $\hat{r}_{LR}$ and $\hat{r}_{W}$ for relationships full sibs whose parents are full sibs (FSFS), full sibling (FS), first cousin (FC) and unrelated (UR) as a function of the number of unrelated individuals (x axis) in a sample. The total number of individuals in a sample is 2 related (or unrelated in the case of UR) individuals plus the number of unrelated individuals shown on x-axis. The expected θ values for FSFS, FS, FC and UR are 0.375, 0.25, 0.0625 and 0 shown by the horizontal dotted lines. A number of $L = 100{,}000$ single nucleotide polymorphisms (SNPs) with allele frequencies in uniform distribution are genotyped for each individual.

4 | DISCUSSION

Genetic relatedness and inbreeding coefficients are relative quantities defined and measured with a reference population (Wang, 2016; Weir et al., 2006) in which all homologous alleles are assumed non-IBD or equivalently all individuals are assumed non-inbred and unrelated. When measured from pedigree data, the reference is the set of founders of the pedigree who have no parental information and are thus assumed unrelated with each other and non-inbred. When calculated from marker data, the reference is the population whose allele frequencies are used in inferring relatedness (Wang, 2014; Weir et al., 2006). However, allele frequencies in real-world populations are rarely known but are frequently estimated from the same genotype data that are used in calculating relatedness. Invariably in the literature, all sampled individuals are assumed non-inbred and unrelated a priori in estimating allele frequencies. When a sample contains a substantial proportion of close relatives, however, the assumption is violated, leading to poor allele frequency estimates and thus poor relatedness inferences as demonstrated by this study. Because some alleles in close relatives are IBD but assumed non-IBD, the frequency of an allele tends to be overestimated and underestimated when it is and is not found in the genotypes of the relatives, respectively. These distorted allele frequency estimates lead, in turn, to an underestimation of relatedness of close relatives and an overestimation of relatedness of unrelated individuals. These biased relatedness estimates can cause serious problems in a downstream analysis. For example, first-degree relatives might be inferred as second-degree ones. Unfortunately, using more markers does not help to ease the problem, and the poor relatedness inferences persist even when genomic markers are used. In fact, more markers make the problem more acute as bias (rather than precision) could become the dominating factor in determining the overall estimation accuracy.

Two approaches can be used to reduce or eliminate the problem. Experimentally, one should avoid non-probability sampling such as convenience sampling. This is especially important for sampling individuals of early-life stages in highly fecund species, such as eggs or fries of fish and tadpoles of frogs (Hansen et al., 1997). Avoiding clusters of close relatives included in a sample will ensure unbiased allele frequency estimation and thus unbiased relatedness estimation from the sampled genotype data. Statistically, one should estimate allele frequencies by accounting for the genetic structure (inbreeding and relatedness) of the sample. The composite likelihood method proposed in this study is the first to estimate allele frequencies and relatedness jointly without imposing any ad hoc pedigree structure to a sample. The EM algorithm developed in this study updates allele frequencies and IBD coefficients in alternative iterations such that each is estimated by accounting for the other. The method ($\hat{r}_{EM1}$) yields less biased and more accurate relatedness estimates than other methods when a high proportion of close relatives are included in a sample, as shown by simulations and analysis of several real datasets.

In the same spirit, relatedness estimation for admixed or non-admixed individuals in a heterogeneous population with subpopulation structure warrants special attention when we are interested in the most recent familial relationships within a subpopulation. In such a situation, subpopulation-specific or individual-specific allele frequencies must be calculated from the heterogenous sample of individuals, which are then used in estimating relatedness (e.g. Conomos et al., 2016; Manichaikul et al., 2010; Moltke & Albrechtsen, 2014; Thornton et al., 2012). The methods developed so far invariably take a two-step approach. First, a population admixture analysis is conducted, using STRUCTURE (Pritchard et al., 2000) like programs, to obtain estimates of the admixture proportions of each sampled individual and estimates of subpopulation allele frequencies. Second, individual-specific allele frequencies, calculated from individual admixture estimates and subpopulation allele frequency estimates, are then used for calculating relatedness by a moment or likelihood estimator (Conomos et al., 2016; Manichaikul et al., 2010; Moltke & Albrechtsen, 2014; Thornton et al., 2012). All of these estimators assume the absence of inbreeding to estimate $\{\Delta_7, \Delta_8, \Delta_9\}$ only. In principle, however, it is possible to remove the assumption from the likelihood estimator (e.g. Moltke & Albrechtsen, 2014) to estimate all nine IBD coefficients, and thus both inbreeding and relatedness.

Also in the same spirit, Bayesian (Ayres & Balding, 1998; Vogl et al., 2002) and likelihood (Hall et al., 2012) methods were developed to estimate allele frequencies $\mathbf{p}$ and inbreeding coefficients F jointly from the multilocus genotype data of a sample of unrelated individuals. However, the simulation by Hall et al. (2012) showed that it did not matter much to estimate F with $\mathbf{p}$ estimated by either sophisticated methods (i.e. estimating F and $\mathbf{p}$ jointly by likelihood or Bayesian methods) or simple allele counting under the assumption of no inbreeding. This conclusion is not too surprising as inbreeding is expected to have a minor effect on $\mathbf{p}$ estimates except when sample size N is extremely small. Consider a sample of N diploid individuals. The number of within-individual allele pairs is N, while the number of between-individual allele pairs is $2N(2N-1)/2 - N = 2N(N-1)$. Even with a small sample of $N = 20$ individuals, only 20 out of 780 pairs of alleles reside within individuals and are thus potentially affected by inbreeding in estimating $\mathbf{p}$, while 760 out of 780 pairs of alleles are from between individuals and are thus potentially affected by relatedness in estimating $\mathbf{p}$. When individuals are unrelated as assumed in these studies, therefore, inbreeding has little effect on $\mathbf{p}$ estimation and thus on F estimation, as found by Hall et al. (2012).

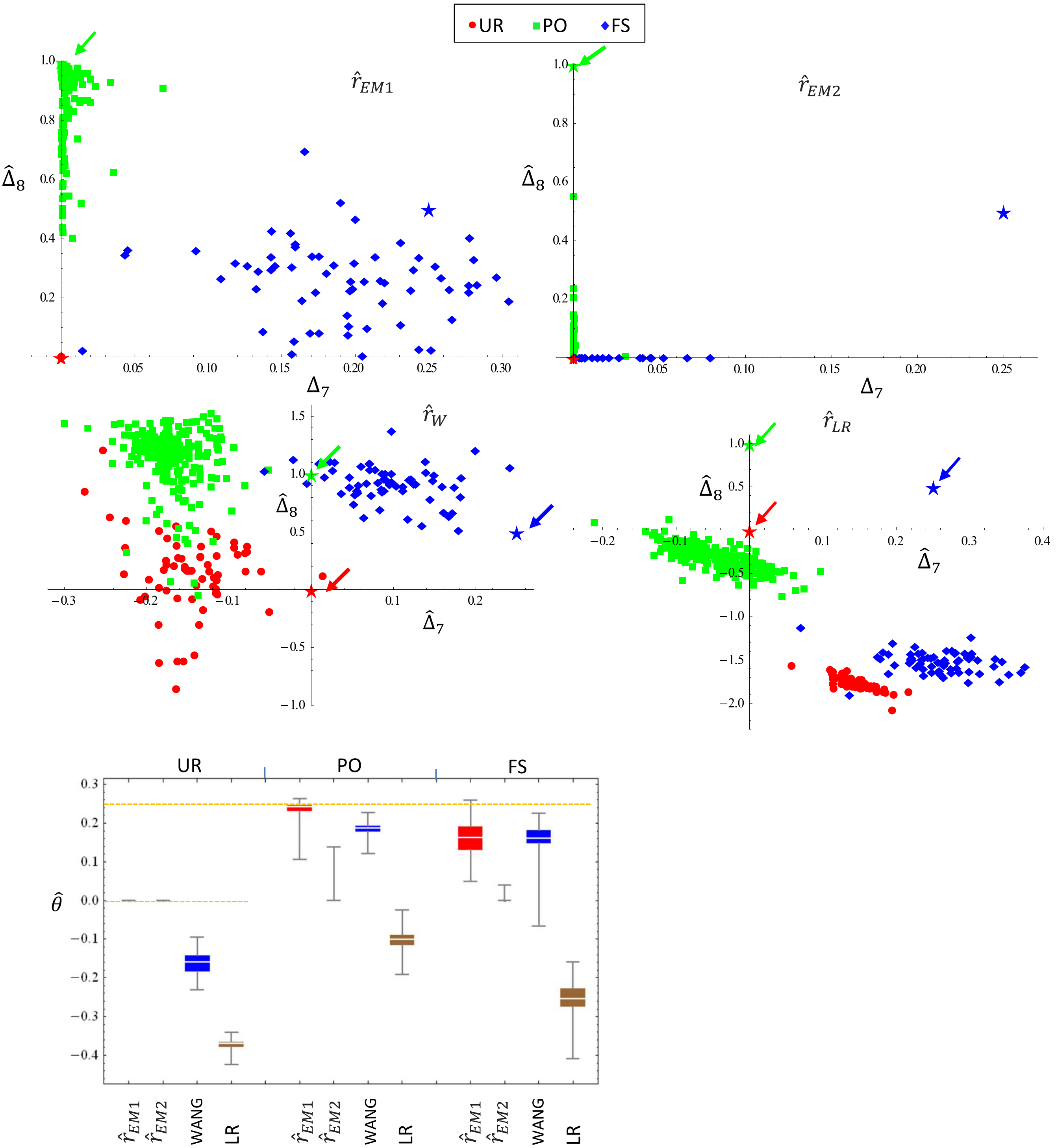

FIGURE 4 Estimated Δ_7 (*x* axis) and Δ_8 (*y* axis) of each of 6 dyads in each of 65 nuclear (2 parents with two full-sib children) families of the CEPH data by estimators $\hat{r}_{EM1}$, $\hat{r}_{EM2}$, $\hat{r}_{LR}$ and $\hat{r}_{W}$. The expected values of Δ_7 and Δ_8 are {0, 0} for unrelated (UR, denoted by red star and arrow), {0, 1} for parent–offspring (PO, denoted by green star and arrow), and {0.25, 0.5} for full sibs (FS, denoted by blue star and arrow). The bottom is the box whisker chart of θ estimates by the four estimators. The orange horizontal lines indicate the expected values of θ for UR (0), and PO and FS (0.25).

In contrast, the above analysis also shows that relatedness could potentially have a large impact on the estimation of **p**. How to estimate **p**, either by simple allele counting (by assuming non-inbred and unrelated individuals) or by estimating **p** and IBD jointly, can be thus important in estimating IBD and summary statistics like relatedness and inbreeding.

The above argument about the differential effects of inbreeding and relatedness on the estimation of **p** can also be understood by considering the effective sample size, ESS. For a sample of *N* diploid

TABLE 4 Estimates of Δ by $\hat{r}_{EM1}$ for dyads in 65 nuclear families of the CEPH data

Relationship	Statistics	Δ_1	Δ_2	Δ_3	Δ_4	Δ_5	Δ_6	Δ_7	Δ_8	Δ_9
UR	$\bar{X}$	0.000	0.001	0.001	0.002	0.001	0.004	0.001	0.003	0.987
	SD	0.001	0.001	0.002	0.004	0.003	0.014	0.001	0.006	0.031
PO	$\bar{X}$	0.001	0.001	0.003	0.002	0.004	0.003	0.006	0.895	0.087
	SD	0.003	0.002	0.008	0.004	0.017	0.012	0.019	0.141	0.127
FS	$\bar{X}$	0.002	0.001	0.002	0.005	0.004	0.002	0.185	0.276	0.522
	SD	0.002	0.004	0.003	0.025	0.008	0.004	0.063	0.140	0.146

Notes: $\bar{X}$ and *SD* are the mean and standard deviation of the estimates. Without background IBD, relationship UR, PO and FS are expected to have $\{\Delta_1, \Delta_2, \dots, \Delta_9\} = \{0,0,0,0,0,0,0,0,1\}$, $\{0,0,0,0,0,0,0,1,0\}$ and $\{0,0,0,0,0,0,0.25,0.5,0.25\}$, respectively.

Abbreviations: FS, full sibling; IBD, identical by descent; PO, parent–offspring; UR, unrelated.

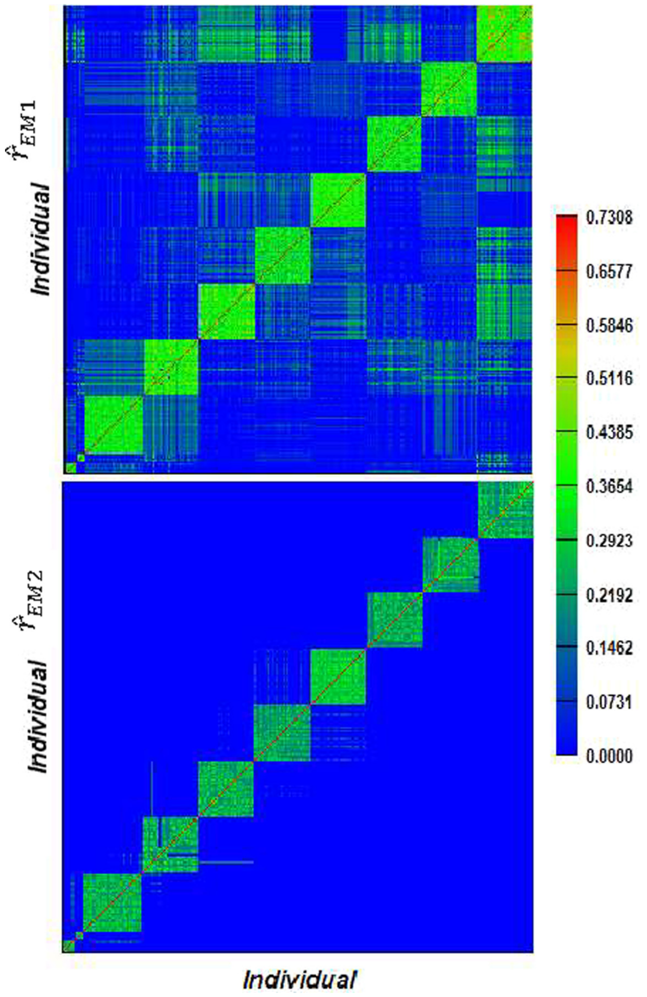

FIGURE 5 Relatedness of 377 ant workers. The upper panel shows relatedness estimated by updating both relatedness and allele frequencies by the expectation maximization (EM) algorithm ($\hat{r}_{EM1}$), while the lower panel shows relatedness estimated by updating relatedness only by the EM algorithm ($\hat{r}_{EM2}$) with allele frequencies estimated by assuming unrelated and non-inbred individuals in the sample. Both *x* and *y* axes show ordered individuals 1–377, with individuals 1–9 from colony 1, 10–16 from colony 2, 17–63 from colony 3, 64–108 from colony 4, 109–153 from colony 5, 154–198 from colony 6, 199–243 from colony 7, 244–288 from colony 8, 289–332 from colony 9, and 333–377 from colony 10.

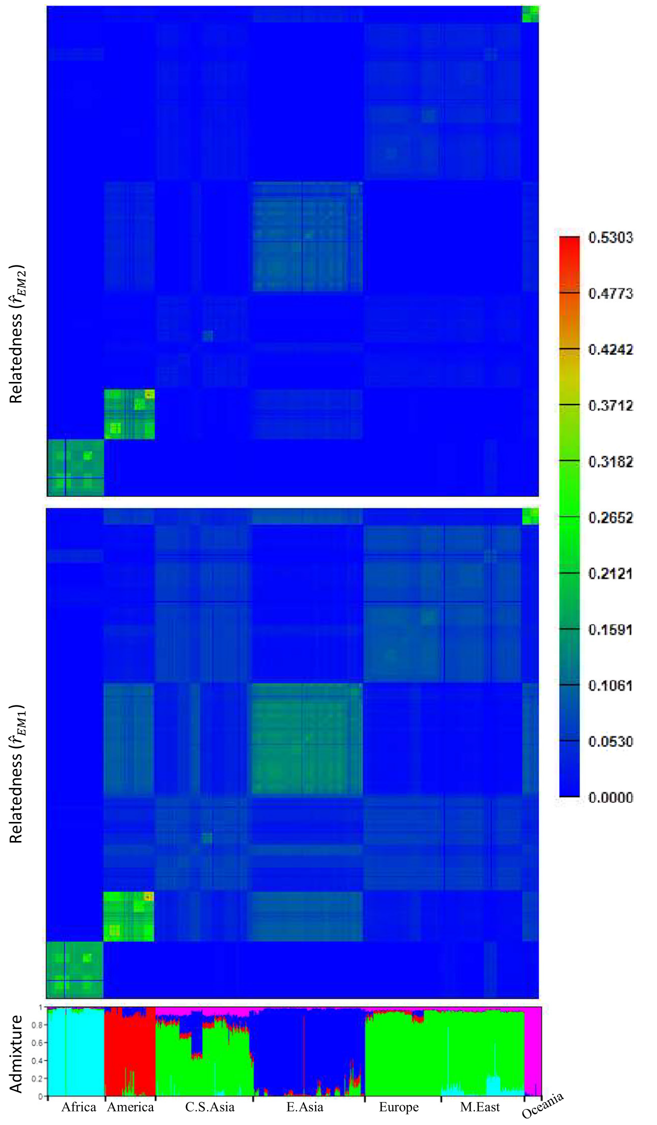

FIGURE 6 Genetic structure of 1,043 human individuals genotyped at 644,258 single nucleotide polymorphisms (SNPs) (644,199 polymorphic) of the Human Genome Diversity Panel (HGDP). The upper and middle panels show heatmaps of relatedness estimated by not updating (by $\hat{r}_{EM2}$) and updating (by $\hat{r}_{EM1}$) allele frequencies, respectively. The lower panel shows the admixture at $K = 5$ estimated by PopCluster. The *x*-axes of the upper, middle and lower panels and the *y*-axes of the upper and middle panels list ordered individuals.

individuals, the maximum ESS is 2*N* when all individuals are non-inbred and unrelated. The minimum ESS is *N* when all individuals are completely inbred ($F = 1$) but are unrelated, and is only 2 when all individuals are completely related but are non-inbred. For a sample of *N* (≥2) FS whose parents are non-inbred and unrelated, the ESS is only 4, despite of a potentially large *N* value. Indeed, Figures 2 and 3 show that, in the presence of relatedness structure in a sample of individuals, how to estimate **p** has a large impact on the estimation of IBD.

Another problem in estimating relatedness stems from a small sample size (Wang, 2017). When only a small number of individuals are sampled and genotyped for relatedness analysis, allele frequencies can still be estimated unbiasedly albeit with low precision (i.e. with a large sampling variance). However, all relatedness estimators use powers and products of allele frequencies in the estimation, which are unfortunately biased if calculated directly from the unbiased allele frequency estimates when sample size is small (Wang, 2017; Weir, 1996). As shown before (Wang, 2017) and herein, estimators assuming known allele frequencies or a large sample size to ignore sampling effects (say, $N > 100$, Ritland, 1996) can lead to severely biased estimates of relatedness between individuals in a small sample. The likelihood estimators developed herein, $\hat{r}_{EM1}$ and $\hat{r}_{EM2}$, allow for small sample sizes by using unbiased estimators of the powers and products of allele frequencies (Appendix 1). They are demonstrated to work better than other estimators and are almost unbiased even when only two individuals are sampled for analysis of relatedness (Figures 2 and 3). Small samples in the real world are surprisingly common, such as ancient samples (e.g. museum samples, excavated fossil bones), mixed samples of unknown sources (e.g. confiscated animal products, victims of a disaster) and samples from highly endangered species. Samples for genomic studies can also be small due to cost and other restrictions. With a high rate of missing data typical of next-generation sequencing, the effective sample sizes can even be smaller because it is the number of complete genotypes (not individuals) that counts.

It is encouraging that by estimating IBD and allele frequencies jointly and by accounting for small sample sizes, the likelihood estimator $\hat{r}_{EM1}$ can yield reasonably accurate estimates of relatedness for individuals of various genealogical relationships in a small sample or in a sample containing a high proportion of close relatives (Figures 2–4; Supplementary Appendix 2). $\hat{r}_{EM1}$ works well even in the extreme case of a sample of two close relatives, from which both allele frequencies and relatedness must be deduced. The estimator is thus especially useful when genomic data are available for just a few individuals sampled from, say, museum or excavated fossil bones. However, the estimator is also valuable for a large sample with hundreds of individuals when the sample is highly genetically structured, as demonstrated by the ant and the human datasets (Figures 5 and 6).

Except for a couple of exceptions (e.g. Wang, 2007), previous likelihood estimators of IBD or relatedness assume the absence of inbreeding (i.e. $\Delta_1 = \Delta_2 = \Delta_3 = \Delta_4 = \Delta_5 = \Delta_6 = 0$) and estimate $\{\Delta_7, \Delta_8, \Delta_9\}$ only. The estimation complexity and computation are greatly reduced because only 3 out of the 9 IBD coefficients for a dyad need to be estimated. However, inbreeding is likely to occur whenever relatedness occurs. This is true even when mating is at random with respect to ancestry. For small populations or large populations with regular close relative mating such as selfing, inbreeding is frequent and should be taken into account in measuring and estimating relatedness. Furthermore, genomic marker data provide sufficient information to delineate the detailed IBD relationships between a pair of individuals. Both estimators developed in this study, $\hat{r}_{EM1}$ and $\hat{r}_{EM2}$, allow for inbreeding in estimating relatedness, and can estimate accurately the nine detailed IBD coefficients of a dyad as well when markers have more than three alleles. In the same situations of a small sample or a highly structured sample, $\hat{r}_{EM1}$ yields good estimates of *F* (Supplementary Appendix 4).

Estimators $\hat{r}_{EM1}$ and $\hat{r}_{EM2}$ are developed for estimating the nine detailed IBD coefficients of a dyad from their genotype data. However, a locus with 2 or 3 alleles (i.e. $n_l < 4$) has fewer IIS modes than IBD modes, and therefore does not afford sufficient information to distinguish and estimate all nine IBD coefficients, as proved by Csűrös (2014). The issue persists even when many two- or three-allele loci are used in an analysis. Fortunately, the most widely applied summary IBD statistics, including coancestry θ and inbreeding coefficient *F*, are still estimable from diallelic (Csűrös, 2014) or triallelic markers. The simulation results in Tables 2 and 3 confirm that, for inbred relationships (i.e. $\Delta_1 + \Delta_2 + \Delta_3 + \Delta_4 + \Delta_5 + \Delta_6 > 0$) such as FSFS, some IBD coefficients are poorly estimated by $\hat{r}_{EM1}$ from diallelic or triallelic marker data. However, the most commonly used summary IBD statistics, θ and *F*, are still accurately calculated by $\hat{r}_{EM1}$ (Tables 2 and 3). The much higher accuracy of $\hat{r}_{EM1}$ over other estimators for estimating θ from a small sample of individuals (Figures 2 and 3) further consolidate the conclusion. For non-inbred relationships (i.e. $\Delta_7 + \Delta_8 + \Delta_9 \equiv 1$) such as FS, all nine IBD coefficients are well estimated by $\hat{r}_{EM1}$, regardless of the number of alleles per locus (Table 3). Therefore, it can be concluded in general that estimators $\hat{r}_{EM1}$ and $\hat{r}_{EM2}$ are applicable to diallelic as well as multiallelic markers for estimating both θ and *F*, although they may provide poor estimates of some of the nine IBD coefficients for inbred relationships when markers are diallelic or triallelic.

Why diallelic markers do not afford reliable estimates of all nine IBD coefficients but these estimates could still be combined to yield accurate estimates of summary IBD statistics such as θ and *F*, as demonstrated by the present study? Fundamentally, diallelic markers provide sufficient information only for some summary statistics but not for the detailed components of the summary statistics. Although some Δ_j (for $j = 1, 2, ..., 9$) may not be accurately estimated by $\hat{r}_{EM1}$ from diallelic markers, these estimates are inherently constrained, as verified by the complicated correlation structures found in simulated data (Supplementary Appendix 1). Some of these Δ_j estimates are highly negatively correlated that a linear combination of them leads to a much better estimate of a summary IBD statistic. Let us take inbreeding coefficient, *F*, as an example. Diallelic markers afford accurate delineation of individual *F*. The relative homozygosity across diallelic loci of an individual X lends a good estimate of F_X. For two individuals X and Y, $\hat{r}_{EM1}$ (implementing likelihood Equation (4)) may not give good estimates of Δ_1, Δ_2,

Δ_3 and Δ_4. However, these estimates are inherently correlated when $F_X > 0$, especially between Δ_2 and Δ_4 and between Δ_3 and Δ_4 which show negative correlations (Supplementary Appendix 1). Therefore, when estimated by (1) as $\Delta_1 + \Delta_2 + \Delta_3 + \Delta_4$, F_X is still accurately inferred.

For a pairwise relatedness estimator, the number of dyads and thus computational time increase quadratically with sample size. Using an EM algorithm to estimate iteratively both allele frequencies and pairwise IBD coefficients, the $\hat{r}_{EM1}$ estimator is much more complicated and runs much slower than other estimators, such as likelihood estimator $\hat{r}_{EM2}$. For genomic marker data with hundreds of thousands of loci, it could handle a sample with hundreds of individuals if MPI and openMP parallel runs are conducted on a cluster with many cores. It is developed for and suits to the analysis of a small sample or a sample containing a high proportion of close relatives. For a large sample deemed to contain a small proportion of close relatives, it is better to use $\hat{r}_{EM2}$ or moment estimators. I used Dempster et al.'s (1977) EM algorithm in both $\hat{r}_{EM1}$ and $\hat{r}_{EM2}$ estimators. The algorithm can be further accelerated by various methods, such as the squared iterative method (Varadhan & Roland, 2008) and Aitken acceleration method (Chow & Kay, 1984). How to improve the EM algorithm in $\hat{r}_{EM1}$ estimator by adopting these accelerations deserves further studies.

AUTHOR CONTRIBUTIONS

Jinliang Wang conceived and conducted the study and wrote the manuscript.

ACKNOWLEDGEMENT

I thank Senior Editor Natalie Cooper, an anonymous editor and two anonymous referees for valuable comments that are helpful for improving earlier versions of this manuscript.

CONFLICT OF INTEREST

The author has no conflicts of interest to declare.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13963

DATA AVAILABILITY STATEMENT

No new empirical data were generated in this study. The simulated data can be reproduced using the software package (Windows version) available on Zoological Society of London at https://www.zsl.org/science/software/EMIBD9 and on Zenodo at https://doi.org/10.5281/zenodo.6672390 (Wang, 2022b).

ORCID

Jinliang Wang https://orcid.org/0000-0002-8467-5448

REFERENCES

Anderson, A. D., & Weir, B. S. (2007). A maximum-likelihood method for the estimation of pairwise relatedness in structured populations. *Genetics*, *176*(1), 421–440.

Ayres, K. L., & Balding, D. J. (1998). Measuring departures from Hardy–Weinberg: A Markov chain Monte Carlo method for estimating the inbreeding coefficient. *Heredity*, *80*(6), 769–777.

Cannings, C., Thompson, E. A., & Skolnick, M. H. (1978). Probability functions on complex pedigrees. *Advances in Applied Probability*, *10*(1), 26–61.

Choi, Y., Wijsman, E. M., & Weir, B. S. (2009). Case-control association testing in the presence of unknown relationships. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, *33*(8), 668–678.

Chow, Y. K., & Kay, S. (1984). On the Aitken acceleration method for nonlinear problems. *Computers & Structures*, *19*(5–6), 757–761.

Conomos, M. P., Reiner, A. P., Weir, B. S., & Thornton, T. A. (2016). Model-free estimation of recent genetic relatedness. *The American Journal of Human Genetics*, *98*(1), 127–148.

Cowell, R. G. (2009). Efficient maximum likelihood pedigree reconstruction. *Theoretical Population Biology*, *76*(4), 285–291.

Csűrös, M. (2014). Non-identifiability of identity coefficients at biallelic loci. *Theoretical Population Biology*, *92*, 22–29.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22.

Elston, R. C., & Stewart, J. (1971). A general model for the genetic analysis of pedigree data. *Human Heredity*, *21*(6), 523–542.

Goudet, J., Kay, T., & Weir, B. S. (2018). How to estimate kinship. *Molecular Ecology*, *27*(20), 4121–4135.

Hall, N., Mercer, L., Phillips, D., Shaw, J., & Anderson, A. D. (2012). Maximum likelihood estimation of individual inbreeding coefficients and null allele frequencies. *Genetics Research*, *94*(3), 151–161.

Hammond, R. L., Bourke, A. F. G., & Bruford, M. W. (2001). Mating frequency and mating system of the polygynous ant, *Leptothorax acervorum*. *Molecular Ecology*, *10*(11), 2719–2728.

Hansen, M. M., Nielsen, E. E., & Mensberg, K. L. (1997). The problem of sampling families rather than populations: Relatedness among individuals in samples of juvenile brown trout *Salmo trutta* L. *Molecular Ecology*, *6*(5), 469–474.

Harris, D. L. (1964). Genotypic covariances between inbred relatives. *Genetics*, *50*(6), 1319–1348.

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

Li, C. C., & Horvitz, D. G. (1953). Some methods of estimating the inbreeding coefficient. *American Journal of Human Genetics*, *5*(2), 107–117.

Li, C. C., Weeks, D. E., & Chakravarti, A. (1993). Similarity of DNA fingerprints due to chance and relatedness. *Human Heredity*, *43*(1), 45–52.

Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., Cann, H. M., Barsh, G. S., Feldman, M., Cavalli-Sforza, L. L., & Myers, R. M. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, *319*(5866), 1100–1104.

Loiselle, B. A., Sork, V. L., Nason, J., & Graham, C. (1995). Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). *American Journal of Botany*, *82*(11), 1420–1425.

Lynch, M. (1988). Estimation of relatedness by DNA fingerprinting. *Molecular Biology and Evolution*, *5*(5), 584–599.

Lynch, M., & Ritland, K. (1999). Estimation of pairwise relatedness with molecular markers. *Genetics*, *152*(4), 1753–1766.

Malécot, G. (1948). *Les mathématiques de l'hérédité*. Masson et Cie.

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

Milligan, B. G. (2003). Maximum-likelihood estimation of relatedness. *Genetics*, *163*(3), 1153–1167.

Moltke, I., & Albrechtsen, A. (2014). RelateAdmix: A software tool for estimating relatedness between admixed individuals. *Bioinformatics*, *30*(7), 1027–1028.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1996). *Numerical Recipes in Fortran 77* (2nd ed.). Cambridge University Press.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–959.

Ritland, K. (1996). Estimators for pairwise relatedness and individual inbreeding coefficients. *Genetical Research*, *67*(2), 175–186.

Thornton, T., Tang, H., Hoffmann, T. J., Ochs-Balcom, H. M., Caan, B. J., & Risch, N. (2012). Estimating kinship in admixed populations. *The American Journal of Human Genetics*, *91*(1), 122–138.

Varadhan, R., & Roland, C. (2008). Simple and globally convergent methods for accelerating the convergence of any EM algorithm. *Scandinavian Journal of Statistics*, *35*(2), 335–353.

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

Vogl, C., Karhu, A., Moran, G., & Savolainen, O. (2002). High resolution analysis of mating systems: Inbreeding in natural populations of *Pinus radiata*. *Journal of Evolutionary Biology*, *15*(3), 433–439.

Wang, J. (2002). An estimator for pairwise relatedness using molecular markers. *Genetics*, *160*(3), 1203–1215.

Wang, J. (2004). Sibship reconstruction from genetic data with typing errors. *Genetics*, *166*(4), 1963–1979.

Wang, J. (2007). Triadic IBD coefficients and applications to estimating pairwise relatedness. *Genetics Research*, *89*(3), 135–153.

Wang, J. (2011a). Unbiased relatedness estimation in structured populations. *Genetics*, *187*(3), 887–901.

Wang, J. (2011b). COANCESTRY: A program for simulating, estimating and analysing relatedness and inbreeding coefficients. *Molecular Ecology Resources*, *11*(1), 141–145.

Wang, J. (2011c). A new likelihood estimator and its comparison with moment estimators of individual genome-wide diversity. *Heredity*, *107*(5), 433–443.

Wang, J. (2012). Computationally efficient sibship and parentage assignment from multilocus marker data. *Genetics*, *191*(1), 183–194.

Wang, J. (2014). Marker-based estimates of relatedness and inbreeding coefficients: An assessment of current methods. *Journal of Evolutionary Biology*, *27*(3), 518–530.

Wang, J. (2016). Pedigrees or markers: Which are better in estimating relatedness and inbreeding coefficient? *Theoretical Population Biology*, *107*, 4–13.

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

Wang, J. (2022a). Fast and accurate population admixture inference from genotype data from a few microsatellites to millions of SNPs. *Heredity*, *129*, 79–92. https://doi.org/10.1038/s41437-022-00535-z

Wang, J. (2022b). A joint likelihood estimator of relatedness and allele frequencies from a small sample of individuals. Version 1.0.0.0. https://doi.org/10.5281/zenodo.6672390.

Wang, J., & Santure, A. (2009). Parentage and sibship inference from multilocus genotype data under polygamy. *Genetics*, *181*(4), 1579–1594.

Weir, B. S. (1996). *Genetic data analysis II*. Sinauer Associates.

Weir, B. S., Anderson, A. D., & Hepler, A. B. (2006). Genetic relatedness analysis: Modern data and new challenges. *Nature Reviews Genetics*, *7*(10), 771–780.

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

Wright, S. (1922). Coefficients of inbreeding and relationship. *The American Naturalist*, *56*(645), 330–338.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics*, *11*(1), 95–103.

Zhang, Q. S., Goudet, J., & Weir, B. S. (2022). Rank-invariant estimation of inbreeding coefficients. *Heredity*, *128*(1), 1–10.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Wang, J. (2022). A joint likelihood estimator of relatedness and allele frequencies from a small sample of individuals. *Methods in Ecology and Evolution*, *00*, 1–20. https://doi.org/10.1111/2041-210X.13963

APPENDIX 1

Expectations of allele frequency powers and products

Suppose N diploid individuals were sampled at random from a large random mating population and genotyped at a locus with n observed codominant alleles $\{A_1, A_2, \ldots, A_n\}$. The frequency of allele A_i ($i = 1, 2, \ldots, n$) in the population is p_i. We are interested in estimating allele frequency powers, such as p_1^2, and allele frequency products, such as $p_1^2 p_2^2$, from the sample of N genotypes. As an example, consider the estimation of p_1^2. Under the above assumption, the number of copies of A_1 in the sample of $2N$ genes, x_1, follows a binomial distribution with parameters 2 N and p_1. The expected value of $\left(\frac{x_1}{2N}\right)^2$ is

$$E\left(\frac{x_1}{2N}\right)^2 = \sum_{j=0}^{2N} \frac{(2N)!}{j!(2N-j)!} p_1^j (1-p_1)^{2N-j} \left(\frac{j}{2N}\right)^2 = \frac{p_1 + (2N-1)p_1^2}{2N}. \quad \text{(A1.1)}$$

When p_1 is replaced by its unbiased estimator, $\hat{p}_1 = x_1/(2N)$, and $E\left(\frac{x_1}{2N}\right)^2$ is replaced by its observation $\left(\frac{x_1}{2N}\right)^2$, (A1.1) reduces to the estimator of p_1^2,

$$\hat{p}_1^2 = \frac{x_1(x_1-1)}{2N(2N-1)}. \quad \text{(A1.2)}$$

Similarly, the expected value of $\left(\frac{x_1}{2N}\right)^3$ is

$$E\left(\frac{x_1}{2N}\right)^3 = \sum_{j=0}^{2N} \frac{(2N)!}{j!(2N-j)!} p_1^j (1-p_1)^{2N-j} \left(\frac{j}{2N}\right)^3 = \frac{p_1 + 3(2N-1)p_1^2 + (2N-1)(2N-2)p_1^3}{(2N)^2}. \quad \text{(A1.3)}$$

When p_1 and p_1^2 are replaced by their unbiased estimators, $\hat{p}_1 = x_1/(2N)$ and (A1.2), and $E\left(\frac{x_1}{2N}\right)^3$ is replaced by its observation $\left(\frac{x_1}{2N}\right)^3$, (A1.3) reduces to the estimator of p_1^3,

$$\hat{p}_1^3 = \frac{x_1(x_1-1)(x_1-2)}{2N(2N-1)(2N-2)}. \quad \text{(A1.4)}$$

Other estimators of allele frequency moments listed in (3) can be derived similarly.

APPENDIX 2

EM algorithm for estimating p and Δ

To facilitate the description and implementation of the EM algorithm, I introduce an auxiliary indicator variable Z. The indicator variables for the IBD modes for individuals X and Y at locus l are $\mathbf{Z}_{XYl} = \{Z_{XYli}\}$ for i = 1, 2, ..., 9, where $Z_{xyli} = 1$ when the IBD mode for X and Y at locus l is D_i and $Z_{xyli} = 0$ when otherwise. As the nine IBD modes are mutually exclusive, we have $\sum_{i=1}^{9} Z_{XYli} \equiv 1$. We are not interested in these $9LN^2$ auxiliary variables but use them in an EM algorithm to infer both $\mathbf{p}$ and $\boldsymbol{\Delta}$. If $\mathbf{Z}_{XYl}$ were known, the MLEs of the IBD coefficients between X and Y would simply be $\widehat{\Delta}_{XYi} = \frac{1}{L}\sum_{l=1}^{L} Z_{XYli}$ for i = 1, 2, ..., 9. However, it is impossible to know for sure whether $Z_{XYli} = 0$ or $Z_{XYli} = 1$ from the genotype data S_{XYl} and allele frequencies $\mathbf{p}_l$ in general. What we can do is to infer the expected value of Z_{XYli}, $E[Z_{XYli}]$ or ε_{XYli} for simple notation, which is equivalent to the probability that the IBD mode for individuals X and Y at locus l is D_i. Therefore, ε_{XYli} can act as a single locus estimate of Δ_{XYi}. Given ε_{XYli}, the estimate of Δ_{XYi} for i = 1, 2, ..., 9 is

$$\widehat{\Delta}_{XYi} = (1/L)\sum_{l=1}^{L} \varepsilon_{XYli}. \tag{A2.1}$$

Allele frequencies can also be refined using ε_{XYli}. Suppose individuals X and Y have genotypes A_iA_i and A_iA_j, respectively, at a locus l. The number of non-IBD copies is 1 for allele j, irrespective of ε_{XYli}. However, it is a random variable for allele i, varying between 1 and 3 when all three copies of allele i are IBD and non-IBD, respectively. Conditional on ε_{XYli}, the expected number of non-IBD copies of allele i is

$$q_{XYli} = \varepsilon_{XYl3} + 2\varepsilon_{XYl4} + 2\varepsilon_{XYl8} + 3\varepsilon_{XYl9}$$

as calculated from Table 1. Similarly, the expected number of non-IBD copies of an allele in any pair of genotypes of X and Y given ε_{XYli} (for i = 1, 2, ..., 9) can be derived, as listed in Table A2.1.

For an individual X with itself (i.e. X with X), the expected number of non-IBD copies of allele A_i is 1 when X is a heterozygote of A_i, and is $\varepsilon_{XXl1} + 2\varepsilon_{XXl7}$ when X is a homozygote of A_i at locus l.

The total number of non-IBD copies of allele A_i at locus l across the N^2 pairs of individuals is

$$q_{li} = \sum_{X=1}^{N}\sum_{Y=1}^{N} q_{XYli}, \tag{A2.2}$$

where q_{XYli} is listed in Table A2.1 for each IIS mode (type of pairs of genotypes of X and Y). q_{XXli} is calculated as described above.

For locus l with n_l alleles, the frequency of allele A_i is estimated by

$$\widehat{p}_{li} = q_{li} / \left(\sum_{j=1}^{n_l} q_{lj}\right). \tag{A2.3}$$

The key to inferring both $\mathbf{p}$ and $\boldsymbol{\Delta}$ lies therefore in calculating ε_{XYli}, the expected value of Z_{XYli} (for i = 1, 2, ..., 9) for a pair of individuals X and Y at a locus l (=1, 2, ..., L). ε_{XYli} can be estimated from estimates of $\mathbf{p}$ and $\boldsymbol{\Delta}$ and genotype data, using Table 1. For individuals X and Y with genotypes A_jA_j and A_jA_k, respectively (i.e. IIS mode S_3), at locus l, for example, $\varepsilon_{XYli} = C_i / C$ where $C_i = 0$ for i = 1, 2, 5, 6 and 7, $C_3 = p_jp_k\Delta_{XY3}$, $C_4 = 2p_j^2p_k\Delta_{XY4}$, $C_8 = p_j^2p_k\Delta_{XY8}$, $C_9 = 2p_j^3p_k\Delta_{XY9}$ and $C = \sum_{i=1}^{9} C_i$. Since allele frequencies are unknown, the expected values of the powers and products of estimated allele frequencies, as calculated by (3) from sample allele counts, should be used in calculating ε_{XYli}. For other IIS modes or genotype pairs, ε_{XYli} can be calculated similarly, using Table 1.

Based on the above, I propose the following EM algorithm for MLEs of both $\mathbf{p}$ and $\boldsymbol{\Delta}$ from a sample of N multilocus genotypes.

1. Initialization

By assuming non-inbred and unrelated individuals, I calculate the initial non-IBD allele counts $\mathbf{q}^{(0)} = \{\mathbf{q}_l^{(0)}\}$ for each locus l (=1, 2, ..., L) by simple allele counting in genotype data, where $\mathbf{q}_l^{(0)} = \left\{q_{li}^{(0)}\right\}$ with $q_{li}^{(0)}$ being the count of non-IBD copies of allele i at locus l. These initial and updated

TABLE A2.1 The expected number of non-IBD alleles in a pair of genotypes

		Expected number of non-IBD copies of allele			
IIS mode	Allelic state	i	j	k	l
S_1	A_iA_i,A_iA_i	$\varepsilon_1+2(\varepsilon_2+\varepsilon_3+\varepsilon_5+\varepsilon_7)$ $+3(\varepsilon_4+\varepsilon_6+\varepsilon_8)+4\varepsilon_9$	0	0	0
S_2	A_iA_i,A_jA_j	$\varepsilon_2+\varepsilon_4+2\varepsilon_6+2\varepsilon_9$	$\varepsilon_2+2\varepsilon_4+\varepsilon_6+2\varepsilon_9$	0	0
S_3	A_iA_i,A_iA_j	$\varepsilon_3+2\varepsilon_4+2\varepsilon_8+3\varepsilon_9$	1	0	0
S_4	A_iA_i,A_jA_k	$\varepsilon_4+2\varepsilon_9$	1	1	0
S_5	A_iA_j,A_iA_i	$\varepsilon_5+2\varepsilon_6+2\varepsilon_8+3\varepsilon_9$	1	0	0
S_6	A_jA_k,A_iA_i	$\varepsilon_6+2\varepsilon_9$	1	1	0
S_7	A_iA_j,A_iA_j	$\varepsilon_7+p_j\varepsilon_8/(p_i+p_j)+2\varepsilon_9$	$\varepsilon_7+p_i\varepsilon_8/(p_i+p_j)+2\varepsilon_9$	0	0
S_8	A_iA_j,A_iA_k	$\varepsilon_8+2\varepsilon_9$	1	1	0
S_9	A_iA_j,A_kA_l	1	1	1	1

Notes: ε_i, shortened from ε_{XYli}, is the expected value of Z_{XYli}, which is the auxiliary variable for the ith (i = 1, 2, ..., 9) IBD mode for individuals X and Y (Y ≠ X) at locus l. Z_{XYli} takes value 1 when the IBD mode is D_i and value 0 when otherwise.

Abbreviation: IBD, identical by descent.

(below) allele counts are used in calculating expected values of allele frequency powers and products as detailed in (3), which are used in all computations in the EM algorithm. Under the same assumption, the initial IBD coefficients are $\boldsymbol{\Delta}^{(0)} = \{\boldsymbol{\Delta}_{XY}^{(0)}\}$ for X,Y = 1, 2, ..., N, with $\boldsymbol{\Delta}_{XY}^{(0)} = \left\{ \Delta_{XYi}^{(0)} \right\}$ for $i = 1, 2, \ldots, 9$. For two different individuals X and Y, we set $\Delta_{XYi}^{(0)} = u$ for $i = 1, 2, \ldots, 8$ and $\Delta_{XY9}^{(0)} = 1 - 8u$, where u is a small positive value such as 0.001. For an individual X with itself, we set $\Delta_{XXi}^{(0)} = u$ for $i = 1, 2, \ldots, 6$ and 8 and 9, and $\Delta_{XY7}^{(0)} = 1 - 8u$. This means, essentially, we assume all individuals are non-inbred and unrelated initially in the EM process.

2. Iteration

We use values of $\mathbf{q}^{(t)}$ and $\boldsymbol{\Delta}^{(t)}$ at the current iteration, t (≥0), to calculate those at the next iteration, $t+1$.

(2.1) Expectation, $\varepsilon_{XYli}^{(t+1)}$. For the pair of genotypes of X and Y at locus l, calculate the expected value, $\varepsilon_{XYli}^{(t+1)}$, of auxiliary variable Z_{XYli}. The calculation uses $\mathbf{q}_l^{(t)}$ and $\boldsymbol{\Delta}_{XY}^{(t)}$ as well as the observed genotypes (IIS mode) of X and Y at locus l, as described and exemplified above. $\varepsilon_{XYli}^{(t+1)}$ is calculated for each of N^2 pairs of individuals at each of L loci.

(2.2) Update of IBD coefficients, $\boldsymbol{\Delta}^{(t+1)}$. For individuals X and Y, the IBD coefficients are updated to $\Delta_{XYi}^{(t+1)} = \frac{1}{L} \sum_{l=1}^{L} \varepsilon_{XYli}^{(t+1)}$ for $i = 1, 2, \ldots, 9$.

(2.3) Update of non-IBD allele counts, $\mathbf{q}^{(t+1)}$. For a pair of genotypes of individuals X and Y at locus l, the expected count of non-IBD copies of allele A_i, $q_{XYli}^{(t+1)}$, can be calculated from $\boldsymbol{\varepsilon}_{XYl}^{(t+1)} = \left\{ \varepsilon_{XYlj}^{(t+1)} \right\}$ for $j = 1, 2, \ldots, 9$, as listed in Table A2.1. The total number of non-IBD copies of allele A_i of locus l across the N^2 pairs of individuals is $q_{li}^{(t+1)} = \left(\frac{1}{2N} \right) \sum_{X=1}^{N} \sum_{Y=1}^{N} q_{XYli}^{(t+1)}$.

3. Termination

The iteration described above is repeated until convergence is reached. At the end of each iteration, I calculate the quantity

$$\tau = \frac{1}{9N^2} \sum_{X=1}^{N} \sum_{Y=1}^{N} \sum_{i=1}^{9} \frac{\mid \Delta_{XYi}^{t+1} - \Delta_{XYi}^{t} \mid}{\left(\Delta_{XYi}^{t+1} + \Delta_{XYi}^{t} \right) / 2}. \qquad \text{(A2.4)}$$

to measure the extent of convergence. The EM is deemed converged when $\tau < \tau_t$, where τ_t is the threshold value of error tolerance. In this study, I use $\tau_t = 0.00001$. On convergence, the iteration is terminated and the final values $\boldsymbol{\Delta}^{(T)}$ are taken as the MLEs of $\boldsymbol{\Delta}$, and the final values $\mathbf{q}^{(T)}$ are used to calculate the MLEs of allele frequencies, $\mathbf{p}$. For locus l with n_l alleles, for example, the MLE of the frequency of allele i, p_{li}, is calculated by $\hat{p}_{li} = q_{li}^{(T)} / \sum_{j=1}^{n_l} q_{lj}^{(T)}$.

Because of the complexity of the likelihood function, Equation (4), with so many variables, and the nature of the EM algorithm (e.g. Wu, 1983), the above described iterations could converge to a local maximum rather than a global one. It is therefore suggested to repeat the EM algorithm several times using different initial values of $\boldsymbol{\Delta}$. Random values of $\boldsymbol{\Delta}_{XY}^{(0)}$ for a random set of individuals (or all individuals) can be generated and used to initiate the EM iterations. Multiple EM repeats serve two purposes. One is to check whether the algorithm converges to a global maximum or not. When it does, all repeats with different $\boldsymbol{\Delta}_{XY}^{(0)}$ values should lead to the same maximum likelihood and the same MLE of both $\boldsymbol{\Delta}$ and $\mathbf{p}$. The other is that, in case the algorithm converges to different local maxima, we can choose the results from the repeat with the best local maximum likelihood. They should be better than results from a single random repeat.

The EM algorithm is slow when N and L are large. The number of parameters, $9N^2$ IBD coefficients and $\sum_{l=1}^{L} n_l$ allele frequencies, increases quadratically with N and linearly with L. When N is large and the relatives are few in the sample such that allele frequencies are well estimated by assuming non-inbred and unrelated individuals, we can use likelihood function (2) in place of (4) to estimate $\boldsymbol{\Delta}$ only. The EM algorithm runs much faster as allele frequencies are not updated and IBD coefficients are estimated independently for each pair of individuals. To speed up the computation no matter allele frequencies are updated (likelihood function (4)) or not (likelihood function (2)), I use openMP and MPI to employ hyperthreads and multiple cores of multiple nodes in a computer cluster with shared and distributed memories for parallel computation.