Genet. Res., Camb. (2007), **89**, *pp.* 135–153.
doi:10.1017/S0016672307008798 Printed in the United Kingdom

Triadic IBD coefficients and applications to estimating pairwise relatedness

JINLIANG WANG*
Institute of Zoology, Zoological Society of London, Regent's Park, London NW1 4RY, UK

(*Received 15 September 2006 and in revised form 23 April and 10 June 2007*)

Summary

Knowledge of the genetic relatedness among individuals is essential in diverse research areas such as behavioural ecology, conservation biology, quantitative genetics and forensics. How to estimate relatedness accurately from genetic marker information has been explored recently by many methodological studies. In this investigation I propose a new likelihood method that uses the genotypes of a triad of individuals in estimating pairwise relatedness (r). The idea is to use a third individual as a control (reference) in estimating the r between two other individuals, thus reducing the chance of genes identical in state being mistakenly inferred as identical by descent. The new method allows for inbreeding and accounts for genotype errors in data. Analyses of both simulated and human microsatellite and SNP datasets show that the quality of r estimates (measured by the root mean squared error, RMSE) is generally improved substantially by the new triadic likelihood method (TL) over the dyadic likelihood method and five moment estimators. Simulations also show that genotyping errors/mutations, when ignored, result in underestimates of r for related dyads, and that incorporating a model of typing errors in the TL method improves r estimates for highly related dyads but impairs those for loosely related or unrelated dyads. The effects of inbreeding were also investigated through simulations. It is concluded that, because most dyads in a natural population are unrelated or only loosely related, the overall performance of the new triadic likelihood method is the best, offering r estimates with a RMSE that is substantially smaller than the five commonly used moment estimators and the dyadic likelihood method.

1. Introduction

A number of estimators have been developed to use genetic marker data in estimating pairwise relatedness (r) between individuals (e.g. Lynch, 1988; Queller & Goodnight, 1989; Li *et al.*, 1993; Ritland, 1996; Lynch & Ritland, 1999; Wang, 2002; Milligan, 2003; Thomas, 2005; Oliehoek *et al.*, 2006), and have been applied to diverse research areas such as behavioural ecology, conservation biology, quantitative genetics and forensics (reviewed by Blouin, 2003). The statistical properties and performances of these estimators were recently investigated by several studies using both simulated and empirical datasets (e.g. Lynch & Ritland, 1999; Van de Casteele *et al.*, 2001; Wang, 2002; Milligan, 2003; Csilléry *et al.*, 2006). Conclusions drawn from these studies are, however, somewhat disappointing. First, no single estimator is universally superior to the others in terms of performance evaluated by estimation bias and variance (Van de Casteele *et al.*, 2001; Wang, 2002; Milligan, 2003). The performance rank order of the estimators depends, in particular, on the true relatedness value being estimated, the informativeness of markers (number of loci, number and frequencies of the alleles at each locus) utilized in an analysis, and the size of the sample in estimating allele frequencies. Although data can be examined regarding marker informativeness and sample size, an investigator generally has no idea of the true relatedness among the sampled individuals and thus the selection of the best estimator is virtually impossible in practice. Second, with the amount of marker information typically available in practice, the sampling variance of these estimators is high. For full-sibs ($r=0{\cdot}5$) as an example, the standard deviation of different estimators varies from 0·14

* Corresponding author. Telephone: +44 20 74496620. Fax: +44 20 75862870. e-mail: jinliang.wang@ioz.ac.uk

to 0·22 when 10 markers, each having 10 alleles with known frequencies in a uniform distribution, are used in the estimation (Lynch & Ritland, 1999; Wang, 2002). The high coefficient of variance (standard deviation/mean) of the estimators, 28–45%, is partly caused by the large inherent variance in IBD (identical by descent) among loci due to Mendelian inheritance (Wang, 2006). Third, moment estimators may yield relatedness estimates outside the legitimate range of [0,1], leading to difficulties in interpreting the estimates and in utilizing the estimates in certain subsequent analyses. One may ask, for example, whether or not a dyad with $\hat{r} = -10$ is less related than a dyad with $\hat{r} = -5$; and, if the answer is yes, how much less related for the first dyad than the second.

Ritland (1996), Lynch & Ritland (1999) and Wang (2002), in developing their moment estimators, all tried likelihood methods for estimating relatedness from marker data. These likelihood estimators showed poor performance and became equivalent to or superior to moment estimators only when a very large number of loci (say, >70) were used in the estimation. As a result, these authors dismissed the utility of likelihood methods for pairwise relatedness in practice. In contrast, Milligan (2003) showed by simulations that in the majority of cases the likelihood estimator has the lowest root mean squared error (RMSE), a measurement of accuracy incorporating both estimation bias and sampling variance. The apparently conflicting results arise mainly because estimates were constrained to the biologically meaningful parameter space in Milligan's likelihood estimator but not in the others. From both biological and statistical points of view, it seems more plausible to restrict the parameters to their legitimate range of values. However, Milligan's (2003) likelihood estimator overestimates relatedness slightly, especially when marker information is scarce and true relatedness is close to 0. Partly as a result of this overestimation, the RMSE of his likelihood estimator is not always smaller than the moment estimators. Increasing the number of loci can reduce the bias when alleles at each locus have an even frequency distribution. It becomes ineffective, however, with uneven allele frequency distributions such as when all alleles except one at a locus are rare (Milligan, 2003).

In this paper, I propose a new likelihood estimator of pairwise relatedness based on the estimation of IBD coefficients among a triad of individuals. By using trios rather than pairs of individuals, this triadic likelihood estimator controls for the background similarity in genotypes of a dyad introduced by sporadic identity-in-state (IIS) rather than IBD. It can therefore reduce the overestimation of IBD (and thus relatedness) characterized by the dyadic likelihood estimator. I also fitted a genotype-error model into the new triadic likelihood estimator to cope with genotyping errors and mutations of data, and incorporated inbreeding into the estimator to deal with inbred populations. The performance of the new likelihood estimator is compared with that of the dyadic likelihood method and five moment estimators using both simulations and two large human datasets.

Table 1. *Dyadic (four-gene) IBD states and their probabilities*

IBD state	Genes IBD	Pr(S_i)
S_1	(abcd)	Δ_1
S_2	(ab,cd)	Δ_2
S_3	(abc), (abd)	Δ_3
S_4	(ab)	Δ_4
S_5	(acd), (bcd)	Δ_5
S_6	(cd)	Δ_6
S_7	(ac,bd), (ad,bc)	Δ_7
S_8	(ac), (ad), (bc), (bd)	Δ_8
S_9	None	Δ_9

Homologous genes a and b are in individual X, and genes c and d are in individual Y. Genes not specified are not IBD with any of those listed in column 2. Alternative identity configurations for a given IBD state are listed in separate sets of parentheses.

2. Methods

(i) *Dyadic IBD coefficients and relatedness*

Homologous genes are identical by descent (IBD) if they are copies descended from the same gene of an ancestor (Malécot, 1948). Inbreeding coefficient can thus be defined as the probability that the two homologous genes within an individual are IBD (Malécot, 1948), while relatedness (r) or coancestry coefficient (θ) between two individuals can be characterized by the possibility of finding such identical genes in their genotypes (Harris, 1964; Jacquard, 1972). Estimating the relatedness (r_{XY}) or coancestry (θ_{XY}) coefficient between individuals X and Y is solved, therefore, by finding the IBD coefficients between genes from X and genes from Y, as implemented in several moment (e.g. Lynch & Ritland, 1999; Wang, 2002) and likelihood estimators (Milligan, 2003).

Among the two genes from X and two genes from Y, there exist 15 mutually exclusive and exhaustive IBD states (Jacquard, 1972; Weir, 1996). When paternal and maternal genes are not distinguished, the 15 IBD states reduce to nine condensed identity states (Harris, 1964; Jacquard, 1972; Lynch & Walsh, 1998) as defined in Table 1. The symbols for the probabilities of the nine identity states are also listed in the table. Note that in general it is impossible to determine whether two genes are IBD or not, even if the pedigree of the two individuals from whom the two genes come are known. Only under certain special

Table 2. *Triadic (six-gene) IBD states and their probabilities*

IBD state	Genes IBD	Pr(s_i)
s_1	(ace, bdf), (acf, bde), (ade, bcf), (adf, bce)	δ_1
s_2	(ac,be,df), (ac,bf,de), (ad,be,cf), (ad,bf,ce), (ae,bc,df), (af,bc,de), (ae,bd,cf), (af,bd,ce)	δ_2
s_3	(ace,bd), (acf,bd), (ade,bc), (adf,bc), (bce,ad), (bcf,ad), (bde,ac), (bdf,ac)	δ_3
s_4	(ace,bf), (ade,bf), (acf,be), (adf,be), (bce,af), (bde,af), (bcf,ae), (bdf,ae)	δ_4
s_5	(ace,df), (bce,df), (acf,de), (bcf,de), (ade,cf), (bde,cf), (adf,ce), (bdf,ce)	δ_5
s_6	(ac,bd), (ad,bc)	δ_6
s_7	(ae,bf), (af,be)	δ_7
s_8	(ce,df), (cf,de)	δ_8
s_9	(ac,be), (ac,bf), (ad,be), (ad,bf), (bc,ae), (bc,af), (bd,ae), (bd,af)	δ_9
s_{10}	(ac,de), (ac,df), (bc,de), (bc,df), (ad,ce), (ad,cf), (bd,ce), (bd,cf)	δ_{10}
s_{11}	(ae,cf), (ae,df), (be,cf), (be,df), (af,ce), (af,de), (bf,ce), (bf,de)	δ_{11}
s_{12}	(ace), (acf), (ade), (adf), (bce), (bcf), (bde), (bdf)	δ_{12}
s_{13}	(ac), (ad), (bc), (bd)	δ_{13}
s_{14}	(ae), (af), (be), (bf)	δ_{14}
s_{15}	(ce), (cf), (de), (df)	δ_{15}
s_{16}	None	δ_{16}

I assume individuals X, Y and Z are all non-inbred so that only 16 of the 66 possible IBD states are necessary. Homologous genes a and b are in individual X, genes c and d are in individual Y, and genes e and f are in individual Z. Genes not specified are not IBD with any of those listed in the 'Genes IBD' column. Alternative identity configurations for a given IBD state are listed in separate sets of parentheses.

circumstances do we know the IBD states of two or more genes. For example, a parent with genotype A_1A_2 and a child with genotype A_1A_3 have their A_1 genes IBD and all other pairs of genes non-IBD. However, if both the parent and child have genotype A_1A_2, we cannot ascertain the IBD states between the two A_1 or A_2 genes in the two individuals. We can, though, always determine the probability of genes that are IBD or non-IBD (or more generally the nine IBD coefficients in Table 1), with or without pedigree information. The likelihood estimator of relatedness is actually based on the estimation of IBD coefficients (probabilities) from genetic marker data of individuals without pedigree (see below).

Knowing the nine IBD coefficients, it is easy to obtain the inbreeding coefficients (F) of, and coancestry (θ) and relatedness (r) coefficients between individuals X and Y. By definition, we have from Table 1 that (Jacquard, 1972; Lynch & Walsh, 1998; Milligan, 2003)

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

It should be noted that while F is a sufficient statistic describing the IBD status between genes within an individual, neither θ nor r is a sufficient statistic for the relationship between genes from two individuals. By transforming the complete set of nine Δ parameters into θ or r, some information is lost so that, for example, the probability of the joint genotypes of individuals X and Y or the probability of the genotype of X given that of Y cannot be derived using either θ or r. These summary statistics are nevertheless useful in quantifying the overall similarity in gene descent between individuals and find wide applications in areas such as quantitative genetics (Lynch & Walsh, 1998) and behaviour ecology (Hamilton, 1964). I will therefore concentrate on the estimation of r rather than the nine IBD coefficients throughout this investigation.

The nine gene IBD states for individuals X and Y cannot be observed, but their probabilities (Table 1) can be inferred from the genotypes of X and Y at a number of marker loci. The probability of the joint genotypes of X and Y conditional on the nine IBD coefficients and allele frequencies was derived by Harris (1964) and was given in a more convenient form by Milligan (2003). One can obtain estimates of the nine IBD coefficients by maximizing the probability of the joint genotypes given allele frequencies, and then using (1) one gets maximum likelihood estimates of θ or r. Following previous moment estimators of relatedness, Milligan (2003) assumed non-inbred individuals so that $\Delta_i \equiv 0$ for $i = 1 \sim 6$ and he estimated Δ_7, Δ_8 and Δ_9 only. The full dyadic likelihood method estimating the nine IBD coefficients jointly can be implemented similarly to estimate r between inbred individuals.

(ii) *Triadic IBD coefficients*

Among the six genes at an autosomal diploid locus of individuals X, Y and Z, there are 203 mutually exclusive and exhaustive IBD states, which reduce to

https://doi.org/10.1017/S0016672307008798 Published online by Cambridge University Press

66 condensed IBD states when maternal and paternal genes are not distinguished (Thompson, 1974). For the purpose of estimating relatedness between non-inbred individuals, only 16 of the 66 condensed IBD states are necessary. Table 2 lists these 16 IBD states and symbols for their corresponding probabilities.

From the 16 six-gene IBD coefficients δ, one can easily calculate the four-gene IBD coefficients Δ_i for $i=7,8,9$, and thus the coancestry (θ) and relatedness (r) coefficients between any two of the three individuals. Without loss of generality, I assume the first two individuals, X and Y, constitute our focal dyad and the third individual, Z, serves as a reference. The coefficients Δ_7 and Δ_8 ($\Delta_9 \equiv 1-\Delta_7-\Delta_8$) of the focal dyad are

$$\Delta_7 = \delta_1 + \delta_3 + \delta_6,$$
$$\Delta_8 = \delta_2 + \delta_4 + \delta_5 + \delta_9 + \delta_{10} + \delta_{12} + \delta_{13} \quad (2)$$

and the coancestry (θ) and relatedness (r) coefficients are then calculated as $\theta = \frac{1}{2}\Delta_7 + \frac{1}{4}\Delta_8$ and $r = 2\theta$, respectively.

To obtain likelihood estimates of δ from genotype data, we need the probability of triadic genotypes given allele frequencies and δ. Thompson (1974) derived a general formula for this probability, but its calculation is quite complicated. Here I give an explicit expression for the probability of a trio of genotypes. Assuming X, Y and Z are random members of the subpopulation composed of trios of individuals having the same IBD coefficients (δ_i, $i=1, 2, \ldots, 16$), the genotypic array of such trios is

$$\begin{aligned}
&\delta_1 \sum_{i,j} p_i p_j \{A_iA_j, A_iA_j, A_iA_j\} + \delta_2 \sum_{i,j,k} p_i p_j p_k \{A_iA_j, A_jA_k, A_iA_k\} \\
&+ \delta_3 \sum_{i,j,k} p_i p_j p_k \{A_iA_j, A_iA_j, A_iA_k\} + \delta_4 \sum_{i,j,k} p_i p_j p_k \{A_iA_j, A_iA_k, A_iA_j\} \\
&+ \delta_5 \sum_{i,j,k} p_i p_j p_k \{A_iA_k, A_iA_j, A_iA_j\} + \delta_6 \sum_{i,j,k,l} p_i p_j p_k p_l \{A_iA_j, A_iA_j, A_kA_l\} \\
&+ \delta_7 \sum_{i,j,k,l} p_i p_j p_k p_l \{A_iA_j, A_kA_l, A_iA_j\} + \delta_8 \sum_{i,j,k,l} p_i p_j p_k p_l \{A_kA_l, A_iA_j, A_iA_j\} \\
&+ \delta_9 \sum_{i,j,k,l} p_i p_j p_k p_l \{A_iA_j, A_iA_k, A_jA_l\} + \delta_{10} \sum_{i,j,k,l} p_i p_j p_k p_l \{A_iA_k, A_iA_j, A_jA_l\} \\
&+ \delta_{11} \sum_{i,j,k,l} p_i p_j p_k p_l \{A_iA_k, A_jA_l, A_iA_j\} + \delta_{12} \sum_{i,j,k,l} p_i p_j p_k p_l \{A_iA_j, A_iA_k, A_iA_l\} \\
&+ \delta_{13} \sum_{i,j,k,l,m} p_i p_j p_k p_l p_m \{A_iA_j, A_iA_k, A_lA_m\} + \delta_{14} \sum_{i,j,k,l,m} p_i p_j p_k p_l p_m \{A_iA_j, A_lA_m, A_iA_k\} \\
&+ \delta_{15} \sum_{i,j,k,l,m} p_i p_j p_k p_l p_m \{A_lA_m, A_iA_j, A_iA_k\} + \delta_{16} \sum_{i,j,k,l,m,n} p_i p_j p_k p_l p_m p_n \{A_iA_j, A_kA_l, A_mA_n\}
\end{aligned} \quad (3)$$

where i, j, k, l, m, n index the alleles of a locus, p_i is the frequency of allele A_i, and a trio of genotypes such as $\{A_iA_j, A_iA_k, A_iA_l\}$ refers to a trio of individuals $\{X, Y, Z\}$. A few examples help in understanding (3). If the two non-IBD genes of X are IBD with those of Y and Z (i.e. IBD state s_1 in Table 2), then the triad will have the same genotype A_iA_j with frequency p_ip_j (first term). If X, Y and Z have one gene IBD and furthermore X and Y have another gene IBD (i.e. IBD state s_3), then the first trio of genes will be A_i with frequency p_i, the second pair of genes will be A_j with frequency p_j, and the other gene (in Z) will be A_k with frequency p_k, resulting in a frequency of $p_i\,p_j\,p_k$ of the trio of genotypes $\{A_iA_j, A_iA_j, A_iA_k\}$ (third term).

To obtain the probability of a specific trio of genotypes for $\{X, Y, Z\}$, one needs to collect the appropriate frequencies from the various terms in (3). For the genotype trio $\{A_iA_j, A_iA_k, A_iA_l\}$ as an example, where i, j, k, l index four different alleles in this particular case, the probability can be found to be $p_i\,p_j\,p_k\,p_l(\delta_{12}+2(\delta_{13}+\delta_{14}+\delta_{15})p_i+8\delta_{16}p_i^2)$. For a locus with six or more co-dominant alleles, there are 66 distinct patterns of IIS for the six genes in a trio of individuals, which correspond to the 66 condensed IBD states. The probabilities of the 66 patterns of IIS (available upon request) as a function of δ and allele frequencies are derived from (3) and are used in the likelihood estimation of the 16 six-gene IBD coefficients in Table 2 and thus of relatedness as described below.

For simplicity in describing the methodology and in accordance with previous moment and likelihood estimators, I have assumed non-inbred individuals so that only 16 of the 66 condensed IBD states are required in the triadic likelihood estimation of r. If desirable, however, inbreeding can be incorporated in the triadic likelihood method to allow for the simultaneous estimation of inbreeding coefficients of each individual and more importantly for the more accurate estimation of r between inbred individuals. An equation similar to (3) but in terms of the probabilities of the 66 condensed IBD states can be derived (available upon request). A likelihood function can then be constructed from the equation, which is used in the estimation of the 66 IBD coefficients and thus of F and r. Because previous r estimators assume non-inbred individuals, I will concentrate on the

https://doi.org/10.1017/S0016672307008798 Published online by Cambridge University Press

simple 16-IBD-state version of the triadic likelihood estimator throughout this manuscript, except when I explicitly investigate the effect of inbreeding.

(iii) *Triadic likelihood estimation of relatedness*

Suppose that a sample of individuals is drawn from a population. Each individual is examined at some marker loci, and we are interested in knowing how pairs of individuals are genetically related based solely on the marker data. I assume that inbreeding is negligible in the population so that only the set of IBD coefficients, $\delta = \{\delta_1, \delta_2, \ldots, \delta_{16}\}$, is relevant in describing the relationship for a trio of individuals. This assumption of non-inbreeding is followed implicitly or explicitly by all previous relatedness estimators. I also assume that the markers are in linkage equilibrium and are unlinked so that the overall likelihood is simply the product of the likelihoods across loci.

For a trio of individuals $\{X, Y, Z\}$, the probability (obtained above) of observing their genotypes at a locus given δ and allele frequencies is the likelihood of δ. By maximizing the product of the likelihoods across loci over the legitimate parameter space (i.e. $\delta_i \geqslant 0$ for $i = 1, 2, \ldots, 16$, subject to constraint $\sum_{i=1}^{16} \delta_i \equiv 1$), one obtains the maximum likelihood estimates (MLEs) of δ. In general, an analytical solution is impossible for the maximum likelihood estimation, and a numerical approach has to be adopted. I choose to use Powell's quadratically convergent method (Press *et al.*, 1996) with slight modifications to solve this 16-dimentional constrained optimization problem. Tests using numerous simulated and empirical datasets with a large number of initial points (a point specifies the 16 parameter values of δ) indicate that the method is fast and converges reliably. Therefore, in the analyses shown below, a single randomly chosen starting point is used for each triad to initiate the search for the MLEs of δ.

For a given dyad of X and Y, I simulate a number of M non-inbred individuals unrelated among themselves and unrelated to either X or Y, with their multilocus genotypes generated by simulations using the allele frequencies of the dataset. Each of the M simulated individuals serves as a reference to obtain the triadic likelihood estimates of δ and then of Δ and r coefficients between X and Y. We therefore face the problem of how to summarize these M estimates into a single best estimate of r_{XY}. In principle, one might use one of many summary statistics, such as the harmonic, arithmetic, geometric means or the median or mode of the distribution of M estimates, as the estimate of r_{XY}. It turns out that the mode estimate recovers the dyadic likelihood estimate (Milligan, 2003) in all cases (different degrees of true relatedness, and different amounts of marker information) investigated (results not shown). Considering that most dyads in natural populations are hardly related (Ritland, 1996; Lynch & Ritland, 1999) and likelihood methods usually overestimate relatedness (e.g. Milligan, 2003), I choose to use the harmonic mean of the estimates in the first percentile of the M estimates as the best estimate of r_{XY}. Analyses using simulated and empirical data show that this relatedness estimator has generally smaller biases and standard deviations than the mean, mode or median estimators for dyads related to different degrees and for different allele frequency distributions. They also show that estimates of r stabilize when M becomes large (say, $M > 500$; see results below). In the results shown below, a value of $M = 500$ is adopted.

Using the same algorithm but relaxing the assumption of non-inbred individuals, I can implement the full version of the triadic likelihood method to estimate F and r coefficients from the 66 IBD coefficients.

(iv) *Accounting for genotype errors*

Previous relatedness estimators invariably ignored genotyping errors in data. This is plausible because typing errors should have a small effect on relatedness estimation except when the rate of errors is exceptionally high (say, 10%). Furthermore, accounting for typing errors may incur a cost, lowering the power of a relatedness analysis (Morrissey & Wilson, 2005). However, in the case of a large number of markers, a high genotyping error rate and highly related dyads, typing errors could reduce the quality of relatedness estimates substantially if they are not accounted for. Fortunately, unlike moment estimators, it is relatively easy to incorporate a model of genotyping errors into the triadic likelihood estimator of relatedness.

I adopt a simple genotyping error model as detailed in Wang (2006). The model assumes that all gene copies at a locus are independently and equally likely to be incorrectly observed, and that an allele, if incorrectly genotyped, is observed to be any (including itself) of the alleles at the locus with an equal probability. Using this error model, we can obtain the probability of a trio of *observed* genotypes (or phenotypes) given allele frequencies, the 16 triadic IBD coefficients and the error rate of the locus, following the same approach as utilized by Wang (2006) in the case of two individuals. Maximizing this probability yields MLEs of the 16 triadic IBD coefficients and thus the pairwise relatedness.

(v) *Comparison with previous relatedness estimators*

As mentioned in Section 1, although a number of relatedness estimators have been developed and used in practice (Blouin, 2003), none of them is clearly superior in performance to the others in all circumstances.

https://doi.org/10.1017/S0016672307008798 Published online by Cambridge University Press

Any new estimator must be carefully examined for its performance and statistical properties in comparison with previous estimators under various situations before it is released for applications.

Six estimators are chosen to compare with the current triadic likelihood estimator, denoted as TL hereafter. These six estimators are described by Queller & Goodnight (1989; denoted by QG), Ritland (1996; denoted by R), Lynch & Ritland (1999; denoted by LR), Lynch (1988) and Li *et al.* (1993; denoted by LL), Wang (2002; denoted by W) and Milligan (2003; denoted by M), and have been compared by several authors recently (e.g. Lynch & Ritland, 1999; Van de Casteele *et al.*, 2001; Wang, 2002; Milligan, 2003; Csilléry *et al.*, 2006). Because several variants of some estimators exist in the literature, I describe each estimator briefly to avoid confusion. For a dyad with genotypes $\{A_iA_j\}$ and $\{A_kA_l\}$ at a locus, the QG estimator is calculated by $(S_{ik}+S_{il}+S_{jk}+S_{jl}-H)/(2+S_{ij}+S_{kl}-H)$, where $S_{ac}=1$ if $a=c$ and $S_{ac}=0$ if otherwise $(a,c=i,j,k,l)$, $H=p_i+p_j+p_k+p_l$. With multiple loci, the sums of the denominator and numerator over loci are obtained and the division of the sums gives the multi-locus estimate. The LL and R estimators are obtained by averaging (across loci) $\hat{r}_{XY}$ calculated by equations (8–9) in Lynch & Ritland (1999). LR and W estimators are implemented as described in the original papers (Lynch & Ritland, 1999; Wang, 2002). The M estimator is implemented following Milligan (2003), except that the algorithm used is Powell's quadratically convergent method (Press *et al.*, 1996), which is more powerful than the simplex method used by Milligan (2003).

(vi) *Simulations and measurements of performance*

To evaluate and compare the performances of different estimators, one has to apply them to simulated or empirical datasets, because it is mathematically intractable to investigate the statistical properties of several estimators analytically. Following previous studies, I simulate data in a range of sampling conditions that it is hoped are typical of or embrace practical applications in relatedness estimation.

Four different allele frequency distributions are used in the simulations: one in which all alleles have an equal frequency (EF), one in which a single allele occurs with a frequency of 0·8 and the remaining alleles are equally frequent (common allele frequency distribution, CF), one in which allele i $(=1, 2, \ldots, k)$ at a k-allele locus has a frequency of $i/(k(k+1)/2)$ (triangular frequency distribution, TF) and one in which allele frequencies are drawn independently from the same Dirichlet distribution with all parameters set to 1 (uniform allele frequency distribution, UF). Five representative genetic relationships found in natural populations were considered: parent–offspring (PO), full-sibs (FS), half-sibs (HS), first cousins (FC) and unrelated individuals (UR). These relationships vary in the extent and pattern of relatedness and have different inherent (Mendelian inheritance) variances of r among loci (Wang, 2006).

To investigate the robustness to inbreeding of the estimators, dyads with a certain inbreeding coefficient (F) and genetic relationship were also simulated and their relatedness estimated by these estimators. In comparison, the same data were also analysed by the full versions of the dyadic and triadic likelihood methods that account for inbreeding. For a PO dyad, the genotype of the parent (X) in the dyad is generated with F ($F<1/3$), and the genotype of the other parent (S) is generated by sampling at random one gene from the population and the other gene from the population and from X with probabilities $1-4F/(1+F)$ and $4F/(1+F)$, respectively. The offspring in the dyad (Y) is then generated from those of the parents following Mendelian segregation. It can be shown that, with this simulation procedure, the nine IBD coefficients for the PO dyad X and Y are $\Delta_1=2F^2/(1+F)$, $\Delta_i=(1-F)\ F/(1+F)$ for $i=3,5,7$, $\Delta_i=0$ for $i=2,4,6,9$, and $\Delta_8=(1-F)^2/(1+F)$. The inbreeding coefficient of both X and Y is F, and the actual relatedness of the dyad is $\frac{1}{2}(1+3F)$. For a FS dyad, genotypes are simulated independently from those of their parents between whom the IBD coefficients are $\Delta_8=4F$ ($F<1/4$) and $\Delta_9=1-4F$, respectively. The inbreeding coefficients of and relatedness between the FS individuals thus generated are F and $\frac{1}{2}+F$, respectively. For a FC dyad, two FS individuals, S and T, can be generated from non-inbred and unrelated parents. An individual, U, related to S with IBD coefficients $\Delta_8=4F$ ($F<1/4$) and $\Delta_9=1-4F$ is generated and an offspring (X) is generated from S and U as parents. Similarly, an offspring (Y) is generated from T and V as parents, where individual V is related to T with IBD coefficients $\Delta_8=4F$ and $\Delta_9=1-4F$. FC dyads thus simulated have inbreeding coefficient F and relatedness $\frac{1}{8}(1+2F)^2$. For a UR dyad, the genotype of an individual is generated independently from that of the other individual with a probability of F that the two genes at a locus are IBD. Therefore, individuals in a UR dyad have inbreeding coefficient F and relatedness 0.

To investigate the effect of data quality, genotyping errors are introduced at a given rate into the genotypes of the individuals. The data are then analysed by different estimators to investigate the impact of typing errors on relatedness estimates, and the robustness of the estimators to typing errors.

The quality of an estimator is evaluated by its bias and the root mean squared errors, calculated as $\mathrm{RMSE}=\sqrt{\frac{1}{R}\sum_{i=1}^{R}(\hat{r}_i-r)^2}$, where $\hat{r}_i$ is the relatedness

https://doi.org/10.1017/S0016672307008798 Published online by Cambridge University Press

Table 3. *Effect of the number of reference individuals on the triadic likelihood estimator*

Frequency distribution	Dyads	$M=200$		$M=400$		$M=600$	
		Correlation	RMSD	Correlation	RMSD	Correlation	RMSD
EF	PO	0·9581	0·0108	0·9849	0·0061	0·9933	0·0039
	FS	0·9977	0·0095	0·9991	0·0058	0·9997	0·0031
	FC	0·9984	0·0068	0·9992	0·0045	0·9997	0·0028
	UR	0·9977	0·0052	0·9989	0·0034	0·9996	0·0019
UF	PO	0·9700	0·0156	0·9892	0·0093	0·9958	0·0057
	FS	0·9967	0·0126	0·9988	0·0075	0·9996	0·0042
	FC	0·9970	0·0092	0·9988	0·0057	0·9996	0·0031
	UR	0·9964	0·0068	0·9978	0·0052	0·9991	0·0032
CF	PO	0·9891	0·0306	0·9963	0·0172	0·9989	0·0092
	FS	0·9912	0·0304	0·9953	0·0213	0·9984	0·0095
	FC	0·9915	0·0194	0·9949	0·0143	0·9986	0·0073
	UR	0·9957	0·0082	0·9980	0·0055	0·9988	0·0043

Correlation coefficients and RMSDs are calculated between estimates using $M=200$, 400, 600 and estimates using $M=800$.

estimate of the ith dyad ($i=1, 2, \ldots, R$) by a given estimator and r is the parameter value of relatedness used in generating the R simulated dyads. RMSE captures an estimator's bias (measured by $B=\bar{r}-r$, where $\bar{r}=\frac{1}{R}\sum_{i=1}^{R}\hat{r}_i$) and precision (measured by variance $V=\frac{1}{R}\sum_{i=1}^{R}(\hat{r}_i-\bar{r})^2$), because obviously $RMSE^2=B^2+V$. For any parameter combination, a number of $R=10\,000$ replicate dyads with a given relationship are simulated and analysed by each estimator. Because the likelihood estimators are constrained to the parameter range of [0,1] while all moment estimators are not, the latter are disadvantaged in comparing sampling variance or RMSE with likelihood methods. It is well known that moment estimators, especially those of Ritland (1996) and Lynch & Ritland (1999), have rather skewed distributions with extreme values far outside of the range [0,1]. To be fair, therefore, all moment estimators are truncated to the range [0,1] before being used in calculating their biases and RMSEs.

(vii) *A human dataset*

The performances of the seven estimators were also compared by analysing a human dataset, CEPH (Centre d'Etude du Polymorphisme Humain), maintained by the Fundation Jean Dausset laboratory. The dataset, in its current version V10 available online (http://www.cephb.fr/cephdb/php/), contains genotypes of individuals from 65 families at 32 356 genetic marker loci. These include 9900 microsatellite markers and 21 480 bi-allelic markers, of which 17 512 are SNPs. Within each family, genotypes are available for the father, mother and a variable number of full-sib children. Some families also have a variable number of grandparents (1–4) genotyped as well. For this dataset, therefore, we have up to four known dyadic relationships: parent–offspring, full-sib, grandparent–grandchild (GG) and unrelated individuals. The pairwise relatedness among individuals within each family was analysed using a variable number of microsatellites or SNPs with known allele frequencies provided by the CEPH dataset.

3. Results

(i) *Number of reference individuals in the triadic likelihood estimator*

As described in Section 2, the triadic likelihood method uses a number (M) of reference individuals in estimating the relatedness of a focal dyad. The triadic likelihood estimates quickly stabilize with an increasing M. Some numerical examples are shown in Table 3, where correlation coefficients and root mean squared differences (RMSDs) are calculated between r estimates using $M=200$, 400 or 600 and those using $M=800$. Estimates were made for PO, FS, FC and UR dyads using 10 loci, each having eight alleles with frequencies in one of three distributions: equal frequency (EF), uniform Dirichlet (UF), and one common with the remaining rare (CF). It is clear from the table that, for all relationships and allele frequency distributions, r estimates using $M=400$ or 600 are very close to those using $M=800$, with correlation coefficients usually larger than 0·99 and RMSDs usually smaller than 0·01. Similar results are obtained using other numbers of loci, and other numbers of alleles per locus. Based on these results, $M=500$ seems to be sufficient, and is adopted in all the analyses shown below.

(ii) *Effects of marker information*

Although all relatedness estimators generally improve with an increasing amount of marker information,

https://doi.org/10.1017/S0016672307008798 Published online by Cambridge University Press

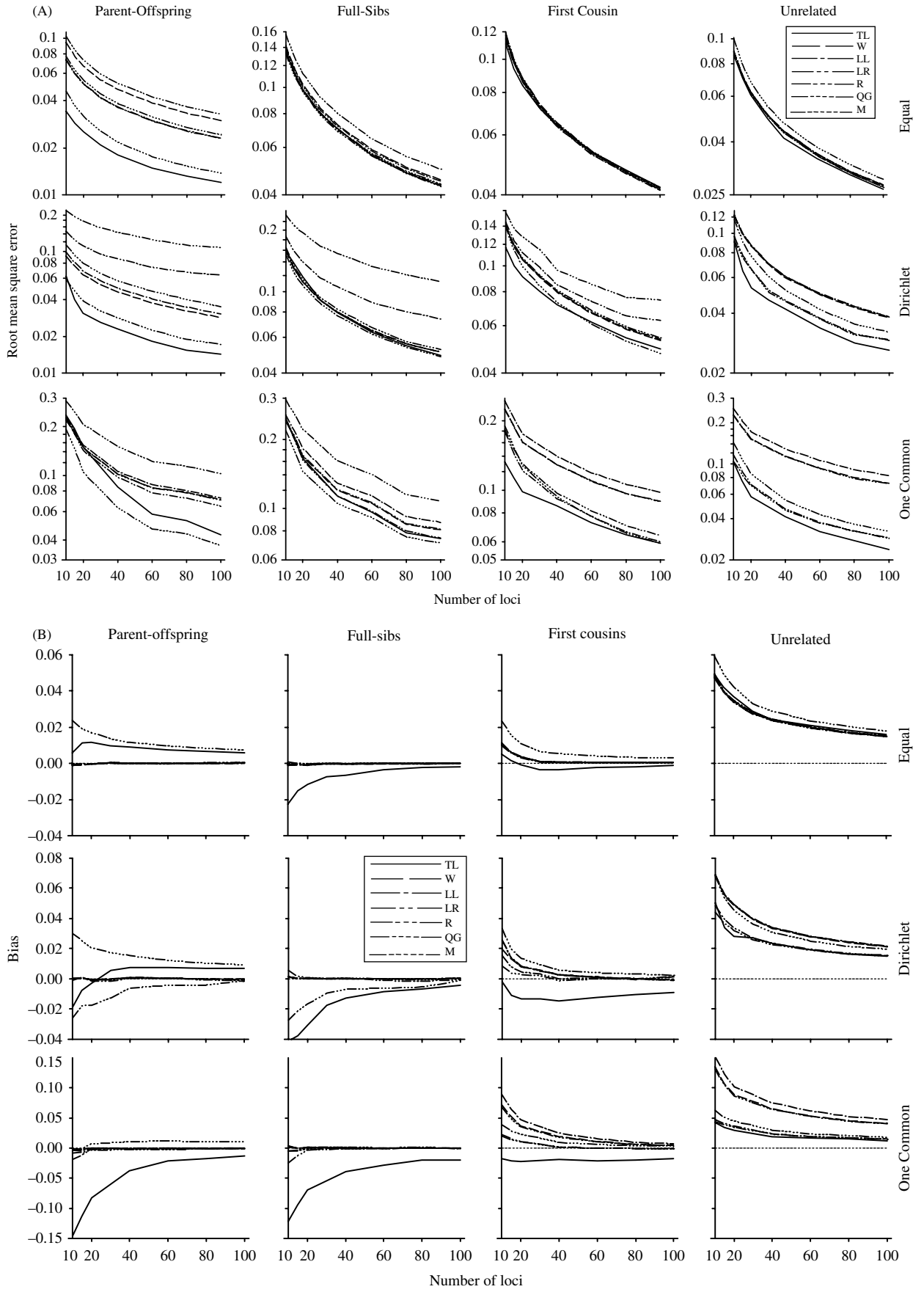

Fig. 1. For legend see opposite page.

https://doi.org/10.1017/S0016672307008798 Published online by Cambridge University Press

determined mainly by the number of loci and number and frequencies of alleles per locus, the relative performance of the estimators changes with both the amount and the pattern of marker information (Wang, 2002). Fig. 1 compares the biases and RMSEs of the seven estimators for parent–offspring, full-sib, first cousin and unrelated individuals as a function of the number of loci. Each locus is assumed to have eight alleles with frequencies in one of three distributions: equal frequency (EF), uniform Dirichlet (UF) and one common with the remaining rare (CF).

In the case of EF, all estimators are little biased for PO, FS, HS (not shown) and FC relationships, but overestimate r substantially for unrelated dyads (Fig. 1*B*). The overestimation diminishes with an increasing number of loci, but is still about 0·02 even when 100 loci are used. The dyadic likelihood yields the highest overestimation across the range of loci. All moment estimators are unbiased when they are not truncated to the value range of [0,1], in agreement with previous work (e.g. Lynch & Ritland, 1999; Wang, 2002; Milligan, 2003), but overestimate r to the same degree as the triadic likelihood method when they are truncated. In terms of RMSE, which takes both bias and sampling variance into account (Fig. 1*A*), the triadic likelihood method is the best, giving the lowest RMSE values across the four relationships and for different numbers of loci. The dyadic likelihood outperforms moment estimators when the actual relatedness is high, but becomes less accurate than the moment estimators when the actual relatedness is zero or close to zero. Except for PO dyads, however, the magnitude of differences in RMSE among the seven estimators is general small.

Similar to the case of EF, the UF distribution of allele frequencies leads to essentially the same patterns of biases and RMSEs of the seven estimators. Allowing allele frequencies to vary within and between loci, however, further differentiates the estimators for all relationships considered (Fig. 1). This is because different moment estimators weigh differently the information from different alleles within a locus and from different loci. Furthermore, under the UF distribution, some alleles may have very low frequencies, causing some estimators (e.g. the Ritland estimator) to yield extreme values. For unrelated dyads, the QG, W and LL estimators (indistinguishable in Fig. 1*B*) give slightly more overestimation than the dyadic likelihood method, while the R, LR and TL estimators (indistinguishable) give the smallest overestimation. For highly related dyads (FS, PO), the R estimator underestimates r because it returns estimates larger than 1 which are truncated to 1. Relative to the true relatedness value, biases are more severe with unrelated dyads for all seven estimators. In terms of RMSE, the triadic likelihood method outperforms other estimators across the four relationships and the range of the number of markers. The R and LR estimators have the highest RMSEs for related dyads (PO, FS, FC) but the second-lowest RMSEs for unrelated dyads. Except for PO dyads, the RMSEs of W, LL and QG estimators are almost indistinguishable.

In the case of one common allele per locus, the triadic likelihood method severely underestimates r for highly related dyads (PO, FS) while the QG, W and LL estimators severely overestimate r for unrelated dyads. The biases decline rapidly, however, with an increasing number of loci. Milligan (2003) compared his dyadic likelihood estimator with five moment estimators in the case of a variable number of loci each having one common allele (with frequency 0·8) and four equally rare alleles (frequency 0·05). Unfortunately, because the moment estimators are not constrained to the same range as the likelihood estimator, the biases and RMSEs are hardly comparable among estimators. He found the dyadic likelihood estimator overestimates r by about 0·10 and 0·14 for FC and UR dyads, respectively, and the overestimation does not decrease with an increasing number of loci. The result is at variance with Fig. 1*B*, which shows both a smaller and a decreasing (with loci) overestimation of r for the dyadic likelihood estimator. It is unclear what causes the differences. However, the dyadic likelihood is a consistent estimator and is thus expected to show a decreasing bias with an increasing amount of information (here number of loci). In terms of RMSE, the dyadic likelihood performs best for highly related dyads (PO, FS) while the triadic likelihood outperforms others for loosely related (FC) or unrelated (UR) dyads.

It should be noted that the two allele frequency distributions, equal allele frequency and one allele common with the remaining rare, represent the extremely informative and uninformative marker cases which are unlikely to be encountered in practice. With one allele at a frequency of 0·8, the expected heterozygosity (h) is always less than 0·36 no matter

Fig. 1. Comparison of seven relatedness estimators using simulated data with different numbers of loci. The root mean squared errors (y-axis, on a logarithmic scale; *A*) and biases (*B*) are plotted as a function of the number of loci (x-axis) for four relationships (indicated by column heads) and three allele frequency distributions (indicated by row heads on the right). Each locus is assumed to have eight alleles with known frequencies in an EF, UF or CF distribution. The seven estimators indicated by different lines are the new triadic likelihood (TL), Wang's (2002) estimator (W), Lynch (1988) and Li *et al.*'s (1993) estimator (LL), Lynch and Ritland's (1999) estimator (LR), Ritland's (1996) estimator (R), Queller and Goodnight's (1989) estimator (QG) and Milligan's dyadic likelihood estimator (M).

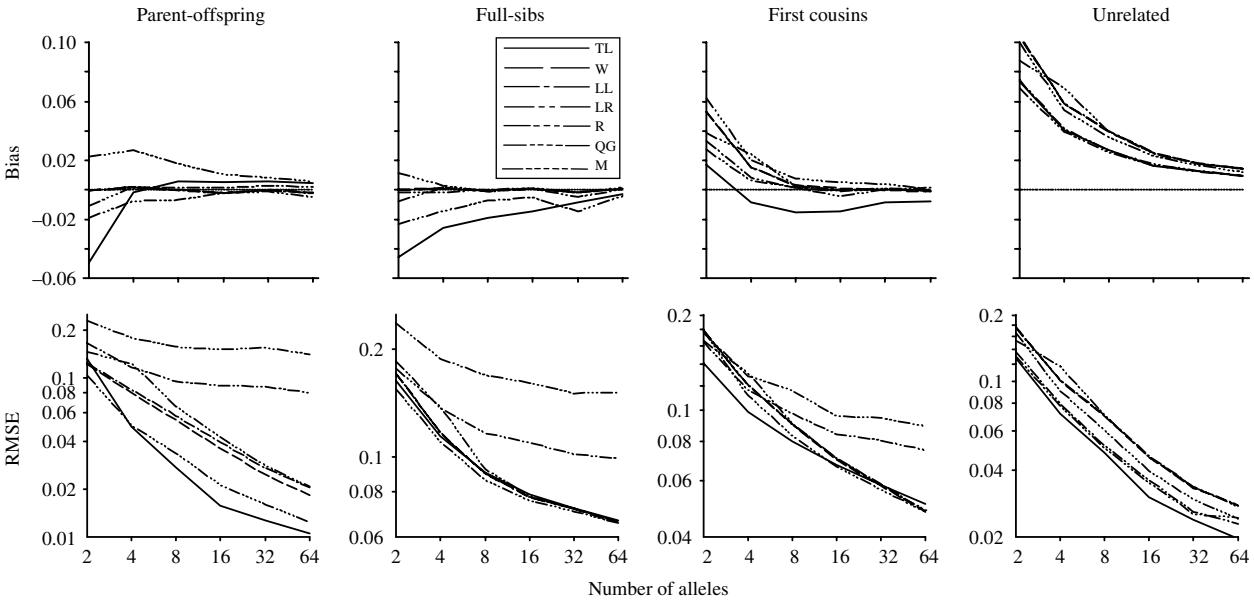

Fig. 2. Comparison of seven relatedness estimators using simulated data with different numbers of alleles per locus. The biases and root mean squared errors (y-axis, on a logarithmic scale) are plotted as a function of the number of alleles per locus (x-axis, on a logarithmic scale) for four relationships (indicated by column heads). The number of loci is fixed at 30, and the alleles at each locus have known frequencies drawn from a uniform Dirichlet distribution. The seven estimators indicated by different lines are specified in Fig. 1 and the text.

how many alleles are present at a locus. A survey of microsatellites in vertebrates and insects shows that only about 8% of the loci display $h \leqslant 0{\cdot}36$ (Aparicio *et al.*, 2006). Among the 10 875 markers with more than two alleles in the CEPH dataset, only 190 markers (1·7%) have $h \leqslant 0{\cdot}36$. Even among the 21 148 biallelic markers in the CEPH dataset, less than 40% have $h \leqslant 0{\cdot}36$. It is understandable that, with extremely uninformative markers, the rank order of estimators in terms of RMSE (or bias) changes depending on true relatedness values. One can imagine the extreme case where each locus has one allele with a frequency close to 1 (i.e. almost fixed). All individuals, irrespective of their relationships, are therefore homozygous for common alleles at all loci. Intuitively, the markers are uninformative about relatedness and any r estimate value in the range [0,1] should be equally probable. It turns out that in such a case the TL, M, W, LL, LR, R, and QG estimators are 1, 1, 1, 1, 1, 0, 0, respectively. The likelihood methods give, however, an almost flat likelihood curve, implying that all linear combinations of Δ_7, Δ_8 and Δ_9 (and thus any value of r between 0 and 1) is equally plausible given the genotype data.

More realistically, with a mixture of allele frequency distributions at different loci, the RMSE and bias patterns become intermediate among those shown in Fig. 1 (data not shown). It is worth noting that the triadic likelihood method yields the lowest RMSEs for loosely related (FC) to unrelated dyads irrespective of the allele frequency distribution. In a real population, most dyads are expected to be either unrelated or only loosed related (Csilléry *et al.*, 2006). This is effectively true even with a small population closed to immigration for many generations, considering that allele frequencies are usually estimated from a sample of individuals taken from the current generation or only a few generations ago and thus the reference population in which the average r value is expected to be zero (Ritland, 1996) refers to either the current or a recent generation.

Fig. 2 compares the seven estimators applied to data simulated using a fixed number of 30 loci and a varying number of alleles per locus. The allele frequencies are drawn from a UF distribution. Relatively, the highest biases occur with unrelated dyads for all estimators. The biases, however, decline rapidly with an increasing number of alleles per locus. Similar to the Dirichlet frequency distribution case in Fig. 1, the TL, R, and LR estimators are indistinguishable in bias and are less biased than the other four estimators. The R and LR estimators show the highest RMSEs for related dyads (FC, FS, PO) but the second-lowest RMSEs for unrelated dyads. Overall, the triadic likelihood method is the most accurate, yielding RMSE values which are either the smallest or close to the smallest among those of the seven estimators for different relationships and different numbers of alleles per locus.

(iii) *Inbreeding*

Previous estimators assume non-inbred individuals (i.e. the homologous genes at a locus within any

https://doi.org/10.1017/S0016672307008798 Published online by Cambridge University Press

individual are always non-IBD). In small populations, however, some individuals are inevitably inbred to varying degrees. To investigate how robust different relatedness estimators are to the violation of no inbreeding, and how much improvement can be gained by incorporating inbreeding into the likelihood estimator, I simulated PO, FS, FC and UR dyads with each individual in a dyad having the same inbreeding coefficient F. Each individual is genotyped at 15 loci, each having eight alleles in a triangular frequency distribution. The results are summarized in Fig. 3*A*, obtained by applying the estimators to the simulated data and assuming known allele frequencies. In comparison, the same data are also analysed by the dyadic (9-dimensional) and triadic (66-dimensional) likelihood methods that account for inbreeding.

For related dyads (PO, FS, FC), all estimators, except for the two likelihood methods that explicitly take inbreeding into account and the R estimator, tend to underestimate relatedness with an increasing F. As a result, the RMSEs of these estimators increase with an increasing F. It is interesting to note that among the seven estimators ignoring inbreeding, the R estimator is the least biased with inbreeding. It, however, has a high sampling variance and thus is the least accurate estimator. The likelihood methods seem to be more susceptible to inbreeding than moment estimators, especially with highly related (PO, FS) dyads. They are the most accurate estimators in the absence of inbreeding but quickly become the least accurate estimators when $F > 0{\cdot}1$.

For unrelated dyads, the actual relatedness value is 0, irrespective of the value of F. Both biases and RMSEs of all estimators assuming non-inbreeding are not affected by the level of inbreeding. The two likelihood methods that account for inbreeding display, however, an increasing bias and RMSE with an increasing F.

Accounting for inbreeding in the likelihood methods does not necessarily lead to a better performance. As can be seen from Fig. 3*A*, estimators accounting for inbreeding result in reduced biases and RMSEs only when F is high (0·15) in closely related dyads (FS, PO), but increased biases and RMSEs when the dyads are loosely related or unrelated or when F is small. Allowing for inbreeding dramatically increases the number of parameters to be estimated in, and thus decreases the precision of, the likelihood estimators. Estimating more parameters jointly requires more data. Fig. 3*B* compares the biases and RMSEs of the nine estimators as a function of the number of markers. Each marker has eight alleles in a triangular frequency distribution. Each individual in a dyad is inbred with $F = 0{\cdot}16$, so that the actual relatedness is 0·7400, 0·6600, 0·2178 and 0 for PO, FS, FC and UR dyads, respectively. At this high level of inbreeding, the triadic and dyadic likelihood estimators that account for inbreeding are hardly distinguishable in performance. For highly related dyads (PO, FS), the likelihood estimators that incorporate inbreeding have both biases and RMSEs declining rapidly with an increasing number of loci, and become increasingly superior to estimators which assume no inbreeding. In contrast, for unrelated dyads, accounting for inbreeding in the likelihood estimators results in a consistent reduction in accuracy, no matter how many loci are used in the estimation. Similar to Fig. 3*A*, Fig. 3*B* also shows that the R estimator is little biased for all relationships and numbers of loci. Its overestimation for the relatedness of UR dyads is caused by truncation. When the number of loci is large and the actual relatedness is high (PO, FS), the R estimator becomes the most accurate (measured by biases and RMSEs) while the two likelihood methods become the least accurate among the seven estimators ignoring inbreeding.

Summarizing the results shown in Fig. 3*A* and *B*, it seems unjustified to take inbreeding into account in the likelihood methods for estimating relatedness, except when a large proportion of dyads in a sample are highly inbred and closely related, and when there is ample marker information (e.g. hundreds of microsatellites).

(iv) *Genotyping errors*

Just like any other type of data, genotype data are usually not perfect and free of errors. Mutations and genotyping errors seem to be inevitable in genotype data (e.g. Bonin *et al*., 2004; Pompanon *et al*., 2005), especially when DNA quantity and quality are limited so that repeated genotyping is either impossible or unhelpful for eliminating typing errors. Although genotyping errors are shown to have dramatic effects on relationship inference (e.g. Wang, 2004*a*), they are perceived as unimportant in relatedness estimation and are thus ignored in all previous estimators. Furthermore, incorporating genotyping errors into the inference may incur a cost, lowering the power of analyses (e.g. Morrissey & Wilson, 2005). It is, however, desirable to understand how much effect genotyping errors have on relatedness estimation, and whether (and when) it is justified to account for genotyping errors in relatedness estimators.

Fig. 4 plots the biases and RMSEs as a function of the actual rate (e) of typing errors at a marker locus used in simulations. Relatedness for PO, FS, FC and UR dyads is estimated by the triadic likelihood estimator assuming an error rate of $\hat{e} = 0$, $\hat{e} = e$, $\hat{e} = 0{\cdot}8e$ and $\hat{e} = 1{\cdot}2e$. In comparison with the case of $\hat{e} = e$, the case of $\hat{e} = 0$ shows the consequence of ignoring errors, while the cases of $\hat{e} = 0{\cdot}8e$ and $\hat{e} = 1{\cdot}2e$ show the effects of sampling errors of e on the estimator's performance. Twenty loci, each having eight alleles in

https://doi.org/10.1017/S0016672307008798 Published online by Cambridge University Press

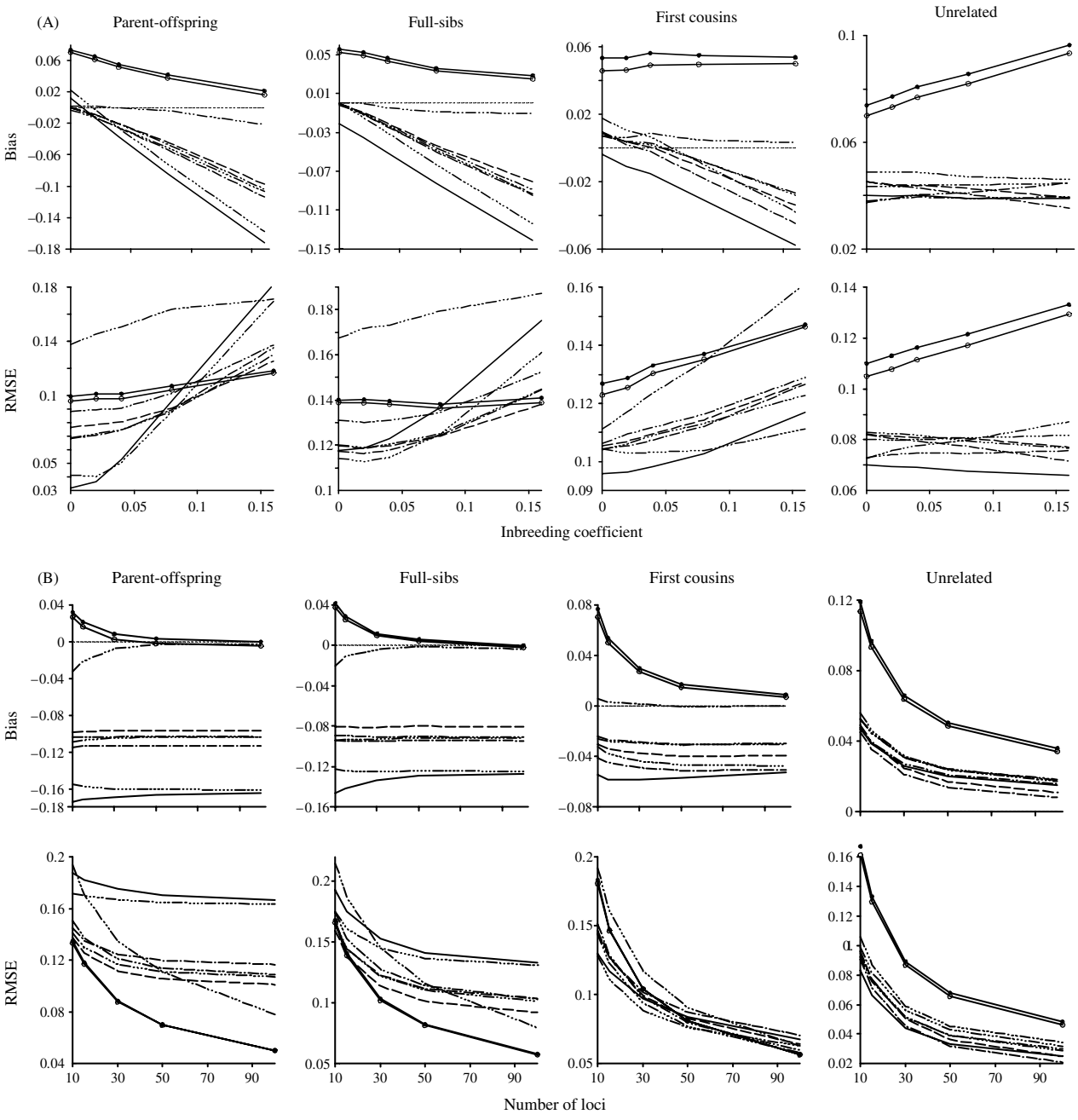

Fig. 3. Effect of inbreeding on the biases and RMSEs of relatedness estimators. Each individual in the PO, FS, FC or UR dyads has the same inbreeding coefficient, *F*. With known allele frequencies, biases and RMSEs are calculated by the seven relatedness estimators (specified in Fig. 1 and the text) assuming non-inbreeding, and by the dyadic (denoted by unbroken lines with filled circles) and triadic (denoted by unbroken lines with open circles) likelihood methods allowing for inbreeding. (*A*) Biases and RMSEs plotted as a function of the inbreeding coefficients of individuals. Each individual is genotyped at 15 loci, with each locus having eight alleles in a triangular frequency distribution. (*B*) Biases and RMSEs plotted as a function of the number of loci, each having eight alleles in a triangular frequency distribution. The *F* value of each individual is fixed at 0·16.

a uniform Dirichlet (UF) frequency distribution, are genotyped for each individual. The genotype at each locus of an individual is changed in simulations according to the error rate and error model described above. As can be seen from Fig. 4, the impact of genotyping errors increases with an increasing error rate. For related dyads, typing errors result in multi-locus genotypes that are less similar than they should be, leading to downward biased and less precise estimates of relatedness, and thus an elevated RMSE. Accounting for typing errors in the estimator improves the estimates (i.e. reducing underestimation and RMSEs) for closely related dyads (e.g. PO, FS), but impairs the estimates for unrelated dyads. This is

https://doi.org/10.1017/S0016672307008798 Published online by Cambridge University Press

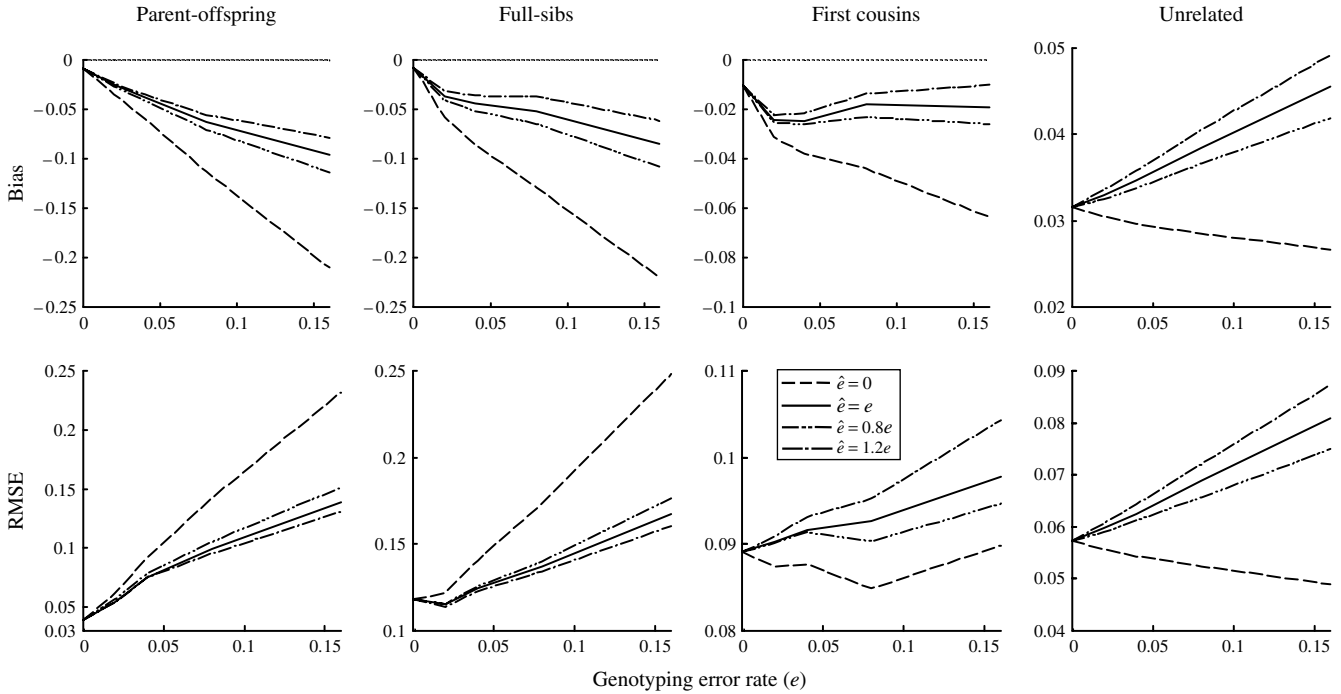

Fig. 4. RMSEs of relatedness estimators as a function of the genotyping error rate (e, x-axis) in generating the simulated data. Each individual in the parent–offspring, full-sib, first cousin and unrelated dyads is genotyped at 20 loci, with each locus having eight alleles in a uniform Dirichlet frequency distribution. RMSEs are calculated for the triadic likelihood estimator assuming known allele frequencies and an error rate of $\hat{e}=0$, $\hat{e}=e$, $\hat{e}=0{\cdot}8e$ or $\hat{e}=1{\cdot}2e$.

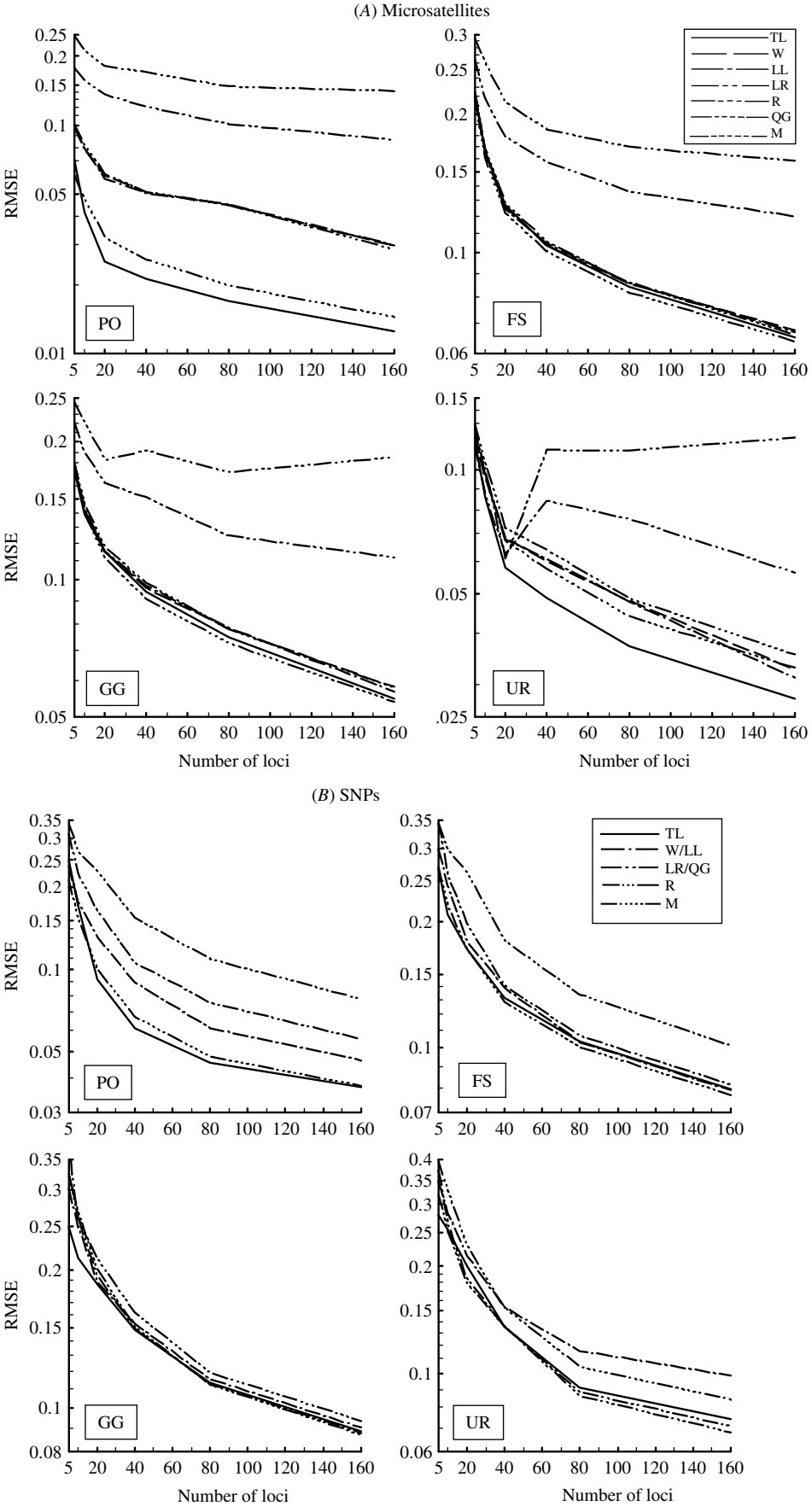

Fig. 5. For legend see opposite page.

https://doi.org/10.1017/S0016672307008798 Published online by Cambridge University Press

understandable because typing errors have little effect on the similarity between the multi-locus genotypes of unrelated individuals in expectation. Therefore accounting for typing errors in the data merely reduces the information of the marker data, which adversely affects the quality of the estimates of unrelated dyads. For loosely related dyads (e.g. FC), accounting for typing errors reduces the downward biases but increases the sampling errors of $\hat{r}$, leading to an increased RMSE.

Another conclusion that can be drawn from Fig. 4 is that the estimator is relatively robust to the sampling effect of e. Estimates obtained by assuming $\hat{e}=e$, $\hat{e}=0{\cdot}8e$ and $\hat{e}=1{\cdot}2e$ have similar biases and RMSEs, especially when the true relatedness is not small. An overestimated typing error rate ($\hat{e}=1{\cdot}2e$) results in a smaller bias and RMSE for closely related dyads but a larger bias and RMSE for unrelated dyads than an underestimated typing error rate ($\hat{e}=0{\cdot}8e$). Note that for unrelated dyads, the bias and RMSE decrease with an increasing typing error rate used in simulating the data. This is an artefact due to the particular combination of allele frequency distribution and error model adopted in the simulations. With a UF distribution of allele frequencies, it is possible that some unrelated dyads display multi-locus genotypes of high similarity, leading to overestimated relatedness. However, typing errors as modelled in the simulation could destroy the similarity and thus reduce the overestimation.

Due to the mixed effects of accounting for typing errors on the biases and RMSEs of closely and loosely related dyads, it is unclear whether or not it is desirable to build a typing error model into relatedness estimator. It depends on the marker data quality (e), the actual proportions of closely and loosely related dyads in a sample, and the particular purpose of the data analysis. When e is substantial for a large number of loci, a two-step procedure might be used to improve relatedness analyses. First, $\hat{r}$ is estimated for each dyad assuming $e=0$. Second, for those dyads with high $\hat{r}$ values obtained in the first step (say, $\hat{r}>0{\cdot}1$), refined estimates can be obtained by account for genotyping errors. Further studies are required to investigate the performance of the two-step procedure.

(v) *Analysis of the CEPH dataset*

Two subsets of the CEPH data were analysed by the seven estimators for pairwise relatedness between individuals within each of 65 families. For the microsatellite subset, a number between five and 160 of the most informative (Wang, 2006) microsatellites are chosen and individuals that are genotyped at $\geqslant 90\,\%$ of the chosen loci are included for relatedness analyses. For the SNP subset, the most informative markers whose minor allele frequencies are $\leqslant 0{\cdot}48$ are chosen, while individuals are screened using the same criteria as the microsatellite subset. SNPs with allele frequencies of 0·5 are the most informative, but are excluded from the subset because they cause the Lynch and Ritland estimator to be undefined.

Fig. 5 plots the RMSEs of relatedness estimates for PO, FS, GG, and UR dyads across families as a function of the number of microsatellites (Fig. 5*A*) or SNPs (Fig. 5*B*) used in the estimation. For microsatellites, the RMSEs of W, LL and QG are hardly distinguishable for all four relationships and different numbers of loci. Compared with the other five estimators, the R and LR estimators have quite poor performance, even for unrelated dyads when the number of loci is larger than 20. The poor performance of the two estimators arises perhaps because a large number of highly polymorphic microsatellites (some having more than 40 alleles) are used in the estimation, and some alleles have very small frequencies which cause extreme values (truncated to the range [0, 1]) of the two estimators. Overall, the new triadic likelihood estimator is the best with the smallest RMSEs. It gives much smaller RMSEs than the other estimators for PO and UR dyads, while it gives RMSEs almost indistinguishable from those of the best estimator for FS and GG dyads.

For the SNP subset (Fig. 5*B*), the LL and W estimators are the same while the LR and QG estimators always have indistinguishable RMSEs for all the four relationships. For clarity, only one estimator in each pair is plotted in Fig. 5*B*. Because of the absence of rare alleles, the R and LR estimators now become better than LL and W estimators for unrelated dyads. They are still the worst estimators, however, for highly related dyads (PO, FS). Evaluated across relationships and numbers of loci, the new triadic likelihood estimator has the best overall performance.

FS and PO relationships have the same expected value of relatedness ($r=0{\cdot}5$), but different expected values of Δ_7 and Δ_8. Some relatedness estimators (e.g. LR, W, likelihood) allow the joint estimates of Δ_7 and Δ_8 and thus can be used to differentiate the relationships. However, the statistical power of such analyses is generally low, because Δ_8 is difficult to estimate accurately (e.g. Lynch & Ritland, 1999). Only when

Fig. 5. RMSEs of relatedness estimators as a function of the number of microsatellite (*A*) or SNP (*B*) loci used in the CEPH dataset. The RMSE of an estimator is calculated for each type and number of markers, and for each of the four relationships (parent–offspring, PO; full-sibs, FS; grandparent–grandoffspring, GG; unrelated, UR). Note that for the case of SNPs, estimators W and LL are the same while LR and QG have indistinguishable RMSEs.

https://doi.org/10.1017/S0016672307008798 Published online by Cambridge University Press

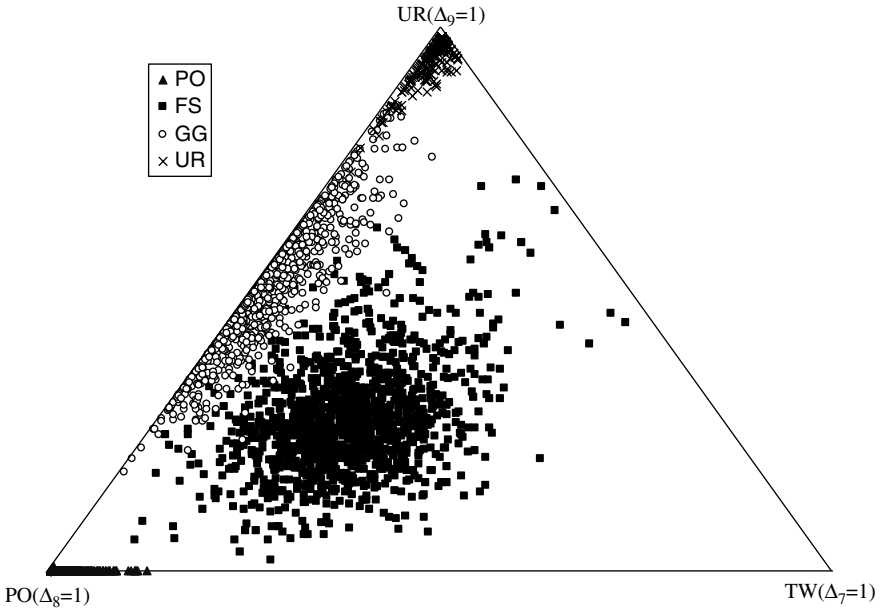

Fig. 6. Triangular plot of IBD coefficients for PO, FS, GG and UR dyads in the CEPH dataset. The numbers of dyads shown in the plot are 795, 1304, 970, and 311 for PO, FS, GG and UR relationships, respectively. Estimates of Δ_i ($i=7$, 8, 9) for each dyad are obtained from the triadic likelihood method using 100 microsatellites with known allele frequencies. The top, left and right points of the triangle have IBD coefficients $\{\Delta_7, \Delta_8, \Delta_9\}=\{0,0,1\}$, $\{0,1,0\}$ and $\{1,0,0\}$, respectively, representing UR, PO and identical twin relationships, respectively.

many polymorphic markers are available can relationships be accurately inferred from the estimates of IBD coefficients. For the CEPH dataset, the Δ_7 and Δ_8 estimates for PO, FS, GG, and UR dyads are shown in Fig. 6. The estimates are obtained from the triadic likelihood method using 100 microsatellites, assuming non-inbreeding. As can be seen, the four relationships are well differentiated. IBD coefficients are accurately estimated for PO and UR dyads, but less so for GG and FS dyads. Estimates of Δ_8 are especially variable for FS dyads, because the FS relationship has the maximal inherent between-locus variance of Δ_8, which is $\frac{1}{4}$ (Wang, 2006).

4. Discussion

Compared with moment estimators of relatedness, the likelihood methods enjoy several attractive features (Milligan, 2003). First, estimates of IBD coefficients and relatedness are naturally constrained to their biologically meaningful ranges [0,1], making the interpretation and application in subsequent analyses (e.g. estimating heritability: Ritland, 2000; Thomas *et al*., 2000) of the results straightforward. One can truncate moment estimates of relatedness to force them to fall in the legitimate range of [0,1], as has been done by the present study. Such truncated estimators have a reduced standard deviation, but are upwardly biased with RMSEs still larger than likelihood estimators in most cases (Milligan, 2003; present study). Second, likelihood methods automatically weigh information among alleles and among loci optimally. The amount of information about pairwise relatedness provided by a locus varies greatly, depending on the number and frequencies of alleles at the locus (Wang, 2006). Microsatellites are generally much more useful than SNPs, for example. More informative markers allow more accurate (reliable) estimates of relatedness, and thus should logically be given more weight in a multi-locus estimator. Unfortunately, however, the informativeness (thus weight) of a marker is also dependent on the true but unknown relatedness of the dyad under consideration. As a result, moment estimators have either to ignore the difference in informativeness among loci by applying an equal weight (e.g. Queller & Goodnight, 1989; Li *et al*., 1993) or to use approximate weights derived assuming an unrelated dyad (e.g. Ritland, 1996; Lynch & Ritland, 1999; Wang, 2002). In general, estimators using equal weights tend to give better estimates for highly related dyads (e.g. FS, PO) and worse estimates for unrelated dyads than estimators using unequal weights (Wang, 2002). Because of the large number of weights that are based on $r=0$, some estimators (Ritland, 1996; Lynch & Ritland, 1999) show the bizarre behaviour that their sampling variance does not decrease or even increases with an increasing number of alleles per locus for highly related dyads (Wang, 2002; Figs. 2 and 5*A*). Third, likelihood methods are more general and flexible than moment estimators in allowing for different kinds of markers, inbreeding and genotyping errors. Some moment

https://doi.org/10.1017/S0016672307008798 Published online by Cambridge University Press

estimators (Lynch & Ritland, 1999; Queller & Goodnight, 1989) have limited use for biallelic markers such as SNPs, because they are undefined (denominator equal 0) under some circumstances. No single moment estimator available can be applied to both dominant (e.g. Ritland, 2005; Wang, 2004*b*) and co-dominant markers. In contrast, likelihood methods apply to markers with any number and frequency distribution of alleles, and can easily be adapted for dominant markers. Similarly, all moment estimators assume the absence of inbreeding and genotyping errors, while likelihood methods can account for inbreeding (Weir *et al*., 2006; present study) and typing errors (present study).

In spite of the above advantages of the likelihood methods, they tend to overestimate relatedness, especially for loosely related or unrelated dyads (Milligan, 2003). As a result, their overall quality of estimates measured by RMSE is not always higher than that of moment estimators. Considering that most dyads in a natural population are probably unrelated (e.g. Csilléry *et al*., 2006), the overall performance of likelihood methods evaluated across all possible dyads in a sample of individuals might be inferior to that of Lynch & Ritland (1999) or Ritland (1996) in some cases. In this investigation, I have shown that the triadic likelihood method can reduce the overestimation and RMSE of relatedness for unrelated or loosely related dyads substantially, compared with the dyadic likelihood method. Effectively, the triadic method uses a third individual as a control in estimating the relatedness of a dyad. Both simulated and empirical data show that the triadic estimator yields relatedness estimates with RMSEs that are the lowest for loosely related dyads and PO dyads, and are close to the lowest for other relationships. Because usually unrelated dyads dominate a random sample of individuals from a natural population, the triadic likelihood estimator offers the best estimates overall.

IBD coefficients and relatedness are all relative measurements with an implicit reference population in which all homologous genes are assumed non-identical by descent (Ritland, 1996). The reference population is defined as specific in both time and space. For a given sample of individuals from a finite population, relatedness of all possible dyads would be increased (decreased) by a similar amount when one moves the reference time point backward (forward), or when one moves the reference space from a subpopulation to a metapopulation (a line within the subpopulation). In practical applications, the reference population is the one whose allele frequencies are used in estimating relatedness. For an unbiased estimator, therefore, the average relatedness estimates across all possible dyads in a sample would be close to zero if allele frequencies were estimated from the same sample, irrespective of the actual relationships among the sampled individuals. This argument is, however, only partially true, because relatedness estimates depend also on the estimators. As an example, I simulated a sample of 50 individuals that were exclusively full-sibs or half-sibs. Each individual was genotyped at 10 loci, with each locus having 10 alleles in a uniform Dirichlet frequency distribution. Estimates from moment estimators were not truncated to the range of [0,1]. Using allele frequencies estimated from the same sample without accounting for relationships, the W, LL, LR, R, QG, TL and M estimators yielded average (among 1225 dyads within a sample, over 1000 replicate samples) estimates of relatedness of 0·17, 0·18, −0·01, −0·01, −0·05, 0·14, and 0·12 for the FS sample, 0·06, 0·05, −0·01, −0·01, −0·01, 0·04, and 0·02 for the HS sample. While the LR, R and QG estimates are close to the expectation of zero, the W, LL, and likelihood methods (TL and M) give higher than expected estimates. In contrast, the average estimates from all estimators are close to 0·5 and 0·25 for FS and HS samples, respectively, when allele frequencies are assumed known or estimated from another sample containing unrelated individuals. Similarly, relatedness estimates are affected by population structures, and by how samples are taken from a structured population and combined in estimating allele frequencies (Oliehoek *et al*., 2006). Although all relatedness estimators are based on a reference population, it seems that the W, LL, TL and M estimators are less sensitive to the reference than the other estimators.

It is shown that relatedness is underestimated when inbreeding is present but ignored by an estimator. Likelihood methods can be made to account for the inbreeding of individuals in estimating their relatedness. However, due to the dramatic increase in the number of parameters to be estimated, incorporating inbreeding into the likelihood methods incurs a cost, resulting in a possible decrease in performance as measured by RMSE over all dyads in a sample. Except for the scenario of highly inbred and closely related individuals and a large number of loci available, the likelihood methods assuming non-inbreeding are recommended.

Genotyping errors have been speculated to have a minor role in relatedness estimation. This study is the first to develop an estimator that allows for genotyping errors and to investigate the impact of data quality through simulations. My simulations suggest that typing errors result in underestimation of relatedness for highly related dyads. Accounting for typing errors improves estimates for closely related dyads but impairs those for loosely related or unrelated dyads. Because most dyads in a sample are unrelated (Ritland, 1996; Csilléry *et al*., 2006), taking typing errors into account in an estimator may well

incur a net cost. However, when one is more concerned with identifying highly related dyads and many markers are available, it is justified to use an estimator that is robust to typing errors.

Following previous estimators, the current triadic likelihood method is developed for unlinked markers. With a large number of markers, however, some of them are inevitably linked. Linkage causes the information from different loci to be correlated, resulting in over-confidence in relatedness estimates obtained assuming independent markers. It should not, however, affect the degree of bias of relatedness estimates. Note that the influence of linkage on the precision of relatedness estimates decreases with a decreasing degree of true relatedness. Linkage is expected to have little effect on the precision and confidence intervals of relatedness estimates for unrelated or loosely related dyads. This is because there are so many generations linking two loosely related individuals and their common ancestors that even tightly linked markers already have recombined. Because most dyads are unrelated, estimators assuming unlinked markers should apply approximately to linked markers. More work is needed to investigate how linkage with and without linkage disequilibrium influences the precision of relatedness estimates.

In this study, relatedness estimators were compared using a limited number of simple relationships (PO, FS, HS, FC, GG, UR) in simulated and empirical datasets. In agreement with previous work (e.g. Lynch & Ritland, 1999; Wang, 2002; Milligan, 2003), I have shown that the rank order of the estimators changes with the true relatedness being estimated. This may raise some doubts that the performance rank of different estimators evaluated using a few simple relationships may not apply to real populations in which numerous complex relationships are present (e.g. Csilléry *et al*., 2006). However, I believe the evaluation procedure adopted in the study is appropriate and the conclusion reached is generally valid for real populations. First, although a real population usually contains myriads of relationships, one has to be satisfied with using just a few representative ones in comparing estimators. Considering all possible kinds and proportions of relationships in a population is impossible even in simulation studies. Second, it is fortunately unnecessary to consider many different relationships in evaluating the estimators. For example, one can compare estimators using UR ($\Delta_7 = \Delta_8 = 0$) and HS ($\Delta_7 = 0$, $\Delta_8 = 0{\cdot}5$) relationships. The results obtained should apply roughly to any relationship that lies between UR and HS, with $\Delta_7 = 0$ and $0 \leqslant \Delta_8 \leqslant 0{\cdot}5$. The relationships (PO, FS, HS, GG, FC, UR) adopted in the study in comparing estimators are reasonably representative of the common ones found in real populations and cover a wide range of degrees and patterns of IBD ($\Delta_7 = 0 \sim 0{\cdot}25$, $\Delta_8 = 0 \sim 1$, $\Delta_9 = 0 \sim 1$). It is also noticeable that in most cases I compared the performances of estimators for each individual relationship independently. It is, however, simple to assess the overall performance of an estimator by assembling the quality measurements (such as RMSE) for separate relationships, provided the proportions of these relationships are known in a population. For the convenience of users in comparing the estimators using their own marker data and specifically interested relationships (with or without inbreeding), I am developing Windows software that implements the seven relatedness estimators and will post it on my website for free downloading.

I thank Bill Hill and two anonymous reviewers for helpful comments on an earlier version of this paper.

References

Aparicio, J. M., Ortego, J. & Cordero, P. J. (2006). What should we weigh to estimate heterozygosity, alleles or loci? *Molecular Ecology* **15**, 4659–4665.

Blouin, M. S. (2003). DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. *Trends in Ecology and Evolution* **18**, 503–511.

Bonin, A., Bellemain, E., Eidesen, P. B., Pompanon, F., Brochmann, C. & Taberlet, P. (2004). How to track and assess genotyping errors in population genetics studies. *Molecular Ecology* **13**, 3261–3273.

Csilléry, K., Johnson, T., Beraldi, D., Clutton-Brock, T. H., Coltman, D., Hansson, B., Spong, G. & Pemberton, J. (2006). Performance of marker-based relatedness estimators in natural populations of outbred vertebrates. *Genetics* **173**, 2091–2101.

Hamilton, W. D. (1964). The genetical evolution of social behaviour: I and II. *Journal of Theoretical Biology* **7**, 1–52.

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

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

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

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

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

Lynch, M. & Walsh, J. B. (1998). *Genetics and analysis of Quantitative Traits*. Sunderland, MA: Sinauer Associates.

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

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

Morrissey, M. B. & Wilson, A. J. (2005). The potential costs of accounting for genotypic errors in molecular parentage analyses. *Molecular Ecology* **14**, 4111–4121.

Oliehoek, P. A., Windig, J. J., van Arendonk, J. A. M. & Bijma, P. (2006). Estimating relatedness between individuals in general populations with a focus on their use in conservation programs. *Genetics* **173**, 483–496.

Pompanon, F., Bonin, A., Bellemain, E., & Taberlet, P. (2005). Genotyping errors: causes, consequences and solutions. *Nature Reviews Genetics* **6**, 847–859.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1996). *Numerical Recipes in Fortran 90*, 2nd edn. Cambridge: Cambridge University Press.

Queller, D. C. & Goodnight, K. F. (1989). Estimating relatedness using molecular markers. *Evolution* **43**, 258–275.

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

Ritland, K. (2000). Marker-inferred relatedness as a tool for detecting heritability in nature. *Molecular Ecology* **9**, 1195–1204.

Ritland, K. (2005). Multilocus estimation of pairwise relatedness with dominant markers. *Molecular Ecology* **14**, 3157–3165.

Thomas, S. C. (2005). The estimation of genetic relationships using molecular markers and their efficiency in estimating heritability in natural populations. *Philosophical Transactions of the Royal Society of London, Series B* **360**, 1457–1467.

Thomas, S. C., Pemberton, J. M. & Hill, W. G. (2000). Estimating variance components in natural populations using inferred relationships. *Heredity* **84**, 427–436.

Thompson, E. A. (1974). Gene identities and multiple relationships. *Biometrics* **30**, 667–680.

Van de Casteele, T., Galbusera, P. & Matthysen, E. (2001). A comparison of microsatellite-based pairwise relatedness estimates. *Molecular Ecology* **10**, 1539–1549.

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

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

Wang, J. (2004*b*) Estimating pairwise relatedness from dominant genetic markers. *Molecular Ecology* **13**, 3169–3178.

Wang, J. (2006). Informativeness of genetic markers for pairwise relationship and relatedness inference. *Theoretical Population Biology* **70**, 300–321.

Weir, B. S. (1996). *Genetic Data Analysis II*. Sunderland, MA: Sinauer Associates.

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