npg

REVIEW

A maximum-likelihood estimation of pairwise relatedness for autopolyploids

K Huang[1,3], ST Guo[1,3], MR Shattuck[2], ST Chen[1], XG Qi[1], P Zhang[1] and BG Li[1]

Relatedness between individuals is central to ecological genetics. Multiple methods are available to quantify relatedness from molecular data, including method-of-moment and maximum-likelihood estimators. We describe a maximum-likelihood estimator for autopolyploids, and quantify its statistical performance under a range of biologically relevant conditions. The statistical performances of five additional polyploid estimators of relatedness were also quantified under identical conditions. When comparing truncated estimators, the maximum-likelihood estimator exhibited lower root mean square error under some conditions and was more biased for non-relatives, especially when the number of alleles per loci was low. However, even under these conditions, this bias was reduced to be statistically insignificant with more robust genetic sampling. We also considered ambiguity in polyploid heterozygote genotyping and developed a weighting methodology for candidate genotypes. The statistical performances of three polyploid estimators under both ideal and actual conditions (including inbreeding and double reduction) were compared. The software package POLYRELATEDNESS is available to perform this estimation and supports a maximum ploidy of eight.

INTRODUCTION

Knowledge of the relatedness between individuals in a population is of central importance to many aspects of biology, including population genetics, conservation and sociobiology (that is, Charpentier *et al.*, 2012; Mattila *et al.*, 2012; Liu *et al.*, 2013). Although the coefficient of relatedness between individuals can be calculated from a known pedigree, in the absence of this information, relatedness can be estimated using genetic marker data. A number of estimators have been developed for this purpose and can generally be classified into two categories: method-of-moment and maximum-likelihood estimators.

Method-of-moment estimators substitute sample moments for the unknown population moment to estimate various population parameters. These methods can generate an unbiased estimation of relatedness, including either *r* directly (that is, Queller and Goodnight, 1989; Li *et al.*, 1993; Ritland, 1996), or both *r* and Δ (four-gene coefficient, the probability that both genes of one individual are identical by descent (IBD) to both genes of another individual) simultaneously (that is, Lynch and Ritland, 1999; Wang, 2002; Thomas, 2010). Here we define *r* as the probability that an allele sampled from one individual at a locus is IBD to one of the alleles from the other individual. Although this IBD definition of relatedness is incompatible with that of Wright (1921), which was based on the correlation between individual allele frequencies (see Supplementary Files for details), the two definitions are identical in the absence of double reduction, inbreeding and selfing. According to IBD, *r* should

range from 0 to 1. However, method-of-moment estimators can produce relatedness values outside of this range. This potential problem can be resolved by truncating the estimators, although this produces bias (Milligan, 2003; Wang, 2011).

The second method, maximum likelihood, was developed by Milligan (2003) and Anderson and Weir (2007) for estimating pairwise relatedness. This was based on the earlier work of Thompson (1975). Maximum likelihood estimates the probability of observing a given pairwise allelic pattern ϕ (two-gene coefficient, the probability that a single allele in one individual is IBD to one in another individual), Δ and the allele frequencies. By searching parameter space for ϕ and Δ values that maximize the probability of the genotype pattern observed, maximum-likelihood values can be determined. Because maximization can be limited to the parameter space as defined by probabilities of IBD, invalid values for the parameters are avoided.

Both methods are limited to making estimations based on disomic inheritance. Although some coefficient of coancestry estimators developed for diploids can be extended to polyploids (that is, Loiselle *et al.*, 1995; Ritland, 1996), as has been done in the software SPAGEDI V1.4 (Hardy and Vekemans, 2002), they fail to directly estimate polysomic inheritance. A significant proportion of plant species are autopolyploid, with 30–80% of angiosperms showing polyploidy (Burow *et al.*, 2001) and most lineages showing evidence of paleoploidy (Otto, 2007). Although rare, polyploidy is also present in animals (for example, Salmonidae fish, African clawed frog:

[1]Key Laboratory of Resource Biology and Biotechnology in Western China of Ministry of Education, College of Life Sciences, Northwest University, Xi'an, China and [2]Center of Human Origins, Department of Anthropology, New York University, New York, NY, USA
[3]These authors contributed equally to this work.
Correspondence: Professor BG Li or Professor ST Guo, Key Laboratory of Resource Biology and Biotechnology in Western China of Ministry of Education, College of Life Sciences, Northwest University, Xi'an, ShaanXi 710069, China.
E-mail: baoguoli@nwu.edu.cn or songtaoguo@nwu.edu.cn

Xenopus laevis, Weather Loach: *Misgurnus anguillicaudatus*). There are two distinct mechanisms of genome duplication that result in polyploidy: allopolyploidy and autopolyploidy. In allopolyploidy, chromosomes originate from two species; in autopolyploidy, all chromosomes originate within a single species, often due to unreduced gametes. This paper focuses on autopolyploids.

Because of their importance to agriculture, there has been much scientific investigation of plant autopolyploids (López-Pujol *et al.*, 2004; Luo *et al.*, 2006). In addition, autopolyploids do not exhibit disomic inheritance, whereas allopolyploids do so because of minor differences between chromosomes originating from different species (Luo *et al.*, 2006). Polyploids displaying disomic inheritance can thus be described using normal diploid models once alleles are assigned to the alternative duplicated loci (cf Ritland and Ganders, 1985). However, diploid models cannot be applied to polyploids that display polysomic inheritance, that is, autopolyploids. Thus, few models apply directly to autopolyploids (cf Murawski *et al.*, 1994; Thompson and Ritland, 2006). Here we focus on polyploids displaying polysomic inheritance, and introduce a maximum-likelihood method for estimating coefficients of relatedness for co-dominant markers in panmictic populations.

THEORY AND MODELLING
Identity-by-descent and relationship estimation
Most estimators assume that: (i) populations are large (that is, in the limit of infinite) and panmictic; (ii) there is no inbreeding; and (iii) individuals have autosomal loci with Mendelian inheritance. In diploids, the relatedness coefficient (r) can be calculated from two 'higher-order' coefficients:

$$r = \phi/2 + \Delta. \tag{1}$$

Δ is the probability that two individuals share two alleles that are IBD at any given locus, and ϕ is the probability that they both share one allele that is IBD (Lynch and Ritland, 1999). For example, the probability that parents and offspring share an allele that is IBD is 1, so $\phi = 1$ and $\Delta = 0$; the probability that full-sibs share one or two alleles that are IBD is either 0.5 or 0.25, so $\phi = 0.5$ and $\Delta = 0.25$. The ϕ and Δ for specific relationships are listed in Table 1.

Using the same assumptions and assuming no inbreeding or double reduction, in tetraploids, the relatedness coefficient can be expressed as:

$$r = \sum_{i=0}^{4} i\Delta_i/4, \tag{2}$$

Table 1 Relatedness coefficients for specific relationships in diploids and tetraploids in the absence of selfing, inbreeding or double reduction

Relationship	r	Diploids		Tetraploids			
		Δ	ϕ	Δ_4	Δ_3	Δ_2	Δ_1
Self/clone	1	1	0	1	0	0	0
Parent/offspring	1/2	0	1	0	0	1	0
Full-sib	1/2	1/4	1/2	1/36	2/9	1/2	2/9
Half-sib/grandparent	1/4	0	1/2	0	0	1/6	2/3
Nephew	1/4	0	1/2	0	0	2/9	5/9
Great-grandparent	1/8	0	1/4	0	0	1/36	4/9
First cousin/grand-nephew	1/8	0	1/4	0	0	1/27	23/54
Non-relatives	0	0	0	0	0	0	0

where $\sum_{i=0}^{4} \Delta_i = 1$, and Δ_i is the probability that two tetraploids share i alleles that are IBD at any given locus. For relationships between polyploids in outbred populations, r is equivalent to that of diploids. The values of deltas for tetraploid relatives, assuming no double reduction (in polyploids, the phenomenon in which two chromatids of a single chromosome can pass to a same gamete; Mather, 1936), are shown in Table 1.

For inbred populations, Jacquard (1972) described a set of nine identity-by-descent configurations that fully describe the possible IBD relationships between a set of four alleles possessed by two diploids. These are denoted $d_1,\ldots,d_9$ and are shown in Figure 1. The probability that a pair of individuals are in IBD mode d_i is denoted as δ_i. Therefore, the coefficient of coancestry (denoted as θ, an equivalent parameter measuring the probability that two alleles, one randomly drawn from each individual, are IBD; Jacquard, 1972) is:

$$\theta = \delta_1 + \frac{1}{2}(\delta_3 + \delta_5 + \delta_7) + \frac{1}{4}\delta_8. \tag{3}$$

Here, the coefficient of δ_i is the number of IBD allele dyads between two individuals of d_i (Figure 1). In outbred populations, the two alleles in a single individual cannot be IBD, so the first six IBD configurations are not possible and $\delta_i = 0$ ($i = 1,\ldots,6$), reducing Equation (3) to Equation (1), and $r = 2\theta$.

The possible IBD relationships between tetraploids are more complex, with a total of 109 IBD configurations possibly existing between two individuals (see Supplementary Materials). Because this estimate assumes outbreeding, we do not further consider inbreeding. Thus, only five configurations are possible, denoted D_i ($0 \leqslant i \leqslant 4$, where i is the number of IBD alleles shared by two individuals). The IBD configuration for a pair of individuals cannot be obtained from their genotypes, because alleles with the same allelic type may not be IBD. However, alleles identical by state (IBS) can be determined; these are alleles sharing the same allelic type, which include those that are both IBD and non-IBD.

There are 9 and 109 IBS configurations in diploids and tetraploids, respectively. Denoted as $s_1,\ldots,s_9$ and $S_0,\ldots,S_{108}$, their patterns are similar to IBD configurations. For diploids, the lines in Figure 1 represent alleles with the same allelic type. IBS modes for autotetraploids are given in Supplementary Materials.

Under the assumption that two individuals belong to a single population that conforms to the Hardy–Weinberg equilibrium, the probabilities of observing each IBS configuration (S), conditioned on a particular IBD mode (D), can be calculated. The conditional probabilities of five outbred IBD configurations, in which one genotype is $A_iA_iA_iA_i$, are shown in Table 2. Additional conditional probabilities can be generated by an additional programme (see Supplementary Files). The conditional probability is the sum of the products of the probabilities of three sub-genotypes:

$$\Pr(S|D) = \sum \Pr(G_{ab,i})\Pr(G_{a,i})\Pr(G_{b,i}). \tag{4}$$

Where G_{ab} is the IBD sub-genotype shared by two individuals, and G_a and G_b are the additional two non-IBD sub-genotypes of two individuals a and b, respectively. The sub-genotype is a subset of a genotype if a genotype can be defined as a multiple set, because the sub-genotype of a tetraploid can consist of zero to four alleles. $\Pr(G)$ is the probability of choosing G from sub-genotypes with the same number of alleles as for G. For example, $\Pr(A_iA_i) = p_i^2$ and $\Pr(A_iA_iA_j) = 6p_i^2p_j^2$, where p_i and p_j denote the allele frequencies of A_i and A_j, respectively.

Taking $\Pr(A_iA_iA_iA_i, A_iA_iA_iA_j|D_2)$ as an example, there is only one possible G_{ab}, G_a and G_b: $G_{ab} = A_iA_i$, $G_a = A_iA_i$ and $G_b = A_iA_j$. Using the
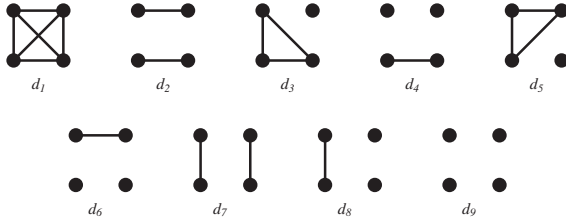
Figure 1 Configurations of identity by descent between two diploids. In each subfigure, the two upper dots represent the two alleles of one individual, whereas the other two represent the alleles of the second individual. The lines indicate alleles that are IBD.

probabilities given in Table 2, the single-locus likelihood of a specific relationship with $\Delta = [\Delta_4,\ldots,\Delta_0]^T$ between two individuals can be calculated. When the IBS mode of those individuals is S, conditioning on the IBD mode yields:

$$L = \Pr(S|\Delta) = \sum \Pr(S|D_i)\Delta_i. \quad (5)$$

Although each locus is characterized by its own set of allele frequencies for multilocus estimation, the degree of relatedness between the two individuals (Δ) is constant across loci because it represents their overall relatedness to each other. Therefore, the multilocus likelihood for unlinked loci is obtained by taking the product of the single-locus likelihoods. The logarithm of likelihoods (L^*) for each loci is computed to simplify the calculations, and their summary is denoted as $\mathcal{L}^*$.

$$\mathcal{L}^* = \sum L_j^*. \quad (6)$$

Parameter space

The maximum-likelihood estimate of Δ is found by searching the parameter space until a maximum is found. In outbred populations, the parameter space of Δ is $\sum \Delta_i = 1$ and $0 \leqslant \Delta_i \leqslant 1$. Another constraint for diploids in outbred populations was given by Thompson (1976): diploid IBD parameters Δ and ϕ are subject to the constraint $4\Delta(1-\Delta-\phi) < \phi^2$, which is applied by Anderson and Weir (2007) but not by Milligan (2003). We considered the situation in which two individuals have fathers who are related and mothers who are also related, but the mother and father of any given individual are unrelated. Under such conditions, p and q were the probabilities that two individuals shared an IBD allele inherited from their fathers and mothers, respectively. As these two events are independent, the diploid IBD parameters Δ and ϕ can be expressed as follows:

$$\Delta = pq$$

and

$$\phi = (1-p)q + (1-q)p,$$

where $0 \leqslant p$ and $q \leqslant 1$. Following the same procedures, we assumed that p_i is the probability that two tetraploids from an outbred population share i IBD alleles inherited from their fathers, and q_i is the probability for i IBD alleles inherited from their mothers. Thus, the tetraploid IBD parameters Δ_i can be expressed as:

$$\Delta_i = \sum_{j=0}^{i} p_j q_{i-j}.$$

Thus, the constraint for Δ can be calculated although complicated to express, and equivalent information is contained within p_i and q_i. Therefore, we used p_i and q_i instead for searching as they are inside the

Table 2 Probability of specific identity-in-state patterns, given the mode of identity by descent if the genotype of one individual is $A_iA_iA_iA_i$

IBS mode	Allelic state	IBD mode				
		D_4	D_3	D_2	D_1	D_0
S_{108}	$A_iA_iA_iA_i,A_iA_iA_iA_i$	p_i^4	p_i^5	p_i^6	p_i^7	p_i^8
S_{107}	$A_iA_iA_iA_i,A_jA_jA_jA_j$	0	0	0	0	$p_i^4 p_j^4$
S_{106}	$A_iA_iA_iA_i,A_iA_iA_iA_j$	0	$p_i^4 p_j$	$2p_i^5 p_j$	$3p_i^6 p_j$	$4p_i^7 p_j$
S_{105}	$A_iA_iA_iA_i,A_iA_jA_jA_j$	0	0	0	$p_i^4 p_j^3$	$4p_i^5 p_j^3$
S_{104}	$A_iA_iA_iA_i,A_jA_kA_kA_k$	0	0	0	0	$4p_i^4 p_j p_k^3$
S_{103}	$A_iA_iA_iA_i,A_iA_iA_jA_j$	0	0	$p_i^4 p_j^2$	$3p_i^5 p_j^2$	$6p_i^6 p_j^2$
S_{102}	$A_iA_iA_iA_i,A_jA_jA_kA_k$	0	0	0	0	$6p_i^4 p_j^2 p_k^2$
S_{101}	$A_iA_iA_iA_i,A_iA_iA_jA_k$	0	0	$2p_i^4 p_j p_k$	$6p_i^5 p_j p_k$	$12p_i^6 p_j p_k$
S_{100}	$A_iA_iA_iA_i,A_jA_jA_jA_k$	0	0	0	$3p_i^4 p_j^2 p_k$	$12p_i^5 p_j^2 p_k$
S_{99}	$A_iA_iA_iA_i,A_jA_jA_kA_l$	0	0	0	0	$12p_i^4 p_j^2 p_k p_l$
S_{98}	$A_iA_iA_iA_i,A_iA_jA_kA_l$	0	0	0	$6p_i^4 p_j p_k p_l$	$24p_i^5 p_j p_k p_l$
S_{97}	$A_iA_iA_iA_i,A_jA_kA_lA_m$	0	0	0	0	$24p_i^4 p_j p_k p_l p_m$

Alleles with the same label are identical by state.

parameter space, making the tetraploid IBD parameters valid. By simulation, we found that Thompson's (1976) constraint can reduce bias of the likelihood estimator.

Genotype ambiguity

A distinct feature of polyploidy population genetics is the formation of partial heterozygotes. Alleles can vary in number from 0 to 4 copies in tetraploids. For example, there are three types of partial heterozygotes (that is, $A_iA_iA_iA_j$, $A_iA_iA_iA_j$ and $A_iA_jA_jA_j$) if two alleles (A_i and A_j) are present in an ambiguous genotype. Although some methods are able to determine tetraploid genotype (that is, Xu et al., 2002; Pfeiffer et al., 2011; Serang et al., 2012; Voorrips et al., 2011; Uitdewilligen et al., 2013), additional instrument or software may be required.

We describe an alternative method to estimate the coefficient of relatedness in scenarios in which heterozygote genotypes are unclear but allele frequencies are known. If two types of alleles, A_i and A_j, are detected in an individual, the probability ratio of the three possible genotypes $A_iA_iA_iA_j$, $A_iA_iA_jA_j$ and $A_iA_jA_jA_j$ is $4p_i^3 p_j : 6p_i^2 p_j^2 : 4p_i p_j^3 = 2p_i^2 : 3p_i p_j : 2p_j^2$. Similarly, if three alleles are detected, A_i, A_j and A_k, the probability ratio of the three genotypes $A_iA_iA_jA_k$, $A_iA_jA_jA_k$ and $A_iA_jA_kA_k$ is $12p_i^2 p_j p_k : 12p_i p_j^2 p_k : 12p_i p_j p_k^2 = p_i : p_j : p_k$. Subsequently, each possible genotype dyad of the two individuals is weighted by its probability, allowing Equation (6) to be modified to:

$$\mathcal{L}^* = \sum_{j,k} P_{j,k} L_{j,k}^*,$$

where $P_{j,k}$ is the probability of kth possible genotype pairs at the jth locus, and $L_{j,k}^*$ is the logarithm of this value. The remaining steps are unchanged, and after the most probable $\hat{\Delta}$ is found, $\hat{r}$ is obtained by Equation (2). In general, an algebraic solution is impossible (Milligan, 2003). As a result, a downhill simplex algorithm is used to search for the $\hat{\Delta}$ that maximizes the likelihood within the parameter space. The simplex consists of $\nu+1$ points (each representing a Δ). If the distance between the points with the minimum and maximum likelihoods is below 0.00001, the algorithm is convergent and the iteration is terminated. A new simplex is then generated by adding a value of the current best point in each dimension, and repeating to prevent the simplex from being trapped in a ridge. An error <0.0001 for

Table 3 The similarity index and coefficients of probability for reference individuals $A_iA_iA_iA_i$

Proband genotype	Similarity index	Coefficient of probability				
		Δ_4	Δ_3	Δ_2	Δ_1	1
$A_iA_iA_iA_i$	1	$1 - p_i^4$	$p_i - p_i^4$	$p_i^2 - p_i^4$	$p_i^3 - p_i^4$	p_i^4
$A_iA_iA_iA_*$	0.75	$-4p_i^3(1 - p_i)$	$(1 - 4p_i^3)(1 - p_i)$	$(2p_i - 4p_i^3)(1 - p_i)$	$(3p_i^2 - 4p_i^3)(1 - p_i)$	$4p_i^3(1 - p_i)$
$A_iA_iA_*A_*$	0.5	$-6p_i^2(1 - p_i)^2$	$-6p_i^2(1 - p_i)^2$	$(1 - 6p_i^2)(1 - p_i)^2$	$(3p_i - 6p_i^2)(1 - p_i)^2$	$6p_i^2(1 - p_i)^2$
$A_iA_*A_*A_*$	0.25	$-4p_i(1 - p_i)^3$	$-4p_i(1 - p_i)^3$	$-4p_i(1 - p_i)^3$	$(1 - 4p_i)(1 - p_i)^3$	$4p_i(1 - p_i)^3$
$A_*A_*A_*A_*$	0	$-(1 - p_i)^4$	$-(1 - p_i)^4$	$-(1 - p_i)^4$	$(1 - p_i)^4$	$(1 - p_i)^4$

tetraploids can be achieved with ~ 600 attempts. Using these methods, this model can be applied to any level of ploidy by replacing the four with v (the level of ploidy) in Equation (2). The conditional probabilities in Table 2 from haploid to octoploid can be generated (see Supplementary Files). However, for species with an odd number of ploidy, Thompson's (1976) constraint cannot be applied.

Polyploid method-of-moment estimator

Huang et al. (2014) developed a method-of-moment estimator for polyploids, which models the probability of each similarity index conditioned on the reference genotype (see also Lynch and Ritland, 1999). The similarity index is defined by the number of alleles that are identical in state between two individuals. However, for this method, unlike for diploid estimators (that is, Wang, 2002; Ritland, 1996), each allele is counted only once. For example, in autotetraploids, the similarity index for each locus has only five values (0, 0.25, 0.5, 0.75 and 1). Table 3 summarizes similarity indices and probabilities of proband genotypes given the allele frequencies and the array of deltas for reference individual $A_iA_iA_iA_i$. By summarizing the expressions with the same similarity index for a reference genotype pattern (Table 3), the following equation is established:

$$\mathbf{P} = \mathbf{E} + \mathbf{M}\Delta,$$

where Δ is a column matrix consisting of all 'higher-order' coefficients from Δ_4 to Δ_1. Each element in $\mathbf{P}$ is the probability of the corresponding similarity index being observed, and $\mathbf{E}$ is the probability that a certain similarity index is observed when relatedness is 0 (the column with the header of 1 in Table 3). $\mathbf{M}$ is the matrix consisting of four columns headed by deltas in Table 3. The moment vector of the similarity index consisting of the first to fourth moments can be expressed as:

$$\mathbf{S} = \mathbf{C}\mathbf{E} + \mathbf{C}\mathbf{M}\Delta.$$

Where $\mathbf{C}$ is a 4×5 matrix with $\mathbf{C}_{ij} = [1 - 0.25(j - 1)]^i$. Equating the observed moments to the expected ($\mathbf{S} = \hat{\mathbf{S}}$) and estimated deltas to the true deltas $\left(\Delta = \hat{\Delta}\right)$ solves the estimator as:

$$\hat{\Delta} = (\mathbf{C}\mathbf{M})^{-1}(\hat{\mathbf{S}} - \mathbf{C}\mathbf{E}). \tag{7}$$

The single-locus $\hat{r}$ can be obtained from Equation (2), whereas in multilocus estimation the locus-specific weight is given by the inverse of the variance of $\hat{r}$. This is calculated numerically by $\text{Var}(X) = E(X^2) - E^2(X)$. The estimate across all loci is the weighted average of each estimate of each locus, with both individuals being used for reference; the final $\hat{r}$ is the arithmetic mean of the two estimates. Huang et al. (2014) also developed a solution to address ambiguous genotypes using this estimator: the matrices $\mathbf{E}$ and $\mathbf{M}$ are weighted by the probability of each reference genotype, and $\mathbf{P}$ is weighted by the probability of each proband–reference genotype pair.

Coefficient of coancestry estimators

Some coefficient of coancestry estimators (θ) developed for diploids can be extended to polyploids (for example, Loiselle et al., 1995; Ritland, 1996). Although θ is alternatively defined as the correlation between the additive values of the two individuals (Ritland, 1996), here we continue with the IBD definition used by other estimators: θ is the probability that a pair of alleles randomly sampled from two individuals at a locus are IBD. In diploid outbred populations, $\theta = 1/4$ for parent–offspring, $\theta = 1/4$ for full-sibs, $\theta = 1/8$ for half-sibs and $\theta = 1/16$ for first-cousins (Jacquard, 1972). The first estimator presented by Ritland (1996) is used as an example and compared with our maximum-likelihood estimator.

Ritland's (1996) estimator assigns a similarity index (S_i) to a genotypic pair for each of n possible alleles. For a diploid, there are four possible values of the ith allele: 0 (one or no individuals contain A_i), 1/4 (both individuals contain a single A_i), 1/2 (one individual contains two and the other individual one A_i) or 1 (both individuals are homozygous for A_i). The single-locus estimator of Ritland (1996) is given by:

$$\hat{\theta} = \frac{\left(\sum_i S_i/p_i\right) - 1}{n - 1}.$$

Hardy and Vekemans (2002) expanded these estimators to higher levels of polyploidy by expanding the definition of the similarity index S_i to a product of the frequency of A_i in the two individuals:

$$\hat{\theta} = \frac{\left(\sum_i S_{ai}S_{bi}/p_i\right) - 1}{n - 1}.$$

Relatedness can be obtained by Equation (12). However, by doing so the estimator becomes biased. To obtain an unbiased estimator, we use the harmonic mean of $\hat{\theta}_{aa}$ and $\hat{\theta}_{bb}$ as the denominator, therefore the single-locus relatedness estimator is:

$$\hat{r} = \frac{1}{2}\hat{\theta}_{ab}\left(\frac{1}{\hat{\theta}_{aa}} + \frac{1}{\hat{\theta}_{bb}}\right). \tag{8}$$

In multilocus estimation, the final estimated relatedness is the weighted average of $\hat{r}$ for each locus, and weight is the inverse of the expected summation of the similarity index across alleles for outbred non-relatives, which is also the allelic richness of this locus.

Using Equation (8), the relatedness coefficient can also be estimated using the method of Loiselle et al. (1995). When the allele frequency within an individual is equal to the population allele frequency (for example, a heterozygote at an uniform biallelic locus), $\hat{\theta}_{aa}$ will be 0. To avoid obtaining an undefined final estimated relatedness, the $\hat{r}$ at such loci is not taken into account for calculating the weighted average of $\hat{r}$.

These coefficients of coancestry estimators consider situations in which inbreeding occurs, but do not incorporate 'higher-order' coefficients ($\hat{\Delta}$) used in this paper's estimator. To handle ambiguous

genotypes, the similarity indices are weighted by the probability of each possible genotype pair.

Calculating relatedness from pedigrees

The coefficient of coancestry can be calculated from pedigree data by a recursive algorithm (Karigl, 1981). For polyploids, the coefficient of coancestry between two individuals is the same as in diploids:

$$\theta_{ab} = \frac{1}{2}(\theta_{fb} + \theta_{mb}). \tag{9}$$

Here, a, b, f and m are individuals, and f and m are the father and mother of a, respectively, where the probability that an allele in f or m is inherited by a is 1/2. It can be inferred that a cannot be an ancestor of b because Equation (9) becomes divergent. If b is an ancestor of a, then another algorithm is needed to calculate the coefficient of coancestry of an individual with itself. This can be achieved using the equation:

$$\theta_{aa} = \frac{1}{2(\nu-1)} + \frac{\nu-2}{4(\nu-1)}(\theta_{ff} + \theta_{mm}) + \frac{1}{2}\theta_{fm}. \tag{10}$$

where ν is the level of ploidy. Clearly, $\theta_{ab} = \theta_{ba}$ and $\theta_{ab} = 0$ if a and b are not related, such that the coefficient of coancestry can be calculated by iteration or recursion for any situation once the ancestry of a and b is known.

In polyploids, multivalent formation can result in double reduction, which occurs when sister chromatids segregate into the same chromosome (Darlington, 1929). Under pure random segregation, the rate of double reduction, α (Fisher and Mather, 1943), assumes a minimum value of 0, and increases to 1/7 (with pure random chromatid segregation) and 1/6 (with complete equational segregation) for tetrasomic inheritance (Muller, 1914).

If double reduction is considered, the Equation (10) to calculate the coefficient of coancestry within the same individual should be modified. For octosomic or decasomic inheritance, there are three rather than two types of origins for the gamete, so an additional parameter is needed for the segregation ratios (Fisher and Mather, 1943). Here, α_i is the probability that double reduction occurred i time(s) in a gamete ($\sum \alpha_i = 1$, $0 \leqslant i \leqslant [\nu/4]$), and θ_{aa} is given by:

$$\theta_{aa} = \frac{1}{2} + \frac{1}{2}\theta_{fm} + \frac{\binom{\nu/2}{2} - \sum i\alpha_i}{\binom{\nu}{2}}(\theta_{ff} + \theta_{mm} - 2). \tag{11}$$

The derivation of Equations (10) and (11) can be found in Supplementary Files. Equation (11) can be inferred from the coefficient of coancestry within the same individual from an outbred population at equilibrium (the genotypic frequencies are equal among generations). This is written as:

$$\theta_{aa} = \frac{8\sum i\alpha_i + \nu}{8\sum i\alpha_i + \nu^2}.$$

Where $\theta_{aa} = 0.3$ for tetraploids under purely random chromatid segregation. In the presence of double reduction, Equation (9) remains unchanged. In tetraploids, for example, although the probability an allele in f or m is inherited by a is 1/4 under double reduction, the number of IBD allele pairs between a and b is also doubled.

In the absence of double reduction, inbreeding or selfing, the relatedness coefficient can be calculated from $r = 4\theta$. However, it cannot be applied because r may exceed 1. By Equations (5) and (6) presented by Hardy and Vekemans (1999), Wright's coefficient of relationship can be calculated from the coefficient of coancestry as:

$$r_{ab} = \frac{\theta_{ab}}{\sqrt{\theta_{aa}\theta_{bb}}}. \tag{12}$$

Which is used as the true relatedness for simulation; the derivation of Equation (12) is given in Supplementary Files.

Statistical behaviour of estimators of relatedness

To investigate the statistical behaviour of estimators, six estimators were compared. These included the maximum-likelihood estimator (ML) described in this paper, the method of moment for polyploids (MOM; Huang et al., 2014) and a coefficient of coancestry estimator (RI; Ritland, 1996), applied when genotypes are known and when genotypes are ambiguous, the latter of which are denoted ML*, MOM* and RI*. All estimators were truncated to the range of 0–1 for ease of comparison.

Four comparisons considering various conditions were performed: (i) the distributions of $\hat{r}$ in a particular relationship; (ii) the performance of multilocus estimations in ideal conditions; (iii) the minimal number of loci needed to achieve specific requirements; and (iv) the robustness of these estimators in a finite population with strong genetic drift, inbreeding and double reduction.

Distribution

For this application, four relationships (parent–offspring, full-sibs, half-sibs and unrelated) were simulated. For each pair of individuals, the genotype of one individual was randomly generated according to the Hardy–Weinberg equilibrium. The other genotype was then obtained from the randomly generated genotype as a reference and
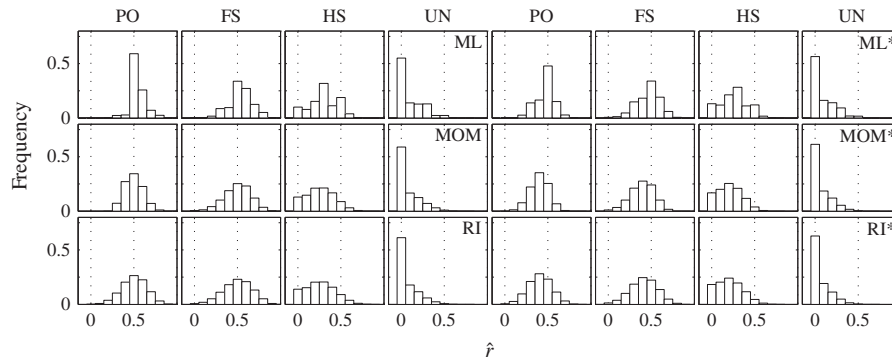

Figure 2 Distribution of $\hat{r}$ estimates between autotetraploids using six different methods (ML, MOM and RI for exact genotypes, ML*, MOM* and RI* for ambiguous genotypes) for four relationships (PO for parent–offspring, FS for full-sibs, HS for half-sibs and UN for unrelated). Each distribution was based on a sample of 200 000 estimates taken from five loci, each segregating for eight alleles with their frequencies drawn from the triangular distribution.

their relationship ($\hat{\Lambda}$). A triangular allele frequency distribution was simulated, where this followed the proportions 1, 2,...,n. Numerical results were obtained from Monte Carlo simulations assuming true allele frequencies were available. The results are given in Figure 2 using five loci, each segregating for eight alleles because eight was the minimum number needed to display all 109 IBS configurations. Their distribution under other settings can be found in Supplementary Materials.

Figure 2 shows that the likelihood estimators exhibited less variance for kin dyads. However, for non-relatives, the RI and MOM estimators gave a higher frequency of 0 estimates. Furthermore, the ML* estimator converged near the true value even when the correct genotype was unavailable. In contrast, the MOM* and RI* estimators both showed negative bias in relatives.

Root mean square error

This application used multiple multiallelic loci to estimate relatedness for specific relationships under ideal conditions: population size was sufficiently large, true allele frequency was available, mating was random, inbreeding was absent and chromosome segregation was purely random.

Because both truncated and maximum-likelihood estimators have a bias, the root mean square error (RMSE), which incorporates bias and sampling variance, was used to measure overall accuracy. RMSE can be calculated by the following equation: $\mathrm{RMSE}(\hat{r}) = \sqrt{\mathrm{Bias}^2(\hat{r}) + \mathrm{Var}(\hat{r})}$.

In this section, the number of loci (l) was simulated from 1 to 100, with each locus segregating five or ten alleles with their frequencies drawn from triangular distribution. The results for each relationship type were calculated for 15 000–100 000 pairs of individuals. The number of dyads simulated was dependent on l; with smaller values of l, more simulations were performed to smoothen the curves. The RMSE of $\hat{r}$ is shown in Figure 3.

Figure 3 shows that different estimators behaved differently statistically. The ML estimator yielded the largest RMSE for non-relatives and half-sibs. Although the RI estimator exhibited a large RMSE for related dyads, the RMSE of these relationships was relatively small for the MOM and ML estimators.

If the correct heterozygous genotypes were unavailable, all the performance of all estimators was reduced. However, when the number of alleles was high, the performance of both the ML* and MOM* estimators improved. For both the RI and RI* estimators, an increase in alleles made little difference to estimator performance.

Confidence intervals

We also evaluated the requirements to distinguish particular relationships. We calculated the minimal number of loci needed to obtain a 95% confidence interval of ± 0.05 units of r (the probability that $\hat{r} \in [r - 0.05, r + 0.05]$ is 0.95), with ploidy levels ranging from diploid to octoploid. Bias was high when ambiguous genotypes were used for the estimations, and $\hat{r}$ did not converge to the true value (Figure 2). As a result, we did not consider ambiguous genotypes for this application. Moreover, because the RMSE of half-sibs was usually the highest among the relationships tested (Figure 3), we only calculated half-sibs for simplicity. Results are shown in Figure 4.

The ML estimator required fewest loci to achieve a high degree of accuracy, nearly 90 and 85% of what was required for the MOM and RI estimators, respectively. Because polyploids have more copies of alleles and genetic information than diploids, the minimal number of loci required for higher levels of ploidy was fewer than that for multiallelic loci. By contrast, in biallelic loci or loci with few alleles, the probability that two non-relatives share IBS alleles was higher for polyploids. Therefore, at higher levels of ploidy, more loci were needed if the loci possessed few alleles, especially for biallelic loci.

Finite populations

Although these estimators performed reasonably well under the given assumptions, real cases often diverge from ideal conditions. To simulate nature, we simulated a finite population with strong genetic
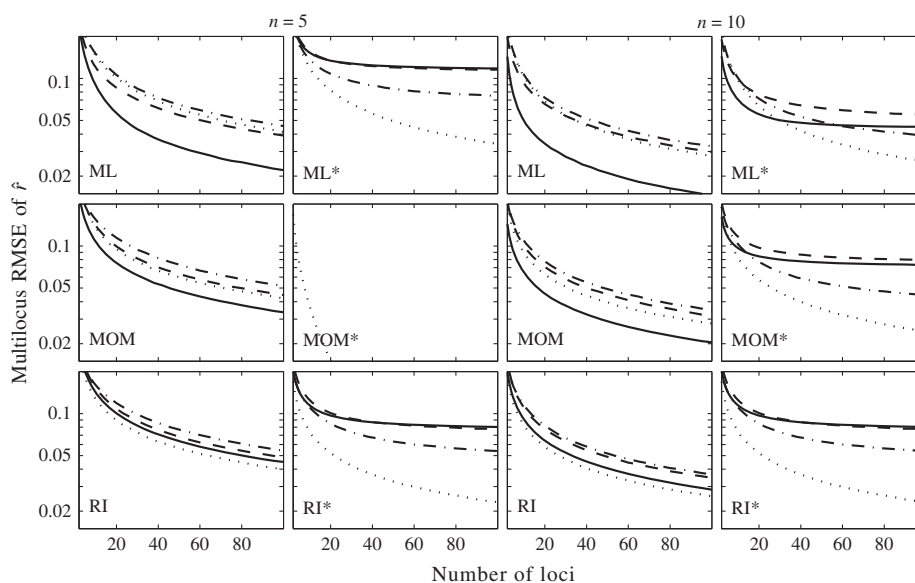

Figure 3 Multilocus RMSE of $\hat{r}$ between autotetraploids as a function of the number of alleles under a triangular allele frequency distribution. Six estimators were compared, including the polyploid maximum-likelihood estimator (ML, first row), the truncated polyploid method-of-moment estimator (MOM, second row), the truncated Ritland (1996) estimator (RI, third row) and their respective modified versions that support ambiguous genotypes (denoted by an asterisk after the estimator abbreviation). Two kinds of loci were simulated: (1) penta-allelic ($n=5$, two leftmost columns) and (2) deca-allelic loci ($n=10$, two rightmost columns). Results were obtained by generating 15 000–100 000 pairs of four relationships including parent–offspring ('—'), full-sibs ('– –'), half-sibs ('– .') and unrelated ('...') using Monte Carlo simulations.
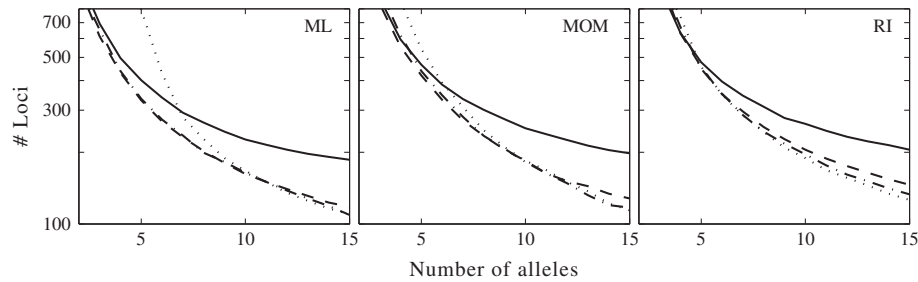
Figure 4 The minimal number of loci required to obtain a 95% confidence interval ± 0.05 units of *r* in half-sibs showed in Figure 3. Three estimators were compared in diploids ('—'), tetraploids ('– –'), hexaploids ('– .') and octoploids ('…'). Results were obtained by the split-half method and 30 000 Monte Carlo simulations per attempt.

drift, inbreeding and selfing. Following Toro *et al.* (2011), the generations originated from 20 founder individuals. The genotypes of the founder individuals were randomly generated according to the Hardy–Weinberg equilibrium.

Ten discrete generations each consisting of 20 individuals were simulated. The parents of an individual were randomly selected from the last generation (some individuals did not reproduce) resulting in a data set of 200 individuals and 20 100 dyads. The true relatedness coefficients were computed from the pedigree using Equation (12), and the estimate of relatedness was obtained using the six estimators (Figure 3). To compute the estimators, we used the observed allele frequencies calculated from the genotypes of the 200 individuals. Because we do not develop a method for estimating allele frequencies from ambiguous genotypes in this paper, the observed allele frequencies were obtained from the true genotypes.

Four kinds of loci were simulated, including two with too few alleles to generate the 109 IBS modes (2 and 5 alleles) and two with enough alleles (10 and 15 alleles). For each simulation, 20 unlinked loci were used with their initial allele frequencies drawn from a triangular distribution. The mean true relatedness ($\bar{r}$) of the simulated populations varied between runs due to random mating. In particular, inbreeding in the first few generations may lead to an increase in ($\bar{r}$). Therefore, for each locus type, the simulation was repeated 40 times to ensure that ($\bar{r}$) for different numbers of alleles was the same. Monomorphic loci were not used because the estimators failed to give a valid estimate. A linear equation $\hat{r} = \beta_1 r + \beta_0$ was used for regression analysis, the weighted least-squares solution was obtained and the coefficient of determination (R^2) and RMSE calculated (Table 4; Figure 5). Additional results from varying numbers of loci and alleles are presented in Supplementary Materials.

We also investigated the effect of double reduction on these estimators. The double reduction rate was assumed to be 1/7 in these simulations (Muller, 1914). To obtain a distribution of genotypes in equilibrium, each individual in the founder population was produced from eight temporary generations of non-relatives, in which individuals in the first temporary generation had a $\theta_{aa} = 0.25$. The true relatedness coefficient was obtained using Equations (9) and (11). Therefore, in the founder population, the coefficient of coancestry for the same individual was 0.299998≈0.3, whereas that between different individuals was 0. The other parameters were the same as in the previous application, and results are shown in Table 4 with $\alpha = 1/7$ in the first column (Figure 5).

The MOM* estimator encountered the singular matrix problem at $n = 5$, and the RI estimator performed worse for biallelic loci, whereas the ML estimator was more stable (Table 4). Although the R^2 of the RI estimator increased as n increased, the RMSE did not change significantly, unlike in the other two estimators. Due to small

population size and strong drift, some alleles were lost in the last few generations, reducing the performance of the estimators compared with results shown in Figure 3, even for a higher initial number of alleles ($n = 15$). Moreover, the range of estimates for all estimators was restricted to (0, 1), so the slopes deviated from 1.

Double reduction did not thus affect the distribution of true relatedness. However, the performance of all estimators was slightly reduced.

In Figure 5, most values of *r* lie in the range (0, 0.5). The points of parent–offspring pairs formed a vertical line at $r = 0.5$ in the MOM and RI estimators. Because these estimators cannot give an accurate estimate for *r*, the longer length of these lines at $r = 0.5$ suggests a larger RMSE. In contrast, a similar line was absent in the ML estimator because the variance of the estimates was too small, resulting in overlapping points. However, for half-sibs and grandparent–offspring pairs ($r = 0.25$), other types of outbred relationships ($r = 0.125$ or 0.0625) or ambiguous genotypes, sampling variance increased, so vertical lines are present. The line of the ML estimator was usually the shortest, suggesting a lower RMSE.

Similar to estimates of 0 or 1, there are two estimates ($\Delta_2 = 1$ or $\Delta_1 = 1$) that also lie on the edge of parameter space. If two individuals share only one or two IBS alleles at all loci, these parameters produce the largest likelihood and give an estimate of $\hat{r} = 0.5$ or 0.25, respectively. As a result, there were two additional horizontal lines in the ML estimators.

The results for the original estimators (without truncation) can be found in Supplementary Materials. Truncation can be expected to cause a reduction in slope and RMSE and an increase in R^2. Nonetheless, the ML estimator still had better statistical values, with the exception of the slope, as the slope of the original MOM estimator was closer to 1.

DISCUSSION

Statistical behaviour

We developed a maximum-likelihood method for estimating the relatedness coefficient for polyploids. The probability of observing an IBS mode conditioned on each IBD mode was calculated by following existing procedures (Thompson, 1975). A numerical algorithm was subsequently applied to find the optimal solution for $\hat{r}$, and the statistical behaviours of various estimators of relatedness for autotetraploids were simulated and compared. Marker-based relatedness estimates typically showed large sampling variance, due to variance in identity by descent among loci and in identity-by-state alleles that are not IBD (Lynch and Ritland, 1999). The RMSE and variance were reduced by increasing the number of loci or by switching to loci that were more polymorphic. Overall, likelihood estimators exhibited lower RMSE than other estimators we examined

Table 4 Statistics of $\hat{r}$ for a finite population of autotetraploids

α	n	$\hat{r}$	Var($\hat{r}$)	ML				MOM				RI			
				β_1	β_0	R^2	RMSE	β_1	β_0	R^2	RMSE	β_1	β_0	R^2	RMSE
0	2	0.162	0.023	0.662	0.024	0.283	0.170	0.687	0.028	0.255	0.185	0.415	0.288	0.021	0.475
	5	0.161	0.023	0.739	−0.037	0.558	0.132	0.727	−0.040	0.554	0.135	0.733	−0.042	0.563	0.135
	10	0.164	0.023	0.749	−0.055	0.646	0.133	0.740	−0.057	0.657	0.134	0.733	−0.059	0.653	0.137
	15	0.161	0.023	0.766	−0.061	0.686	0.131	0.760	−0.062	0.702	0.131	0.743	−0.064	0.685	0.135
	Mean	0.162	0.023	0.729	−0.032	0.543	0.141	0.728	−0.033	0.542	0.146	0.656	0.031	0.481	0.221
0*	2	0.161	0.023	0.331	0.015	0.212	0.167	0.572	0.026	0.212	0.183	0.212	0.011	0.196	0.178
	5	0.162	0.023	0.552	−0.026	0.509	0.144	0.229	−0.024	0.376	0.194	0.521	−0.028	0.523	0.148
	10	0.162	0.023	0.651	−0.045	0.629	0.137	0.582	−0.042	0.631	0.143	0.626	−0.049	0.638	0.142
	15	0.162	0.023	0.688	−0.054	0.672	0.135	0.656	−0.052	0.692	0.136	0.658	−0.056	0.672	0.141
	Mean	0.162	0.023	0.556	−0.028	0.505	0.146	0.510	−0.023	0.478	0.164	0.504	−0.031	0.507	0.152
1/7	2	0.165	0.023	0.649	0.024	0.279	0.171	0.683	0.028	0.258	0.185	0.292	0.296	0.011	0.470
	5	0.162	0.023	0.730	−0.037	0.552	0.134	0.716	−0.039	0.544	0.137	0.729	−0.042	0.559	0.136
	10	0.163	0.023	0.754	−0.055	0.649	0.132	0.745	−0.057	0.661	0.133	0.735	−0.059	0.655	0.136
	15	0.163	0.023	0.758	−0.060	0.681	0.132	0.753	−0.062	0.698	0.132	0.735	−0.063	0.680	0.137
	Mean	0.163	0.023	0.723	−0.032	0.540	0.142	0.724	−0.033	0.540	0.147	0.623	0.033	0.476	0.220
1/7*	2	0.165	0.023	0.320	0.015	0.204	0.172	0.561	0.026	0.209	0.186	0.202	0.013	0.183	0.182
	5	0.164	0.023	0.540	−0.025	0.499	0.147	0.220	−0.024	0.366	0.197	0.512	−0.028	0.512	0.151
	10	0.162	0.023	0.643	−0.044	0.624	0.138	0.575	−0.041	0.627	0.144	0.618	−0.048	0.632	0.143
	15	0.163	0.023	0.690	−0.055	0.672	0.136	0.660	−0.052	0.693	0.136	0.660	−0.057	0.673	0.141
	Mean	0.163	0.023	0.548	−0.027	0.500	0.148	0.504	−0.023	0.474	0.166	0.498	−0.030	0.500	0.154

Twenty unlinked loci each segregating n alleles were simulated and results displayed. The first column shows the rate of double reduction, and ambiguous genotypes are denoted by an asterisk. The slope (β_1), intercept (β_0), coefficient of determination (R^2) and RMSE were calculated for each estimator. Statistics were obtained from 804 000 dyads by the weighted least-squares method.
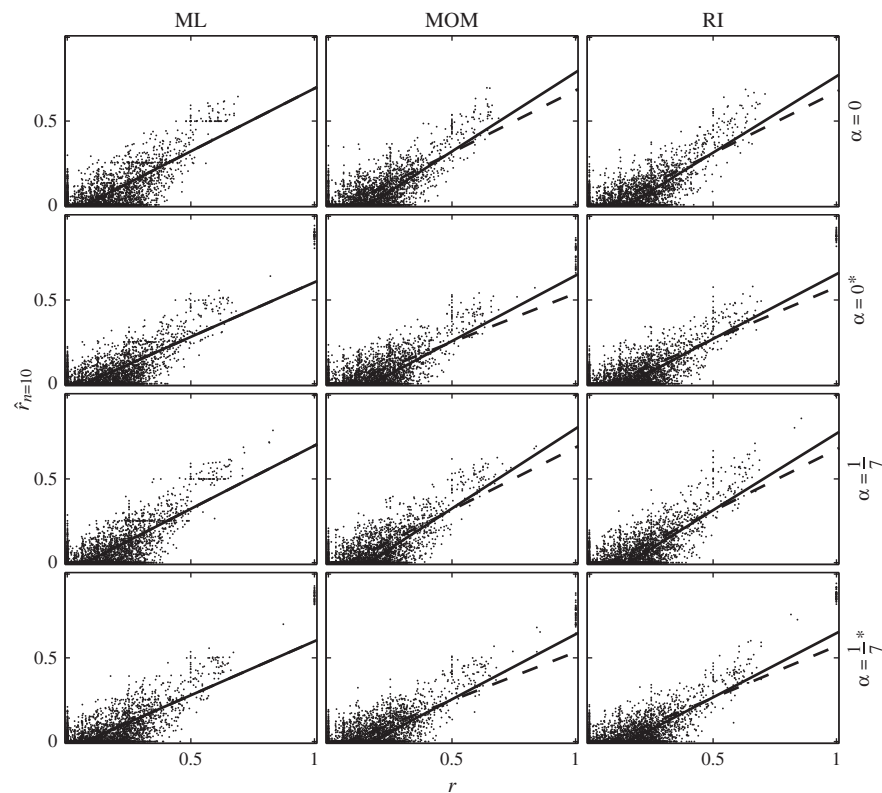

Figure 5 Marker-based estimated relatedness ($\hat{r}$) as a function of pedigree-based true relatedness (r) in a finite population. Twenty unlinked loci were used in the estimation, each initially segregating with 10 alleles under triangular distribution. Each figure shows 4000 points randomly selected from 402 000 dyads. The top two rows show results for a population without double reduction, the bottom two rows show a population with double reduction. The second and fourth rows include genotype ambiguity. The trend lines for truncated ('—') and original ('– –') estimators were obtained by weighted least-squares regression.

for relatives, whereas the ML estimator produced a higher RMSE for non-relatives (Figures 2 and 3). The RMSE of estimators decreased rapidly as a function of l in multilocus estimations (Figure 3).

The RMSE of the likelihood estimator for unrelated dyads was highest because the estimator also generated IBS configurations that shared some IBS alleles by chance. This resulted in a positive estimate of r (Figure 2). Because each locus had eight total alleles that could be sampled between individuals, the probability that unrelated tetraploid dyads shared IBS alleles was higher than that for diploids, especially for biallelic loci. As a result, all estimators performed worse for biallelic loci (Figure 3; Table 4). For this reason, for hexaploids and octoploids, more alleles were required.

The MOM estimator may encounter a singular matrix problem when the number of alleles is too small (Huang et al., 2014). Singularity or near singularity can result from zero or near-zero coefficients in the set of equations, making Equation (7) unsolvable. This problem can also occur for diploid estimators. For example, for biallelic loci with uniformly distributed allele frequency, the Lynch and Ritland (1999) estimator has a sampling variance of infinity. Unfortunately, this scenario is more frequent for polyploids, but the probability of a singular matrix is reduced when the number of alleles is greater or equal to the level of ploidy (Huang et al., 2014). However, this does not guarantee that singularity is avoided, because some combinations of allele frequencies can also result in a singular matrix.

Ambiguous genotypes

When genotyping polyploid heterozygotes, balanced heterozygotes cannot be distinguished from unbalanced heterozygotes. In this case, each candidate genotype pair is weighted by its probability. This situation also brings a negative bias to both the likelihood and moment estimators. The true genotype pair of a kin dyad is diluted by other candidate genotype pairs that are usually less similar. For example, in a pair of tetraploid clonemates, both genotypes are $A_iA_iA_iA_j$. Therefore, each has three possible genotypes ($A_iA_iA_iA_j$, $A_iA_iA_jA_j$ and $A_iA_jA_jA_j$) and there are nine combinations of genotype pairs. Three pairs give accurate estimates. Estimates of the other six pairs are <1, and the final $\hat{r}$ is a maximum-likelihood solution (ML) or a weighted average (MOM or RI). Under such conditions, all estimators are less efficient, but in particular, the RI estimators are unusable if the number of alleles is too few. The ML estimator performs better than other estimators under most conditions (Table 4).

Inbreeding, selfing and double reduction

The new likelihood estimator assumes no inbreeding or double reduction, so that the number of non-zero 'higher-order' coefficients is equal to the level of ploidy and the range of $\hat{r}$ is (0, 1). Because the probabilities of inbreeding, selfing or double reduction IBD configurations are not modelled, some estimates may be inaccurate. For example, the genotype patterns $A_iA_jA_kA_l$ and $A_iA_iA_iA_i$ do not have a higher estimate than $A_iA_jA_kA_l$ and $A_iA_mA_nA_o$. If inbreeding, selfing or double reduction occurs, the former genotypes are more similar and should be assigned a larger $\hat{r}$. Therefore, underestimation occurs for the likelihood estimator (that is, the slope deviates from 1 in Table 4).

There are nine IBD configurations for diploids. By contrast, IBD/IBS models for autotetraploids are more complex, with a total of 109 distinct configurations. The number of IBD configurations increases from haploid to octoploid, with 2, 9, 31, 109, 339, 1043, 2998 and 8405 possibilities, respectively. Because there are too many deltas in polyploids, it is impossible to solve for each delta. As a result, IBD configurations that involve double reduction and inbreeding were omitted from this model. However, a finite population including both

inbreeding and double reduction was simulated, and regression analyses were performed to evaluate the statistics. Although all estimators were less efficient than under ideal conditions (a larger RMSE and greater sensitivity to the initial number of alleles because of the strong role of drift), the likelihood estimator showed greater robustness in simulation and was superior to the other estimators for multiallelic loci regardless of double reduction (Table 4; Figure 5). The MOM estimator performed well when there were multiallelic loci, and the original MOM estimator (without truncation) had the largest slope (>0.8; see Supplementary Material).

Biallelic markers

The likelihood estimator can be applied to a wide range of data, including microsatellites, single nucleotide polymorphisms and other co-dominant markers. However, because fewer alleles reduce accuracy, we suggest that only loci with many alleles be used, particularly if an application requires a high level of accuracy. Loci containing few alleles (for example, biallelic loci) can only achieve high levels of reliability with many. However, the number of unlinked loci is limited in the genome and with many loci, there is increased risk that adjacent loci will be linked and thus not represent independent data points. Although linked loci do not introduce bias, they do not increase reliability. This causes the RMSE to reach an asymptote as the number of loci increases. Furthermore, where genotypes are ambiguous, biallelic loci are also problematic as the bias is too large (Figure 5). However, single nucleotide polymorphism data can still be used, especially with newer genotyping-by-sequencing and haplotype prediction technologies (Xu et al., 2002; Uitdewilligen et al., 2013). These techniques provide means for unambiguous genotyping and also determine the haplotype. The haplotype of adjacent single nucleotide polymorphisms can be treated as an allele of a multiallelic locus, which can largely improve the reliability of estimation.

Properties of polyploids

Alleles in polyploids have more copies, so contain more information than alleles in diploids (Huang et al., 2014). Therefore, for multiallelic loci, fewer loci are required to achieve the same reliability for polyploids (Figure 4). Nevertheless, it is noteworthy that the chance of two non-relatives or less-related individuals sharing IBS alleles is higher under the same conditions, which can interfere with estimation and result in a positive bias. For example, more loci are required for higher levels of ploidy when alleles are few (Figure 4).

In extreme allotetraploids, there are two homologous sets each consisting of two homologous chromosomes. If a chromosome exclusively pairs with its homologue, this leads to disomic inheritance (Stift et al., 2008). For these cases, we can use diploid estimators. Some empirical studies show that many polyploids are actually in the intermediate inheritance (Allendorf and Danzmann, 1997; Jannoo et al., 2004). Intermediate inheritance may be expected in fertile interspecific hybrids, as their parents are usually related and therefore are expected to possess some degree of chromosomal homology (Jannoo et al., 2004), leading to complex mixtures of disomic and polysomic inheritance. Gametic and genotypic frequencies also deviate from expectation, resulting in additional positive bias for the ML and MOM estimators (see Supplementary Material), whereas the RI estimator is not affected.

CONCLUSIONS

Overall, the maximum-likelihood estimator we developed provides several advantages over existing methods. First, it generally exhibits lower RMSE compared with other estimators. Second, all estimates fall

within a biologically meaningful range, and 'higher-order' coefficients can be explained as probabilities. Thus, the biological interpretation of individual estimates is straightforward. Third, it provides a solution for situations in which the allele dosage cannot be determined.

Although the maximum-likelihood estimator of relatedness performed well in simulations, there are conditions under which other estimators performed better, according to specific metrics. There is no single estimator with superior performance under all conditions and by all metrics. For specific applications under specific research conditions, it is possible to identify one optimal estimator. The software package POLYRELATEDNESS provides a simulation function that helps researchers evaluate the performance of each estimator under their given conditions.

DATA ARCHIVING

The software POLYRELATEDNESS V1.4 (Huang K, Northwest University, Xi'an, China), user manual and example data set are available on Google Project (http://polyrelatedness.googlecode.com).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

Allendorf FW, Danzmann RG (1997). Secondary tetrasomic segregation of mdh-b and preferential pairing of homeologues in rainbow trout. *Genetics* **145**: 1083–1092.
Anderson AD, Weir BS (2007). A maximum-likelihood method for the estimation of pairwise relatedness in structured populations. *Genetics* **176**: 421–440.
Burow MD, Simpson CE, Starr JL, Paterson AH (2001). Transmission genetics of chromatin from a synthetic amphidiploid to cultivated peanut (*Arachis hypogaea* L.): broadening the gene pool of a monophyletic polyploid species. *Genetics* **159**: 823–837.
Charpentier MJE, Fontaine MC, Cherel E, Renoult JP, Jenkins T, Benoit L *et al.* (2012). Genetic structure in a dynamic baboon hybrid zone corroborates behavioural observations in a hybrid population. *Mol Ecol* **21**: 715–731.
Darlington CD (1929). Chromosome behaviour and structural hybridity in the tradescantiae. *J Genet* **21**: 207–286.
Fisher RA, Mather K (1943). The inheritance of style length in *Lythrum salicaria*. *Ann Eugen* **12**: 1–23.
Hardy OJ, Vekemans X (1999). Isolation by distance in a continuous population: reconciliation between spatial autocorrelation analysis and population genetics models. *Heredity* **83**: 145–154.
Hardy OJ, Vekemans X (2002). SPAGEDI: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Mol Ecol Notes* **2**: 618–620.
Huang K, Ritland K, Guo ST, Shattuckn M, Li BG (2014). A pairwise relatedness estimator for polyploids. *Mol Ecol Resour* **14**: 734–744.

Jacquard A (1972). Genetic information given by a relative. *Biometrics* **28**: 1101–1114.
Jannoo N, Grivet L, David J, D'Hont A, Glaszmann JC (2004). Differential chromosome pairing affinities at meiosis in polyploid sugarcane revealed by molecular markers. *Heredity* **93**: 460–467.
Karigl G (1981). A recursive algorithm for the calculation of identity coefficients. *Ann Hum Genet* **45**: 299–305.
Li CC, Weeks DE, Chakravarti A (1993). Similarity of DNA fingerprints due to chance and relatedness. *Hum Hered* **43**: 45–52.
Liu ZJ, Huang CM, Zhou QH, Li YB, Wang YF *et al.* (2013). Genetic analysis of group composition and relatedness in white-headed langurs. *Integr Zool* **8**: 410–416.
Loiselle BA, Sork VL, Nason J, Graham C (1995). Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). *Am J Bot* **82**: 1420–1425.
López-Pujol J, Bosch M, Simon J, Blanche C (2004). Allozyme diversity in the tetraploid endemic *Thymus loscosii* (Lamiaceae). *Ann Bot* **93**: 323–332.
Luo ZW, Zhang ZE, Zhang RM, Pandey M, Gailing O, Hattemer HH *et al.* (2006). Modeling population genetic data in autotetraploid species. *Genetics* **172**: 639–646.
Lynch M, Ritland K (1999). Estimation of pairwise relatedness with molecular markers. *Genetics* **152**: 1753–1766.
Mather K (1936). Segregation and linkage in autotetraploids. *J Genet* **32**: 287–314.
Mattila ALK, Duplouy A, Kirjokangas M, Lehtonen R, Rastas P, Hanski I (2012). High genetic load in anold isolated butterfly population. *Proc Natl Acad Sci USA* **109**: E2496–E2505.
Milligan BG (2003). Maximum-likelihood estimation of relatedness. *Genetics* **163**: 1153–1167.
Muller HJ (1914). A new mode of segregation in gregory's tetraploid primulas. *Am Nat* **48**: 508–512.
Murawski DA, Fleming TH, Ritland K, Hamrick JL (1994). The mating system of an autotetraploid cactus, Pachycereus pringlei. *Heredity* **72**: 86–94.
Otto SP (2007). The evolutionary consequences of polyploidy. *Cell* **131**: 452–462.
Pfeiffer T, Roschanski AM, Pannell JR, Korbecka G, Schnittler M (2011). Characterization of microsatellite loci and reliable genotyping in a polyploid plant, *Mercurialis perennis* (Euphorbiaceae). *J Hered* **102**: 479–488.
Queller DC, Goodnight KF (1989). Estimating relatedness using genetic markers. *Evolution* **43**: 258–275.
Ritland K (1996). Estimators for pairwise relatedness and individual inbreeding coefficients. *Genet Res* **67**: 175–185.
Ritland K, Ganders FR (1985). Variation in the mating system of *Bidens menziesii* (Asteraceae) in relation to population substructure. *Heredity* **55**: 235–244.
Serang O, Mollinari M, Garcia AAF (2012). Efficient exact maximum a posteriori computation forBayesian SNP genotyping in polyploids. *PLoS ONE* **7**: e30906.
Stift M, Berenos C, Kuperus P, van Tienderen PH (2008). Segregation models for disomic, tetrasomic and intermediate inheritance in tetraploids: a general procedure applied to *Rorippa* (yellow cress) microsatellite data. *Genetics* **179**: 2113–2123.
Thomas SC (2010). A simplified estimator of two and four gene relationship coefficients. *Mol Ecol Resour* **10**: 986–994.
Thompson EA (1975). The estimation of pairwise relationships. *Ann Hum Genet* **39**: 173–188.
Thompson EA (1976). A restriction on the space of genetic relationships. *Ann Hum Genet* **40**: 201–204.
Thompson S, Ritland K (2006). A novel mating system analysis for modes of self-oriented mating applied to diploid and polyploid arctic Easter daisies (*Townsendia hookeri*). *Heredity* **97**: 119–126.
Toro MÁ, García-Cortés LA, Legarra A (2011). A note on the rationale for estimating genealogical coancestry from molecular markers. *Genet Sel Evol* **43**: 27.
Uitdewilligen JGAML, Wolters AA, D'hoop BB, Borm TJA, Visser RGF, van Eck HJ (2013). A nextgeneration sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. *PLoS ONE* **8**: e62355.
Voorrips R, Gort G, Vosman B (2011). Genotype calling in tetraploid species from bi-allelic marker data using mixture models. *BMC Bioinformatics* **12**: 172.
Wang JL (2002). An estimator for pairwise relatedness using molecular markers. *Genetics* **160**: 1203–1215.
Wang JL (2011). Unbiased relatedness estimation in structured populations. *Genetics* **187**: 887–901.
Wright S (1921). Systems of mating. I. the biometric relations between parent and offspring. *Genetics* **6**: 111.
Xu CF, Lewis K, Cantone KL, Khan P, Donnelly C, White N *et al.* (2002). Effectiveness of computational methods in haplotype prediction. *Hum Genet* **110**: 148–156.

Supplementary Information accompanies this paper on Heredity website (http://www.nature.com/hdy)