MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2014) **14**, 734–744 doi: 10.1111/1755-0998.12217

A pairwise relatedness estimator for polyploids

KANG HUANG,* KERMIT RITLAND,† SONGTAO GUO,* MILENA SHATTUCK‡ and BAOGUO LI*
Key Laboratory of Resource Biology and Biotechnology in Western China of Ministry of Education, and College of Life Sciences, Northwest University, Xi'an, ShaanXi 710069, China, †Department of Forest and Conservation Sciences, University of British Columbia, Vancouver, BC V6T1Z4, Canada, ‡Anthropology Department, University of Illinois at Urbana-Champaign, Urbana-Champaign, IL 61801, USA

Abstract

Studies in genetics and ecology often require estimates of relatedness coefficients based on genetic marker data. Many diploid estimators have been developed using either method-of-moments or maximum-likelihood estimates. However, there are no relatedness estimators for polyploids. The development of a moment estimator for polyploids with polysomic inheritance, which simultaneously incorporates the two-gene relatedness coefficient and various 'higher-order' coefficients, is described here. The performance of the estimator is compared to other estimators under a variety of conditions. When using a small number of loci, the estimator is biased because of an increase in ill-conditioned matrices. However, the estimator becomes asymptotically unbiased with large numbers of loci. The ambiguity of polyploid heterozygotes (when balanced heterozygotes cannot be distinguished from unbalanced heterozygotes) is also considered; as with low numbers of loci, genotype ambiguity leads to bias. A software, PolyRelatedness, implementing this method and supporting a maximum ploidy of 8 is provided.

Keywords: genotype ambiguity, method-of-moments, polyploids, relatedness coefficient

Received 27 June 2013; revision received 10 December 2013; accepted 13 December 2013

Introduction

Studies of population genetics, quantitative genetics, behavioural ecology and sociobiology often require measurements of pairwise relatedness between individuals (i.e. Andrew *et al*. 2011; Carl *et al*. 2011; Charpentier *et al*. 2012; Mattila *et al*. 2012). However, in the absence of pedigree information, researchers often rely on genetic markers to estimate relatedness. Many methods have been developed for estimating relatedness coefficients based on genetic data, and these estimators fall into two general categories: moment estimators and maximum-likelihood estimators.

Method-of-moments estimators equate sample moments with unobservable population moments. They are generally unbiased but suboptimal in terms of statistical efficiency. Such estimators have been developed to estimate the relatedness coefficient, r (the probability that two alleles, one sampled from each individual, are identical-by-descent relative to a reference population) by itself (i.e. Queller & Goodnight 1989; Li *et al*. 1993), or simultaneously both r and Δ (four-gene coefficient, the probability that both genes of one individual are identical-by-descent to both genes of the other individual; i.e. Lynch & Ritland 1999; Wang 2002; Thomas 2010).

Maximum likelihood, in contrast, invokes a probability model for the observed pairwise allele pattern given r, Δ and the allele frequencies. By searching the parameter space for values of r and Δ that maximize the probability of the genotype pattern observed, maximum-likelihood parameter values can be determined. However, the maximum-likelihood estimate may produce values that are highly negative or even negative infinity (when few or no alleles are shared by chance). It has been suggested that biologically impossible estimates that fall outside the interval [0, 1] should be truncated at the edges of this interval (Thompson 1975; Milligan 2003; Anderson & Weir 2007). These authors have suggested that biologically impossible estimates outside the interval [0, 1] should be truncated at the edges of this interval. However, truncating negative estimates of r to zero introduces statistical bias (Ritland 1996). Moment estimators also give estimates outside this range, but do not give estimates of negative infinity (Ritland 1996).

Correspondence: Baoguo Li, Fax: +86-29-88303572;
E-mail: baoguoli@nwu.edu.cn

Regardless of the methods, to date all estimators have been developed for species with disomic inheritance, and no estimator of pairwise relatedness has been developed for species with polysomic inheritance. While estimators developed for diploids can be extended to polyploids, as has been performed in the software SPAGeDi V1.3 (Hardy & Vekemans 2002), they do not directly model polysomic inheritance. SPAGeDi, for example, in the case of heterozygotes with unknown allele dosage, treats every allele as having an equal chance of being present in more than one copy, which is an oversimplified assumption that can lead to statistical bias. Thus, an estimator that explicitly considers the unique features of polyploids in its model is needed.

While polyploids are rare in higher animals (i.e. Salmonidae fish, African clawed frog: *Xenopus laevis*, Weather Loach: *Misgurnus anguillicaudatus*), they are pervasive in plants. Some studies suggest that 30–80% of angiosperm species are polyploid (Burow *et al.* 2001), and most lineages show evidence of palaeopolyploidy in their genomes (Otto 2007). Polyploidy can occur through two distinct mechanisms of genome duplication. In allopolyploidy, interspecific hybridization and chromosome doubling result in a polyploid individual with heterologous chromosomes originating from two different species. In contrast, autopolyploidy is the doubling of the whole genome within a single species, often due the presence of unreduced gametes.

Although allopolyploids are more common, autopolyploids have increasingly become the focus of theoretical and experimental research, in large part due to their significant role in evolutionary biology and agriculture (López-Pujol *et al.* 2004; Luo *et al.* 2006; Voorrips & Maliepaard 2012). Furthermore, with allopolyploids, chromosomes from different species are not usually completely homologous, so bivalents form between pairs of chromosomes originating from the same species. Thus, allopolyploids generally display disomic inheritance (Luo *et al.* 2006). For these organisms, normal diploid models can be used once alleles are assigned to the alternative duplicated loci, as in, for example, the study by Ritland & Ganders (1985) for allotetraploids *Bidens menzeisii*. However, diploid models cannot properly be applied to autotetraploids, which display polysomic inheritance. Models for genetic data analysis of autotetraploids have been developed only for selfing rate estimations (cf. Murawski *et al.* 1994; Thompson & Ritland 2006). What is needed is an estimation of r and Δ that deals more generally with polyploids displaying polysomic inheritance (e.g. tetraploidy and higher levels of ploidy). This study addresses this deficiency by introducing a moment estimator for relatedness coefficients using codominant markers in polyploid species.

Theory and modelling

Single-locus estimation

Most estimators assume the following ideal conditions: (i) large, randomly mating populations, (ii) no inbreeding and (iii) autosomal loci and Mendelian inheritance. For diploids, there are two classes of relationship coefficients: (i) 'two-gene' coefficients based upon pairs of genes and (ii) 'four-gene' coefficients based on patterns of gene identity for all four alleles of the two diploid individuals. The relatedness coefficient (r) is the sum of the two coefficients:

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

where Δ is the probability that at any given locus, two individuals share two alleles identical-by-descent (IBD); and ϕ is the probability that they share one allele IBD (Lynch & Ritland 1999). For example, parents and offspring share an allele IBD with a probability of 1, so $\phi = 1$ and $\Delta = 0$; full-sibs share one or two alleles IBD with respective probabilities of 0.5 or 0.25, so $\phi = 0.5$ and $\Delta = 0.25$. Values of ϕ and Δ for specific relationships are listed in Table 1.

Similarly, by the same assumptions and using tetraploids as an example, the relatedness coefficient can be expressed by expanding eqn (1) as follows:

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

where $\sum_{i=0}^{4} \Delta_i = 1$ and Δ_i is the probability that at any given locus, two tetraploid individuals share i alleles identical-by-descent. This statistic assumes that there is no inbreeding or double-reduction, which can cause the alleles within an individual to be identical-by-descent.

Table 1 Relatedness coefficients of specific relationships in diploids and tetraploids

		Diploids		Tetraploids			
Relationship	r	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-sibs	1/2	1/4	1/2	1/36	2/9	1/2	2/9
Half-sibs/ 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-grand parent	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
Nonrelatives	0	0	0	0	0	0	0

Therefore, there are 109 and 48 deltas in inbreeding and double-reduction events, respectively (data not shown), and the probability that the matrices become ill-conditioned or singular is high. The values of the deltas for tetraploid relatives assuming no double-reduction are shown in Table 1. The deltas for hexaploids and octoploids can be found in the Table S1.

Following the method used by Lynch & Ritland (1999), one individual in a pair serves as a 'reference', and the probabilities for locus-specific genotypes in the other 'proband' individual are conditioned on the reference. First, this method reduces the number of potential genotype patterns for individual pairs and yields simpler probability expressions. Unfortunately, there are still numerous genotype patterns in tetraploids. Second, it takes advantage of allele frequencies: once two individuals share a rare allele, the estimator decreases sampling variance and increases locus weight, meaning the coefficients estimated by this locus are more reliable. The estimate of r_{xy} is based on the proband individual y conditioned on reference individual x and expresses the probability that y is observed on the condition of x with the parameter of $\mathbf{\Delta}$, where $\hat{r}_{xy}$ cannot equal $\hat{r}_{yx}$.

Either individual in the pair can be used as the reference individual (Lynch & Ritland 1999). Thus, the reciprocal estimation of r_{xy} and r_{yx}, etc., can be arithmetically averaged to further refine the pairwise relationship estimates for the pair of individuals x and y:

$$\begin{aligned} r &= \left(r_{xy} + r_{yx}\right)/2, \\ \mathbf{\Delta} &= \left(\mathbf{\Delta}_{xy} + \mathbf{\Delta}_{yx}\right)/2. \end{aligned} \tag{3}$$

Where $\mathbf{\Delta}$ is the column vector: $\mathbf{\Delta} = \left[\Delta_1, \Delta_2, \Delta_3, \Delta_4\right]^T$. The alleles in the reference individual are defined as A_i, A_j, A_k and A_l, and the alleles that do not appear in the reference individual are defined as A_*. So there are altogether five genotype patterns for the reference individual, where $A_iA_iA_iA_i$, $A_iA_iA_iA_j$, $A_iA_iA_jA_j$, $A_iA_iA_jA_k$ and $A_iA_jA_kA_l$ have 5, 15, 15, 35 and 70 proband genotype patterns, respectively.

Using this method, the similarity index can be obtained, which is here defined as the number of alleles that are identical-in-state between two individuals. This definition differs from other usages (e.g. Li *et al.* 1993), in that each allele is counted only once. For each locus, the similarity index has only five values (0, 0.25, 0.5, 0.75 and 1).

Table 2 summarizes similarity indices and probabilities for proband genotypes given the allele frequency and the array of deltas for the reference individual $A_iA_iA_iA_i$. The allele frequency of A_i is denoted as p_i. For example, if the reference genotype is $A_iA_iA_iA_i$, the probability of the proband genotype $A_iA_iA_iA_i$ is:

$$\begin{aligned} \Pr(A_iA_iA_iA_i|A_iA_iA_iA_i, \mathbf{\Delta}) = \Delta_4 + p_i\Delta_3 + p_i^2\Delta_2 + p_i^3\Delta_1 \\ + p_i^4\Delta_0. \end{aligned}$$

Because $\Delta_0 = 1-\Delta_1-\Delta_2-\Delta_3-\Delta_4$, the coefficients of Δ_1 to Δ_4 are subtracted by p_i^4. A more detailed explanation of the coefficients of Δ_i is given in the Appendix S1, and the remaining coefficients can also be found in the Table S2.

Summarizing the expressions with the same similarity index for a reference genotype pattern from Table 2, and letting $\mathbf{E}_1$ be the matrix consisting of the four columns headed with deltas, and $\mathbf{E}_2$ be the vector consisting of the column with a header of 1, then the expected moments of the similarity index from first to fourth can be calculated by the following formula:

$$\mathbf{E} = \mathbf{A} + \mathbf{M}\mathbf{\Delta}, \tag{4}$$

where $\mathbf{E}$ is the column vector consisting of moments from first to fourth, $\mathbf{E} = [\mathrm{E}(S), \mathrm{E}(S^2), \mathrm{E}(S^3), \mathrm{E}(S^4)]^T$, and $\mathbf{A} = \mathbf{S}\mathbf{E}_2$, $\mathbf{M} = \mathbf{S}\mathbf{E}_1$, where

$$\mathbf{S} = \begin{bmatrix} 1 & 0.75 & 0.5 & 0.25 & 0 \\ 1 & 0.75^2 & 0.5^2 & 0.25^2 & 0 \\ 1 & 0.75^3 & 0.5^3 & 0.25^3 & 0 \\ 1 & 0.75^4 & 0.5^4 & 0.25^4 & 0 \end{bmatrix}.$$

Equating observed moments to expected ($\mathbf{E} = \hat{\mathbf{E}}$) and estimated deltas to the true deltas ($\mathbf{\Delta} = \hat{\mathbf{\Delta}}$) solves for the estimator from eqn (4):

Table 2 The similarity index and coefficient of probability expressions for reference $A_iA_iA_iA_i$

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

$$\hat{\boldsymbol{\Delta}} = \mathbf{M}^{-1}(\hat{\mathbf{E}} - \mathbf{A}). \quad (5)$$

For a pair of individuals, the observed moment vector $\hat{\mathbf{E}}$ is calculated from the genotype pattern (the 'similarity index' in Table 2). For example, for a genotype pattern of $A_iA_iA_iA_j$-$A_iA_iA_iA_i$, the similarity index is 0.75 and $\hat{\mathbf{E}} = [0.75, 0.75^2, 0.75^3, 0.75^4]^T$. Then, $\hat{r}$ can be obtained by substituting $\hat{\boldsymbol{\Delta}}$ into eqn (2).

Multilocus estimation

To reliably assess relatedness, a number of loci should be used. When loci are unlinked, the locus-specific estimates are independent of each other. In such a case, supposing individuals x and y are observed, then $\hat{r}_{xy}$ will be weighted separately for each locus to generate the least sampling variance, where the weight is the inverse of the sampling variance of $\hat{r}_{xy}$ in each locus:

$$\hat{r}_{xy} = \sum \hat{r}_{xy,i} W_{xy,i} / \sum W_{xy,i},$$
$$W_{xy,i} = 1/\mathrm{Var}(\hat{r}_{xy,i}).$$

Similarly, $\hat{\Delta}_{xy,i}$ is also weighted in the same way. However, these sampling variances are determined by both allele frequency and $\boldsymbol{\Delta}$. Where allele frequency can be estimated from population genotypes, $\boldsymbol{\Delta}$ is one of the parameters to be estimated and is therefore unknown. Here, a common weighting method used in many estimators (i.e. Lynch & Ritland 1999; Wang 2002; Thomas 2010) is employed: without any *a priori* information, individuals x and y are assumed nonrelatives and $\boldsymbol{\Delta} = O$. After weighting, $\hat{r}$ and $\hat{\boldsymbol{\Delta}}$ are obtained by eqn (3).

In practice, however, the mathematical solution for the variance of $\hat{\Delta}_{xy,1}$ to $\hat{\Delta}_{xy,4}$ and $\hat{r}_{xy}$ is far too complicated; it has hundreds of terms and is thus not practical. As a viable alternative, the numerical solution of the equation $\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}^2(X)$ is used for weighting, using standard approximations in numerical analysis.

Ill-conditioned and singular matrices

If the number of alleles at a locus is less than or equal to 4, the condition number (*cond*) of $\mathbf{M}$ may be very large (i.e. over 10 000), in which case $\hat{r}$ and $\hat{\Delta}$ will be vulnerable to the estimation variance of the allele frequency. That is to say, even a minor change in allele frequency can cause the outcome to vary greatly. Such a matrix is called an ill-conditioned matrix. In some extreme cases, *cond* will approach infinity and the matrix $\mathbf{M}$ becomes singular, making eqn (5) unsolvable.

Singularity or near-singularity results from probabilities near or equal to zero, resulting in an unsolvable set of equations. For example, consider a locus that has four alleles whose frequencies are all 0.25 and the reference genotype is $A_iA_jA_kA_l$. For any proband genotype, the coefficients of Δ_1 are zero and $\mathbf{M}$ is singular. At the same time, Δ_1 can be any value and eqn (3) is still valid. This problem also occurs with diploid estimators. For example, when the allele number is two, the estimator of Lynch & Ritland (1999) has a high sampling variance when the allele frequency is near 0.5.

Mathematicians have developed several methods to solve this problem, including singular value decomposition, LU decomposition and Moore–Penrose pseudoinverse (Horn 1990). However, although $\mathbf{M}$ is inverted, the solution is far from normal and $\hat{r}$ can be either below -50 or above 50. As an alternative, a conservative method to minimize the sampling variance would be to set $\hat{\boldsymbol{\Delta}}$ to O and locus weight to zero if any element of $\hat{r}$ exceeds a predefined range. While this method produces bias in single-locus estimations, the bias can be reduced by using additional loci in multilocus estimations.

Genotyping ambiguity

A distinct feature in the population genetics of polyploids is the formation of partial heterozygotes. In an individual, more than two alleles can be present at the same locus; the dosage of an allele can vary from 0 to 4 copies. For example, when two alleles (A_i and A_j) segregate at a locus in an autotetraploid population, there are three types of partial heterozygotes: $A_iA_iA_iA_j$, $A_iA_iA_jA_j$ and $A_iA_jA_jA_j$. All three produce similar electrophoresis signals and cannot be distinguished from each other. This typically occurs with polymerase chain reaction-based markers when the reaction is not limited by initial concentration. Although there are some methods that can to determine the genotype of heterozygous tetraploids (i.e. Xu *et al.* 2002; Gidskehaug *et al.* 2011; Voorrips *et al.* 2011; Serang *et al.* 2012; Uitdewilligen *et al.* 2013), extra equipment or software may be needed.

Here, a simple alternative method is proposed to estimate relatedness when the correct heterozygote genotypes cannot be scored, but the allele frequency is available. First, the probabilities of each genotype are determined. If a genotype has two alleles A_i and A_j, the probability ratio of the three candidate genotypes $A_iA_iA_iA_j$, $A_iA_iA_jA_j$ and $A_iA_jA_jA_j$ is $4p_i^3p_j : 6p_i^2p_j^2 : 4p_ip_j^3 = 2p_i^2 : 3p_ip_j : 2p_j^2$. Similarly, if a genotype has three alleles, denoted by A_i, A_j and A_k, the probability ratio of three candidate genotypes $A_iA_iA_jA_k$, $A_iA_jA_jA_k$ and $A_iA_jA_kA_k$ is $12p_i^2p_jp_k : 12p_ip_j^2p_k : 12p_ip_jp_k^2 = p_i : p_j : p_k$.

Second, all possible genotypes of both reference and proband individuals are combined. $\hat{\mathbf{E}}$ is weighted by the probability of the combined genotypes, while $\mathbf{A}$ and $\mathbf{M}$ are weighted by the probability of the reference genotype:

$$\mathbf{M} = \sum \mathbf{M}_i \Pr(G_{x,i}),$$
$$\mathbf{A} = \sum \mathbf{A}_i \Pr(G_{x,i}),$$
$$\hat{\mathbf{E}} = \sum \hat{\mathbf{E}}_{ij} \Pr(G_{x,i}) \Pr(G_{y,j}).$$

Where $G_{x,i}$ and $G_{y,j}$ represent the i^{th} and j^{th} possible genotype of individual x and y, respectively.

The remaining steps for single-locus and multilocus estimation are unchanged. Note, however, that this method of handling ambiguous genotypes introduces bias. For example, if the parent genotype is $A_iA_iA_iA_i$ and the offspring genotype is $A_iA_iA_jA_k$, the correct genotype is clouded by the consideration of the other two candidate genotypes. As $A_iA_jA_jA_k$ and $A_iA_jA_kA_k$ produce a smaller similarity index, $\hat{r}$ is underestimated.

Model extension

The moment estimator described here can be extended to other levels of ploidy, including polyploids with an odd number of ploidy or haploids, with slight modifications. First, eqn (2) should be modified as follows:

$$r = \sum_{i=0}^{m} i\Delta_i / m, \tag{6}$$

where m is the ploidy of the target organism

Second, the vectors and matrices are also modified: $\mathbf{E}$ is the column vector consisting of the expected moments of the similarity index from first to m^{th}, and $\hat{\mathbf{E}}$ is the observed similarity indices vector: $\hat{\mathbf{E}} = [\mathrm{E}(S), \cdots, \mathrm{E}(S^m)]^T$, and $\mathbf{S}$ is redefined as a $m \times (m+1)$ matrix, where $\mathrm{S}_{i,j} = (m+1-j)^i m^{-i}$. The remaining steps including weighting and handling ambiguous genotypes are the same as in the tetraploid estimator described above.

Diploid coefficient of coancestry estimators adapted to polyploids

While the moment estimator established here is designed specifically for polyploids, there are two estimators originally developed for diploids that can be extended to polyploids (Loiselle *et al.* 1995; Ritland 1996). These estimates measure the coefficient of coancestry between two individuals x and y. While this quantity, denoted θ, is the probability that two alleles, one randomly sampled from each individual, are identical-by-descent (Jacquard 1972). This quantity θ is alternatively defined as the correlation between the additive values of the two individuals (Ritland 1996); for the purposes of this study, we follow the first definition. This coefficient increases with the level of relationship. 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).

These estimators factor in the differing information provided by each allele by assigning a similarity index for genetic pairs for each of n possible alleles. The similarity index has four possible values for a diploid, where the i^{th} element (S_i) is equal to 0 (the allele is not shared), 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 A_iA_i). The single-locus estimator of Ritland (1996) is given by:

$$\hat{\theta}_{xy,\text{Ritland}} = \frac{\sum_i S_i/p_i - 1}{n-1}.$$

Hardy & Vekemans (2002) expanded these estimators to higher levels of ploidy by extending the definition of the similarity index S_i to a product of two frequencies: $S_i = p_{i,x}p_{i,y}$, where $p_{i,x}$ and $p_{i,y}$ are the frequencies of A_i in x and y, respectively. Thus, the multilocus estimators of Ritland (1996) and Loiselle *et al.* (1995) are given by:

$$\hat{\theta}_{xy,\text{Ritland}} = \frac{\sum_j \left(\sum_i S_{i,j,x}S_{i,j,y}/p_{i,j} - 1\right)}{\sum_j (n_j - 1)},$$

$$\hat{\theta}_{xy,\text{Loiselle}} = \frac{\sum_{j,i}(S_{i,j,x} - p_{i,j})(S_{i,j,y} - p_{i,j})}{\sum_{j,i} p_{i,j}(1 - p_{i,j})}.$$

Here, $p_{i,j}$ is the frequency of A_i of the j^{th} locus and n_j denotes the number of alleles of the j^{th} locus. Both estimators are symmetrical, so that $\hat{\theta}_{xy} = \hat{\theta}_{yx}$. The difference between these two estimators is in how loci and alleles are weighted, but they give the same estimates in multilocus estimations with biallelic loci. The $\hat{r}$ can be obtain by the equation $\hat{r} = m\hat{\theta}$. Neither estimator incorporates 'higher-order' coefficients ($\hat{\mathbf{\Delta}}$), which we have in the polyploid moment estimator.

Simulations and comparisons

The bias and variance of the above estimators were compared to evaluate the properties of the new polyploid estimator for tetraploids and hexaploids. For the two-gene coefficient of relatedness estimators of Ritland (1996) and Loiselle *et al.* (1995), which can be extended to higher levels of ploidy, the relatedness coefficient can be obtained by $\hat{r} = m\hat{\theta}$, given above. Because these estimators are unbiased and have similar statistical variances, the new polyploid estimator was compared only to Ritland's first estimator (Ritland 1996). In simulation, the valid range of $\hat{r}$ of the polyploid estimator was $[-16, 1]$.

The statistical properties of these two parameters were obtained from Monte Carlo simulations based on given allele frequencies. The role of genotypic ambiguity of polyploid heterozygotes was also considered. During

simulations, for each pair of individuals, the genotype of one individual was randomly generated according to Hardy–Weinberg expectations; the other genotype was then obtained from the conditional genotype distributions given the reference genotype and the pair's relationship (**Δ**). Two types of allele frequency distributions were simulated: triangular and uniform distributions. The allele frequency meets the proportions 1, 2, …, n in the former distribution; for the latter, the frequencies of all alleles at a locus are equal. Two typical applications were simulated: one using a single multiallelic locus and the other using multiple biallelic loci, assuming unlinked loci. To gauge the performance of the new polyploid estimator with multiple multiallelic loci, the minimal number of loci that gave a variance less than a threshold was computed.

Single-locus estimation

In single-locus simulations, n (number of alleles) was considered from 2 to 15. For each n, both the triangular and uniform allele frequencies were estimated, and the data were simulated for four different relationships. Each run simulated 100 000 pairs of individuals. The results in terms of bias and variance are shown in Figs 1 and 2, respectively. The Ritland (1996) estimator is unbiased, so the bias of it is not shown.

With the exception of nonrelatives, the polyploid moment estimator has a negative bias, and the absolute bias of $\hat{r}$ decreases as n increases, becoming nearly zero (Fig. 1). For either triangular or uniformly distributed allele frequencies, the biases of $\hat{r}$ for parent-offspring and full-sib pairs are similar (Fig. 1), while their variances differ.

The variances of $\hat{r}$ for parent-offspring pairs is usually the smallest among the relationships tested, except for some cases when n is low. The variance seems to be less sensitive to the gene frequency, and the ranking of statistical variance for the four types of pairwise relationships remains the same. In contrast, the variances of the Ritland (1996) estimator under different allele frequencies are quite different, with the ranking of variances of the four relationships changing.

In part due to the ill-conditioned or singular matrix solution, in nonrelatives, the variance is multimodal and bias in nonrelatives as a function of n is not monotonic (dotted lines in Figs 1 and 2). When n is low, or under certain allele frequency distributions, the bias reaches a peak at around $n = 5$. Between different levels of ploidy, their variances are similar, but their biases differ. The hexaploid estimator requires a greater n, especially when heterozygous genotypes are ambiguous (Fig. 1).

Multilocus estimation

To examine the properties of a multilocus estimation, biallelic loci whose number ranged from 1 to 40 were simulated. Relatedness coefficients were estimated between 30 000 pairs of individuals, and four relationships were considered. Results of simulations for bias and variance of $\hat{r}$ are given in Figs 3 and 4, respectively.

The bias of $\hat{r}$ in parent–offspring and full-sib pairs are similar; however, with heterozygote ambiguity, they

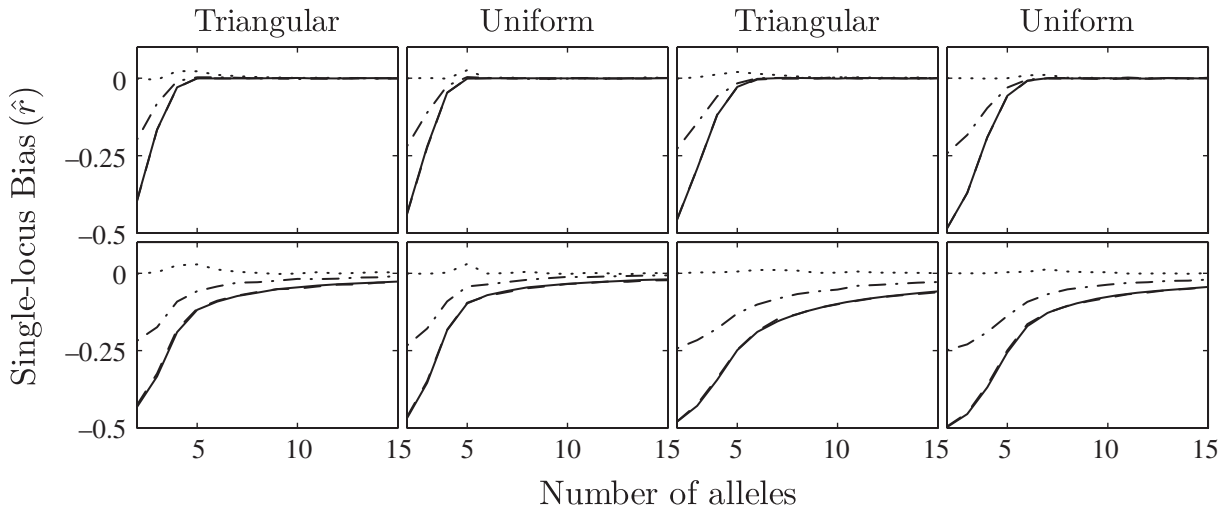

Fig. 1 Single-locus bias of $\hat{r}$ as a function of the number of alleles under triangular and uniform allele frequency distributions. Two models were compared: the top row shows the polyploid estimator with known genotypes, and the bottom row shows the polyploid estimator with ambiguous heterozygote genotypes. For each model, the case of tetraploids (leftmost two columns) and hexaploids (rightmost two columns) was considered, and four relationships were simulated: the solid line '—' denotes parent–offspring, the dashed line '—' denotes full-sibs, the dash-dot line '– ·' denotes half-sibs, and the dotted line '···' denotes nonrelatives. For each case, 100 000 Monte Carlo simulations were performed.

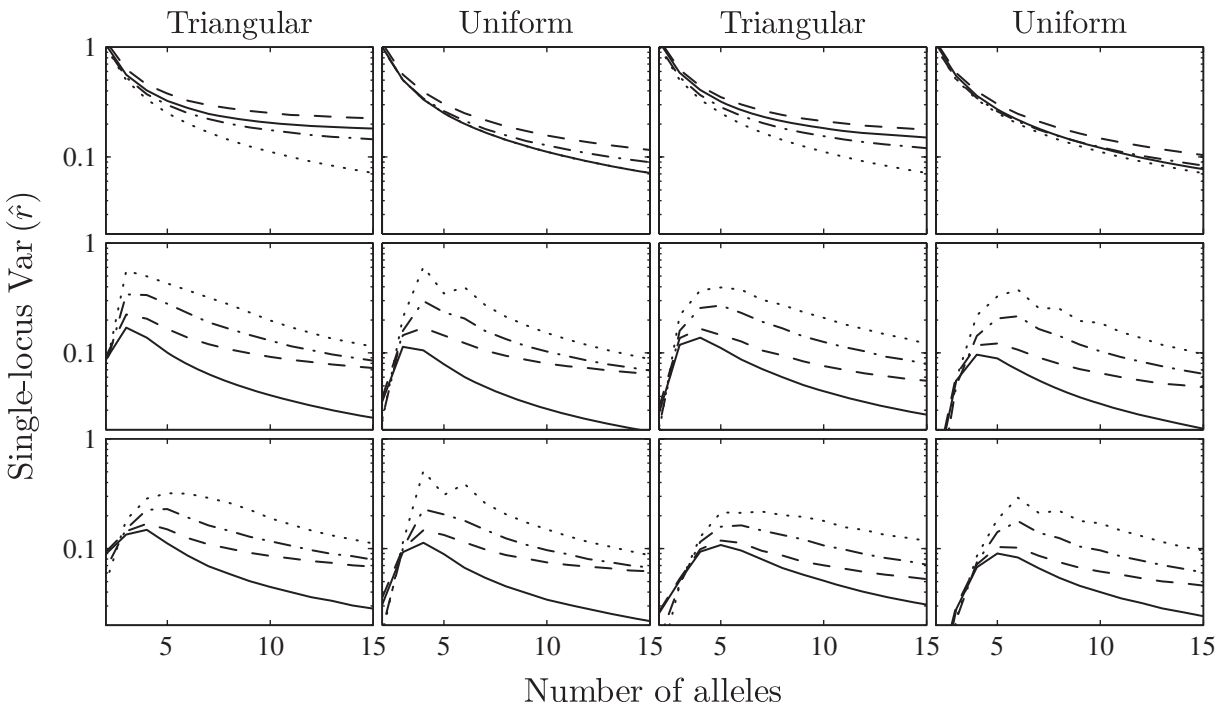

Fig. 2 Single-locus variance of $\hat{r}$ as a function of the number of alleles at loci with triangular and uniform allele frequency distributions. Three models for tetraploids (leftmost two columns) and hexaploids (rightmost two columns) were compared: the top row shows the Ritland (1996) estimator, the middle row shows the polyploid estimator developed in this study, and the bottom row shows the effect of heterozygote genotype ambiguity for the polyploid estimator. Results were obtained from 100 000 simulated pairs of four relationships, as in Fig. 1.

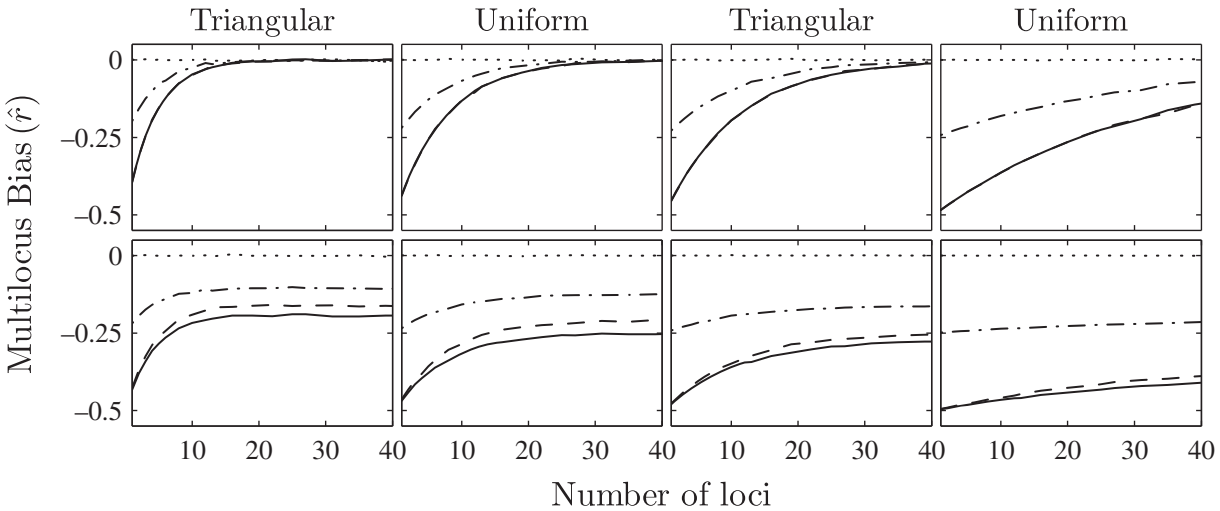

Fig. 3 Bias of $\hat{r}$ using multiple biallelic loci with triangular and uniform allele frequency distributions. The estimators, ploidy levels, relationships and curve shapes are the same as in Fig. 1. Results were obtained from 30 000 Monte Carlo simulations.

begin to differ, but by less than 0.04. The bias for tetraploid estimators is reduced to a negligible level at 20 loci, and loci with triangular allele frequencies give better estimates. As ploidy increases, more loci are needed to minimize bias (Fig. 3). As before, estimators are biased when heterozygotes are ambiguous, but bias does decline gradually to an asymptote with an increasing number of loci, especially for hexaploids exhibiting a uniform allele frequency distribution.

The Var($\hat{r}$) of the Ritland (1996) estimator are monotonic, and each relationship produces similar declines in variance with increasing number of loci. For polyploid

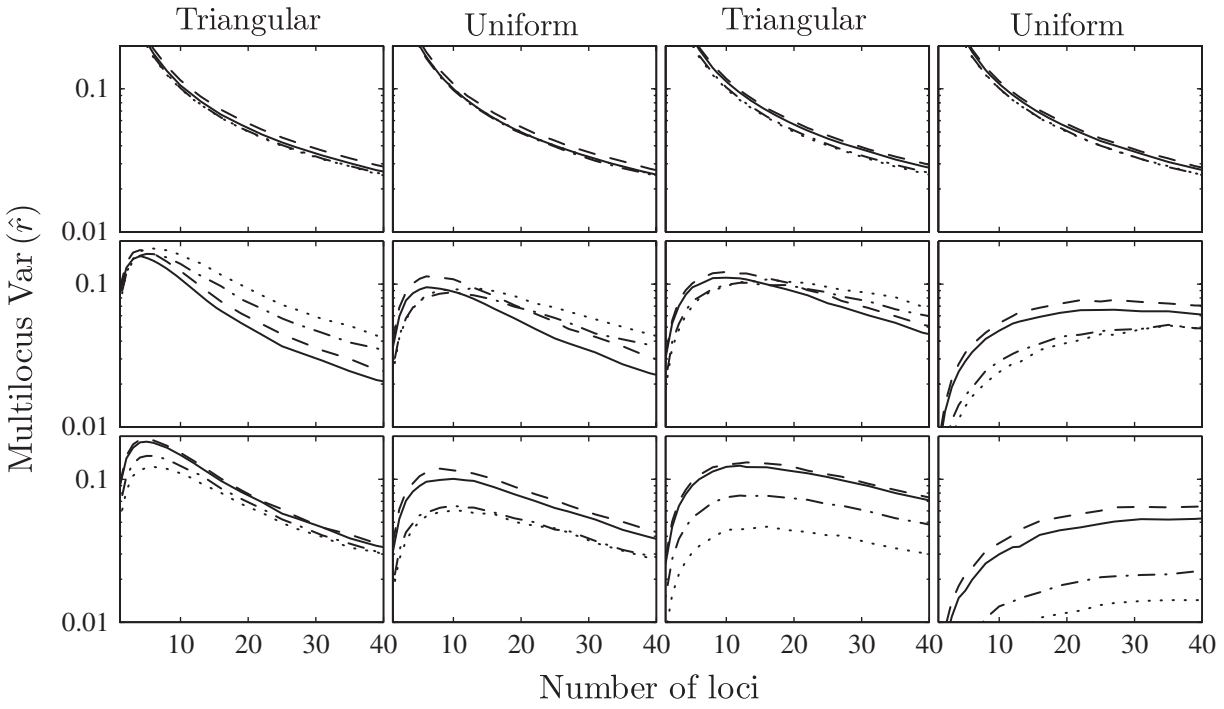

Fig. 4 Multilocus variance of $\hat{r}$ for biallelic loci with triangular and uniform distributed allele frequencies. The estimators, ploidy levels, relationships, and curve shapes are the same as in Fig. 2. Results were obtained from 30 000 Monte Carlo simulations.

estimators, the Var($\hat{r}$) initially increases, plateaues and then decreases; the variances between different relationships are quite different.

Multilocus estimation with multiallelic loci

To demonstrate the performance of the estimators with multiallelic loci, the minimum number of loci that gave a variance of $\hat{r} < 0.01$ was compared, as shown in Fig. 5. Two estimators were compared, including Ritland (1996) and the polyploid estimator, both of which were simulated for diploids to octoploids. Two main things can be seen from the comparisons. First, uniform allele frequency distributions give higher variances than triangular distributions for biallelic loci; with multiallelic loci, uniform distributions perform better (Fig. 5). Second, with increasing ploidy, the estimators become more efficient, thus requiring fewer loci for the same level of accuracy. For the Ritland (1996) estimator, the number of loci required for Var($\hat{r}$) < 0.01 in octoploids is usually less than that in diploids, while in the polyploid estimator presented in this article, more loci are required in octoploids compared with tetraploids when there are fewer than eight alleles.

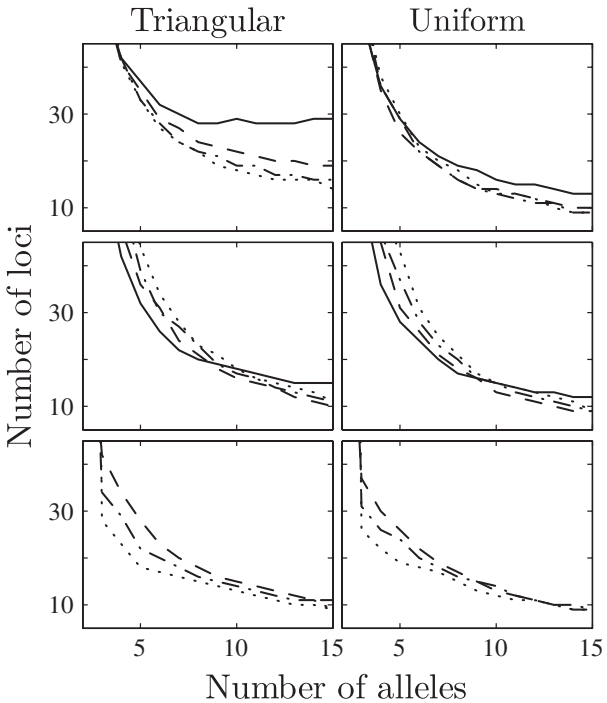

Fig. 5 The minimum number of multiallelic loci *l* that gives a statistical variance of less than 0.01 for Max(Var($\hat{r}$)) across four types of pairwise relationships. The estimators were the same as in Fig. 2; diploids ('—'), tetraploids ('—'), hexaploids ('– ·') and octoploids ('···') were simulated for each model.

Discussion

This study describes the development and validation of an estimator of relatedness for species exhibiting poly-

somic inheritance based on codominant markers. Its performance was compared with existing estimators in three applications with two types of allele frequencies by Monte Carlo simulation.

Properties of the polyploid estimator

Marker-based relatedness estimates typically show a large sampling variance due to the relatively small number of loci used and the presence and variance of identity-in-state for alleles that are not identical-by-descent (Lynch & Ritland 1999). In addition, many relationships, particularly full-siblings, have variance of identity-by-descent between loci. The variance can be reduced by increasing the number of loci or by using more polymorphic loci (Figs 2 and 4).

Because of nonlinear treatment in solving ill-conditioned or singular matrices, the variance of the polyploid estimator is multimodal in single-locus estimations (Fig. 2). The variance becomes approximately monotonic with increasing number of alleles, particularly when the number of alleles is at least 5. The variance is nonmonotonic when a low number of loci are used in multilocus estimations (Fig. 4). Low numbers of loci frequently cause **M** to become ill-conditioned, or give estimates beyond the valid range. In this situation, the estimator gives a zero estimate but with zero weight for this locus. However, this problem can be reduced by increasing the number of loci, as when many loci are used, some loci will give valid estimates.

In single-locus estimations, with the exception of unrelated pairs, the variances of the polyploid estimator are smaller than those of the Ritland (1996) estimator for loci with at least six alleles (Fig. 2). In multilocus estimations, this property also occurs (Fig. 4), even when the correct heterozygous genotype is not known. However, the polyploid estimator can perform worse than the Ritland (1996) estimator if multiple biallelic loci are used (Fig. 4).

The polyploid estimator exhibits bias with a finite number of alleles and loci. The bias has two origins: (i) nonlinear solutions of ill-conditioned or singular matrices and (ii) heterozygote genotype ambiguity. Case (i) happens when the number of alleles is low. This bias can be reduced to a low level by increasing the number of loci used or by using loci with greater allele numbers, or both. The number of loci and/or alleles used need not be very high. For example, in single-locus estimation, 8 alleles practically eliminate bias (Fig. 1), and in multilocus estimation, 13 tri-allelic, five tetra-allelic or two penta-allelic loci reduce bias to 0.01 for tetraploids, hexaploids and octoploids, respectively (data not shown). Case (ii) occurs when genotyping tetraploid heterozygotes, given that balanced heterozygotes cannot be distinguished from unbalanced heterozygotes. In this case, **M** and **A** are weighted by the probability of heterozygote dosage, and $\hat{\mathbf{E}}$ is weighted by the combined probability of candidate reference-proband genotype patterns. While this procedure is fairly straightforward for tetraploids, the number of possible heterozygous genotypes for hexaploids and octoploids reaches 10 and 35, respectively. This large number of possible heterozygous patterns results in greater bias for higher ploidy estimators.

The bias for $\hat{r}$ when heterozygote genotype is ambiguous asymptotically approaches zero with an increasing number of alleles (Fig. 1). However, this bias does not approach zero with increasing number of loci; for tetraploids, the bias reaches an asymptote at about 20 loci, and higher ploidies require even more loci to reach an asymptote (Fig. 3). The biases for parent–offspring and full-sib pairs in polyploids are similar in single-locus estimations (Fig. 1). However, for multilocus estimation, the bias for full-sib pairs is somewhat higher than for parent-offspring pairs (Fig. 3).

With regard to allele frequency, the uniform distribution exhibits less variance than the triangular distribution for single-locus estimation (Fig. 2). By contrast, uniform distribution exhibits more variance for multilocus estimation (Fig. 4). A uniform distribution shows the same level of bias for single-locus estimation, but more bias than the triangular distribution for multilocus estimation (Fig. 3). Compared with the Ritland (1996) estimator, the new polyploid estimator described here is less sensitive to allele frequency distribution (Fig. 2).

The locus-specific weighting scheme follows Lynch & Ritland (1999): $\hat{r}_{xy}$ and $\hat{r}_{yx}$ of each locus, and their weighted averages are calculated first, then $\hat{r}$ is obtained by eqn (3). The weighting scheme for deltas is similar. The advantage is that the four 'higher-order' coefficients are all unbiased. An alternative weighting scheme from Wang 2002 would be to calculate the weighted averages of **M** and **A** first and then substitute them into eqn (5). This approach reduces the occurrence of ill-conditioned matrices in multilocus estimations; however, the deltas obtained by this method are biased.

Limitations of the polyploid estimator

One potential limitation of the new polyploid estimator is that it assumes no inbreeding or double-reduction so that the 'higher-order' coefficients are available, and the range of $\hat{r}$ is $(-\infty, 1)$. In contrast, coefficient of coancestry estimators allows inbreeding and double-reduction, and $\hat{r} \in (-\infty, \infty)$. Therefore, the $\hat{r}$ for clonemates using the polyploid estimator is strictly one, but ranges around one in coefficient of coancestry estimators. These effects can result in the underestimation of the relatedness coef-

ficient for two reasons. (i) The true value of r when considering inbreeding and double-reduction can be greater than one (i.e. $r = 4$ for autotetraploid homogeneous individuals; the maximum r for full-sibs considering double-reduction is two), but the upper limit of $\hat{r}$ of the polyploid estimator is one. (ii) The polyploid estimator counts each pair of IBS alleles only once. Consider $A_iA_iA_iA_i$-$A_iA_jA_kA_l$ and $A_iA_jA_kA_l$-$A_iA_mA_nA_o$: their similarity indices are all 1/4. If inbreeding or double-reduction is considered, the former genotypes are more similar and should be assigned a larger $\hat{r}$. If inbreeding or double-reduction is considered, there will be 109 and 48 deltas in autotetraploids. Until recently, estimation of so many deltas has been impossible.

A second limitation of the new polyploid estimator is in its range of application. Specifically, while this estimator may be used for a wide variety of data, including microsatellites, SNPs and other codominant markers, when applications require high reliability, the use of biallelic loci is not suggested. This is for several reasons. First, the number of biallelic loci required to achieve the same level of reliability is higher than for multi-allelic loci (Fig. 5). Second, the number of unlinked loci in the genome is limited. As the number of loci used increases, there is a greater likelihood that adjacent loci are linked. Although linked loci do not introduce a bias, they also do not increase the reliability. This causes the variance to plateau as the number of loci increases. Third, when there are ambiguous genotypes, biallelic loci produce a bias that is too large for them of use (Fig. 3). All of these features tend to make microsatellite data more appropriate for this estimator. However, SNP data can still be used, especially with newer genotyping-by-sequencing and haplotype prediction (Xu *et al.* 2002; Uitdewilligen *et al.* 2013) technologies. These do provide a method of unambiguous genotyping and also make the haplotype known. The haplotype of adjacent SNPs can be treated as an allele of a multi-allelic locus, which can largely improve the reliability of estimation.

Additionally, this estimator is somewhat sensitive to the presence of null-alleles (alleles that do not amplify or otherwise cannot be detected). Two effects brought by null-alleles can cause the estimation to become biased. (i) The null-allele will make the observed genotypes appear more or less similar than they really are, which causes an overestimation or underestimation of relatedness. (ii) The summation of the observed frequencies of the amplifiable alleles is extended to one, resulting in an underestimation. Further studies are needed to fully understand the effects of null-alleles, but some simple simulations for the new polyploid estimator indicate that in tetraploids with ambiguous genotypes, the bias is always negative and the bias of $\hat{r}$ is 1.3 to 1.8 times the original bias (data not shown).

Properties of ploidy level

Finally, while this estimator can theoretically be extended to any level of ploidy, at higher levels of ploidy, this estimator becomes computationally difficult. For diploids, there are two different types of reference genotypes: homozygote and heterozygote. For tetraploids, hexaploids and octoploids, there are 5, 11 and 22 distinct reference genotype modes, respectively. The number of distinct reference genotype modes is also the number of partitions of m.

The total number of proband-reference modes equals $\sum_{i=1}^{m} f(m,i)\binom{i+m}{m}$, where $\binom{i+m}{m}$ is the number of proband genotypes if i kinds of IBS alleles are observed in the reference genotype, and $f(m, i)$ is the number of reference genotypes that have i kinds of IBS alleles:

$$f(a,b) = \begin{cases} \sum_{i=1}^{b} f(a-b,i) & \text{if } 1 \le b \le a-1, \\ 1 & \text{if } a = b \text{ or } b = 1. \end{cases}$$

The total numbers of proband-reference modes from haploids to dodecaploids are 2, 9, 34, 140, 538, 2149, 8318, 32 661, 126 895, 495 693, 1 929 303 and 7 531 849. This numerical sequence increases by about 3.90 times for each level, with the number of expressions becoming extremely cumbersome at higher levels of ploidy. Therefore, the generation of symbolic expressions in Table 2 is not practical for levels of ploidy higher than 8, so the software PolyRelatedness supports a maximum ploidy of 8.

There are m copies of a gene at a polyploid locus, and just two copies of a gene at a diploid locus. The larger copy number in higher level of polyploids gives more information about relatedness, but to date, there has been no established methodology for estimating relatedness in polyploids. The estimator outlined here addresses that issue, and simulations under various conditions show that, indeed, relatedness can be more accurately estimated in polyploids than in diploids (Fig. 5). However, there are limitations to this approach; if few loci are sampled and allele numbers at each locus are low, the matrix in the equation for estimating relatedness in polyploids becomes ill-conditioned and is often noninformative about relatedness. In this case, the polyploid estimator performs worse, especially for biallelic loci (Fig. 3). Thus, this method is best employed where there are data from multiple, moderately polymorphic loci.

Acknowledgements

We thank Dr. Da-qing Wang, Stephanie Chen and two anonymous reviewers for providing comments on this article, and Prof. Olivier J. Hardy for providing help about extending coefficient of coancestry estimators. This study was supported by the

National Nature Science Foundation of China (31130061, 30970379, 31270441) and the University of British Columbia (KR).

References

Anderson AD, Weir BS (2007) A maximum-likelihood method for the estimation of pairwise relatedness in structured populations. *Genetics*, **176**, 421–440.

Andrew JK, Fay EC, Guy C (2011) The dining etiquette of desert baboons: the roles of social bonds, kinship, and dominance in cofeeding networks. *American Journal of Primatology*, **73**, 768–774.

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.

Carl V, Joachim M, Deborah AD, Viki V, Luc L (2011) Spatial heterogeneity in genetic relatedness among house sparrows along an urban-rural gradient as revealed by individual-based analysis. *Molecular Ecology*, **20**, 4643–4653.

Charpentier MJE, Fontaine MC, Cherel E *et al.* (2012) Genetic structure in a dynamic baboon hybrid zone corroborates behavioural observations in a hybrid population. *Molecular Ecology*, **21**, 715–731.

Gidskehaug L, Kent M, Hayes BJ, Lien S (2011) Genotype calling and mapping of multisite variants using an Atlantic salmon iSelect SNP array. *Bioinformatics*, **27**, 303–310.

Hardy OJ, Vekemans X (2002) SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology Notes*, **2**, 618–620.

Horn RA (1990) *Matrix Analysis*. Cambridge University Press, Cambridge.

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

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

Loiselle BA, Sork VL, Nason J, Graham C (1995) Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). *American Journal of Botany*, **82**, 1420–1425.

López-Pujol J, Bosch M, Simon J, Blanche C (2004) Allozyme diversity in the tetraploid endemic *Thymus loscosii* (Lamiaceae). *Annals of Botany*, **93**, 323–332.

Luo ZW, Zhang ZE, Zhang RM *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.

Mattila ALK, Duplouy A, Kirjokangas M *et al.* (2012) High genetic load in an old isolated butterfly population. *Proceedings of the National Academy of Sciences*, **109**, E2496–E2505.

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

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.

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. *Genetical Research*, **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 for Bayesian SNP genotyping in polyploids. *PLoS ONE*, **7**, e30906.

Thomas SC (2010) A simplified estimator of two and four gene relationship coefficients. *Molecular Ecology Resources*, **10**, 986–994.

Thompson EA (1975) The estimation of pairwise relationships. *Annals of Human Genetics*, **39**, 173–188.

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.

Uitdewilligen JGAML, Wolters AA, D'hoop BB *et al.* (2013) A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. *PLoS ONE*, **8**, e62355.

Voorrips RE, Maliepaard CA (2012) The simulation of meiosis in diploid and tetraploid organisms using various genetic models. *BMC Bioinformatics*, **13**, 248.

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.

Xu CF, Lewis K, Cantone KL *et al.* (2002) Effectiveness of computational methods in haplotype prediction. *Human genetics*, **110**, 148–156.

S.T.G. and B.G.L. designed the research, K.H. performed research and wrote the draft, M.R.S. edited the manuscript, and K.R. edited the manuscript for grammar and provided earlier additions.

Data Accessibility

The computer program POLYRELATEDNESS V1.2, user manual and an example data set are available as a Google Project at http://polyrelatedness.googlecode.com.

The simulation program, data, derivation of the coefficient of deltas in Table 2, and the details of Tables 1 and 2 are provided as supplemental files through this journal.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1 Relatedness coefficients of specific relationships from diploids to octoploids.

Table S2 The similarity index and coefficient of probability expressions for reference.

Appendix S1 The coefficient of deltas in Table 2.

Appendix S2 data.txt: Result file of simulations.

Appendix S3 Plot_Figures.m: Matlab program to plot the figures in the manuscript.

Appendix S4 Simulate.exe: Simulation program, x86 executable for Windows.