Article Type: Resource Article

Title

Estimating pairwise relatedness between individuals with different levels of ploidy

Authors

Kang Huang∗, Kermit Ritland#, Songtao Guo∗, Derek W. Dunn∗, Dan Chen∗, Yi Ren∗†, Xiaoguang Qi∗, Pei Zhang∗, Gang He∗, and Baoguo Li∗†.

Addresses

∗: Key Laboratory of Resource Biology and Biotechnology in Western China of Ministry of Education, and College of Life Sciences, Northwest University, Xi'an, ShaanXi, China, 710069.

#: Department of Forest and Conservation Sciences, University of British Columbia, Vancouver, British Columbia, Canada, V6T 1Z4.

†: Institute of Zoology, Shaanxi Academy of Sciences, Xi'an, ShaanXi, China, 710032.

Keywords

relatedness coefficient, method-of-moment, maximum-likelihood, arrhenotoky.

Corresponding author

Name: Baoguo Li. Address: Key Laboratory of Resource Biology and Biotechnology in Western China of Ministry of Education, and College of Life Sciences, Northwest University, Xi'an, ShaanXi, China, 710069. Telephone: +8613572209390. Fax: +86 29 88303572. E-mail: baoguoli@nwu.edu.cn.

Running title

Estimating relatedness in arrhenotoky

Estimating pairwise relatedness between individuals with different levels of ploidy

December 1, 2014

Abstract

Estimates of relatedness coefficients, based on genetic marker data, are often necessary for studies of genetics and ecology. While many estimates based on method-of-moment or maximum-likelihood methods exist for diploid organisms, no such estimators exist for organisms with multiple ploidy levels, which occur in some insect and plant species. Here, we extend five estimators to account for different levels of ploidy: one relatedness coefficient estimator, three coefficient of coancestry estimators, and one maximum-likelihood estimator. We use arrhenotoky (when unfertilized eggs develop into haploid males) as an example in evaluations of estimator performance by Monte-Carlo simulation. Also, three virtual sex-determination systems are simulated to evaluate their performances for higher levels of ploidy. Additionally, we used two real datasets to test the robustness of these estimators under actual conditions. We make available a software package, PolyRelatedness, for other researchers to apply to organisms that have various levels of ploidy.

Keywords: relatedness coefficient, method-of-moments, haplodiploid system, arrhenotoky.

Introduction

Quantifying relatedness between two individuals is critical to studies of population genetics, quantitative genetics, behavioral ecology and sociobiology (i.e. Charpentier *et al.* 2012; Liu *et al.* 2013; Guo *et al.* Inpress). The investigation of pairwise relatedness is also important in characterizing inbreeding and genetic diversity in natural populations, in studies of kin selection in social insect populations, and in inferring population structures (Wang 2004). While a relatedness coefficient can be easily calculated from a pedigree, a lack of a pedigree means relatedness must be estimated from genetic marker data using method-of-moments or maximum-likelihood methods.

Method-of-moment estimators equate unobservable population moments to mathematical sample moments. While generally unbiased, they are usually non-optimal in terms of statistical efficiency and have higher sampling variance than likelihood estimators. Different method-of-moment estimators can estimate the relatedness coefficient, *r*, alone, (i.e. Queller & Goodnight 1989; Li *et al.* 1993; Ritland 1996; Loiselle *et al.* 1995), or both *r* and Δ (a four-gene coefficient: the probability that both genes of one individual are identical-by-descent (IBD) to both genes of the other individual) simultaneously (i.e. Lynch & Ritland 1999; Thomas 2010; Wang 2002; Huang *et al.* 2014b). The estimators of Huang *et al.* (2014b), Ritland (1996) and Loiselle *et al.* (1995) can also be used for polyploids by comparing all allele pairs between two individuals, although allele dosage must be known.

In contrast, maximum likelihood estimators model the probability of observing a specific pairwise allele pattern given ϕ (two-gene coefficient: the probability that two individuals share one pair of IBD alleles), Δ and the allele frequencies. Estimates are found by searching the parameter space of ϕ and Δ for values that maximize the probability of the observed genotype pattern. Since values are limited to the parameter space, invalid parameter values (outside [0, 1]) are not possible. Previous work with maximum-likelihood estimators include that of Milligan (2003) and Anderson & Weir (2007), based on earlier work by Thompson (1975). We have also developed a maximum likelihood estimator for polyploids (Huang *et al.* 2014a).

Differences of ploidy within a population can occur both naturally and by artificial inducement, such that related individuals have different levels of ploidy. Arrhenotoky is a form of parthenogenesis in which haploid males are produced from unfertilized eggs. This is best known to occur in some insect orders such as the Hymenoptera (bees, ants, and wasps) (Grimaldi 2005), Hemiptera (mealybugs) (Gullan & Cook 2007), and the Thysanoptera (thrips) (White 1984). Pseudo-arrhenotoky, which occurs in mites (Sabelis & Nagelkerke 1988), some Coleoptera (Borsa & Kjellberg 1996), and some mealybugs (Bongiorni *et al.* 2001), is the phenomenon by which males develop from fertilized eggs but the paternal genome is heterochromatinized or lost in the somatic cells and not passed on to offspring. Dihaploids and polyhaploids are formed by the haploidisation of polyploids, i.e. by the halving of the chromosomal constitution. These processes are important for the selective breeding of tetraploid crop plants (Behnke 1980; Hermsen *et al.* 1981), because selection is faster with diploids than with tetraploids. Tetraploids can be reconstituted from diploids by somatic fusion (Deimling *et al.* 1988). The ploidy of gametophytes in some organisms such as lichens and mosses are also halved relative to the diploid state (Shaw & Beer 1999; Gerrienne & Gonez 2011).

A problem with existing estimators is they cannot estimate the relatedness between individuals with different levels of ploidy, and are applicable only to diploids (Queller & Goodnight 1989; Li *et al.* 1993; Ritland 1996; Loiselle *et al.* 1995; Weir 1996; Lynch & Ritland 1999; Wang 2002; Milligan 2003) perhaps with unbiased correction for population subdivision (Anderson & Weir 2007; Wang 2011), or are applicable just to polyploids (Huang *et al.* 2014ab). Therefore, in this paper, we extend five estimators applicable to polyploids: Huang *et al.* (2014b) (abbreviation for MOM), Ritland (1996) (RI), Loiselle *et al.* (1995) (LO), Weir (1996, Equation 2.28) (WE) and Huang *et al.* (2014a) (ML) to estimate relatedness coefficients at different levels of ploidy with co-dominant markers.

Theory and modelling

The definition of relatedness

We use two relatedness definitions for this study: (i) the IBD (identical by descent) definition and (ii) Wright's definition. The IBD definition is the probability that an allele sampled from one individual at a locus is IBD to one of the alleles from the other individual. For monozygotic twins or clonemates, relatedness $r = 1$; for parent-offspring and full-sib relationships, $r = 1/2$; and for second- and third-order relationships, $r = 1/4$ and 1/8, respectively. The relatedness coefficient for two individuals (x and y) in a diploid outbred population can therefore be expressed by two "higher-order" coefficients:

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

Here, Δ and ϕ are the four- and two-gene coefficients as described in the introduction. For example, parent-offspring pairs share an IBD allele with a probability of 1, so $\phi = 1$ and $\Delta = 0$; full-sibs share one or two alleles IBD with respective probabilities of 1/2 or 1/4, so $\phi = 1/2$ and $\Delta = 1/4$.

In this definition, IBD relatedness is symmetric in outbred diploid populations and in polyploid populations, in the absence of inbreeding, selfing or double-reduction (these can cause an individual to carry IBD alleles; for simplicity, we use the term inbreeding to include all three events). The relatedness coefficient is asymmetric between inbred individuals or individuals with different levels of ploidy, which can be written as $r_{xy} \neq r_{yx}$. Using an inbred diploid dyad as an example, $x - y = A_iA_i - A_iA_j$, where alleles with the same subscript denote IBD alleles, $r_{xy} = 1$ (x to y, all alleles in x is IBD to at least one allele in y) and $r_{yx} = 1/2$.

In haplodiploid systems, IBD relatedness is also asymmetric. For example $A_iA_j - A_i$, $r_{xy} = 1/2$ and $r_{yx} = 1$. However, the number of alleles in one individual IBD to any allele from another individual (denoted N_{IBD}) in both individuals is equal in an outbred population, so $r_{xy} = N_{\mathrm{IBD}}/v_x$ and $r_{yx} = N_{\mathrm{IBD}}/v_y$, where v_x and v_y are the levels of ploidy of x and y, respectively. Therefore, $r_{xy} = r_{yx}v_y/v_x$. The IBD relatedness coefficients for an outbred population are listed in Table 1. For simplicity, we present only the relatedness of higher ploidy individuals to lower ploidy individuals (r_{HL}) for female-male dyads. If $v_x = v_y$, then r_{HL} is equivalent to r.

Wright's coefficient of relationship is defined as the correlation of allele frequencies between individuals (Cockerham 1969). It can be calculated using the coefficient of coancestry and the inbreeding coefficient (Cockerham 1969; Hardy & Vekemans 1999). Here, we expand the principle to dyads with different levels of ploidy (see supplementary materials for details):

$$r = \frac{\theta_{xy}}{\sqrt{\theta_{xx}\theta_{yy}}}. \qquad (2)$$

Wright's definition is symmetric because it is independent of direction. Using the two previous examples from the IBD definition, for the diploid inbred dyads, A_iA_i-A_iA_j, Wright's relatedness is $\sqrt{2}/2$; for a haploid-diploid outbred dyad $A_iA_j - A_i$, Wright's relatedness is $\sqrt{2}/2$. Since the value is a geometric mean of IBD relatedness, we can unify IBD relatedness in outbred populations or inbred populations with $v \leq 2$. Unfortunately, for inbred polyploid populations, the unified IBD relatedness can be different from Wright's relatedness. For example, for an inbred tetraploid dyad $A_iA_iA_jA_k - A_iA_lA_lA_l$, the unified IBD relatedness is $1/\sqrt{8}$, whereas Wright's relatedness is $1/\sqrt{15}$.

Previous studies have preferentially used directional relatedness (i.e. Barron *et al.* 2001) for haploiddiploid dyads. Therefore we can split Wright's relatedness into two directional coefficients, regardless of inbred or outbred populations:

$$r_{xy} = \theta_{xy}/\theta_{yy}. \qquad (3)$$

This is equivalent to the IBD relatedness for outbred polyploids, inbred diploids, inbred haploid-diploid and haploid dyads. In outbred populations, if $v_x = v_y$, then $\vartheta_{xx} = \vartheta_{yy} = 1/v$, reducing Equation (2) to $r = v\vartheta$. If $v_x \neq v_y$, we get $r_{xy} = N_{\mathrm{IBD}}/v_x$ and $\vartheta = N_{\mathrm{IBD}}/(v_xv_y)$, and Equation (3) can be modified to

$$r_{xy} = v_y\theta_{xy}. \qquad (4)$$

Polyploid moment estimator

This estimator assumes that individuals are drawn from an outbred, panmictic population so $r_{xy} = r_{yx}v_y/v_x$. For simplicity, we only consider the relatedness of higher ploidy individuals to lower ploidy individuals (r_{HL}). The relatedness coefficient and coefficient of coancestry can be expressed by generalizing Equations (1) and (4):

$$r_{\mathrm{HL}} = v^{*}\theta = \sum_{i=0}^{v^{*}} i\Delta_i / v. \qquad (5)$$

Here *v* denotes the higher ploidy, v^* the lower ploidy, and $\sum_{i=0}^{v^*}\Delta_i = 1$ and Δ_i is the probability that at any given locus two individuals share i alleles that are IBD. Therefore, in diploids, the two-gene coefficient ϕ is identical to Δ_1, the four-gene coefficient Δ is equivalent to Δ_2, and Equation (5) is reduced to Equation (1). In haploid-diploid dyads, Δ_2 is undefined.

We demonstrate the values of deltas with higher levels of ploidy, by using three virtual sex-determining systems. In the virtual diplotetraploid system, males develop from unfertilized eggs and are diploid, whereas females develop from fertilized eggs and are tetraploid. For the other two virtual systems, the deltas in the virtual outbreeding diplotetraploid system, assuming no double-reduction, are shown in Table 1, where the relationships shown in the first column are divided into three categories: (i) female-female, (ii) male-female and (iii) male-male. The deltas for the virtual triplohexaploid and tetraplooctoploid systems are given in the supplementary files.

Table 1 is about here.

This estimator follows the method used by Lynch & Ritland (1999): one individual of a pair serves as a "reference", and the probabilities of the locus-specific genotypes in the other "proband" individual are conditioned on the reference genotype. There is no particular reason to use one individual of a pair as the reference rather than the other (Lynch & Ritland 1999). The reciprocal estimates $\hat{r}_{\mathrm{HL},xy}$ and $\hat{r}_{\mathrm{HL},yx}$ can be arithmetically averaged to give an overall pairwise relationship estimate for the pair of individuals *x* and *y*:

$$\hat{r}_{\mathrm{HL}} = \left(\hat{r}_{\mathrm{HL},xy} + \hat{r}_{\mathrm{HL},yx}\right)/2,$$

$$\ddot{\boldsymbol{\Delta}} = \left(\ddot{\boldsymbol{\Delta}}_{xy} + \ddot{\boldsymbol{\Delta}}_{yx}\right)/2.$$

Here, **Δ** is a column vector consisting of deltas from v^* to 1: $\boldsymbol{\Delta} = [\Delta_{v^*}, \ldots, \Delta_1]^T$. $\hat{r}_{xy}$ and $\hat{r}_{yx}$ are obtained by Equation (5). The alleles in the reference individual are defined as $A_i, A_j, \ldots$, with respective population frequencies $p_i, p_j, \ldots$. The alleles that do not appear in the reference individual are defined as A_*, and p_* denotes the corresponding frequencies of A_*. The similarity index (*S*) is the number of identical-by-state (IBS) allele pairs divided by *v*, where each IBS allele is counted only once. For example, there are one and two IBS alleles for genotype pairs $A_iA_j - A_iA_i$ and $A_iA_i - A_iA_i$, respectively. For a haploid-diploid dyad, *S* can take the values 0 or 1/2. Therefore the probability that a pair of individuals has a specific similarity index given a specific relationship (**Δ**) can be calculated. For example, if the two individuals are both diploid, the probability of *S* = 1 is:

$$\Pr(S = 1 | A_iA_i, \boldsymbol{\Delta}) = \Delta_2 + p_i\Delta_1 + p_i^2\Delta_0.$$

For a haploid-diploid dyad, the two probabilities of $S = 1/2$ using each of two individuals as a reference respectively are:

$$\Pr(S = 1/2 | A_iA_i, \boldsymbol{\Delta}) = \Delta_1 + p_i\Delta_0,$$

$$\Pr(S = 1/2 | A_i, \boldsymbol{\Delta}) = \Delta_1 + p_i^2\Delta_0.$$

The remaining coefficients of deltas for other dyads are given in the supplementary files. For the deltas subjected to the constraint $\sum_{i=0}^{v^*}\Delta_i = 1$, for simplicity the coefficients of Δ_1 to Δ_{v^*} are subtracted by the coefficient of Δ_0. Table 2 lists the distribution of similarity indices given the allele frequencies and the array of deltas in the haplodiploid system. Taking the first row of Table 2 as an example, in diploid dyads, $\Pr(S = 1|A_iA_i, \mathbf{\Delta}) = \Delta_2 + p_i\Delta_1 + p_i^2\Delta_0$can be simplified as $(1 - p_i^2)\Delta_2 + (p_i - p_i^2)\Delta_1 + p_i^2$ because $\Delta_0 = 1 - \Delta_1 - \Delta_2$. With the distribution of similarity indices given in Table 2, the vector consisting of the first to $v^{*\mathrm{th}}$ moments of *S* can be expressed with deltas (denoted **E**). By equating observed moments to expected ($\mathbf{E} = \hat{\mathbf{E}}$), the $\hat{\mathbf{\Delta}}$ is obtained (see supplementary materials for details).

Table 2 is about here.

For there is a matrix inversion operation in calculating $\hat{\mathbf{\Delta}}$, the inverted matrix may be ill-conditioned or singular under some conditions, and numerical calculation may bring additional errors (i.e. 10^{-16}), leading to $\hat{r}_{\mathrm{HL}}$ being sensitive to allele frequency. The $\hat{r}_{\mathrm{HL}}$ may be greater than 100 or smaller than −100. We made several attempts to obtain the best estimation, i.e. by setting a valid range for $\hat{r}$, $\hat{\Delta}_i$ or the condition number of **M**, at a specific locus. However, we found a correction abandoning the loci giving estimates of $\hat{r}$ outside of an empirical range of [−16*,* 1] yields the least bias at a locus with few alleles. This can result in some bias if the loci are few (i.e. 5), at least this estimator is asymptotically unbiased with increasing numbers of loci.

Coefficient of coancestry estimators

The three following estimators (RI, WE and LO) aim at estimating the coancestry coefficient presented previously (ϑ, the probability that two alleles, one randomly sampled from each individual, are IBD Jacquard 1972). This quantity ϑ is alternatively defined as the correlation between the additive values of the two individuals (Ritland 1996). 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).

Ritland's (1996) estimator assigns a similarity index (S_i) to a genotypic pair for each of n possible alleles, and we described here an example. S_i is a product of two similarity indices (Hardy & Vekemans 2002): $S_i = S_{xi}S_{yi}$, where S_{xi} and S_{yi} are the frequencies of A_i in individuals x and y, respectively. The multilocus estimators of Ritland (1996), Loiselle *et al.* (1995) and Weir (1996) are:

$$\hat{\theta}_{xy,\mathrm{RI}} = \frac{\sum_j\left(\left(\sum_i S_{xij}S_{yij}/p_{ij}\right) - 1\right)}{\sum_j(n_j - 1)},$$

$$\hat{\theta}_{xy,\mathrm{LO}} = \frac{\sum_j(S_{xij} - p_{ij})(S_{yij} - pij\ )}{\sum_j p_{ij}(1 - p_{ij})},$$

$$\hat{\theta}_{xy,\mathrm{WE}} = \frac{\sum_j(S_{xij}S_{yij} - p_{ij}^2)}{\sum_j(1 - \sum_i p_{ij}^2)}. \tag{6}$$

Here j is the loop variable of the locus, and S_{xij} and S_{yij} are the frequencies of A_i at *j*th locus in x and y, respectively, and n_j is the number of alleles at *j*th locus.

The directional relatedness of the RI and WE estimators, $\hat{r}_{xy}$, can be calculated by using Equation (3), but not for LO, because $\ddot{\theta}_{xx}$ may be negative. Like the ill-conditioned matrix problem for the MOM estimator, when allele frequency in *x* is equal to the population allele frequency, $\ddot{\theta}_{xx} = 0$ and cannot thus be used as the denominator in Equation (3). Similarly, $\hat{r}_{xy}$ and $\hat{r}_{yx}$ cannot be unified by using their geometric mean, because both values may be negative. Alternatively, we calculated their arithmetic mean, because in outbred populations, $r_{xy} = r_{yx}$. Therefore the single-locus relatedness converter of RI and WE estimators are:

$$\hat{r} = \frac{1}{2}\hat{\theta}_{xy}\left(\frac{1}{\hat{\theta}_{xx}} + \frac{1}{\hat{\theta}_{yy}}\right). \tag{7}$$

The $\hat{r}_{\mathrm{HL}}$ is then obtained by substituting Equation (4) into Equation (7):

$$\hat{r}_{\mathrm{HL}} = \frac{2v^*}{v + v^*}\hat{r}.$$

For multilocus estimation, the estimated relatedness is a weighted average of $\hat{r}$ for each locus. The locus specific weight is the inverse of the sum of the expected similarity indexes across all alleles for nonrelatives, which is also the allelic richness at the locus ($1/\sum p_i^2$), and is valid for any levels of ploidy. Using Equation (7) for dyads with the same levels of ploidy, $\hat{r} \leq 1$. For dyads with different levels of ploidy, $\hat{r}_{\mathrm{HL}}$ may exceed one. Furthermore, for biallelic loci the variance of Equation (7) may be high.

In outbred populations, the relatedness coefficient can also be calculated by Equation (5). To make it compatible for different levels of ploidy and to obtain r_{HL}, the genotype of the lower ploidy is extended to the higher by adding $v - v^*$ dummy alleles. The extra allele is randomly generated according to the allele frequency and is uncorrelated to the other alleles or individual, and the expected value of the similarity index is taken into account. Therefore, the expected similarity index S'_{xi} is given by

$$S'_{xi} = S_{xi} + \frac{v - v_x}{v}p_i.$$

By replacing the S_{xi} in Equation (6) by S'_{xi}, the $\hat{r}_{\mathrm{HL}}$ can be solved by Equation (5), and the biallelic locus with $p_i = p_j = 0.5$ can be used. However, the variance of $\hat{r}_{\mathrm{HL}}$ in a multiallelic locus is usually higher than the estimates obtained by Equation (7).

Equation (5) can be used by all three coefficient of coancestry estimators, so we have two approaches for converting the relatedness for RI and WE estimators. After comparing their robustness for empirical data, we use Equation (7) for the RI estimator, and Equation (5) for the WE and LO estimators.

Maximum-likelihood estimator

Jacquard (1972) described a set of nine identical-by-descent modes that describe the possible IBD relationships between the set of four alleles possessed by two diploids. These are denoted $d_0, \ldots, d_8$ and are shown in the first two rows of Figure 1. The probability that a pair of individuals will be in IBD mode d_i is denoted δ_i. Therefore, the coefficient of coancestry is:

$$\theta = \delta_8 + \frac{1}{2}(\delta_8 + \delta_4 + \delta_2) + \frac{1}{4}\delta_1. \qquad (8)$$

δ_1 and δ_2 are both equivalent to the two- and four-gene coefficient. In outbred populations, the two alleles within an individual cannot be IBD, so the last 6 IBD modes are not valid and $\delta_i = 0 \ \ (i = 3, \dots, 8)$, reducing Equation (8) to Equation (1).

For a haplodiploid system, these IBD modes are shown in Figure 1. There are 4 IBD modes in haploid-diploid dyads, and 2 IBD modes between haploids. The coefficient of coancestry in haploiddiploid dyads is $\theta = \delta_3 + \delta_1/2$; similarly, the relatedness coefficient between haploids is $\theta = \delta_1$. In outbred populations, r_{HL} can be obtained by Equation (5).

Following our previously proposed method (Huang *et al.* 2014a), the probabilities of observing each IBS mode, conditioned on the IBD mode between the different levels of ploidy, are listed in Table 3. The probabilities for a haplodiploid system are also shown in Table 3, with the probabilities for higher levels of ploidy given in the supplementary files. However, for higher levels of ploidy, there are many IBD modes when inbreeding is present. For example, there are 109, 1043 and 8405 IBD modes for tetraploids, hexaploids and octoploids, and 29, 162 and 815 IBD modes for diploid-tetraploid, triploid-hexaploid and tetraploid-hexaploid dyads. It would be impractical to compute so many deltas, so that inbreeding or double-reduction is not considered in our maximum-likelihood estimator. If so, the probabilities of different levels of ploidy can be generated from the same levels of ploidy: the corresponding IBS (denoted s') can be found by adding the male genotype with a virtual allele with a frequency of one. This probability can be calculated by the following equation:

$$\Pr(s|d_i) = \Pr(s'|d_i)/\binom{v-i}{v-v^*}.$$

Here, i is the number of IBD allele pairs shared by the two individuals and $0 \le i \le v^*$.

Table 3 is about here.

The single-locus likelihood of a specific relationship (**Δ**) between two individuals, given the observation of IBS mode (*s*), is:

$$\mathrm{L} = \Pr(s|\boldsymbol{\Delta}) = \sum \Pr(s|d_i)\,\delta_i.$$

The multilocus likelihood for unlinked loci is obtained by taking the product of the single-locus likelihoods. The likelihood is then maximized by searching over the parameter space. In inbred populations, the parameter space of **Δ** is $\sum \delta_i = 1$, $0 \le \delta_i \le 1$, whilst in outbred populations, IBD alleles cannot appear in an individual, so $\sum_{i=3}^{8} \delta_i = 0$ in diploids.

Thompson (1976) proposed a constraint for diploids in outbred populations. We extend this to polyploids. However, in arrhenotoky, this constraint only applies to female-female dyads (see supplementary files for details).

To find the maximum likelihood, an algebraic solution is impossible (Milligan 2003). As a result, Nelder & Mead's (1965) Simplex Algorithm is applied to search for $\widehat{\boldsymbol{\Delta}}$ that maximizes the likelihood in parameter space. Then, the relatedness coefficient is obtained by Equation (5).

Simulations and comparisons

Despite the ill-conditioned or singular matrix problem, the MOM estimator is asymptotically unbiased with increasing numbers of loci or alleles. If the number of loci or alleles are few or finite, there is a bias for the MOM or ML estimators. Therefore, the mean square error (MSE) is employed to evaluate the efficiency of the estimators, where $\mathrm{MSE}(\hat{r}_{\mathrm{HL}}) = \mathrm{Bias}^2(\hat{r}_{\mathrm{HL}}) + \mathrm{Var}(\hat{r}_{\mathrm{HL}})$. Therefore, for the three unbiased coefficient of coancestry estimators, $\mathrm{MSE}(\hat{r}_{\mathrm{HL}}) = \mathrm{Var}(\hat{r}_{\mathrm{HL}})$. In simulations, the valid range of $\hat{r}$ of MOM estimator is [-16, 1]. To show the performances of these estimators for higher levels of ploidy, we simulated the haplodiploid and virtual diplotetraploid systems.

The numerical results of these two parameters are obtained from Monte-Carlo simulations assuming allele frequency is known. In these simulations, for each pair of individuals the genotype of one individual is randomly generated according to the Hardy-Weinberg Equilibrium (HWE); the other genotype is then obtained condition upon the reference genotype and their relationships (deltas in Table 1). For example, for diploid dyads, for each locus, we generate a random number t uniformly distributed from 0 to 1: if $0 \leq t \leq \Delta_2$, the second genotype at this locus will be equal to the first; if $\Delta_2 \leq t \leq \Delta_1 + \Delta_2$, one allele is randomly taken from the first genotype, the other is randomly generated according to the allele frequency; if $\Delta_1 + \Delta_2 \leq t \leq 1$, the second genotype is randomly generated according to HWE.

Two types of allele frequency distributions are simulated: (i) triangular and (ii) uniform. These allele frequencies are in proportions 1, 2, ⋯, *n* in the former distribution; for the latter, the frequencies of all alleles at a locus are equal. Two typical applications are simulated: one using ten multiallelic loci, and the other using multiple biallelic loci, assuming the loci are unlinked. To evaluate the performance of the estimators using multiple multiallelic loci, we computed the minimal number of loci that enabled the MSEs to be less than a threshold of 0.01. Finally, two real published datasets are used to evaluate the robustness and efficiency of these estimators under actual conditions.

Estimation with ten multiallelic loci

Here, estimates are obtained from 10 multiallelic loci, and the number of alleles (*n*) range from 2 to 15. For each *n*, both the triangular and uniform allele frequencies are used. When there are more than eleven types of relationships (Table 1), a single figure with all relationships plotted would be difficult to understand. Therefore, five typical relationships are shown, including mother-son or father-daughter, female full-sibs, female-male sibs, female nonrelatives and female-male nonrelatives. The results in terms of MSE are shown in Figure 2, the MSE curves of LO and WE estimators are identical, so are shown in the same row.

In estimating mother-son or father-daughter relationships, the MOM and ML estimators can give accurate estimates with $\mathrm{MSE}(\hat{r}_{\mathrm{HL}}) = 0$, as well as for male or female clonemates (data not shown). Only 4 curves are visible for these two estimators. As mentioned in the methods section, the RI estimator performs worse for biallelic loci, so there is a point of inflexion at *n* = 3 for the RI estimator. For multiallelic loci, the estimates from the RI estimator are better than those of the other two estimators. For some relationships, MSE begins to increase for the RI estimator. The MSEs of the ML estimator are usually the lowest, and quickly reach an asymptote (Figure 2).

Estimation with multiple biallelic loci

To examine the properties of multilocus estimation, we simulated biallelic loci whose number of loci (*l*) ranged from 1 to 100. We did not perform simulations with thousands of loci because all estimators assume independent loci, and unlinked loci are limited in the genome. Like the multiallelic case, allele frequencies exhibiting a triangular distribution or a uniform distribution are considered. Relatedness coefficients were estimated between 30,000 pairs of individuals for 5 relationships, and the same for multiallelic estimation. The resulting MSE of $\hat{r}_{\mathrm{HL}}$ is displayed in Figure 3.

The RI estimator performs worst under the uniform distribution of allele frequencies. A locus cannot give valid estimates when there is a heterozygote in the dyads, so the curves for mother-son or fatherdaughter and female full-sibs relationships are flat for the RI estimator.

For MOM and ML estimators, the female-male and female-female $\mathrm{MSE}(\hat{r}_{\mathrm{HL}})$ curves did not show properties as above, and the MSE among different relationships are different. Some curves are nonmonotonic in MOM and ML estimators because the MSE has two components. For the MOM estimator, uniform distributed allele frequencies are more prone to the singular matrix problem. Their performance was not significantly better than triangular distributed allele frequencies, and their maxima was at a larger *l* (Figure 3).

Requirement for $\mathrm{MSE}(\hat{r}_{\mathrm{HL}}) < 0.01$

We compared the performances of these estimators using multiallelic loci for different levels of ploidy. The minimum number of loci that allowed the MSE of $\hat{r}_{\mathrm{HL}}$ for all eleven types of relationship (as in Table 1) to remain below 0.01 was used to evaluate the performance, and the results are shown in Figure 4. All these estimators are simulated for four types of ploidy, including haplodiploid systems and three virtual systems: diplotetraploid, triplohexaploid and tetraplooctoploid. The results were obtained from 30,000 simulations for each estimator in each type of relationship.

If the number of alleles are fewer than the female ploidy, the unrelated dyads can share IBS allele(s) with a higher probability, which introduces a large positive bias for the ML estimator, especially under the triangular allele frequency distribution. In such case, the MSE can never be reduced to 0.01 by increasing the number of loci used. However, as the number of alleles increases, the number of loci required for the ML estimator rapidly decreases (Figure 4).

In the four systems, the haplodiploid system requires the least *l* to achieve the same accuracy when *n* is low, but as *n* increases, the Min(*l*) of the haplodiploid system decreases much slower than the other three, and requires more loci for *n* > 6 and even increases after $n = 6$ for the RI estimator. As *n* increases, higher levels of ploidy require fewer loci to achieve the same accuracy.

Empirical data

To evaluate the robustness and efficiency of these estimators under real conditions, we use two datasets, one each from Beekman *et al.* (2009) and Kronauer *et al.* (2004) for two species of social insects.

The study species of Beekman *et al.* (2009) was the eusocial Cape honey bee (*Apis mellifera capensis*). Six microsatellites were typed for drones and queens ($10 \le n \le 26$), and four of those six loci were typed for workers. For workers from different colonies, the typed loci were sometimes unequal. Pre-emergent workers and pre-emergent drones were the offspring of the queens, workers or previous queens. In this dataset, we removed data from individuals that were not offspring of the current queen. This resulted in a dataset of 1289 individuals, including mother-daughter, mother-son, brother-sister, sister-sister, brother-brother, and nonrelative (between colony) relationships. Because allelic diversity was high among workers within a same colony, we considered them to be half-sibs.

Kronauer *et al.* (2004) studied African army ants (*Dorylus molestus*), sampling four microsatellites for all 769 individuals ($3 \le n \le 10$). Males from other colonies mate with the queen in a colony to produce workers. By performing a parentage analysis, we consider the male-worker dyads to be 'father-daughter' if the genotypes of the trio (male-queen-worker) match, so a worker may have several 'fathers'. The mismatched male-worker, male-queen and individuals from different colonies are considered to be nonrelatives. Therefore, there are mother-daughter, father-daughter, sister-sister, and nonrelative relationships in the dataset.

In these datasets, each dyad in the two datasets is classified into a relationship, their relatedness is estimated by all five estimators. Results of each relationship in term of mean and standard deviation of $\hat{r}_{\mathrm{HL}}$ are shown in Table 4. The allele frequency is simply calculated by counting alleles in each individual. Table 4 is about here.

The results of these estimators are similar to those of our simulations, with the estimated values similar to the real values. However, bias also appeared for unbiased moment estimators. Judging by the sum of the MSE for all relationships, estimator performance ranked from best to worst is: ML, MOM, RI, LO and WE.

Discussion

Differences in ploidy between individuals in a population can occur both in nature and by artificial inducement, for example in arrhenotoky (Grimaldi 2005; Gullan & Cook 2007; White 1984), pseudo-arrhenotoky (Sabelis & Nagelkerke 1988; Borsa & Kjellberg 1996; Bongiorni *et al.* 2001), polyhaploids (Behnke 1980; Hermsen *et al.* 1981), and gametophytes vs. sporophytes (Shaw & Beer 1999; Gerrienne & Gonez 2011). In this paper, five relatedness coefficient estimators are extended to various mixtures of ploidy. Arrhenotoky is used as an example to compare their performance by Monte-Carlo simulations under ideal conditions, and two real datasets are used to evaluate the robustness of the estimators and their efficiency under real conditions.

MSE

Relatedness estimators typically show high sampling variance due to variance in identity-by-descent among loci and variance in identity-by-state for alleles that are not identical-by-descent (Lynch & Ritland 1999). The MSE can be reduced by increasing the number of loci included in the analysis or by using more polymorphic loci (Figure 2 and 3).

The MSE in the three coefficients of coancestry estimators can be divided into two classes: female-male and female-female. The MSEs as a function of *l* in LO (Loiselle *et al.* 1995) and WE (Weir 1996) estimators for estimating female-male relatedness are similar, and often overlap, whilst the female-female MSEs are also similar (Figure 3). In our modifications of coancestry estimators, the male genotype is padded to have the same alleles as females. The additional alleles are randomly generated, and the expected value for the similarity index is used for the calculation, so they can be treated as fixed. However, using a mother-daughter pair as an example, they share a pair of IBD alleles, but other pairs of alleles are not IBD and are variable. They can be either identical-by-state or not, resulting in variance from the similarity index and the $\hat{r}_{\mathrm{HL}}$. As a result, the MSEs for female-female dyads are larger than for female-male dyads.

The MOM (method of moments) estimator uses a different procedure that subdivides the r_{HL} into many deltas, calculating the probability of each similarity index given the reference genotype and the relationship between the two individuals. The $\widehat{\boldsymbol{\Delta}}$ can be obtained by substituting the observed similarity index moments, and the $\hat{r}_{\mathrm{HL}}$ calculated from Equation (5). The MSEs in the MOM estimator have no such property stated previously.

Bias

The bias is often positive for likelihood relatedness estimators (i.e. Milligan 2003; Anderson & Weir 2007). However, a negative bias appeared in our maximum-likelihood estimators, especially when the number of alleles was fewer than that of the ploidy. Actually, a negative bias occurs when the less related IBD modes mimic related IBS modes with a higher probability. For example, in the haploid-diploid IBD modes, $\Pr(s_1|d_0)/\Pr(s_1|d_1)$ when $p_i > 1/2$ (Table 3). Taking mother-offspring as an example, Table 5 lists the allelic state and the corresponding probability and bias of maximum-likelihood estimation for mother-daughter and mother-son dyads with a biallelic locus assuming the IBD alleles between a mother and her offspring is A_i. The probability that a mother-daughter dyad produces the IBS modes with a positive bias is higher than that with negative bias because $p_i^2 + p_j^2 \geq 2p_ip_j$. On the contrary, the bias in the mother-son dyads is negative except when $p_i = p_j$, because their relatedness lies on the edge of parameter space. This bias can be eliminated by increasing the number of loci used for estimation.

Real conditions

In practice, there are many factors that can reduce the reliability of relatedness estimation such as inbreeding, selection or hitchhiking, linked loci and population subdivision. All of these, except linked loci, will result in deviation in the distributions of genotypes from the Hardy-Weinberg equilibrium. Using linked loci will not increase reliability, and will not result in extra bias, assuming two identical loci are used and the weight of the locus is actually doubled. If each locus is unbiased, no matter the weight, the final weighted average is also unbiased. However, the sampling variance is larger than expected. Furthermore, some genotypes may be mistyped as the result of allelic dropout, false alleles (Taberlet *et al.* 1996), null alleles (Schlötterer & Tautz 1992) or incorrect assignment of alleles.

When there is differentiation between subpopulations, the $\hat{r}_{\mathrm{HL}}$ of the moment estimators will be positively biased for relatives, but negatively biased for individuals from different subpopulations (Wang 2011). Using the number of chromosome sets as the weight for the expected heterozygosity of each colony (Clark & Jasieniuk 2011), Nei's (1973) G_{ST} is 0.361 (0.283 for females only) and 0.115 for data from Beekman *et al.* (2009) and Kronauer *et al.* (2004), respectively. The $\hat{r}_{\mathrm{HL}}$ for nonrelatives estimated by moment estimators is thus negatively biased, whilst the $\hat{r}_{\mathrm{HL}}$ for brother-

brother in Beekman *et al.*'s (2009) data, and sister-sister and father-daughter pairs in Kronauer *et al.*'s (2004) data, are positively biased.

In contrast, the estimates for brother-sister, sister-sister, and mother-daughter pairs in Table 4 are negatively biased, with bias reaching −0.078 for the RI estimators. We noticed that heterozygosity in female bees was high in Beekman *et al.*'s (2009) samples. For females, the observed heterozygosity was 0.830, while the weighted expected heterozygosity across the colony was 0.614. We inferred that the excess of heterozygosity may be due to sex-biased dispersal and outcrossing. A negative bias for relatives occurs because the distribution of offspring genotypes deviates from expected values (homozygote deficiency).

In contrast, males have no father so for the mother-son and brother-brother pairs from Beekman *et al.*'s (2009) data, the estimates were nearly unbiased. The mother-daughter pair estimates from Kronauer *et al.*'s (2004) data were also negatively biased. However, sister-sister estimates were positively biased. We suggest that there were a number of full-sibs among sisters. By a simple correction (the $\hat{r}_{\mathrm{HL}}$ of sister-sister is multiplied by the $\hat{r}_{\mathrm{HL}}$ of the mother daughter pair, then divided by 0.5), we can obtain the percentage of full-sibs as between 4.5% to 12% ($2\hat{r}_{\mathrm{HL}} - 0.5$).

Properties of ploidy

In simulations, the haplodiploid system requires more loci to achieve the same level of accuracy than other systems when the number of alleles is more than six (Figure 4). However, higher levels of ploidy have more copies of a gene, which can present an obstacle if the loci are not sufficiently polymorphic. A significant limitation of the MOM estimator is when few loci are sampled and allele numbers are also few, the matrix in the equation for estimating relatedness in polyploids then becoming ill-conditioned and often non-informative about relatedness. In this case, the MOM estimator performs worse, especially for biallelic loci (Figure 4). Similar problems can also occur in the ML estimator, if the alleles are fewer than the level of ploidy, the resulting bias for unrelated dyads being high. In some cases, a plateau is encountered when the likelihood surface is flat and the value obtained by the likelihood function is indistinguishable from the value returned for nearby regions. In such cases, the numerical algorithms may be unable to determine in which direction it should step, and may give solutions of no improvement.

The estimators used different assumptions, the MOM estimator and the maximum-likelihood estimator, assumed no inbreeding or double-reduction. Whilst the coefficient of coancestry estimators do not assume inbreeding, the "higher-order" coefficients cannot be obtained by them. The moment estimators are essentially unbiased, but due to the ill-conditioned matrices problem in higher levels of ploidy, the MOM estimator shows some bias. The estimates of moment estimators (MOM, RI, LO and WE) may lie out of the range of [0*,* 1]: the MOM estimator can give negative estimates, and the estimates of coefficient of coancestry estimators can be either negative or above unity. Here, their performances are compared in three applications with two types of allele frequency by Monte-Carlo simulations, and they are quite different (Figure 2 and 3).

Although one estimator performs well in some cases, there are conditions under which others perform better according to specific metrics. There is no single estimator that has superior performance under all conditions and by all metrics: the MSE depended on the degree of ploidy, allele frequency, the number of alleles and loci, and the type of relationship between individuals (Figure 2 and 3).

In our study of the two empirical datasets, we found that whilst the ML estimator had the smallest MSE, it was too positively biased for nonrelatives. When loci are few, we suggest using moment estimators instead. For moment estimators, the RI estimator was sensitive to allele frequency (Figure 3) and population subdivision (Table 4), and performed worse for biallelic loci. The variance of the WE and LO estimators were highest (Table 4), but less biased; the bias and variance of the MOM estimator were both small, but may have suffered from the ill-condition matrix problem if the number of alleles was less than the level of ploidy.

The simulations and comparisons we present in this paper are likely to help researchers in choosing the most suitable estimator for their applications. Furthermore, for specific applications under specific genetic conditions, it is possible to identify one optimal estimator. The simulation function is also implemented in our software, so that other researchers can themselves determine the best estimator to use.

Acknowledgement

We thank Prof. Olivier J. Hardy for help in extending the coefficient of coancestry estimators to polyploids and converting the relatedness from the coefficient of coancestry. We also thank Stephanie T Chen for improvements in the manuscript grammar. This study was supported by National Natural Science Foundation of China (NSFC, 31130061, 31270441, 30970379, 31201721); Opening Foundation of Key Laboratory of Resource Biology and Biotechnology in Western China (Northwest University), Ministry of Education (ZS12016), Scientific Research Foundation of the Education Department of Shaanxi Province, China (12JK0837, 14JK1741).

References

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

Barron AB, Oldroyd BP, Ratnieks FL (2001) Worker reproduction in honey-bees (Apis) and the anarchic syndrome: a review. *Behavioral Ecology and Sociobiology*, **50**, 199–208.

Beekman M, Allsopp MH, Jordan LA, Lim J, Oldroyd BP (2009) A quantitative study of worker reproduction in queenright colonies of the Cape honey bee, *Apis mellifera capensis*. *Molecular Ecology*, **18**, 2722–2727.

Behnke M (1980) Selection of dihaploid potato callus for resistance to the culture filtrate of *Fusarium oxysporum*. *Zeitschrift fur Pflanzenzuchtung*, **85**, 254–258.

Bongiorni S, Mazzuoli M, Masci S, Prantera G (2001) Facultative heterochromatization in parahaploid male mealybugs: involvement of a heterochromatin-associated protein. *Development*, **128**, 3809–3817.

Borsa P, Kjellberg F (1996) Experimental evidence for pseudo-arrhenotoky in *Hypothenemus hampei* (Coleoptera: Scolytidae). *Heredity*, **76**, 130–135.

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.

Clark LV, Jasieniuk M (2011) polysat: an R package for polyploid microsatellite analysis. *Molecular Ecology Resources*, **11**, 562–566.

Cockerham CC (1969) Variance of gene frequencies. *Evolution*, pp. 72–84.

Deimling S, Zitzlsperger J, Wenzel G (1988) Somatic fusion for breeding of tetraploid potatoes. *Plant breeding*, **101**, 181–189.

Gerrienne P, Gonez P (2011) Early evolution of life cycles in embryophytes: a focus on the fossil evidence of gametophyte/sporophyte size and morphological complexity. *Journal of Systematics and Evolution*, **49**, 1–16.

Grimaldi D (2005) *Evolution of the Insects*. Cambridge University Press.

Gullan PJ, Cook LG (2007) Phylogeny and higher classification of the scale insects (Hemiptera: Sternorrhyncha: Coccoidea). *Zootaxa*, **1668**, 413–425.

Guo ST, Huang K, Ji WH, Garber PA, Li BG (Inpress) The role of kinship in the formation of a primate multilevel society. *American Journal of Physical Anthropology*, **Inpress**, Inpress.

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. *Molecular Ecology Notes*, **2**, 618–620.

Hermsen JGT, Ramanna MS, Helsop-Harrison J, den Jijs APM (1981) Haploidy and Plant Breeding. *Philosophical Transactions of the Royal Society of London*, **292**, 499–507.

Huang K, Guo ST, Qi XG, Zhang P, Li BG (2014a) A maximum-likelihood estimation of pairwise relatedness for autotetraploids. *Heredity*, **Online**, doi:10.1038/hdy.2014.88.

Huang K, Ritland K, Guo ST, Shattuckn M, Li BG (2014b) A pairwise relatedness estimator for polyploids. *Molecular Ecology Resources*, **14**, 734–744.

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

Kronauer DJC, Sch¨oning C, Pedersen JS, Boomsma JJ, Gadau J (2004) Extreme queen-mating frequency and colony fission in African army ants. *Molecular Ecology*, **13**, 2381–2388.

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

Liu ZJ, Huang CM, Zhou QH *et al.* (2013) Genetic analysis of group composition and relatedness in white-headed langurs. *Integrative Zoology*, **8**, 410–416.

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.

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

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

Nei M (1973) Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences*, **70**, 3321–3323.

Nelder JA, Mead R (1965) A simplex method for function minimization. *Computer Journal*, **7**, 308–313.

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.

Sabelis MW, Nagelkerke CJ (1988) Evolution of pseudo-arrhenotoky. *Experimental & Applied Acarology*, **4**, 301–318.

Schlötterer C, Tautz D (1992) Slippage synthesis of simple sequence DNA. *Nucleic Acids Research*, **20**, 211–215.

Shaw J, Beer SC (1999) Life history variation in gametophyte populations of the moss *Ceratodon purpureus* (Ditrichaceae). *American Journal of Botany*, **86**, 512–521.

Taberlet P, Griffin S, Goossens B *et al.* (1996) Reliable genotyping of samples with very low DNA quantities using PCR. *Nucleic Acids Research*, **24**, 3189–3194.

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 EA (1976) A restriction on. the space of genetic relationships. *Annals of Human Genetics*, **40**, 201–204.

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

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

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

Weir BS (1996) *Genetic data analysis II: methods for discrete population genetic data*. Sinauer Associates.

White MJD (1984) Chromosomal mechanisms in animal reproduction. *Italian Journal of Zoology*, **51**, 1–23.

Data Accessibility

The software POLYRELATEDNESS V1.4, user manual and example dataset are available on Google Project (http://polyrelatedness.googlecode.com).

The simulation program, data, detail of Table 1, the program use to generate the results presented in Table 2 and 3 are detailed in the online supporting information.

The genotypes of data used by Beekman *et al.* (2009) were obtained from Table S1 and S2 in doi: 10.1111/j.1365-294X.2009.04224.x, and the genotype data files of data used by Kronauer *et al.* (2004) were obtained from Appendix S1 in doi: 10.1111/j.1365-294X.2004.02262.x.

Author Contributions

KH, STG and BGL designed the project, KH, DC, YR and XGQ performed the research and wrote the draft, GH and PZ provided the tools and data, KR and DWD checked the model and edited the manuscript.

Figure Legends

Figure 1: Modes of identity-by-descent in a haplodiploid system. The first two rows show the IBD modes for diploids, and the third row and the bottom row show the IBD modes for haploid-diploid and haploid dyads, respectively. In each subfigure, the upper dot(s) represent the allele(s) in one individual, whilst the bottom dot(s) represent the allele(s) in the other individual. The lines indicate alleles that are identical-by-descent.

Figure 2: Multiallelic MSE of $\hat{r}_{\mathrm{HL}}$ as a function of the number of alleles at loci with triangular and uniform allele-frequency distributions. The leftmost two columns display the MSE($\hat{r}_{\mathrm{HL}}$) of a haplodiploid system, whilst the rightmost two columns show the same for diplotetraploid systems. Five estimators are compared, including MOM, RI, LO, WE and ML. Each row shows an estimator. RI and WE estimators give the same estimates for single-locus estimation, so they are shown on a same row. For each estimator, five relationships for haplodiploid and virtual diplohexaploid system were simulated: the solid line "—" denotes mother-son or father-daughter, the dashed line "– –" denotes female full-sibs, the dashed-dot line "– ·" denotes female-male sibs, the dotted line "···" denotes female nonrelatives, and the grey thick line denotes female-male nonrelatives. Results were generated from 100,000 Monte-Carlo simulations.

Figure 3: Multilocus MSE of $\hat{r}_{\mathrm{HL}}$ for biallelic loci with triangular or uniform distributed allele frequencies. The columns, estimators, relationships, and curve shapes are as for Figure 2.

Figure 4: The fewest loci using multiple multiallelic loci that allows for the MSE($\hat{r}_{\mathrm{HL}}$) < 0.01 for the total of eleven relationships listed in Table 1. Four estimators are the same as in Figure 2, for each estimator, the haplodiploid ("—"), diplotetraploid ("– –"), triplohexaploid ("– ·") and tetraplooctoploid ("···") are simulated. Results were obtained from 30,000 Monte-Carlo simulations.

Tables and Figures

Table 1: Relatedness coefficients and deltas of relationships in haplodiploid and virtual diplotetraploid systems

Relationship	r_{HL}	Haplodiploid		Diplotetraploid			
		Δ	ϕ	Δ_4	Δ_3	Δ_2	Δ_1
Female-female:	1	1	0	1	0	0	0

Female clonemates							
Female full-sibs	3/4	1/2	1/2	1/6	2/3	1/6	0
Mother-daughter	1/2	0	1	0	0	1	0
Female half-sibs	1/4	0	1/2	0	0	1/6	2/3
Unrelated females	0	0	0	0	0	0	0
Female-male:							
Mother-son, daughter-father	1/2		1			1	0
Female-male sibs	1/4		1/2			1/6	2/3
Unrelated female-male	0		0			0	0
Male-male:							
Male clonemates	1		1			1	0
Male sibs	1/2		1/2			1/6	2/3
Unrelated males	0		0			0	0

Table 2: The coefficients of deltas in the reference genotype in haplodiploid system

Proband ploidy-Reference genotype	Similarity index	Coefficients of deltas		
		Δ_2	Δ_1	1
Diploid-A_iA_i	1	$1 - p_i^2$	$p_i - p_i^2$	p_i^2
Diploid-A_iA_i	1/2	$-2p_ip_*$	$p_*(1 - 2p_i)$	$2p_ip_*$
Diploid-A_iA_i	0	$-p_*^2$	$-p_*^2$	p_*^2
Diploid-A_iA_j	1	$1 - 2p_ip_j$	$(p_i + p_j)/2 - 2p_ip_j$	$2p_ip_j$
Diploid-A_iA_j	1/2	$-p_i(p_i + 2p_*)$	$p_* + p_i(1/2 - p_i - 2p_*) + p_j(1/2 - p_j - 2p_*$	$p_i(p_i + 2p_*)$

Diploid-A_iA_j	0	$-p_*^2$	$-p_*^2$	p_*^2
Haploid-A_iA_i	1/2		$1 - p_i$	p_i
Haploid-A_iA_i	0		$-p_*$	p_*
Haploid-A_iA_j	1/2		$1 - p_i - p_j$	$p_i + p_j$
Haploid-A_iA_j	0		$-p_*$	p_*
Diploid-A_i	1/2		p_*^2	$p_i(p_i + 2p_*)$
Diploid-A_i	0		$-p_*^2$	p_*^2
Haploid-A_i	1		$1 - p_i$	p_i
Haploid-A_i	0		$-p_*$	p_*

Table 3: Probability of patterns of identity-in-state given modes of identity-by-descent

IBS mode	Allelic state									
		d_8	d_7	d_6	d_5	d_4	d_3	d_2	d_1	d_0
Diploid-diploid:										
s_8	A_iA_i-A_iA_i	p_i	p_i^2	p_i^2	p_i^3	p_i^2	p_i^3	p_i^2	p_i^3	p_i^4
s_7	A_iA_i-A_jA_j	0	p_ip_j	0	$p_ip_j^2$	0	$p_i^2p_j$	0	0	$p_i^2p_j^2$
s_6	A_iA_i-A_iA_j	0	0	p_ip_j	$2p_i^2p_j$	0	0	0	$p_i^2p_j$	$2p_i^3p_j$
s_5	A_iA_i-A_jA_k	0	0	0	$2p_ip_jp_k$	0	0	0	0	$2p_i^2p_jp_k$
s_4	A_iA_j-A_iA_i	0	0	0	0	p_ip_j	$2p_i^2p_j$	0	$p_i^2p_j$	$2p_i^3p_j$

s_3	A_iA_j-A_kA_k	0	0	0	0	0	$2p_ip_jp_k$	0	0	$2p_ip_jp_k^2$
s_2	A_iA_j-A_iA_j	0	0	0	0	0	0	$2p_ip_j$	$p_i^2p_j+p_ip_j^2$	$4p_i^2p_j^2$
s_1	A_iA_j-A_iA_k	0	0	0	0	0	0	0	$p_ip_jp_k$	$4p_i^2p_jp_k$
s_0	A_iA_j-A_kA_l	0	0	0	0	0	0	0	0	$4p_ip_jp_kp_l$
Diploid-haploid:										
s_3	A_iA_i-A_i						p_i	p_i^2	p_i^2	p_i^3
s_2	A_iA_i-A_j						0	p_ip_j	0	$p_i^2p_j$
s_1	A_iA_j-A_i						0	0	p_ip_j	$2p_i^2p_j$
s_0	A_iA_j-A_k						0	0	0	$2p_ip_jp_k$
Haploid-haploid:										
s_1	A_i-A_i								p_i	p_i^2
s_0	A_i-A_j								0	p_ip_j

Table 4: Mean and standard deviation of $\hat{r}_{\mathrm{HL}}$ in dataset from Beekman *et al*. (2009) and Kronauer *et al*. (2004)

Relation	Count	r_{HL}	MOM	RI	LO	WE	ML
Beekman *et al*.'s (2009) data							
MD	714	0.5	0.476±0.068	0.412±0.153	0.440±0.157	0.470±0.139	0.526±0.052
MS	565	0.5	0.500±0.000	0.489±0.061	0.489±0.097	0.499±0.065	0.500±0.000
SS	35192	0.25	0.245±0.234	0.209±0.236	0.272±0.280	0.264±0.282	0.264±0.219
BS	32613	0.25	0.233±0.178	0.215±0.177	0.233±0.190	0.242±0.194	0.237±0.162
BN	19000	0.5	0.510±0.277	0.510±0.268	0.503±0.269	0.514±0.261	0.513±0.258
FF-NR	225820	0	−0.057±0.186	−0.044±0.173	−0.047±0.199	−0.068±0.217	0.057±0.133
FM-NR	375882	0	−0.029±0.132	−0.029±0.138	−0.032±0.142	−0.027±0.162	0.041±0.094

MM-NR	140330	0	−0.048±0.153	−0.048±0.185	−0.057±0.169	−0.046±0.182	0.044±0.123
Kronauer et al.'s (2004) data							
MD	459	0.5	0.484±0.116	0.478±0.202	0.461±0.232	0.461±0.279	0.537±0.092
FD	571	0.5	0.500±0.000	0.503±0.116	0.514±0.160	0.496±0.189	0.500±0.000
SS	7574	0.25	0.264±0.294	0.275±0.289	0.286±0.325	0.283±0.365	0.320±0.249
FF-NR	83344	0	−0.033±0.251	−0.030±0.230	−0.030±0.248	−0.035±0.329	0.096±0.165
FM-NR	145377	0	−0.006±0.168	−0.005±0.184	−0.003±0.182	−0.027±0.249	0.064±0.110
MM-NR	57971	0	0.007±0.234	−0.014±0.275	0.012±0.267	−0.032±0.286	0.097±0.168

Description of ‘Relation’ column: ‘MD’: mother-daughter; ‘MS’: mother-son; ‘SS’: sister-sister; ‘BS’:

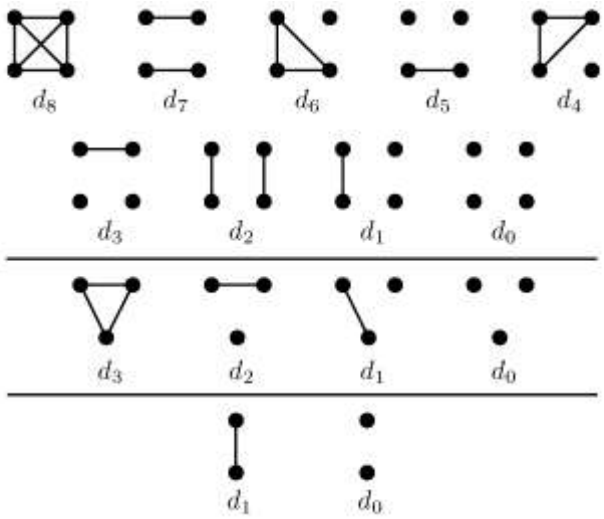
brother-sister; ‘BB’: brother-brother; ‘FF-NR’: female-female nonrelatives; ‘FM-NR’: female-male

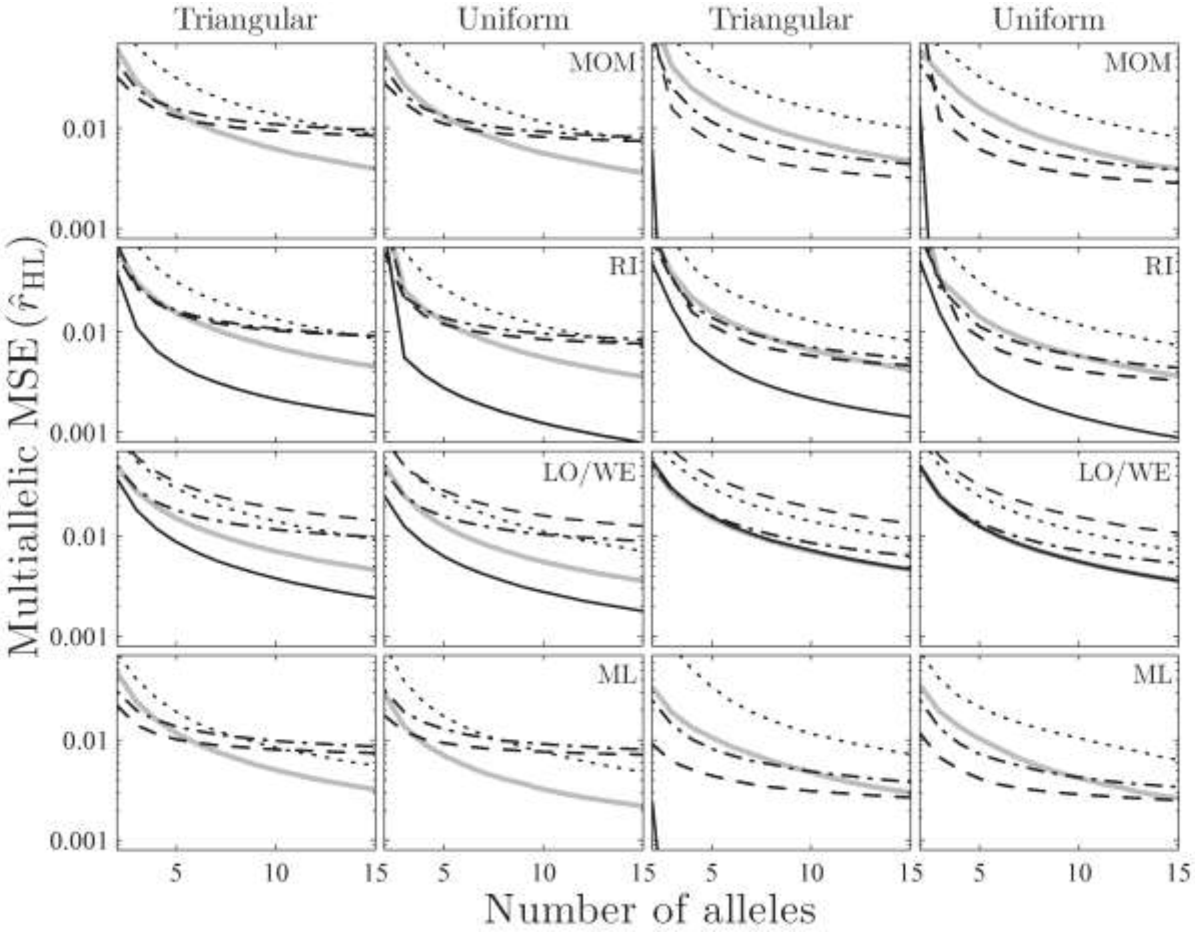
nonrelatives; ‘MM-NR’: male-male nonrelatives, and ‘FD’ denotes father-daughter.

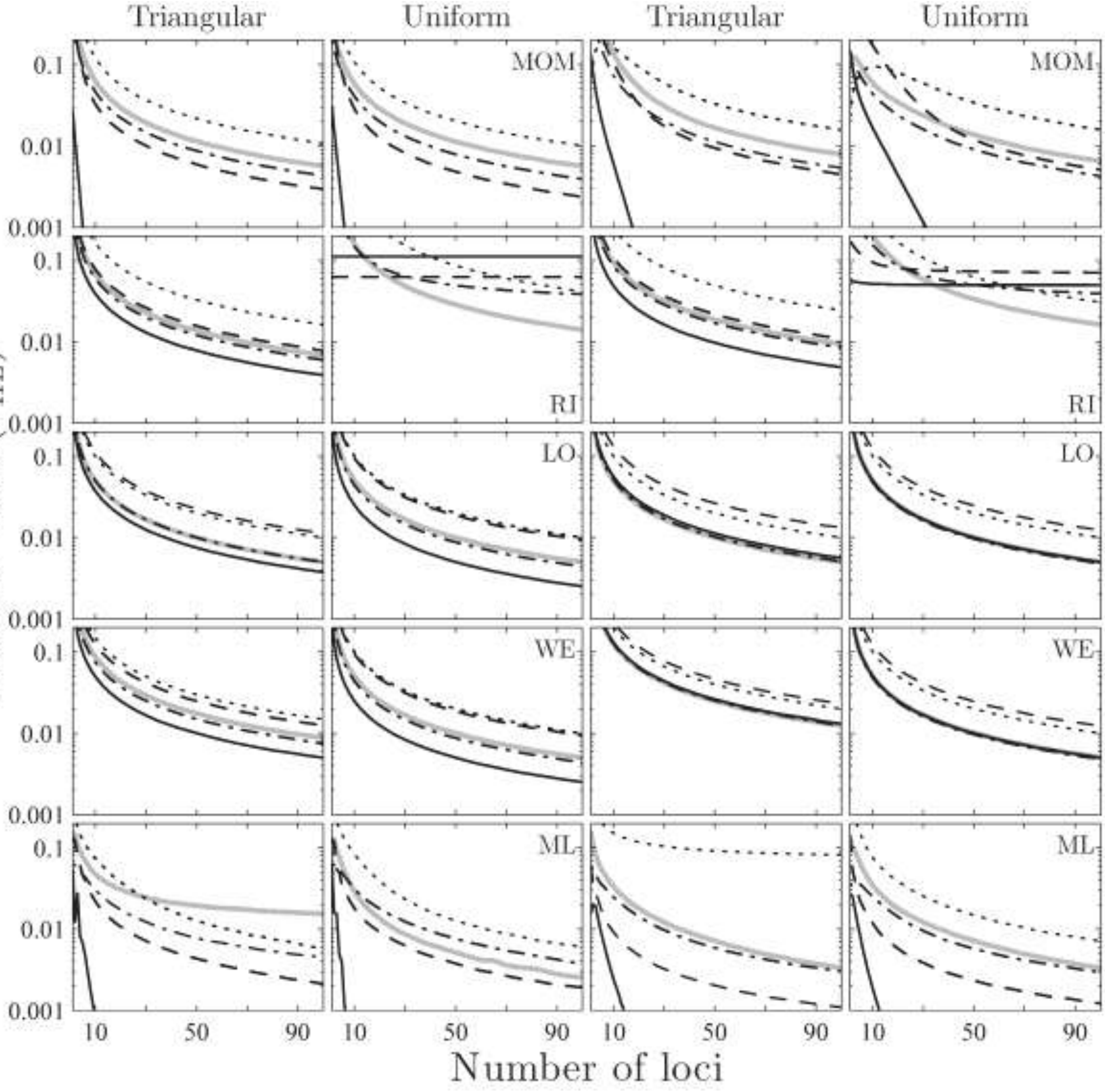
Table 5: Bias of maximum-likelihood estimation for mother-offspring dyads with a biallelic locus, where

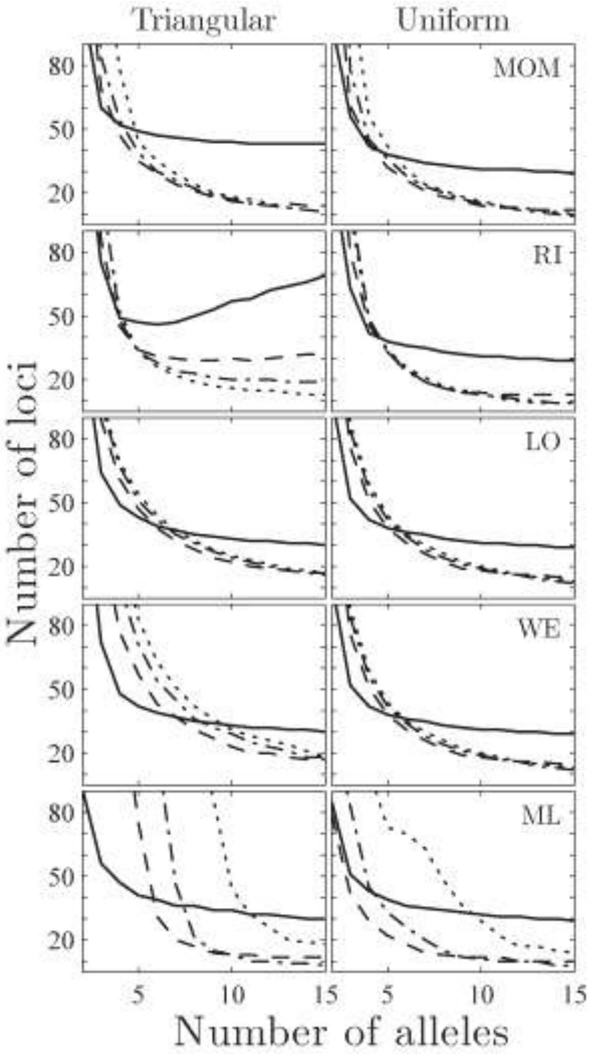
A_i is the IBD alleles between the mother and the offspring

Allelic state	Probablity	Bias
Mother-daughter:		
A_iA_i-A_iA_i	p_i^2	Positive
A_iA_i-A_iA_j	p_ip_j	Negative if $p_i > 0.5$
A_iA_j-A_iA_i	p_ip_j	Negative if $p_i > 0.5$
A_iA_j-A_iA_j	p_j^2	Positive
Mother-son:		
A_iA_i-A_i	p_i	Zero
A_iA_j-A_i	p_j	Negative if $p_i > 0.5$


d_8
d_7
d_6
d_5
d_4
d_3
d_2
d_1
d_0
d_3
d_2
d_1
d_0
d_1
d_0


Triangular
Uniform
Triangular
Uniform
MOM
RI
LO/WE
ML
0.01
0.001
Multiallelic MSE ($\hat{r}_{\mathrm{HL}}$)
5
10
15
Number of alleles


Triangular
Uniform
Triangular
Uniform
MOM
MOM
RI
RI
LO
LO
WE
WE
ML
ML
0.1
0.01
0.001
Biallelic MSE $(\hat{r}_{\mathrm{HL}})$
10
50
90
Number of loci


Triangular
Uniform
MOM
RI
LO
WE
ML
80
50
20
5
10
15
Number of loci
Number of alleles