Performing Parentage Analysis in the Presence of Inbreeding and Null Alleles

Kang Huang,*,[1] Rui Mi,*,[1] Derek W. Dunn,* Tongcheng Wang,* and Baoguo Li*,†,[2]
*Shaanxi Key Laboratory for Animal Conservation, College of Life Sciences, Northwest University, Xi'an 710069, China and
†Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China
ORCID IDs: 0000-0002-8357-117X (K.H.); 0000-0001-5909-1224 (D.W.D.); 0000-0001-7430-3889 (B.L.)

ABSTRACT Parentage analysis is an important method that is used widely in zoological and ecological studies. Current mathematical models of parentage analyses usually assume that a population has a uniform genetic structure and that mating is panmictic. In a natural population, the geographic or social structure of a population, and/or nonrandom mating, usually leads to a genetic structure and results in genotypic frequencies deviating from those expected under the Hardy-Weinberg equilibrium (HWE). In addition, in the presence of null alleles, an observed genotype represents one of several possible true genotypes. The true father of a given offspring may thus be erroneously excluded in parentage analyses, or may have a low or negative LOD score. Here, we present a new mathematical model to estimate parentage that includes simultaneously the effects of inbreeding, null alleles, and negative amplification. The influences of these three factors on previous model are evaluated by Monte-Carlo simulations and empirical data, and the performance of our new model is compared under controlled conditions. We found that, for both simulated and empirical data, our new model outperformed other methods in many situations. We make available our methods in a new, free software package entitled PARENTAGE. This can be downloaded via http://github.com/huangkang1987/parentage.

KEYWORDS Parentage analysis; inbreeding; population subdivision; null alleles; LOD score; negative amplification

THE use of genetic markers to investigate the relationships between individuals is common in studies of animal populations (Goodnight and Queller 1999), and various methods have provided much insight into animal reproductive biology and population structure that would be difficult or impossible to obtain from observation alone (Kalinowski *et al.* 2007). The most common of these techniques, parentage analyses, enables researchers to obtain data on mating systems (Monteiro *et al.* 2017), social organization (Garber *et al.* 2016), reproductive success (Gerzabek *et al.* 2017), multi-generational survival (Cremona *et al.* 2017), sexual selection (Johannesson *et al.* 2016), and kin selection (Dias *et al.* 2017).

Current parentage analysis methods assume that population genotypic frequencies accord with those of the Hardy-Weinberg equilibrium (HWE) (Marshall *et al.* 1998; Kalinowski *et al.* 2007). Such an assumption implies the absence of both close inbreeding (due to mating between relatives, such as siblings) and pervasive inbreeding (due to genetic drift in a finite population or population subdivision) (Wang 2011). Therefore, current methods only allow the frequency of a genotype, or the transitional probability from a parental genotype to an offspring's genotype, to be calculated based on the HWE and basic Mendelian inheritance, but do not allow for the inclusion of both inbreeding factors.

Both artificial and natural populations are finite and are usually genetically structured. Mating is also usually confined to a subset of individuals within a population (Wang 2011). Thus, both types of inbreeding (close and pervasive) may exist, and more extreme forms of close inbreeding, such as back-crossing, may also be present. Depending on the mating system and other ecological factors, a false father may be a potential mate of the true mother. For close inbreeding, the false father may be related to the true mother. For pervasive inbreeding, the false father may come from the same population as the true mother. Hence, a false parent may share identical-by-descent (IBD) alleles with the offspring, and may thus be mistakenly identified as the true parent.

doi: https://doi.org/10.1534/genetics.118.301592
Manuscript received April 11, 2018; accepted for publication October 16, 2018; published Early Online October 18, 2018.
Available freely online through the author-supported open access option.
Supplemental material available at Figshare: https://doi.org/10.25386/genetics.7221965.

[1] These authors contributed equally to this work.
[2] Corresponding author: Northwest University No.229 Taibai North Ave., Xi'an 710069, China. E-mail: baoguoli@nwu.edu.cn

In addition, microsatellites are the most frequently used genetic marker for parentage analyses, but null alleles are pervasive in microsatellite markers (Kalinowski *et al.* 2006; Ravinet *et al.* 2016). Such alleles cause two types of genotyping problems: (i) a homozygote A_yA_y fails to be amplified, where A_y is a null allele; (ii) a heterozygote A_iA_y is mistyped as a homozygote A_iA_i, where A_i is a visible allele (Wagner *et al.* 2006). These incorrect genotypes can be problematic for parentage analyses, because such genotyping errors can mistakenly reject a true parent due to an observed lack of the shared alleles with the offspring (Blouin 2003). Moreover, negative amplification reduces the accuracy of parentage analysis because of the loss of genotypic data. When the genotype of an individual fails to be amplified, all genotypes at this locus in a duo or a trio will be discarded from the analysis.

In this paper, we consider the effects of inbreeding, null alleles, and negative amplification in a parentage analysis. We will first extend the model of Kalinowski *et al.* (2007) to an alternative model, so as to accommodate the effects of these three factors. Second, we use a simulated dataset to evaluate the influences of these three factors on the model of Kalinowski *et al.* (2007), and the performance is also compared with those of our alternative model and another model presented by Wang (2016). Finally, we use a real microsatellite genotyping dataset to test and compare four applications using the models of Kalinowski *et al.* (2007), Wang (2016), and our new model in natural situations. Our model can be applied to any codominant markers that may be affected by inbreeding and/or null alleles. We make available a free software package entitled PARENTAGE v1.0, which can be downloaded via http://github.com/huangkang1987/parentage.

Theory and Modeling

Genotypic frequencies

Under the HWE, alleles appear randomly within a genotype according to their frequencies. The frequency of a genotype G can be expressed as a piecewise function:

$$\Pr(G) = \begin{cases} p_i^2 & \text{if } G = A_iA_i, \\ 2p_ip_j & \text{if } G = A_iA_j, \end{cases} \tag{1}$$

where A_i and A_j denote the i^{th} and j^{th} alleles, respectively, which are different identical-by-state (IBS) alleles, and p_i and p_j are their frequencies.

If the inbreeding in a population is more frequent than random, the homozygosity of a population is increased. We use the inbreeding coefficient f (also known as Wright's F_{IS}) to measure the degree of inbreeding, which is defined as the correlation between the frequencies of two alleles within an individual. According to Equation 1, the frequency of G in the presence of inbreeding is given by

$$\Pr(G|f) = \begin{cases} fp_i + (1-f)p_i^2 & \text{if } G = A_iA_i, \\ 2(1-f)p_ip_j & \text{if } G = A_iA_j. \end{cases} \tag{2}$$

In the presence of null alleles, for an observed genotype (denoted by $\mathcal{O}$) A_iA_i, the actual genotype may be a heterozygote A_iA_y or a homozygote A_iA_i, where A_y is a null allele. If an observed genotype has no any detected alleles, it is termed *negative*, denoted by $\varnothing$. Let p_y be the frequency of the null allele A_y. According to Equation 2, the frequency of an observed genotype $\mathcal{O}$ in the presence of both inbreeding and null alleles is

$$\Pr\left(\mathcal{O}|f,p_y\right) = \begin{cases} fp_i + (1-f)\left(2p_ip_y + p_i^2\right) & \text{if } \mathcal{O} = A_iA_i, \\ 2(1-f)p_ip_j & \text{if } \mathcal{O} = A_iA_j, \\ fp_y + (1-f)p_y^2 & \text{if } \mathcal{O} = \varnothing. \end{cases} \tag{3}$$

Furthermore, in the presence of null alleles and negative amplification, a negative observed genotype $\varnothing$ may arise from either a null allele homozygote A_yA_y or a negative amplification (Kalinowski *et al.* 2006). Let β be the negative amplification rate. Then, under the three factors: inbreeding, null alleles, and negative amplification, Equation 3 should be modified as follows:

$$\Pr\left(\mathcal{O}|f,p_y,\beta\right) = \begin{cases} (1-\beta)\left[fp_i + (1-f)\left(2p_ip_y + p_i^2\right)\right] & \text{if } \mathcal{O} = A_iA_i, \\ 2(1-\beta)(1-f)p_ip_j & \text{if } \mathcal{O} = A_iA_j, \\ \beta + (1-\beta)\left[fp_y + (1-f)p_y^2\right] & \text{if } \mathcal{O} = \varnothing. \end{cases} \tag{4}$$

Procedures of parentage analysis

There are three typical categories of parentage analysis: (i) identifying the father when the mother is unknown; (ii) identifying the father when the mother is known; and (iii) identifying the father and mother jointly. The procedures of a parentage analysis are roughly as follows:

For each of the first two categories, two hypotheses are established: the first hypothesis is that the alleged father is the true father, denoted by H_1; the alternative hypothesis is that the alleged father is not the true father, denoted by H_2. For the third category, "father" needs to be altered to "parents" in hypotheses H_1 and H_2.

Given a hypothesis H, the likelihood is defined as the probability of some observed data given H, written as $\mathcal{L}(H)$. Returning to H_1 and H_2 in the previous paragraph, we refer to the logarithm of the ratio of $\mathcal{L}(H_1)$ to $\mathcal{L}(H_2)$ as the LOD score (abbreviated to LOD); symbolically $\text{LOD} = \ln \mathcal{L}(H_1)/\mathcal{L}(H_2)$, in other words, $\text{LOD} = \ln \mathcal{L}(H_1) - \ln \mathcal{L}(H_2)$. Moreover, a positive LOD score means that the first hypothesis H_1 is more likely to be true than the second hypothesis H_2. Similarly, a negative LOD score means that H_2 is more likely to be true than H_1.

Marshall *et al.* (1998) provided a statistic Δ for resolving paternity. Let LOD_1 and LOD_2 be the LOD scores of the most-likely and the next most-likely alleged fathers, respectively, and let n be the number of all alleged fathers. Then, Δ is defined as follows:

$$\Delta = \begin{cases} \text{LOD}_1 - \text{LOD}_2 & \text{if } n \geq 2, \\ \text{LOD}_1 & \text{if } n = 1, \\ \text{undefined} & \text{if } n = 0. \end{cases}$$

A separate statistic Δ has to be calculated for each individual offspring.

Monte-Carlo simulations are subsequently used to assess the confidence level of each value of Δ. The symbol $\Delta_{0.99}$ represents the threshold of Δ to reach the correct assignment rate of 99%, in the sense that, if $\Delta \geq \Delta_{0.99}$, it implies up to a confidence level of 99%. In other words, the proportion 99% of assignments is correct if $\Delta \geq \Delta_{0.99}$.

The Ka-model

Kalinowski *et al.* (2007) developed a model of parentage analyses, called the Ka-model for short, which accommodates the effect of genotyping errors. This model consists of two likelihood formulas (see Equation 5 below), together with the rules and methods for a general parentage analysis.

As stated in the previous section, the procedures of using Ka-model to conduct a parentage analysis are as follows: (i) calculating $\mathcal{L}(H_1)$ and $\mathcal{L}(H_2)$, (ii) calculating LOD and Δ, (iii) finding the thresholds of Δ, and (iv) using the values obtained in the previous three steps to determine the significance of the parentage analysis.

We will here use the first category in a parentage analysis (*i.e.*, identifying the father when the mother is unknown) as an example to show how to calculate the likelihoods $\mathcal{L}(H_1)$ and $\mathcal{L}(H_2)$ with the consideration of genotyping errors. The two likelihoods in Ka-model are expressed as the following formulas:

$$\begin{aligned} \mathcal{L}(H_1) &= \Pr(\mathcal{O}_A)\big[(1-e)^2 T(\mathcal{O}_O|\mathcal{O}_A) \\ &\qquad + 2e(1-e)\Pr(\mathcal{O}_O) + e^2\Pr(\mathcal{O}_O)\big], \\ \mathcal{L}(H_2) &= \Pr(\mathcal{O}_A)\big[(1-e)^2 \Pr(\mathcal{O}_O) \\ &\qquad + 2e(1-e)\Pr(\mathcal{O}_O) + e^2\Pr(\mathcal{O}_O)\big], \end{aligned} \tag{5}$$

where e is the genotyping error rate, $\mathcal{O}_O$ and $\mathcal{O}_A$ are respectively the observed genotypes of the offspring and the alleged father, $\Pr(\mathcal{O}_O)$ and $\Pr(\mathcal{O}_A)$ are their frequencies, and $T(\mathcal{O}_O|\mathcal{O}_A)$ is the transitional probability from $\mathcal{O}_A$ to $\mathcal{O}_O$.

Denote G_O, G_A, and G_F for the genotypes of the offspring, the alleged father and the true father, respectively. For the term $(1-e)^2 T(\mathcal{O}_O|\mathcal{O}_A)$, both genotypes of the offspring and the alleged father are assumed to be correctly genotyped, then $\mathcal{O}_O = G_O$ and $\mathcal{O}_A = G_A$. Therefore, the expression $T(\mathcal{O}_O|\mathcal{O}_A)$ can be rewritten as $T(G_O|G_F)$ when H_1 holds.

Under the assumption of the HWE, for the genotype G_O, one allele is randomly inherited from the parent, and the other is randomly sampled from the population according to the allele frequencies. Then, the transitional probability $T(G_O|G_F)$ can be expressed as

$$T(G_O|G_F) = \begin{cases} p_i & \text{if } G_O = A_iA_i \text{ and } G_F = A_iA_i, \\ p_j & \text{if } G_O = A_iA_j \text{ and } G_F = A_iA_i, \\ (p_i + p_j)/2 & \text{if } G_O = A_iA_j \text{ and } G_F = A_iA_j, \\ \dfrac{1}{2}p_k & \text{if } G_O = A_iA_k \text{ and } G_F = A_iA_j, \\ 0 & \text{otherwise,} \end{cases}$$

where A_i, A_j, and A_k are non-IBS alleles, p_i, p_j, and p_k are their frequencies. According to the above analyses, the final formula can be used to calculate the value of $T(\mathcal{O}_O \mid \mathcal{O}_A)$ in Equation 5.

The genotyping error can be considered as the replacement of the true genotype with a random genotype at a probability of e. The conditional probability of an observed genotype $\mathcal{O}$ given a genotype G is given by

$$\Pr(\mathcal{O}|G) = \begin{cases} (1-e) + e\Pr(\mathcal{O}) & \text{if } G = \mathcal{O}, \\ e\Pr(\mathcal{O}) & \text{if } G \neq \mathcal{O}. \end{cases} \tag{6}$$

Thus, the genotyping error does not change the observed genotypic frequencies; in other words, $\Pr(\mathcal{O}) = \Pr(G = \mathcal{O})$. Because any null alleles, negative amplification and inbreeding are not considered in the Ka-model, $\Pr(G)$ can be directly calculated by Equation 1, and so the values of $\Pr(\mathcal{O}_A)$ and $\Pr(\mathcal{O}_O)$ in Equation 5 can be obtained.

Alternative forms of likelihoods

The likelihoods $\mathcal{L}(H_1)$ and $\mathcal{L}(H_2)$ in Equation 5 can be obtained by taking the weighted sum of products of the corresponding frequencies of $\mathcal{O}_O$ and $\mathcal{O}_A$ conditional on their genotypes, with their genotypic frequencies as the weights. Then, Equation 5 can be rewritten to the following alternative forms:

$$\begin{aligned} \mathcal{L}(H_1) &= \sum_{OFM} \Pr(G_F, G_M) T(G_O|G_F, G_M) \Pr(\mathcal{O}_O|G_O) \Pr(\mathcal{O}_A|G_F), \\ \mathcal{L}(H_2) &= \sum_{OA} \Pr(G_O) \Pr(G_A) \Pr(\mathcal{O}_O|G_O) \Pr(\mathcal{O}_A|G_A), \end{aligned} \tag{7}$$

where G_O, G_F, G_M, and G_A are, respectively, taken from all possible genotypes of the offspring, the true father, the true mother, and the false father; $\Pr(G_F, G_M)$ is the joint distribution of G_F and G_M; and $T(G_O|G_F, G_M)$ is the transitional probability from G_F and G_M to G_O. Additionally, the three conditional probabilities in Equation 7 can be calculated by Equation 6.

In the absence of inbreeding, if matings are random, then the genotypes G_F and G_M will be independent to each other. Thus

$$\Pr(G_F, G_M) = \Pr(G_F)\Pr(G_M).$$

Additionally, if the genotypes of both parents are known, then the distribution of the genotype of offspring can be derived by Mendelian segregation in the sense that each parent randomly contributes one allele to the genotype of an offspring. Thus,

$$T(G_O|G_F, G_M) = \frac{1}{4}\sum_{i=1}^{2}\sum_{j=1}^{2} K_{G_O, a_ia_j}, \tag{8}$$

where a_ia_j is a possible offspring's genotype, in which a_i is the i^{th} allele copy in G_F and a_j is the j^{th} allele copy in G_M; and K_{G_O, a_ia_j} is a Kronecker operator, such that $K_{G_O, a_ia_j} = 1$ if $G_O = a_ia_j$, and $K_{G_O, a_ia_j} = 0$ if $G_O \neq a_ia_j$.

Schemes of our model

Our model will be established by giving several likelihood formulas based on Equation 7. We use the first category in a parentage analysis as an example to describe the scheme of establishing our model. For the second and third categories, the schemes are presented in Appendices A and B.

In order to simultaneously accommodate the effects of both inbreeding and null alleles together with negative amplification, Equation 7 needs to be modified by replacing the probabilities with those under these effects, including the genotypic frequencies (*e.g.*, $\Pr(G_O)$), the transitional probability $T(G_O|G_F,G_M)$, and the conditional probabilities (*e.g.*, $\Pr(\mathcal{O}_O|G_O)$). It is noteworthy that the alternative hypothesis H_2 should be modified if inbreeding is involved. For close inbreeding, a false father may be a relative of the true mother; for pervasive inbreeding, a false father may be sampled from a same population as the true mother. The genotype of a false father may therefore be either dependent or independent of the true mother in both of these inbreeding scenarios. For identifying an alleged father, the hypothesis H_2 will thus imply two possibilities: (i) he is unrelated to the true parents and to the offspring, denoted by $H_{2,1}$; (ii) he is a relative of the true mother, denoted by $H_{2,2}$.

Because we lack *a priori* information (*e.g.*, information about the pedigree, mating system, population origin, or allele frequency of each population), we cannot determine which of the alternatives $H_{2,1}$ and $H_{2,2}$ is most likely. We thus define the likelihood of H_2 as the geometrical mean of $\mathcal{L}(H_{2,1})$ and $\mathcal{L}(H_{2,2})$. Then, using Equation 7, the likelihoods of H_1 and H_2 can be calculated by the following formulas:

$$\begin{aligned}
\mathcal{L}(H_1) &= \sum_{OFM}\Pr(G_F,G_M|f)T(G_O|G_F,G_M)\\
&\quad \Pr\Big(\mathcal{O}_O|G_O,f,p_y,\beta,e\Big)\\
&\quad \Pr\Big(\mathcal{O}_A|G_F,f,p_y,\beta,e\Big),\\
\mathcal{L}(H_2) &= \sqrt{\mathcal{L}(H_{2,1})\mathcal{L}(H_{2,2})},\\
\mathcal{L}(H_{2,1}) &= \sum_{OA}\Pr(G_O|f)\Pr(G_A|f)\\
&\quad \Pr\Big(\mathcal{O}_O|G_O,f,p_y,\beta,e\Big)\\
&\quad \Pr\Big(\mathcal{O}_A|G_A,f,p_y,\beta,e\Big),\\
\mathcal{L}(H_{2,2}) &= \sum_{OMA}\Pr(G_A,G_M|f)T(G_O|G_M,f)\\
&\quad \Pr\Big(\mathcal{O}_O|G_O,f,p_y,\beta,e\Big)\\
&\quad \Pr\Big(\mathcal{O}_A|G_A,f,p_y,\beta,e\Big),
\end{aligned} \tag{9}$$

where $\Pr(G_F,G_M|f)$ and $\Pr(G_A,G_M|f)$ are the joint distributions of genotypes of mates, and $T(G_O|G_M,f)$ is the transitional probability from G_M to G_O and is conditional on an inbreeding coefficient of f. These joint distributions and the transitional probability will be derived in the next section. Moreover, we can calculate $T(G_O|G_F,G_M)$ by Equation 8, and $\Pr(G_O|f)$ and $\Pr(G_A|f)$ by Equation 2. Additionally, the values of $\Pr(\mathcal{O}_O|G_O,f,p_y,\beta,e)$, $\Pr(\mathcal{O}_A|G_F,f,p_y,\beta,e)$ and $\Pr(\mathcal{O}_A|G_A,f,p_y,\beta,e)$ can all be calculated by the following formula:

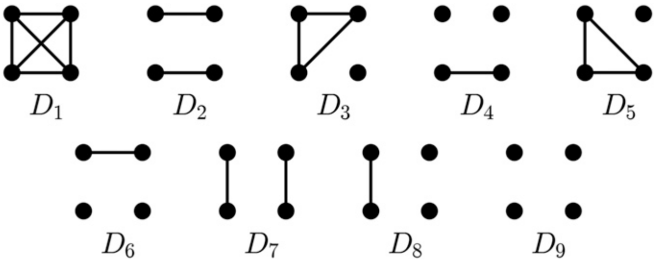

Figure 1 Configurations of IBD alleles between two diploids. For each configuration, we denote the upper two dots for the two alleles of one individual, and the lower two dots for those of the other individual. Moreover, two dots connected by a line indicate that two alleles are IBD.

$$\begin{aligned}
\Pr\left(\mathcal{O}|G,f,p_y,\beta,e\right) = \beta K_{\mathcal{O},\varnothing} + (1-\beta)\left[K_{\mathcal{O},\mathcal{O}^*}(1-e)\right.\\
\left. + e\Pr\left(\mathcal{O}|f,p_y\right)\right],
\end{aligned}$$

where $K_{\mathcal{O},\varnothing}$ and $K_{\mathcal{O},\mathcal{O}^*}$ are two Kronecker operators, in which $\mathcal{O}^*$ is the observed genotype of G accounting for the effect of null alleles (without accounting for the effects of genotyping error and negative amplification), whose expression is

$$\mathcal{O}^* = \begin{cases} A_iA_i & \text{if } G = A_iA_i \text{ or } A_iA_y,\\ A_iA_j & \text{if } G = A_iA_j,\\ \varnothing & \text{if } G = A_yA_y. \end{cases} \tag{10}$$

Joint distributions of genotypes

In the presence of close inbreeding, both parents will be related to each other. Their genotypes will thus be correlated. For example, G_F always shares an IBD allele with G_M in backcrossing. It is noteworthy that there are several forms of mating, *e.g.*, self-fertilization, and matings between parents and offspring (backcrossing), full-siblings, half-siblings, or other relatives, such that the joint distributions of genotypes of a parent pair among these forms of mating will differ, even if their inbreeding coefficients are equal.

Jacquard (1972) defined nine configurations of IBD alleles between two individuals (denoted by $D_1, D_2, \cdots, D_9$, see Figure 1), and used a vector $\boldsymbol{\delta}$ to measure the degree of relationship between two individuals in the presence of inbreeding. Where $\boldsymbol{\delta}$ consists of nine elements, whose n^{th} element δ_n $(n = 1,2,\cdots,9)$ represents one probability that the four alleles at a single locus in two diploid individuals share the configuration D_n of IBD alleles (Milligan 2003). Table 1 summarizes the values of elements in $\boldsymbol{\delta}$ for various mating forms. For example, if the mating form is selfing, then $\boldsymbol{\delta}$ consists of the elements in the SE row (the top row) in Table 1, denoted by $\boldsymbol{\delta}_{\text{SE}}$, *i.e.*, $\boldsymbol{\delta}_{\text{SE}} = [f,0,0,0,0,0,1-f,0,0]$

We will use the symbols $\boldsymbol{\delta}_{\text{SE}}, \boldsymbol{\delta}_{\text{PO}}, \boldsymbol{\delta}_{\text{FS}}, \cdots, \boldsymbol{\delta}_{\text{NR}}$ to denote the vector $\boldsymbol{\delta}$ consisting of the elements in the rows from the top to bottom in Table 1, respectively. Similarly, the symbols $s_{\text{SE}}, s_{\text{PO}}, s_{\text{FS}}, \cdots, s_{\text{NR}}$ are used to denote the proportions of offspring in an inbred population that are the results the

Table 1 The values of δ between mates in different mating forms

Relation	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	s
SE	f						λ_3			$\frac{2f}{1+f}$
PO	f^2		λ_1		λ_1			λ_2		$\frac{4f}{(1+f)^2}$
FS	$\frac{1}{4}f+\frac{1}{2}f^2$	$\frac{1}{4}f^2$	$\frac{1}{2}\lambda_1$	$\frac{1}{4}\lambda_1$	$\frac{1}{2}\lambda_1$	$\frac{1}{4}\lambda_1$	$\frac{1}{4}\lambda_3$	$\frac{1}{2}\lambda_2$	$\frac{1}{4}\lambda_2$	$\frac{8f}{2+3f+f^2}$
HS	$\frac{1}{2}f^2$	$\frac{1}{2}f^2$	$\frac{1}{2}\lambda_1$	$\frac{1}{2}\lambda_1$	$\frac{1}{2}\lambda_1$	$\frac{1}{2}\lambda_1$		$\frac{1}{2}\lambda_2$	$\frac{1}{2}\lambda_2$	$\frac{8f}{(1+f)^2}$
FC	$\frac{1}{4}f^2$	$\frac{3}{4}f^2$	$\frac{1}{4}\lambda_1$	$\frac{3}{4}\lambda_1$	$\frac{1}{4}\lambda_1$	$\frac{3}{4}\lambda_1$		$\frac{1}{4}\lambda_2$	$\frac{3}{4}\lambda_2$	$\frac{16f}{(1+f)^2}$
HFC	$\frac{1}{8}f^2$	$\frac{7}{8}f^2$	$\frac{1}{8}\lambda_1$	$\frac{7}{8}\lambda_1$	$\frac{1}{8}\lambda_1$	$\frac{7}{8}\lambda_1$		$\frac{1}{8}\lambda_2$	$\frac{7}{8}\lambda_2$	$\frac{32f}{(1+f)^2}$
SC	$\frac{1}{16}f^2$	$\frac{15}{16}f^2$	$\frac{1}{16}\lambda_1$	$\frac{15}{16}\lambda_1$	$\frac{1}{16}\lambda_1$	$\frac{15}{16}\lambda_1$		$\frac{1}{16}\lambda_2$	$\frac{15}{16}\lambda_2$	$\frac{64f}{(1+f)^2}$
NR		f^2		λ_1		λ_1			λ_2	

Where $\lambda_1 = f(1-f)$, $\lambda_2 = (1-f)^2$, $\lambda_3 = 1-f$ and s are the proportion of offspring produced by the corresponding inbreeding form under the equilibrium state (assuming that only the outcrossing can happen except for this inbreeding form). SE: self; PO: parent-offspring; FS: full-sibs; HS: half-sibs; FC: first-cousins; HFC: half first-cousins; SC: second-cousins; NR: nonrelatives.

corresponding mating forms. Now, the degree of relationship between mates can be measured by the weighted average $\overline{\boldsymbol{\delta}}$:

$$\overline{\boldsymbol{\delta}} = s_{\mathrm{SE}}\boldsymbol{\delta}_{\mathrm{SE}} + s_{\mathrm{PO}}\boldsymbol{\delta}_{\mathrm{PO}} + s_{\mathrm{FS}}\boldsymbol{\delta}_{\mathrm{FS}} + \cdots + s_{\mathrm{NR}}\boldsymbol{\delta}_{\mathrm{NR}} + \cdots. \quad (11)$$

Let $\overline{\boldsymbol{\delta}} = [\overline{\delta}_1, \overline{\delta}_2, \cdots, \overline{\delta}_9]$ Then, the inbreeding coefficient f' in the next generation can be expressed as

$$f' = \overline{\delta}_1 + \frac{1}{2}(\overline{\delta}_3 + \overline{\delta}_5 + \overline{\delta}_7) + \frac{1}{4}\overline{\delta}_8. \quad (12)$$

If only one inbreeding form occurs at the proportion s, then Equation 11 can be simplified. For example, if there is only the occurrence of selfing or backcrossing with $s_{\mathrm{SE}} = s$ or $s_{\mathrm{PO}} = s$, then $s_{\mathrm{NR}} = 1 - s$, and Equation 11 becomes $\overline{\boldsymbol{\delta}} = s\boldsymbol{\delta}_{\mathrm{SE}} + (1-s)\boldsymbol{\delta}_{\mathrm{NR}}$ or $\overline{\boldsymbol{\delta}} = s\boldsymbol{\delta}_{\mathrm{PO}} + (1-s)\boldsymbol{\delta}_{\mathrm{NR}}$. Under such condition, Equation 12 can be written as

$$f' = s\left[f + \frac{1}{2}(1-f)\right] \text{or} f' = s\left[f^2 + f(1-f) + \frac{1}{4}(1-f)^2\right].$$

Under the equilibrium state, $f' = f$, such that the proportion s can be solved. The solutions of s for each inbreeding form are listed in the right-most column in Table 1.

Table 1 reveals that the elements in the five rows determined by PO, HS, FC, HFC, and SC are proportional with the ratio $1 : 1/2 : 1/4 : 1/8 : 1/16$. This shows that there are many similarities among the five inbreeding forms: parents-offspring, half-siblings, first-cousins, half first-cousins, and second-cousins. We therefore chose backcrossing as the representative form (which ensures $0 \leq s \leq 1$), and add the value $1/2s_{\mathrm{HS}} + 1/4s_{\mathrm{FC}} + 1/4s_{\mathrm{HFC}} + 1/16s_{\mathrm{SC}} = 0$ to s_{PO}. Hence, the last equation becomes

$$f = \frac{1}{2}s_{\mathrm{SE}}(1+f) + \frac{1}{4}s_{\mathrm{PO}}(1+f)^2 + \frac{1}{8}s_{\mathrm{FS}}(1+f)(2+f). \quad (13)$$

Unfortunately, there are still three unknown inbreeding proportions s_{SE}, s_{PO} and s_{FS} in Equation 13, whose solutions are not unique (f is regarded as a constant). In order to obtain a unique solution, some constraints have to be added to this equation according to relevant *a priori* knowledge of the focal population.

If $\overline{\boldsymbol{\delta}}$ is known, the expression of the joint distribution of G_F and G_M is

$$\Pr(G_F, G_M|f) = \sum_{n=1}^{9} \Pr(G_F, G_M|D_n)\overline{\delta}_n,$$

where, for every n, the value of $\Pr(G_F, G_M|D_n)$ is listed in Table 2.

Assuming that the false parents are relatives of the true parents of opposite sexes, and let various joint distributions of genotypes of parent pairs be the same as $\Pr(G_F, G_M|f)$, then

$$\Pr(G_A, G_M|f) = \Pr(G_F, G_B|f) = \Pr(G_A, G_B|f) = \Pr(G_F, G_M|f).$$

Here, G_B is the genotype of the false mother.

Thus far, the joint distributions of genotypes in Equation 9 have been derived. Next, we derive the transitional probability $T(G_O|G_M,f)$ in Equation 9. By the generalized product rule of probabilities, $\Pr(G_F, G_M|f)$ can be rewritten as

$$\Pr(G_F, G_M|f) = \Pr(G_F|G_M,f)\Pr(G_M|f). \quad (14)$$

Therefore, the conditional probability $\Pr(G_F|G_M,f)$ is made available, and will be used to calculate the transitional probability:

$$T(G_O|G_M,f) = \sum_F \Pr(G_F|G_M,f)T(G_O|G_F, G_M).$$

Allele frequency estimator

In this section, we develop a novel estimator to estimate the allele frequencies in the presence of inbreeding and negative amplification, which is a modification of Summers and Amos (1997) estimator.

Suppose that there are altogether k visible alleles. Denote N_i for the number of observed genotypes consisting of the i^{th} visible allele A_i $(i = 1, 2, \cdots, k)$, and N_{vis} for the number of

Table 2 Joint distribution of genotypes under different IBD configurations

IBD configuration	Genotypic template	$\Pr(G_F, G_M \mid D_n)$
D_1	A_iA_i, A_iA_i	p_i
D_2	A_iA_i, A_jA_j	p_ip_j
D_3	A_iA_i, A_iA_j	p_ip_j
D_4	A_iA_i, A_jA_k	$(2-K_{j,k})p_ip_jp_k$
D_5	A_iA_j, A_iA_i	p_ip_j
D_6	A_iA_j, A_kA_k	$(2-K_{i,j})p_ip_jp_k$
D_7	A_iA_j, A_iA_j	$(2-K_{i,j})p_ip_j$
D_8	A_iA_j, A_iA_k	$p_ip_jp_k$
D_9	A_iA_j, A_kA_l	$(2-K_{i,j})(2-K_{k,l})p_ip_jp_kp_l$

For each genotypic template, if the alleles are with the same subscript, then they are IBD alleles, otherwise they are IBS or non-IBS alleles. If A_i and A_j are IBS alleles, then $K_{i,j} = 1$, otherwise $K_{i,j} = 0$.

visible observed genotypes. Because N_i and N_{vis} can be obtained directly from the observed genotypic data, their ratio N_i/N_{vis} is a constant. Also, according to Equation 3, and assuming that the rate of negative amplification is independent of the genotype, this ratio can be expressed as

$$\frac{N_i}{N_{\text{vis}}} = \frac{fp_i + (1-f)p_i^2 + 2(1-f)p_i(1-p_i)}{1 - fp_y - (1-f)p_y^2}, \tag{15}$$

where f and p_y are known, in which the inbreeding coefficient f is an *a priori* value, these are obtained by the average estimate of the inbreeding coefficients from Nei (1977) estimator from all polymorphic loci. Then, Equation 15 is a quadratic equation, with p_i as the unknown, whose solution is

$$p_i = \frac{1}{2}\left[f - 2 + \sqrt{(f-2)^2 + 4c_i(f-1)}\right] \text{ or,}$$
$$p_i = \frac{1}{2}\left[f - 2 - \sqrt{(f-2)^2 + 4c_i(f-1)}\right]$$

where $c_i = N_i/N_{\text{vis}}[1 - fp_y - (1-f)p_y^2]$ The latter solution should be excluded because it is outside of the range $[0, 1]$.

A half-interval search algorithm is used to estimate the allele frequencies, the procedure of which is described as follows.

1. Set the initial minimum and maximum values of p_y at $p_{y,\min} = 0$, and $p_{y,\max} = 1$, respectively.
2. Substitute $\hat{p}_y$ with $(p_{y,\min} + p_{y,\max})/2$ in the former solution to Equation 15 (where p_y and p_i in this solution will be regarded as $\hat{p}_y$ and $\hat{p}_i$), and then find the value of $\hat{p}_i$, $i = 1, 2, \cdots, k$.
3. Test the value of $1 - \sum_{i=1}^{k} \hat{p}_i - \hat{p}_y$. If this is greater than, or equal to, zero, then update $p_{y,\min}$ with $\hat{p}_y$; otherwise update $p_{y,\max}$ with $\hat{p}_y$.
4. Repeat steps (ii) and (iii) until the difference $p_{y,\max} - p_{y,\min}$ is less than a threshold, *e.g.*, 10^{-12}.

The final values of $\hat{p}_y, \hat{p}_1, \hat{p}_2, \cdots, \hat{p}_k$ in the above procedures are the estimates of allele frequencies.

We now consider the estimation of the negative amplification rate β. Denote N_{true} for the true sample size (*i.e.*, the number of observed genotypes excluding those with negative amplification). By Equation 3, the ratio of N_{vis} to N_{true} is $N_{\text{vis}}/N_{\text{true}} = 1 - fp_y - (1-f)p_y^2$, then

$$N_{\text{true}} = \frac{N_{\text{vis}}}{1 - fp_y - (1-f)p_y^2}.$$

Thus, the estimate $\hat{N}_{\text{true}}$ can be obtained by substituting $\hat{p}_y$ for p_y in the final expression. The estimate of β can therefore be calculated by $\hat{\beta} = \max(0, 1 - \hat{N}_{\text{true}}/N_{\text{tot}})$ where N_{tot} is the total number of individuals.

Data availability

Genotyping data used to test the model's efficiency may be found at doi: 10.5061/dryad.689v4.

The software PARENTAGE V1.0, user manual and example dataset are available on GitHub (http://github.com/huangkang1987/parentage). Supplemental material available at Figshare: https://doi.org/10.25386/genetics.7221965.

Results

Evaluation

In this study, we use Monte-Carlo simulations to generate the observed genotypic data and to perform parentage analyses for four typical applications. The influences of the following three factors on the Ka-model are evaluated: inbreeding and null alleles either each singly or in unison. The performance of both our model and an additional model, named the Wa-model (Wang 2016), under the same conditions are compared with that of the Ka-model. We also use the empirical data published by Nietlisbach *et al.* (2015) to test and compare the accuracy of all three models under natural conditions.

Simulated data

In order to evaluate the influences of the three factors under scrutiny (inbreeding, null alleles, and negative amplification), we first set some levels for the inbreeding coefficient f or for the null allele frequency p_y. For null alleles, we set $f = 0$ and $p_y = 0$, 0.05, 0.15, or 0.3, where the four values of p_y represent the minimum, low, medium, and high levels, respectively. For inbreeding, we set $p_y = 0$ and $f = 0$, 0.05, 0.15, or 0.3. For inbreeding and null alleles jointly, we set $p_y = f = 0$, 0.05, 0.15, or 0.3.

For the first two categories in a parentage analysis (*i.e.*, identifying the father when the mother is either unknown or known), each is designated its own application [named Application (i) or (ii)], with 100 alleged fathers randomly generated for each offspring. For the third category (*i.e.*, identifying the father and mother jointly), we also designate two applications [named Applications (iii) and (iv)]. For Application (iii), the sexes of the alleged parents are known, and 100 alleged fathers and 100 alleged mothers are randomly generated for each offspring; for Application (iv), the sexes of the alleged parents are unknown, and 100 alleged parents with the predefined sex ratio of 1 : 1 are generated for each offspring.

For each application, 1000 offspring and their true and alleged parents are simulated. The observed genotypes of all individuals are generated at 4–16 unlinked loci. Based on these observed genotypes, parentage analyses are performed by either the Ka-model or by our model with three different thresholds (0, $\Delta_{0.80}$, and $\Delta_{0.99}$) of Δ, or by the Wa-model with three different thresholds (0, 0.80, and 0.99) of posterior probability. The performance of each of these three models are presented in two graphical formats. For the Ka-model, the graphs of the correct assignment rate as a function of the number of loci under four applications and under different levels of p_y and/or f are shown in Figure 2. For these three

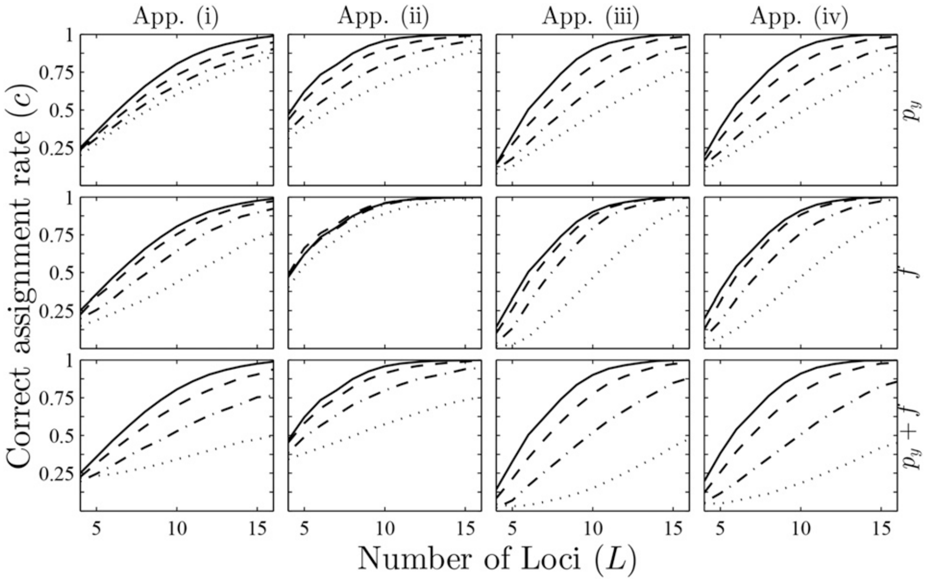

Figure 2 The influence of inbreeding and null alleles on the Ka-model in each of four applications. Each column denotes one application. The top (or middle) row shows the effect of null alleles (or inbreeding). Each factor has four levels (*i.e.*, p_y or $f = 0$, 0.05, 0.15, or 0.3) and the corresponding results are shown by solid, dashed, dash-dotted, and dotted lines, respectively. Each line is a graph of the correct assignment rate as a function of the number L of loci. The bottom row shows the influence of the effect of both null alleles and inbreeding acting simultaneously. Each factor also has four levels (p_y and f are set as equal, *i.e.*, $p_y = f = 0$, 0.05, 0.15, or 0.3), with the line styles of different levels the same as for the previous rows.

models, each correct assignment rate is shown by a part of the overlapped bar charts (see Figure 3 in detail, and Supplemental Material, Figures S1 and S2 shows the results with the thresholds $\Delta_{0.80}$ and $\Delta_{0.99}$). Here, a correct assignment means that the true parents have been assigned correctly and there is either a Δ value or a posterior probability above the corresponding threshold.

The procedures used to generate the observed genotypes are as follows. First, L unlinked loci are created, and the allele frequencies at all loci are equal. In order to accelerate the simulation, we reduce the number of alleles at a locus and modify their frequencies used in Kalinowski *et al.* (2007). We then use the loci with six visible alleles to perform our simulation, and the vector of visible allele frequencies is set as $[0.25, 0.25, 0.2, 0.15, 0.10, 0.05]$. In the presence of null alleles, each frequency of visible alleles is multiplied by $1 - p_y$ to unify the allele frequencies.

In order to simulate inbreeding, the true parents should be regarded as relatives, so their genotypes are not independent. Hence the genotypes of true parents are generated via Equation 12. Thereafter, the genotypes of offspring are generated by Equation 8. The false parents may be related to the true parents of the opposite sex, and their genotypes are generated by Equation 14. The three proportions s_{SE}, s_{PO}, and s_{FS} are assumed to be equal, *i.e.*, their ratio is $s_{\mathrm{SE}} : s_{\mathrm{PO}} : s_{\mathrm{FS}} = 1 : 1 : 1$. When our model is applied to perform parentage analysis, we will also use this ratio as the relative ratio among the corresponding three mating forms. In other words, we will use this ratio as a constraint for Equation 13, then there is a unique solution of Equation 13 as follows:

$$s_{\mathrm{SE}} = s_{\mathrm{PO}} = s_{\mathrm{FS}} = \frac{8f}{8 + 11f + 3f^2}. \qquad (16)$$

Finally, the generated genotypes are converted into observed genotypes. Each generated genotype is randomly replaced with a false genotype according to Equation 2 at a probability e to simulate the genotyping error. Next, to account for the presence of null alleles, the genotype obtained after the previous step is converted to an observed genotype according to Equation 10. Furthermore, this observed genotype is randomly set as $\varnothing$ at a probability β to simulate the effect of negative amplification. The negative amplification rate β and the genotyping error rate e are set as 0.05 and 0.01, respectively. All alleged parents are sampled in our simulation.

The generated observed genotypes are used to perform parentage analysis. Unfortunately, because the false parents are assumed to be relatives of the true parents, the alleles carried by the true parents will appear at a higher frequency than their true frequencies, which will bias the allele frequency estimation. In order to avoid this bias, 100 nonrelatives are generated according to Equation 2, and their observed genotypes are converted by using the same method described above, and are used to estimate the allele frequencies.

For both the Ka-model and the Wa-model, the allele frequencies are estimated by counting the numbers of alleles without considering the effects of both null alleles and negative amplification. For our model, these frequencies are estimated by our new allele frequency estimator, and the three proportions $\hat{s}_{\mathrm{SE}}$, $\hat{s}_{\mathrm{PO}}$, and $\hat{s}_{\mathrm{FS}}$ are estimated from the inbreeding coefficient f according to Equation 16. Because we do not develop an estimator to estimate the inbreeding coefficient f under null alleles and negative amplification, f is estimated by the Nei (1977) estimator. Additionally, the true values of both genotyping error rate e and sampling rate of true parents are used in all models.

For the Wa-model, we write the individual observed genotypes and the allele frequency estimates together with other necessary parameters into a file, named *.dat, according to the input file format of COLONY V2.0.6.4. After calling colony2p.exe by a command-line mode, we read the results from the output files. COLONY uses a different algorithm to perform parentage analysis: by evaluating the likelihood of pedigrees, it searches the optimal full- and half-sibs families (Wang 2016). This algorithm neither performs a simulation to obtain the thresholds of Δ, nor calculates the LOD scores. Instead, it uses the posterior probability as an indicator of confidence. Therefore, three thresholds (0, 0.8, and 0.99) of the posterior probability are used to denote three levels of confidence, where a threshold of posterior probability equal to 0 means that the alleged parent(s) with the highest posterior probability is chosen. The mating system for both sexes is assumed to be polygamous, and allele frequencies are not updated during iteration. The rates of two genotyping errors (allelic dropout, and all other errors involved in genotyping) are both assumed to be equal to the true value of 0.01.

In addition, to evaluate the performance of the Nei (1977) estimator relative to our allele frequency estimator, an extra 100,000 simulations are performed. In each simulation, the observed genotypes at 10 loci of 100 nonrelatives are generated by Equation 2. These observed genotypes are used to estimate the inbreeding coefficient, the negative amplification rate, and the null allele frequency. We use bias and SD to evaluate the accuracy of each inbreeding coefficient, negative amplification rate and null allele frequency.

The estimation of allele frequency uses the estimate of the inbreeding coefficient, which may introduce some errors. To account for possible effects of the inbreeding coefficient estimator on the accuracy of the estimated parameters, and to explore the potential of our allele frequency estimator, we also use the true value of f to perform simulation. The corresponding results are used for comparison.

Simulated results

The influences of both inbreeding and null alleles on Ka-model in the four applications with a $\Delta > 0$ are shown in Figure 2. Because the distribution of genotypes of the alleged parents are deviated from the HWE, the threshold of $\Delta_{0.80}$ or $\Delta_{0.99}$ cannot ensure a confidence level of 80 or 99%. Therefore, we do not show the results involving these thresholds in Figure 2. Although these results are shown in Figure 3 for reference, these are not analyzed nor discussed. It is clear that both inbreeding and null alleles significantly affect the accuracy of our parentage analysis.

In the presence of null alleles, the curves inside each subfigure are nearly equally spaced. Application (i) is relatively less affected, the values of correct assignment rate (denoted by c) are decreased at most 0.2 when $p_y = 0.3$, while those values are decreased at most 0.3 for Application (ii), and at most 0.4 for Applications (iii) and (iv).

In the presence of inbreeding, Application (ii) is barely affected, although the values of the correct assignment rate c are slightly increased (at most 0.03) when $f = 0.05$. For the remaining applications, the values of c are slightly decreased when $f = 0.05$, but they are greatly reduced when f increases from 0.15 to 0.3.

When null alleles and inbreeding are both present, the influences of both factors are cumulated, and the performance are greatly affected. The curves in all applications become increasingly flat as f and p_y also increase.

The results of the Ka-model, the Wa-model and our model for the four applications are presented in Figure 3. In order to compare the performance of all models, we consolidate the results of each of the three models under the same conditions, and we use bar charts to show the various correct assignment rates.

In the absence of both inbreeding and null alleles, our model and the Ka-model perform similarly, although some small changes in the estimated allele frequencies result in small differences between their correct assignment rates. The Wa-model performs well in all applications when $L \geq 14$. However, when L is small, the Wa-model performs a little worse than our model and the Ka-model in Applications (i) and (ii), with the highest differences between the correct assignment rates being 0.03 and 0.06, respectively. The Wa-model performs even worse in Applications (iii) and (iv), with the highest differences increasing to 0.12 and 0.14, respectively.

In the presence of null alleles, our model outperforms the Ka-model, especially when $p_y = 0.3$ and for Applications (iii) and (iv), which resulted in maximum differences between correct assignment rates of up to 0.15. In Applications (i) and (ii), both Wa- and Ka-models perform similarly. However, the Wa-model performs worse than the Ka-model in Applications (iii) and (iv), with differences being $\sim$0.1–0.2.

In the presence of inbreeding, our model and the Ka-model perform similarly if $f = 0.05$. Our model performs a little worse than the Ka-model in Application (ii), but more or less better in Applications (i) and (iv) if $f \geq 0.15$, or in Application (iii) if $f = 0.3$. The Wa- and Ka-models perform similarly in Applications (i) and (ii). However, the Wa-model performs worse than Ka-model in Applications (iii) and (iv), for example, if $f = 0.3$, the maximum differences between the correct assignment rates being $\sim$0.35 and 0.5, respectively.

In the presence of null alleles and inbreeding, our model outperforms the other two models in most cases, with the Ka- and Wa-models performing similarly in Applications (i) and (ii). However, both Ka- and Wa-models are strongly negatively affected in Application (iii) and (iv). For instance, if $p_y = f = 0.3$, the correct assignment rates for both models are $\sim$75 and 35% of those for our model, respectively.

The evaluation results of our allele frequency estimator are shown in Table S1. The presence of null alleles introduces an overestimation of the inbreeding coefficient using Nei (1977) estimator. The SD of $\hat{f}$, $\hat{\beta}$ and $\hat{p}_y$ are slightly increased as both f and p_y also increase, respectively. The bias of $\hat{p}_y$ becomes extremely large as p_y increases, and reaches 0.25 when $p_y = 0.3$. The bias of $\hat{\beta}$ is also affected by p_y, and reaches 0.1 at $f = p_y = 0.3$. If the true value of the inbreeding coefficient is used in simulation, these biases are considerably reduced (at most 0.02, Table S1).

Empirical data

We used microsatellite genotyping data for a population of song sparrows (*Melospiza melodia*) on Mandarte Island, Canada (Nietlisbach *et al.* 2015), to test the efficiency of our model. These data are available at doi: 10.5061/dryad.689v4.

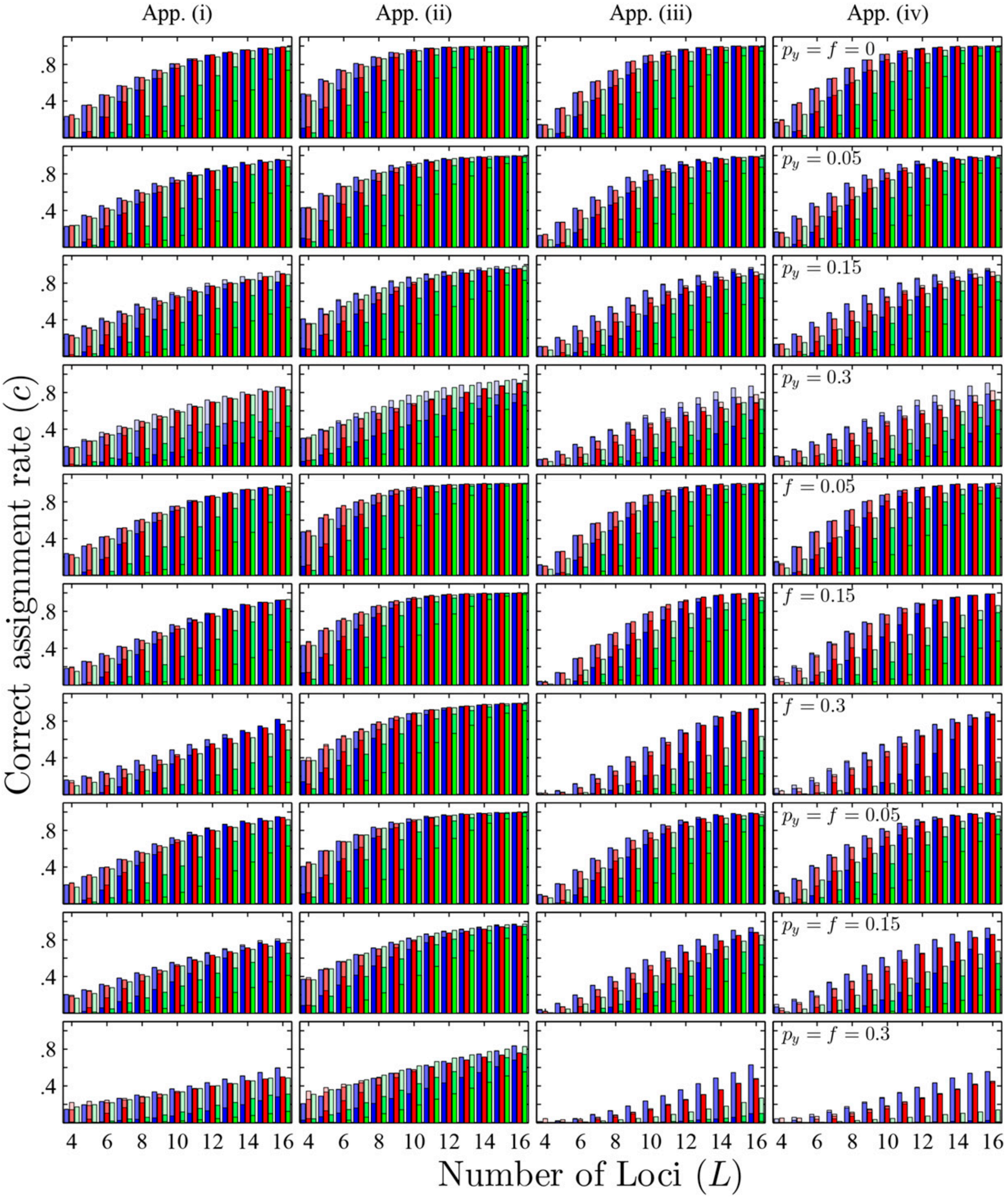

Figure 3 The correct assignment rates of our model, the Ka-model and the Wa-model as a function of number L of loci under 10 different levels of null allele and inbreeding. Each column denotes an application. Each row shows all correct assignment rates with the same level (the value representing this level is listed in the subfigure located in the rightmost column). Every correct assignment rate is shown by a part of the overlapped bar charts. The results of the Ka-model are shown by the red bars, and those of both the Wa-model and our model by the green and blue bars, respectively. The bars with light, medium, and bright colors denote the correct assignment rates with the thresholds 0, $\Delta_{0.80}$, and $\Delta_{0.99}$ for our model and the Ka-model, or with the thresholds 0, 0.8, and 0.99 for the Wa-model, respectively.

The song sparrow is a medium-sized passerine bird, native to North America. The past breeding density of this population has fluctuated 18-fold due to two major population crashes (Keller *et al.* 1994). In 1988–1989, only four adult females and seven adult males were present (Keller *et al.* 1994). This resulted in inbreeding in this population due to a bottle-neck effect, and across the 2364 birds whose all four grandparents were genetically verified during 1993–2013, the mean inbreeding coefficient was 0.087 (Nietlisbach *et al.* 2015).

The dataset of Nietlisbach *et al.* (2015) contains the genotypes of 3301 individuals at 209 microsatellite loci. There were 3186 individuals whose father or mother was genotyped, and these were used as offspring in the parentage analysis. The fathers and mothers of these individuals were recorded from long-term pedigree data, but the sexes of the offspring are not given. The average inbreeding coefficient estimated by the Nei (1977) estimator based on 199 autosomal loci is 0.074, which is used as an *a priori* inbreeding coefficient in our model.

Because the microsatellites used by Nietlisbach *et al.* (2015) were less polymorphic than those used in the simulation, and because their genotyping ratios were also lower, we used more loci to perform the parentage analysis. We scanned the 199 autosomal microsatellites and chose two subsets of loci. Subset one consisted of the loci with the highest estimated null allele frequencies. Subset two consisted of haphazardly selected loci, that were chosen by first ranking all loci alphabetically, and then selecting the top 40 name-ranked loci. The indices of the genetic diversity of the selected loci are shown in Tables S2 and S3. The average numbers of visible alleles are 8.175 and 8.725, respectively, the average genotyping ratios are 68.28 and 72.98%, respectively, the average estimated inbreeding coefficients are 0.333 and 0.097, respectively, and the average estimated null allele frequencies are 0.200 and 0.053, respectively.

Following the definition of the above applications for the simulated data, four similar applications are considered as follows: (I)–(II) identifying one parent when the other is unknown (6284 cases, including 3162 for identifying father and 3122 for identifying mother), or when the other parent is known (6196 cases); (III)–(IV) identifying jointly the father and mother in which the sexes of the candidate parents are both known (3098 cases), or unknown (also 3098 cases). Here, the hypotheses H_1 and H_2 in Applications (I) and (II) need to be modified as follows: the alleged parent is the true parent (H_1), or is not the true parent (H_2).

In Applications (I)–(IV), all males or females are included as either the alleged fathers or the alleged mothers, whereas an offspring itself is excluded from the pool of alleged parents in each case. The average numbers of candidate parents for each case are 290 in Applications (I) and (II), and 581 in Application (IV). For Application (III), the average numbers of candidate fathers and mothers in each case are 297 and 284, respectively.

We use 5–40 loci to perform our parentage analysis and use the correct assignment rate to measure the efficiencies of each of the three models. For Applications (III) and (IV), the identification is considered as correct when both parents are correctly identified.

For the Ka- and Wa-models, the allele frequencies are estimated by counting the numbers of alleles without considering the effects of both null alleles and negative amplification. For our model, an *a priori* inbreeding coefficient is set as 0.074, and the allele frequencies (including the null allele frequency) and the negative amplification rate are both estimated simultaneously.

Three thresholds (0, $\Delta_{0.80}$, and $\Delta_{0.99}$) of Δ are obtained by using a Monte-Carlo simulation (Marshall *et al.* 1998). In each application, 100,000 offspring are generated, and the number of alleged parents for each offspring is taken from the average number of alleged parents. For the Ka-model, the false parents are nonrelatives of the true parents. For our model, inbreeding is assumed to be present due to backcrossing (because this mating form represents many inbreeding forms with reduced relatedness between mates, it is probably the common form of inbreeding), the false parents are related to the true parents of the opposite sex, the genotyping rate is equal to the average genotyping rate among the loci currently being used, the sampling rate is equal to one, and the genotyping error rate is assumed to be 0.01. For the Wa-model, the configuration of COLONY is identical to that of the simulated data.

Empirical results

In the four applications, the results of the parentage analysis of all three models for subset 1 are shown in Figure 4 (Figures S3 and S4 show the results with the thresholds $\Delta_{0.80}$ and $\Delta_{0.99}$), and those for subset 2 are shown in Figure S5 (Figures S6 and S7 show the results with the thresholds $\Delta_{0.80}$ and $\Delta_{0.99}$). Because some individuals are typed at only a few loci (*e.g.*, 783 individuals are typed at four loci), the true parents cannot easily be identified. Therefore, each curve for the correct assignment rates in the range from 0.5 to 0.7 reaches near to an asymptote.

Figure 4 also shows that our model outperforms both the Ka- and Wa-models in all four applications, especially in Applications (III) and (IV). In Applications (I) and (II), for ~65–80% of loci, our model achieves similar levels of accuracy as the Ka-model. Similarly, according to the simulation results, Applications (III) and (IV) are more sensitive to the presences of both inbreeding and null alleles, and, for ~55–70% of loci, our model achieves similar levels of accuracy as the Ka-model.

For subset 2, the average estimated null allele frequency (0.053) and the average estimated inbreeding coefficient (0.097) are both low (Table S3). Therefore, the performance of our model for subset 2 is not so good as that for subset 1, which is consistent with our simulated data. However, our model still a little outperforms both the Ka- and the Wa-models. For example, the correct assignment rates in Applications (I) and (IV) for our model are at most 3.2 and 7.4% higher than for the Ka-model (Figure S5), respectively.

Discussion

Impacts of inbreeding and null alleles

Both inbreeding and null alleles can cause serious problems in parentage analyses (Wagner *et al.* 2006; Wang 2011). In the presence of null alleles, the genotypes of parents and

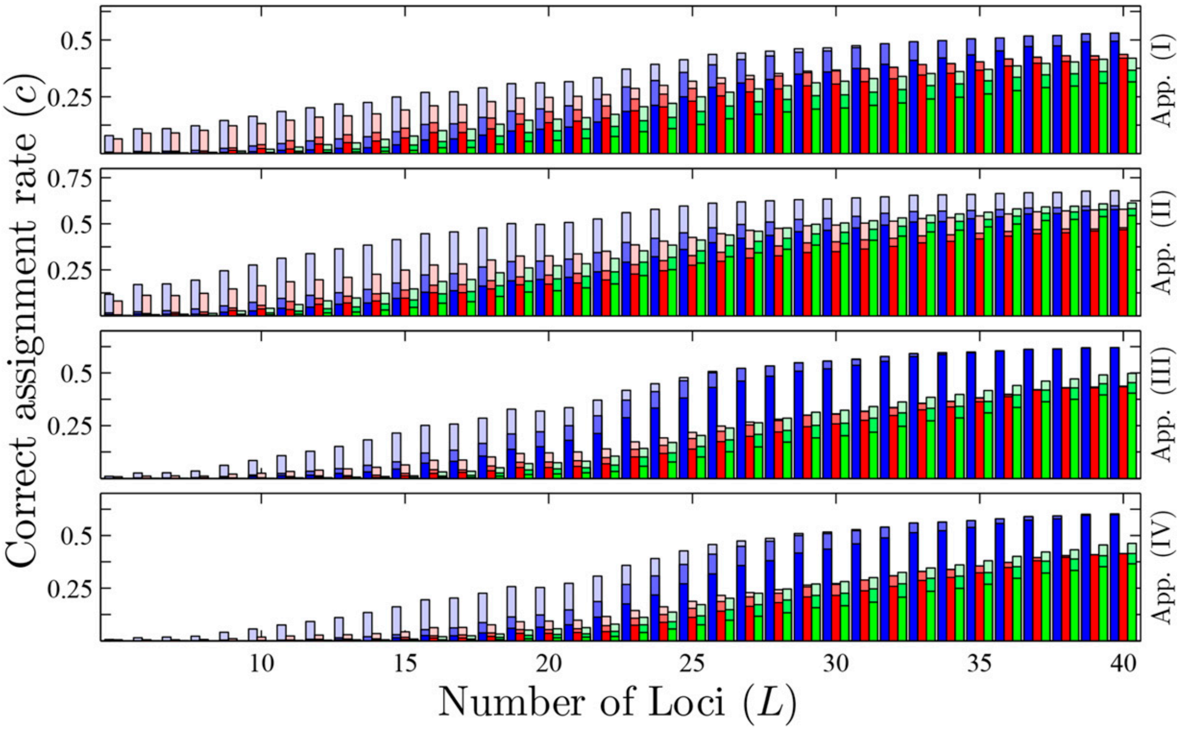

Figure 4 Results of the parentage analysis using the dataset of Nietlisbach *et al.* (2015), in which the loci chosen are with the highest estimated null allele frequency. Each row denotes an application. The definitions of bars together with their colors are as for Figure 3.

offspring may be mismatched (Brookfield 1996). In addition, a null allele homozygote may also be treated as a negative observed genotype, and hence it is omitted from any likelihood calculations (Marshall *et al.* 1998; Kalinowski *et al.* 2007). In the presence of inbreeding, genotypic frequencies become deviated from the HWE and the genotypes of both parents are not independent. Moreover, the false parents may also be potential mates and relatives of the true parents of the opposite sex, who may also share IBD alleles with the offspring.

With our computer simulations, we found that even a small inbreeding coefficient (*e.g.*, 0.05) or a small null allele frequency can result in a large reduction in the correct assignment rate (up to 0.15; Figure 2). In the presence of inbreeding and/or null alleles, the information given by the genotyping data are reduced, and so more loci should be used in order to reach the same level of accuracy. For example, 180% additional loci are required to reach the correct assignment rate of 50% for the Ka-model if the inbreeding coefficient and null allele frequency are both 0.3 (Figure 2).

Corrections for inbreeding and null alleles

In the process of establishing our model, we made several modifications to the Ka-model. These included the actual genotypic and observed genotypic frequencies, joint distribution of parental genotypes, conditional probability of parental genotypes, alternative hypotheses, alternative forms of likelihood calculation, and allele frequency estimations.

The performance of our model using both computer generated and empirical data were also evaluated under the same conditions as the Ka-model. The results showed almost ubiquitous improvement except for some situations with only few alleles. Although our model is still affected negatively by either the presence of inbreeding and/or null alleles, it is still able to recover much information. Importantly, our new method requires at least 55% of all loci to attain an equal degree of accuracy as the Ka-model (Figure 3).

Compared with the effects of inbreeding, our model performs better in the presence of null alleles. In the Ka-model, negative amplification is not considered so any negative observed genotypes are ignored in the calculation of the likelihoods. For example, in Application (i), if the observed genotype of either the offspring or the alleged father at a locus is negative, and if such an observed genotype is ignored, then the likelihood at this locus is omitted. This omission is equivalent to the likelihoods of H_1 and H_2 at this locus being set to one, which will result in the overestimation of both likelihoods and a bias of the LOD score.

A negative observed genotype is similar to a visible allele homozygote, representing one of several possible genotypes. In our model, negative amplification is considered and each negative observed genotype is treated as a normal observed genotype. In our alternative forms of likelihoods, each possible genotype is weighted according to its probability (either conditional, prior or joint), such that the likelihoods considering any negative amplification and any negative observed genotypes can be calculated.

Alternative hypothesis

Among the four applications used to test the efficiency of the Ka-model, the first two (identifying the father when the mother is either known or unknown) are both affected to a relatively lesser degree (Figure 2). The latter two (identifying jointly the two parents when the sexes of the candidate parents are either known or unknown) are more sensitive to the effects of inbreeding and/or null alleles (Figure 2).

The scheme of the Ka-model also contributes to the relatively poor performance for Applications (iii) and (iv) [or (III) and (IV)]. Here, the hypothesis H_1 that the alleged parents are the true parents, is evaluated relative to the alternative hypothesis H_2 that the alleged parents are unrelated to the offspring. However, in this scheme, the scenario that one alleged parent is a true parent while the other is not is not considered.

We give two additional events to this scheme, and use the geometrical mean of the corresponding likelihoods as the likelihood of H_2. From the validation of both simulations (Figure 3) and the empirical data (Figure 4), the performances after the scheme has made the appropriate correction are significantly improved.

Allele frequency estimator

In this paper, we develop a novel estimator to estimate the allele frequencies in the presence of both inbreeding and negative amplification, which is a modification of Summers and Amos (1997) estimator. This estimator estimates the allele frequency and negative amplification rate separately. The allele frequencies are first estimated, with only the visible observed genotypes being used. This approach can eliminate the impact of negative amplification, because the ratio N_i/N_{vis} has nothing to do with β under the assumption that the negative amplification is independent of the genotypes. The negative amplification rate is subsequently estimated from the estimates of these allele frequencies.

Our allele frequency estimator assumes that the inbreeding coefficient has an *a priori* value, and we thus use the Nei (1977) estimator to estimate f; however, this estimator does not consider either negative amplification or null alleles, the errors are accumulated during allele frequency estimation. Therefore, the biases of $\hat{\beta}$ and $\hat{p}_y$ are high when f and p_y are also high (0.102 and -0.251, respectively, Table S1). However, our model still works well and outperforms both the Ka- and Wa-models in many cases. If the true value of the inbreeding coefficient is used in simulation, these biases will be largely reduced (at most 0.02, Table S1) which suggests that our allele frequency estimator has considerable potential to improve estimations of both allele frequencies and parentage analysis.

Pervasive inbreeding

Although we only consider close inbreeding, pervasive inbreeding will also have a similar influence on the parentage analysis in two ways: (i) the genotypic frequencies deviate from the HWE, which will bias the likelihood estimate; (ii) the candidate parents may be sampled from the same population as the true parents, who may share the same IBD alleles as the true parents and the offspring, which will result in an overestimation of the LOD score of the false parents and interfere with our analysis.

These problems are solved by a two-step process: (i) estimating the allele frequency for each population, and then (ii) using the local allele frequencies to calculate the likelihoods. Unfortunately, due to dispersal among populations, for an individual, the natal population may not be the same as the sampled population. Although the natal population can be calculated by Bayesian clustering (Pritchard *et al.* 2000) or by population assignment (Peakall and Smouse 2012), the results may be unreliable in cases in which the population structure is weak (*e.g.*, Wright's $F_{ST} < 0.05$). Moreover, the estimation of local allele frequencies will also be inaccurate due to the limited sample size. Hence, this two-step process can result in an accumulation of errors.

The level of pervasive inbreeding can be measured by Wright's F_{ST} (Wang 2011). An alternative approach is to consolidate both F_{ST} and F_{IS} into a single parameter. Using the formula $F_{IT} = 1 - (1 - F_{IS})(1 - F_{ST})$, the effects of pervasive inbreeding can be incorporated into our model. The genotypic frequency $\Pr(G)$ in the total population and by incorporating both types of inbreeding can be expressed as

$$\Pr(G) = \begin{cases} F_{IT}p_i + (1 - F_{IT})p_i^2 & \text{if } G = A_iA_i, \\ 2(1 - F_{IT})p_ip_j & \text{if } G = A_iA_j. \end{cases} \tag{17}$$

Wright's F_{IT} is a measure of the correlation of gene frequencies among all individuals in the total population. Comparing Equation 17 with Equation 2, both have the same form, but the applied range of Equation 17 is wider.

Similar to the method for applying Equation 2, the joint distribution of parental genotypes or the conditional probability of the alleged parental genotypes in the total population, can now be derived. This alternative method only makes a slight change to our model, but it can be applied without involving both an estimation of the local allele frequencies and an identification of the natal population of each individual. This will thus prevent the accumulation of errors. However, when the population structure is strong (*i.e.*, F_{ST} is large), Equation 17 cannot accurately predict the genotypic frequency. Meanwhile, if the sample size is large, the natal population of each individual can be accurately obtained and the initial approach will perform better than the alternative approach.

Acknowledgments

This study was funded by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB31020000), the National Natural Science Foundation of China (31770411, 31501872, 31730104, and 31370408), the National Key Programme of Research and Development, the Ministry of Science and Technology of China (2016YFC0503200), the

Young Elite Scientists Sponsorship Program by CAST (2017QNRC001), and the Natural Science Foundation of Shaanxi Province (2018JM3024).

Author contributions: K.H. and B.L. designed the project, K.H. and R.M. constructed the model and wrote the draft, K.H. designed the software, T.W. performed the simulations and analyses, and D.W.D. checked the model and helped to write the manuscript.

Literature Cited

Blouin, M. S., 2003 DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. Trends Ecol. Evol. 18: 503–511. https://doi.org/10.1016/S0169-5347(03)00225-8

Brookfield, J. F. Y., 1996 A simple new method for estimating null allele frequency from heterozygote deficiency. Mol. Ecol. 5: 453–455. https://doi.org/10.1046/j.1365-294X.1996.00098.x

Cremona, T., P. Spencer, R. Shine, and J. K. Webb, 2017 Avoiding the last supper: parentage analysis indicates multi-generational survival of re-introduced 'toad-smart' lineage. Conserv. Genet. 18: 1475–1480. https://doi.org/10.1007/s10592-017-0973-3

Dias, R. I., M. S. Webster, and R. H. Macedo, 2017 Parental and alloparental investment in campo flickers (*Colaptes campestris campestris*): when relatedness comes first. Behav. Ecol. Sociobiol. 71: 139. https://doi.org/10.1007/s00265-017-2368-3

Garber, P. A., L. M. Porter, J. Spross, and A. Di Fiore, 2016 Tamarins: insights into monogamous and non-monogamous single female social and breeding systems. Am. J. Primatol. 78: 298–314. https://doi.org/10.1002/ajp.22370

Gerzabek, G., S. Oddou-Muratorio, and A. Hampe, 2017 Temporal change and determinants of maternal reproductive success in an expanding oak forest stand. J. Ecol. 105: 39–48. https://doi.org/10.1111/1365-2745.12677

Goodnight, K. F., and D. C. Queller, 1999 Computer software for performing likelihood tests of pedigree relationship using genetic markers. Mol. Ecol. 8: 1231–1234. https://doi.org/10.1046/j.1365-294x.1999.00664.x

Jacquard, A., 1972 Genetic information given by a relative. Biometrics 28: 1101–1114. https://doi.org/10.2307/2528643

Johannesson, K., S. H. Saltin, G. Charrier, A.-K. Ring, C. Kvarnemo *et al.*, 2016 Non-random paternity of offspring in a highly promiscuous marine snail suggests postcopulatory sexual selection. Behav. Ecol. Sociobiol. 70: 1357–1366. https://doi.org/10.1007/s00265-016-2143-x

Kalinowski, S. T., A. P. Wagner, and M. L. Taper, 2006 ML-RELATE: a computer program for maximum likelihood estimation of relatedness and relationship. Mol. Ecol. Notes 6: 576–579. https://doi.org/10.1111/j.1471-8286.2006.01256.x

Kalinowski, S. T., M. L. Taper, and T. C. Marshall, 2007 Revising how the computer program cervus accommodates genotyping error increases success in paternity assignment. Mol. Ecol. 16: 1099–1106. https://doi.org/10.1111/j.1365-294X.2007.03089.x

Keller, L. F., P. Arcese, J. N. Smith, W. M. Hochachka, and S. C. Stearns, 1994 Selection against inbred song sparrows during a natural population bottleneck. Nature 372: 356–357. https://doi.org/10.1038/372356a0

Marshall, T. C., J. Slate, L. E. B. Kruuk, and J. M. Pemberton, 1998 Statistical confidence for likelihood-based paternity inference in natural populations. Mol. Ecol. 7: 639–655. https://doi.org/10.1046/j.1365-294x.1998.00374.x

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

Monteiro, N. M., D. Carneiro, A. Antunes, N. Queiroz, M. N. Vieira *et al.*, 2017 The lek mating system of the worm pipefish (*Nerophis lumbriciformis*): a molecular maternity analysis and test of the phenotype-linked fertility hypothesis. Mol. Ecol. 26: 1371–1385. https://doi.org/10.1111/mec.13931

Nei, M., 1977 *f*-Statistics and analysis of gene diversity in subdivided populations. Ann. Hum. Genet. 41: 225–233. https://doi.org/10.1111/j.1469-1809.1977.tb01918.x

Nietlisbach, P., G. Camenisch, T. Bucher, J. Slate, L. F. Keller *et al.*, 2015 A microsatellite-based linkage map for song sparrows (*Melospiza melodia*). Mol. Ecol. Resour. 15: 1486–1496. https://doi.org/10.1111/1755-0998.12414

Peakall, R., and P. E. Smouse, 2012 GenAIEx 6.5: genetic analysis in excel. population genetic software for teaching and research-an update. Bioinformatics 28: 2537–2539. https://doi.org/10.1093/bioinformatics/bts460

Pritchard, J. K., M. Stephens, and P. Donnelly, 2000 Inference of population structure using multilocus genotype data. Genetics 155: 945–959.

Ravinet, M., A. Westram, K. Johannesson, R. Butlin, C. Andr *et al.*, 2016 Shared and nonshared genomic divergence in parallel ecotypes of *Littorina saxatilis* at a local scale. Mol. Ecol. 25: 287–305. https://doi.org/10.1111/mec.13332

Summers, K., and W. Amos, 1997 Behavioral, ecological, and molecular genetic analyses of reproductive strategies in the Amazonian dart-poison frog, *Dendrobates ventrimaculatus*. Behav. Ecol. 8: 260–267. https://doi.org/10.1093/beheco/8.3.260

Wagner, A. P., S. Creel, and S. T. Kalinowski, 2006 Estimating relatedness and relationships using microsatellite loci with null alleles. Heredity 97: 336–345. https://doi.org/10.1038/sj.hdy.6800865

Wang, J. L., 2011 Unbiased relatedness estimation in structured populations. Genetics 187: 887–901. https://doi.org/10.1534/genetics.110.124438

Wang, J. L., 2016 Individual identification from genetic marker data: developments and accuracy comparisons of methods. Mol. Ecol. Resour. 16: 163–175. https://doi.org/10.1111/1755-0998.12452

Communicating editor: N. Rosenberg

Appendix

A Identifying the Father when the Mother is Known

For the second category of parentage analyses (*i.e.* identifying the father when the mother is known), the hypotheses H_1 and H_2 are as in the first category, and the likelihoods for which can be calculated by modifying Equation 7 and Equation 9 with the consideration of mother's observed genotypes. The expressions are as follows:

$$\begin{aligned}
\mathcal{L}(H_1) &= \sum_{OFM} \Pr(G_F, G_M|f) T(G_O|G_F, G_M) \\
&\quad \Pr\Big(\mathcal{O}_O|G_O,f, p_y, \beta, e\Big)\Pr\Big(\mathcal{O}_M|G_M,f, p_y, \beta, e\Big)\Pr\Big(\mathcal{O}_A|G_F,f, p_y, \beta, e\Big), \\
\mathcal{L}(H_2) &= \sqrt{(\mathcal{L}(H_{2,1})\mathcal{L}(H_{2,2})}, \\
\mathcal{L}\big(H_{2,1}\big) &= \sum_{OFMA} \Pr(G_F, G_M|f) T(G_O|G_F, G_M) \Pr(G_A|f) \\
&\quad \Pr\Big(\mathcal{O}_O|G_O,f, p_y, \beta, e\Big)\Pr\Big(\mathcal{O}_M|G_M,f, p_y, \beta, e\Big)\Pr\Big(\mathcal{O}_A|G_A,f, p_y, \beta, e\Big), \\
\mathcal{L}\big(H_{2,2}\big) &= \sum_{OFMA} \Pr(G_F, G_M|f) \Pr(G_A|G_M,f) T(G_O|G_F, G_M) \\
&\quad \Pr\Big(\mathcal{O}_O|G_O,f, p_y, \beta, e\Big)\Pr\Big(\mathcal{O}_M|G_M,f, p_y, \beta, e\Big)\Pr\Big(\mathcal{O}_A|G_A,f, p_y, \beta, e\Big).
\end{aligned}$$

The definitions of $G_F, G_M, G_O, G_A,$ and so on are as in Equation 7 and Equation 9.

The subsequent procedures are the same as for the first category.

B Identifying Parents Jointly

For the third category of a parentage analysis (*i.e.* identifying the true and mother jointly), the first hypothesis H_1 is that the alleged parents are the true parents, and the alternative hypothesis H_2 is that the alleged parents are not the true parents. In this case, H_2 implies three possibilities: the first is that both of the alleged parents are not the true parents, and the second and third are that the alleged mother (or father) is a true parent whereas the alleged father (or mother) is not a true parent. Note that each possibility represents two scenarios that the alleged parents are either nonrelatives or are relatives of each of the true parents of the opposite sex. There are thus six scenarios for the hypothesis H_2, denoted by $H_{2,1}, H_{2,2}, \cdots, H_{2,6}$. Similarly, the geometrical mean of the likelihoods of all six scenarios is used as $\mathcal{L}(H_2)$. The likelihoods of these hypotheses are given as follows:

$$\begin{aligned}
\mathcal{L}(H_1) &= \sum_{OFM} \Pr(G_F, G_M|f) T(G_O|G_F, G_M) \Pr\Big(\mathcal{O}_O\Big|G_O,f, p_y, \beta, e\Big) \\
&\quad \Pr\Big(\mathcal{O}_A|G_F,f, p_y, \beta, e\Big)\Pr\Big(\mathcal{O}_{AM}|G_M,f, p_y, \beta, e\Big), \\
\mathcal{L}(H_2) &= \sqrt[6]{\mathcal{L}\big(H_{2,1}\big)\mathcal{L}\big(H_{2,2}\big)\mathcal{L}\big(H_{2,3}\big)\mathcal{L}\big(H_{2,4}\big)\mathcal{L}\big(H_{2,5}\big)\mathcal{L}\big(H_{2,6}\big)}, \\
\mathcal{L}\big(H_{2,1}\big) &= \sum_{OAB} \Pr(G_O|f) \Pr(G_A|f) \Pr(G_B|f) \\
&\quad \Pr\Big(\mathcal{O}_O|G_O,f, p_y, \beta, e\Big)\Pr\Big(\mathcal{O}_A|G_A,f, p_y, \beta, e\Big)\Pr\Big(\mathcal{O}_{AM}|G_B,f, p_y, \beta, e\Big), \\
\mathcal{L}\big(H_{2,2}\big) &\approx \sum_{OFMAB} \frac{1}{2} \Pr(G_F, G_M|f) T(G_O|G_F, G_M) \\
&\quad [\Pr(G_A|G_M,f) \Pr(G_B|G_A,f) + \Pr(G_B|G_F,f) \Pr(G_A|G_B,f)] \\
&\quad \Pr\Big(\mathcal{O}_O|G_O,f, p_y, \beta, e\Big)\Pr\Big(\mathcal{O}_A|G_A,f, p_y, \beta, e\Big)\Pr\Big(\mathcal{O}_{AM}|G_B,f, p_y, \beta, e\Big), \\
\mathcal{L}\big(H_{2,3}\big) &= \sum_{OFMA} \Pr(G_F, G_M|f) T(G_O|G_F, G_M) \Pr(G_A|f) \\
&\quad \Pr\Big(\mathcal{O}_O|G_O,f, p_y, \beta, e\Big)\Pr\Big(\mathcal{O}_{AM}|G_M,f, p_y, \beta, e\Big)\Pr\Big(\mathcal{O}_A|G_A,f, p_y, \beta, e\Big),
\end{aligned}$$

$$
\begin{aligned}
\mathcal{L}(H_{2,4}) = & \sum_{OFMA} \Pr(G_F, G_M|f) T(G_O|G_F, G_M) \Pr(G_A|G_M, f) \\
& \Pr\left(\mathcal{O}_O|G_O, f, p_y, \beta, e\right) \Pr\left(\mathcal{O}_{AM}|G_M, f, p_y, \beta, e\right) \Pr\left(\mathcal{O}_A|G_A, f, p_y, \beta, e\right), \\
\mathcal{L}(H_{2,5}) = & \sum_{OFMB} \Pr(G_F, G_M|f) T(G_O|G_F, G_M) \Pr(G_B|f) \\
& \Pr\left(\mathcal{O}_O|G_O, f, p_y, \beta, e\right) \Pr\left(\mathcal{O}_A|G_F, f, p_y, \beta, e\right) \Pr\left(\mathcal{O}_{AM}|G_B, f, p_y, \beta, e\right), \\
\mathcal{L}(H_{2,6}) = & \sum_{OFMB} \Pr(G_F, G_M|f) T(G_O|G_F, G_M) \Pr(G_B|G_F, f) \\
& \Pr\left(\mathcal{O}_O|G_O, f, p_y, \beta, e\right) \Pr\left(\mathcal{O}_A|G_F, f, p_y, \beta, e\right) \Pr\left(\mathcal{O}_{AM}|G_B, f, p_y, \beta, e\right),
\end{aligned}
$$

where $\mathcal{O}_{AM}$ is the observed genotype of the alleged mother, G_B is the genotype of the false mother, with the meanings of the remaining symbols as for the previous section.

For the hypothesis $H_{2,2}$, the genotypes of both the true and the false parents are correlated, because the false mother (or father) is a relative of both the true and the false fathers (or mothers). Therefore, the distribution of their genotypes accords with the joint distribution $\Pr(G_F, G_M, G_A, G_B|f, p_y, \beta, e)$. However, because of the computational difficulties, we use the following expression to approximate this joint distribution:

$$
\frac{1}{2}\Pr(G_F, G_M|f)[\Pr(G_A|G_M, f)\Pr(G_B|G_A, f) + \Pr(G_B|G_F, f)\Pr(G_A|G_B, f)].
$$

The subsequent procedures are the same as for the first category.