

# Estimating relatedness and relationships using microsatellite loci with null alleles

AP Wagner, S Creel and ST Kalinowski

Department of Ecology, 310 Lewis Hall, Montana State University, Bozeman, MT 59717, USA

Relatedness is often estimated from microsatellite genotypes that include null alleles. When null alleles are present, observed genotypes represent one of several possible true genotypes. If null alleles are detected, but analyses do not adjust for their presence (ie, observed genotypes are treated as true genotypes), then estimates of relatedness and relationship can be incorrect. The number of loci available in many wildlife studies is limited, and loci with null alleles are commonly a large proportion of data that cannot be discarded without substantial loss of power. To resolve this problem, we present a new approach for estimating relatedness and relationships from data sets that include null alleles. Once it is recognized that the probability of the observed genotypes is

dependent on the probabilities of a limited number of possible true genotypes, the required adjustments are straightforward. The concept can be applied to any existing estimators of relatedness and relationships. We review established maximum likelihood estimators and apply the correction in that setting. In an application of the corrected method to data from striped hyenas, we demonstrate that correcting for the presence of null alleles affect results substantially. Finally, we use simulated data to confirm that this method works better than two common approaches, namely ignoring the presence of null alleles or discarding affected loci.

*Heredity* advance online publication, 26 July 2006;  
doi:10.1038/sj.hdy.6800865

**Keywords:** null alleles; relatedness; relationship; estimation; parentage; *Hyaena hyaena*

## Introduction

Microsatellite genotypes are useful for estimating the relationship and relatedness between individuals of unknown ancestry. Current relationship/relatedness estimators either assume that genotypes are error-free (Thompson, 1991) or that genotyping error is rare (Boehnke and Cox, 1997; Marshall *et al.*, 1998). Genotype data, however, often do contain errors, resulting in discrepancies between the *observed* individual genotypes and the *true* underlying genotypes (Dakin and Avise, 2004).

A significant source of such genotyping errors that is not accounted for in current methods for estimating relationship/relatedness is the occurrence of null alleles – alleles that fail to amplify during PCR, often due to mutation within a primer site. Null alleles cause two types of genotyping problems. First, if an individual is homozygous for a null allele ( $nn$ , where  $n$  is a null allele), genotyping will fail. Second, if an individual is a heterozygote with one null allele ( $in$ , where  $i$  is an ordinary non-null allele), the observed genotype will be indistinguishable from a true homozygote ( $ii$ ).

Null alleles complicate the interpretation of all data on coancestry, but the problem is most apparent in parentage analysis (Blouin, 2003). Even when they occur at very low frequencies, null alleles may eliminate potential parents as possible candidates. Parents and offspring must share one identical allele at every locus. If

the observed genotypes at one locus show no identical alleles between a potential parent and offspring, the probability of the parent–offspring relationship is zero (Blouin, 2003), regardless of the number of loci considered. For example, a candidate parent with the observed genotype  $ii$  is excluded as the parent of an offspring with the observed genotype  $jj$ . If there is a null allele at this locus, however, the parent and offspring may share a null allele: true genotypes could be  $in$  for the parent, and  $jn$  for the offspring (Paetkau and Strobeck, 1995). These genotypes are consistent with a parent–offspring relationship, so measuring the frequency of null alleles and taking them into account is clearly necessary to avoid false exclusion of a parent in cases such as this. This problem also affects estimates of relatedness where including genotypes with null alleles may cause underestimation of the coefficient of relatedness between individuals.

The occurrence of null alleles is widely acknowledged and many papers report the results of diagnostic tests for the presence of null alleles (Dakin and Avise, 2004), but options for dealing with null alleles are limited. When null alleles are detected, researchers may eliminate them by redesigning the primer for the affected locus or circumvent the problem by developing new primers for alternate loci that do not contain null alleles. However, these solutions involve additional time and expense, and are not readily available to many investigators who seek to apply microsatellite data to questions in behavioral ecology and conservation biology. Dakin and Avise (2004) summarized 233 studies that detected null alleles in microsatellite data, often at frequencies up to 0.25 (and occasionally as high as 0.70–0.75). Through simulations, they demonstrated that dropping loci with null alleles is

better than including them in analyses and recommended that strategy. However, they did not consider the overall number of loci available. A large number of loci may be required to differentiate between relationship categories or to accurately estimate relatedness (Queller *et al*, 1993; Blouin *et al*, 1996; Sancristobal and Chevalet, 1997; Milligan, 2003), but wildlife biologists are often restricted in the number of loci by the availability of pre-existing primers (Blouin, 2003). Dropping data from problem loci may then prove an impractical option as any omission of loci would substantially reduce inferential and discriminatory power (Marshall *et al*, 1998). Consequently, many studies have simply included loci with null alleles in their analyses (Dakin and Avise, 2004) without explicitly considering the consequences.

A better option for correcting for errors caused by null alleles would be to accommodate them in data analysis (Sobel *et al*, 2002). In this paper, we account for null alleles by modifying well-established maximum likelihood approaches for estimating relationship and relatedness ( $r$ ) (Thompson, 1991; Marshall *et al*, 1998; Blouin, 2003). We account for null alleles by distinguishing between an observed genotype and the set of true genotypes that may have produced that observation. We determine the probability of observing the genotype pair  $ii/ii$ , for example, as the sum of the probabilities that the true genotypes are  $ii/ii$ ,  $in/ii$ ,  $ii/in$ , or  $in/in$  – the four true genotypes that would be observed as  $ii/ii$ . In addition to describing these calculations in detail, we use microsatellite genotypes from striped hyenas (*Hyaena hyaena*) to show that ignoring null alleles can have a substantial impact on estimation of relatedness and inferences concerning population biology. Finally, we use a set of simulations to demonstrate that this technique provides more accurate results than the methods most commonly used in recent papers, while utilizing all available data.

## Relationships

Before showing how maximum likelihood estimators of relationship and relatedness are derived for loci with null alleles, we review the maximum likelihood formulae for estimating genealogical relationships and relatedness from genotypic data not affected by null alleles. We begin with estimating relationship.

In practice, estimating relationship usually means identifying the most likely of a small set of potential relationships that might exist between a pair of individuals, for example, parent–offspring, full-siblings, half-siblings, or unrelated. If  $R$  represents a potential relationship between individuals and  $G_1/G_2$  represents the pair of genotypes observed at a homologous locus in two individuals, by definition, the likelihood of  $R$ ,  $L(R)$ , is the probability of observing  $G_1/G_2$  in two individuals having the relationship  $R$ . These probabilities have been described previously (Thompson, 1991), but they are subtly complex and are essential to understand our estimators – so we present their derivation in detail.

The probability of observing  $G_1/G_2$  in two individuals having relationship  $R$  is calculated by conditioning on the number of alleles in the pair that are identical by descent (IBD) (Cotterman, 1941; Thompson, 1975, 1991). Every pair of individuals will have 0, 1, or 2 alleles IBD at each locus. The probability of observing genotypes  $G_1/G_2$

in a pair of individuals is equal to the probability of observing  $G_1/G_2$  if there are zero alleles identical by descent, plus the probability of observing  $G_1/G_2$  if one allele is IBD, plus the probability of observing  $G_1/G_2$  if two alleles are identical by descent. This approach works because the probability that a pair of individuals has either 0, 1, or 2 alleles IBD is determined by the genealogical relationship between the individuals. Let  $m$  represent the number of alleles IBD between individuals and let  $k_m$  represent the probability that the individuals with genealogical relationship  $R$  have  $m$  alleles IBD (Table 1 lists  $k_m$  values for relationships commonly of interest; Cotterman, 1941; Thompson, 1975). If, for example, two individuals are unrelated, all loci within the pair of individuals will have no alleles IBD ( $k_0 = 1$ ,  $k_1 = 0$ ,  $k_2 = 0$ ). If two individuals are parent–offspring, all loci will share one allele IBD ( $k_0 = 0$ ,  $k_1 = 1$ ,  $k_2 = 0$ ). And if two individuals are full-siblings, loci may share 0, 1, or 2 alleles IBD ( $k_0 = 0.25$ ,  $k_1 = 0.5$ ,  $k_2 = 0.25$ ). Where  $k_0$ ,  $k_1$ , and  $k_2$  are the  $k$ -coefficients for the relationship  $R$ , the probability of observing  $G_1/G_2$ , given  $R$ , is calculated by:

$$P(G_1/G_2|k_0, k_1, k_2) = P(G_1/G_2|m=0)k_0 + P(G_1/G_2|m=1)k_1 + P(G_1/G_2|m=2)k_2 \quad (1)$$

All the terms on the right hand side of Equation (1) are straightforward to calculate (eg Thompson, 1991). Three of these depend on the genealogical relationship between the individuals –  $k_0$ ,  $k_1$ , and  $k_2$ . The remaining probabilities in Equation (1) [ $P(G_1/G_2|m=0)$ ,  $P(G_1/G_2|m=1)$ ,  $P(G_1/G_2|m=2)$ ] depend on the genotypes of the individuals and are calculated from the allele frequencies in the population. Expressions for  $P(G_1/G_2|m)$  are provided in Table 2 for all possible genotype pairs, assuming no inbreeding and no null alleles (Thompson, 1975). These probabilities have been presented in two basic forms: one in which the individuals are ordered and one in which they are not ordered (ie,  $G_1/G_2$  is not distinct from  $G_2/G_1$ ). Either approach is valid, but the approach used affects the probabilities and it is necessary to be consistent. Here, we use the ordered approach for individuals, although the positions of alleles within individuals remain unordered.

The derivations of the probabilities in Table 2 differ according to the number of alleles IBD. If  $m = 0$ , the two genotypes being considered are independent, so that the probability of obtaining the pair of genotypes is simply the product of obtaining each of the two individual genotypes:

$$P(G_1/G_2|m=0) = P(G_1)P(G_2) \quad (2a)$$

**Table 1** A list of  $k$ -coefficients for common relationship categories

Relationship	$k_0$	$k_1$	$k_2$
Parent–offspring	0	1	0
Full-siblings	0.25	0.50	0.25
Half-siblings/grandchild–grandparent/ niece/nephew–uncle/aunt	0.50	0.50	0
First cousin	0.75	0.25	0
Unrelated	1	0	0

$k_m$  represents the probability that two individuals share  $m$  alleles IBD under a given relationship.

**Table 2** A list of all possible pairs of observed genotypes and the probability of each pair given the number of alleles identical by descent ( $m$ )

Genotypes	Probability given $m$ genes IBD		
	$m=0$	$m=1$	$m=2$
$ii/ii$	$p_i^4$	$p_i^3$	$p_i^2$
$ii/ij$	$2p_i^3p_j$	$p_i^2p_j$	0
$ij/ii$	$2p_i^3p_j$	$p_i^2p_j$	0
$ii/jj$	$p_i^2p_j^2$	0	0
$ii/jk$	$2p_i^2p_jp_k$	0	0
$jk/ii$	$2p_i^2p_jp_k$	0	0
$ij/ij$	$4p_i^2p_j^2$	$p_i p_j (p_i + p_j)$	$2p_i p_j$
$ij/ik$	$4p_i^2p_jp_k$	$p_i p_j p_k$	0
$ij/kl$	$4p_i p_j p_k p_l$	0	0

The individual genotypes are ordered, so that  $ii/ij$  is distinct from  $ij/ii$ , because ordering affects the probabilities for genotype pairs.  $p_x$  represents the observed frequency of the allele  $x$  in the population. This table assumes null alleles are not present.

If  $m=2$ , the two genotypes are identical and therefore completely dependent, so that the probability of obtaining the genotypes is the probability of obtaining either genotype once:

$$P(G_1/G_2|m=2) = P(G_1) = P(G_2) \tag{2b}$$

Determining the probability of obtaining the observed genotypes under  $m=1$  is more difficult and is best explained by example. The most complex situation occurs when  $m=1$  and both individuals are homozygous for the same allele ( $G_1$  and  $G_2=ii$ ). Let  $p_i$  indicate the frequency of allele  $i$  in the population. For  $m=1$ , the probability of the individuals having the pair of genotypes  $ii/ii$  is given by:

$$\begin{aligned} P(G_1 = ii/G_2 = ii|m=1) &= P(G_1 = ii)[P(G_2 = ii|G_1 = ii/m=1)] \\ &= p_i^2 [p_i \frac{1}{2}(1) + \frac{1}{2}(1)p_i] = p_i^3 \end{aligned} \tag{2c}$$

In Equation (2c), the probability of the first  $ii$  genotype is calculated directly from allele frequencies, but the probability of obtaining a second  $ii$  genotype must then take into account that one allele is IBD to an allele in the first individual. Thus, the probability for the second individual's genotype is the product of the probability of the second individual having one  $i$  allele ( $p_i$ ) and, for the second allele in the second individual, the probability ( $=1$ ) that the IBD allele is an  $i$ , and the probability ( $=\frac{1}{2}$ ) that IBD allele is in the second position. This is the first term within square brackets. The second term within brackets accounts for the alternative possibility that the IBD allele is in the first position. Probabilities are calculated for the IBD allele being in each of the two possible positions in the second individual and then summed, giving  $p_i^2 p_i = p_i^3$  (Table 2).

Using the same approach, the probability of two individuals having the pair of genotypes  $ij/ik$  when  $m=1$  would then consider the probability of getting an  $i$  and  $j$  in the first individual in either configuration ( $p_i p_j + p_j p_i = 2p_i p_j$ ). Probabilities for the second individual are dependant on the probability of having a  $k$  allele ( $=p_k$ ) and there being one allele IBD with the first individual: the probability that the IBD allele is an  $i$  is  $\frac{1}{2}$ , given that an

or  $j$  could be IBD, while the probability of that IBD allele being in either the first or second position is  $\frac{1}{2}$ :

$$\begin{aligned} P(G_1 = ij/G_2 = ik|m=1) &= P(G_1 = ij)[P(G_2 = ik|G_1 = ij/m=1)] \\ &= 2p_i p_j [\frac{1}{2} * p_k] + 2p_i p_j [p_k * \frac{1}{2}] \\ &= p_i p_j p_k \end{aligned} \tag{2d}$$

Similar logic can be used to determine the remaining seven probabilities for  $m=1$  in Table 2, of which four have zero probability because a pair of genotypes with no alleles in common cannot have one allele IBD, given that we are not (yet) allowing for null alleles in 'observed' genotypes.

Once the probabilities of Table 2 are defined, relationships are evaluated using Equation (1), so that the likelihood of the genotypic data is calculated for each candidate relationship. The values for  $P(G_1/G_2|k_0, k_1, k_2)$  are multiplied across loci to yield the likelihood of the relationship,  $L(R)$ . By definition, the maximum likelihood relationship between two individuals is the relationship for which the observed data is most probable.

### Relatedness ( $r$ )

Relatedness ( $r$ ) may be interpreted as the proportion of genes IBD between two individuals or groups of individuals (Cotterman, 1941). For outbred individuals,  $r$  is given by (Thompson, 1975):

$$r = \frac{k_1}{2} + k_2 \tag{3}$$

The maximum likelihood estimate of  $r$ ,  $ML(r)$ , is equal to the maximum likelihood estimate of  $k_1/2$  plus the maximum likelihood estimate of  $k_2$ . Maximum likelihood estimates of  $k_1$  and  $k_2$  can be obtained from genotypic data by varying  $k_0, k_1$ , and  $k_2$  through all possible values (subject to the constraint that they sum to one) to find the set of  $k$ -coefficients that maximize the product of  $P(G_1/G_2|k_0, k_1, k_2)$  (Equation (1)) across all loci. Note the difference between estimates of relationship and estimates of relatedness. When estimating relationship, values for  $k_0, k_1$ , and  $k_2$  are determined by the genealogy of the relationship (Table 1) and then used in Equation (1). When estimating  $r$ , Equation (1) is used to find the optimum values of  $k_1$  and  $k_2$  that are then used in Equation (3). If  $r$  is being calculated for an evaluation of the relatedness of one individual to a group, the individual of interest is first paired with each group member and an average of the pairwise  $r$ -values is used.

### Null alleles

The formulae above show how to estimate relationship and relatedness assuming genotypes have no null alleles. In other words, the above formulae show how to calculate the probability if the true genotypes in two individuals are  $G_1$  and  $G_2$ . If null alleles are present at a locus, however, the probability of observing  $G_1$  and  $G_2$ ,  $P(\text{Observe } G_1/G_2|k_0, k_1, k_2)$ , needs to be determined. Only observed homozygotes may have null alleles. If  $G_1$  or  $G_2$  is an observed heterozygote (eg  $ij$ ), we assume that the observed genotype is correct. However, if  $G_1$  or  $G_2$  is an observed homozygote, it can be a true homozygote (eg  $ii$ ) or a heterozygote with one null and one non-null

allele (eg *in*). If there are no homozygotes observed in the pair,  $G_1/G_2$ , the only possible true genotype pair is identical to the observed pair. However, if one homozygote is observed, there are two possible genotype pairs (eg the observed *ii/ij* may actually be *ii/ij* or *in/ij*). Further, if two homozygotes are observed, there are four possible true genotype pairs (eg the observed *ii/ii* may actually be *ii/ii*, *in/ii*, *ii/in*, or *in/in*). Genotype pairs, therefore, may have either 0, 1, or 2 null alleles, depending on how many homozygotes are observed. As up to four true genotype pairs can have the same observed genotype, the likelihood of an observed genotype pair is calculated by summing the probabilities of all the genotype pairs that have the same observed genotype. For example, the probability of observing *ii/ii* is calculated by summing the probabilities of two individuals actually having genotypes *ii/ii* (no null allele), *in/ii* (null allele in first individual), *ii/in* (null allele in second individual), and *in/in* (null allele in both individuals).

Table 3 lists the true genotypes that may produce each of the nine possible observed genotype pairs and the

**Table 3** A list of all possible observed genotypes for a pair of individuals, the underlying true genotypes that can produce the observed genotypes given the possible number of null alleles ( $n$ ), and the probability of each underlying genotype pair given the number of alleles identical by descent ( $m$ )

Observed genotypes	True genotypes	n	Probability of true genotypes given $m$ genes IBD		
			$m=0$	$m=1$	$m=2$
<i>ii/ii</i>	<i>ii/ii</i>	0	$p_i^4$	$p_i^3$	$p_i^2$
	<i>ii/in</i>	1	$2p_i^3p_n$	$p_i^2p_n$	0
	<i>in/ii</i>	1	$2p_i^3p_n$	$p_i^2p_n$	0
	<i>in/in</i>	2	$4p_i^2p_n^2$	$p_i p_n(p_i + p_n)$	$2p_i p_n$
<i>ii/ij</i>	<i>ii/ij</i>	0	$2p_i^3p_j$	$p_i^2p_j$	0
	<i>in/ij</i>	1	$4p_i^2p_j p_n$	$p_i p_j p_n$	0
<i>ij/ii</i>	<i>ij/ii</i>	0	$2p_i^3p_j$	$p_i^2p_j$	0
	<i>ij/in</i>	1	$4p_i^2p_j p_n$	$p_i p_j p_n$	0
<i>ii/jj</i>	<i>ii/jj</i>	0	$p_i^2 p_j^2$	0	0
	<i>ii/jn</i>	1	$2p_i^2 p_j p_n$	0	0
	<i>in/jj</i>	1	$2p_i p_j^2 p_n$	0	0
	<i>in/jn</i>	2	$4p_i p_j p_n^2$	$p_i p_j p_n$	0
<i>ii/jk</i>	<i>ii/jk</i>	0	$2p_i^2 p_j p_k$	0	0
	<i>in/jk</i>	1	$4p_i p_j p_k p_n$	0	0
<i>jk/ii</i>	<i>jk/ii</i>	0	$2p_i^2 p_j p_k$	0	0
	<i>jk/in</i>	1	$4p_i p_j p_k p_n$	0	0
<i>ij/ij</i>	<i>ij/ij</i>	0	$4p_i^2 p_j^2$	$p_i p_j(p_i + p_j)$	$2p_i p_j$
<i>ij/ik</i>	<i>ij/ik</i>	0	$4p_i^2 p_j p_k$	$p_i p_j p_k$	0
<i>ij/kl</i>	<i>ij/kl</i>	0	$4p_i p_j p_k p_l$	0	0

$p_x$  represents the frequency of allele  $x$  in population corrected for the presence of null alleles (ie,  $p_n$  is considered when summing allele frequencies to 1). The genotypes within a pair are ordered, for example, *ii/ij* is distinct from *ij/ii*. The probability of each observed genotype pair at each locus is obtained by summing the probabilities of the possible underlying true genotypes.

corresponding probabilities under each value of  $m$ . Once these new probabilities are determined, the probability of the observed genotypes is still calculated following Equation (1) by listing all true genotype pairs that would be observed as  $G_1/G_2$  and then summing  $P(\text{Observe } G_1/G_2 | k_0, k_1, k_2)$  values for each possible true genotype. In essence, all this entails is using the multiple probabilities for the true genotype pairs in Table 3, rather than the probabilities in Table 2. For example, if the observed genotypes are *ii/ii*, then the true underlying genotypes are taken from Table 3 and the probability of the observed genotypes, accounting for the possible presence of null alleles at a single locus, is thus:

$$\begin{aligned} P(\text{Observe } G_1 = ii/G_2 = ii | k_0, k_1, k_2) \\ = P(\text{Observe } ii/ii | k_0, k_1, k_2) \\ = P(ii/ii | k_0, k_1, k_2) + P(in/ii | k_0, k_1, k_2) \\ + P(ii/in | k_0, k_1, k_2) + P(in/in | k_0, k_1, k_2) \end{aligned} \quad (4a)$$

The four probabilities listed in the right-hand side of Equation (4a) are calculated using Equation (1). For example,

$$\begin{aligned} P(ii/ii | k_0, k_1, k_2) = P(ii/ii | m = 0)k_0 \\ + P(ii/ii | m = 1)k_1 \\ + P(ii/ii | m = 2)k_2 \end{aligned} \quad (4b)$$

and

$$\begin{aligned} P(in/ii | k_0, k_1, k_2) = P(in/ii | m = 0)k_0 \\ + P(in/ii | m = 1)k_1 \\ + P(in/ii | m = 2)k_2 \end{aligned} \quad (4c)$$

For those true underlying genotypes having null alleles ( $n=1$  or  $2$ ), the probabilities are determined following the same logic used in Table 2 (where  $n=0$ ). For example,  $P(in/ii | m=0)$ ,  $P(in/ii | m=1)$ , and  $P(in/ii | m=2)$  are calculated in the same way as was  $P(ij/ii | m)$  for  $m=0, 1$ , or  $2$ . To be more specific,  $P(in/ii | m=0)$ ,  $P(in/ii | m=1)$ , and  $P(in/ii | m=2)$  are equal to  $2p_i^3 p_n$ ,  $p_i^2 p_n$ , and  $0$  (respectively, Table 3). Note that although  $p_n$  is a total null allele frequency, we make no assumptions about the number of different null alleles at a locus or about whether any null alleles are IBD. We need only to account for the possibilities that there are 0 or 1 or 2 null alleles in the pair (summing along columns in Table 3) and that there are 0 or 1 or 2 null or non-null alleles IBD (summing along rows in Table 3). As is true for all alleles, the probability under  $m=0$  and the partial probability under  $m=1$  account for the possibility that the null allele is not IBD, while the alternative that there is an IBD null allele is accounted for by the partial probability under  $m=1$  and the probability under  $m=2$  (see Appendix for a demonstration that having multiple non-IBD null alleles does not affect the probability of observing any particular genotype).

Calculating  $P(\text{Observe } G_1/G_2 | k_0, k_1, k_2)$  requires knowing the frequency of the null allele,  $p_n$ . In practice,  $p_n$  will not be known, but it can be estimated with several approaches (Chakraborty *et al*, 1992; Brookfield, 1996; Summers and Amos, 1997; Kalinowski and Taper, in press) that have been implemented in programs such as Genepop (Raymond and Rousset, 1995), Cervus (Marshall *et al*, 1998), Micro-Checker (Van Oosterhout *et al*, 2004) and ML-Relate (Kalinowski *et al*, 2006) or can

be programmed into an Excel spreadsheet (Kalinowski and Taper, in press). Frequencies for observed non-null alleles may be corrected accordingly. As a source of typing error, inaccurate estimates of  $p_n$  would affect probabilities of false exclusion in parentage analysis (San Cristobal and Chevalet, 1997; Marshall *et al*, 1998). The probabilities above and those in Table 3 assume genotypes are observed in each case and that there are no homozygotes for a null allele producing 'blank' genotypes. The maximum likelihood approach developed by Kalinowski and Taper (in press) and implemented in ML-Relate (Kalinowski *et al*, 2006) uses an EM algorithm and performs better under this assumption than the approaches of Summers and Amos (1997) and Chakraborty *et al* (1992).

One aspect of Table 3 is particularly note worthy. When parents and offspring are being considered, definitive exclusion (as opposed to the relative considerations made below) of the true parent-offspring ( $PO$ ) relationship occurs when the likelihood of that relationship is zero ( $L(PO) = 0$ ). By this measure, false exclusion of the true relationship would occur when, at any locus, parent and offspring are true heterozygotes with one common null allele and distinct non-null alleles ( $in/jn$ ) but are observed as homozygotes for different alleles ( $ii/jj$ ). The probability of false exclusion is then equal to the probability of having  $in/jn$  when one allele is IBD ( $m = 1$ ). From Table 3, this is  $p_i p_j p_n$ , which is just the probability of having any two different alleles ( $p_i p_j$ ) multiplied by the frequency of the IBD null allele. By definition, this is equivalent to the observed heterozygosity ( $He_{obs}$ ) multiplied by  $p_n$ , so the probability of false exclusion of a parent-offspring relationship at a single locus if null alleles are ignored is  $He_{obs} p_n$ .

### Comparison of analytical methods

As discussed at the outset, there have been two alternative approaches to the analysis of microsatellite data that include null alleles when redesigning existing or developing new primers is not an option. One approach is to drop the data from affected loci. Another approach is to use the data from affected loci and proceed with estimation of relatedness or relationship using Table 2, ignoring the existence of null alleles. Above, we developed a new approach that explicitly accounts for null alleles by using Table 3. We now use empirical data to show how the results of these approaches differ.

Table 4 shows microsatellite genotypes at eight loci from a putative family of striped hyenas (Wagner, 2006), within which the adult female (F09) was thought to be

the mother of the three cubs (cubs 30, 31, and 32). Ignoring null alleles and using Table 2 and Equation (1), we tested the hypothesized parent-offspring relationship for F09 to each of the three cubs. F09 is immediately ruled out as the potential mother for two of the three cubs (cub 30 and cub 31), because the female and cubs share no alleles identical in state (and therefore none IBD) at locus CCR5. At this locus,  $P(G_1/G_2 | k_0, k_1, k_2) = 0$  and, since probabilities are multiplied across loci to determine the probability of the relationship, the entire probability of maternity is 0. This is a good illustration of the general problem that null alleles can easily create observed genotypes at one locus that are impossible under the hypothesized relationship even if genotypes at other loci strongly support that relationship.

In this case, null alleles were detected at three of the evaluated loci (CCRA5, CCRA3, and the critical CCR5) and, for loci where null alleles were detected, adjusted and null allele frequencies were calculated following Kalinowski and Taper (in press). Although the null allele frequency at CCR5 was relatively low (0.074), it appears to have created problems for assigning maternity. The characteristics of this data set illustrate a common problem contributing to the prevalence of studies that include, but do not correct for, loci with null alleles (Dakin and Avise, 2004): in many existing data sets from wildlife studies researchers are restricted to using existing primers, only a limited number of loci are available, null alleles are present, but retention of inferential and discriminatory power requires salvaging those problem loci.

Table 5 summarizes conclusions about the maternity of the cubs using the three approaches. For the female and each cub, the probability of the genotypes for the three adult-cub pairs was calculated for parent-offspring *vs* unrelated relationships, although any hypothesized relationships could be used for comparison. Likelihood ratios were used to evaluate the relative degree of support for the competing relationships. A ratio  $>1$  indicates that the relationship in the numerator is more likely, whereas a ratio  $<1$  favors the denominator. Clearly, support for the parent-offspring hypothesis is highly dependent on the approach that is employed. Accounting for the occurrence of null alleles can give very different results than ignoring them (cubs 30 and 31) or discarding them (cubs 31 and 32). Only our approach, applying a correction for the presence of null alleles, retains enough information and correctly interprets the observed genotypes to indicate that F09 is likely to be the mother of all three cubs.

We also used the three competing approaches to calculate the coefficient of relatedness between female F09 and each of the three cubs, and to the cubs as a group

**Table 4** Observed microsatellite genotypes at eight loci for a group of striped hyenas

Individual ID	Locus							
	CCR4	CCR6	CCROC01	CCRA5*	CCROC05	CCRA3*	CCROC06	CCR5*
Female09	114/130	114/116	199/203	143/149	159/167	143/143	161/169	148/148
Cub30	114/130	114/114	203/203	143/143	159/167	143/143	161/169	152/152
Cub31	114/114	114/116	203/203	149/149	157/159	143/143	169/169	150/150
Cub32	114/130	114/116	203/203	143/143	157/159	143/143	169/169	148/150

Numbers in the table indicate specific observed alleles, expressed as number of base-pairs in the allele. Loci with null alleles are indicated by an asterisk (\*).

**Table 5** Relationship likelihood ( $L(R)$ ) ratios and maximum likelihood calculations of relatedness ( $ML(r)$ ) for the population subset

Individuals	Approach used	L(R) ratio	
		PO/UR	ML(r)
F09 cub30	Nulls not considered	0	0.44
	Loci with nulls not used	14.91	0.70
	Correction for nulls applied	15.09	0.61
F09 cub31	Nulls not considered	0	0.33
	Loci with nulls not used	4.27	0.50
	Correction for nulls applied	20.50	0.50
F09 cub32	Nulls not considered	15.21	0.50
	Loci with nulls not used	3.25	0.39
	Correction for nulls applied	10.17	0.50
F09 Cubs as a group	Nulls not considered		0.43
	Loci with nulls not used	NA	0.53
	Correction for nulls applied		0.54

$L(R)$  ratios are the probability of the observed genotypes given the hypothesized parent-offspring (PO) relationship vs the alternative unrelated (UR) relationship.  $ML(r)$  was determined between F09 and each cub and all cubs as a group. Calculations were made without considering null alleles (Table 2), without using loci for which null alleles were detected, and with all loci corrected for the probabilities of null alleles at each locus (Table 3).

(Table 5). In general, this example shows that correcting for null alleles sometimes gives the same result as dropping loci with null alleles (cub 31), sometimes gives the same result as ignoring null alleles (cub 32), and sometimes differs from both of these methods (cub 30). When the presence of null alleles is not considered and all loci are used, the observed genotype probabilities for the F09-cub 31 pair under-estimates the relatedness of the female to her own cub by more than 20%. When cubs are viewed as a group, support for this female as the mother of this litter would be greatly reduced based on this result. Using the corrected approach, however, the female would still be considered a likely candidate.

In addition to the empirical tests above, we used computer simulation to further test the effectiveness of our method for accommodating null alleles in relationship and relatedness estimation. We did this by repeatedly simulating genotype data for pairs of related individuals and evaluating how accurately our method could estimate the relationship between individuals, in comparison to the commonly employed methods of ignoring null alleles or discarding problem loci. Each iteration of the simulation began by simulating parametric allele frequencies in a randomly mating population, using broken stick random numbers. This method produces allele frequencies that are uniformly distributed in multidimensional space (eg [0.2, 0.2, 0.2, 0.2, 0.2] is as likely as [0.96, 0.01, 0.01, 0.01, 0.01]) (Devroye, 1986). All simulations had five alleles per locus, which resulted in an average observed heterozygosity of 0.66 (which was reduced to 0.59 when null alleles were introduced). With these allele frequencies, genotypes of the adults in the population were simulated by multinomial sampling. We simulated relatedness within the population by forming monogamous mating pairs from the adults and then simulating genotypes for two offspring per mating pair. For example, most of our simulated data had 96 individuals in 24 families, each with a dam, sire, and two offspring. We simulated null alleles by choosing

one allele at a locus to be a null allele. Loci that were homozygous for null alleles were treated as missing data. Note that the frequency of null alleles in our simulated data is a random variable. Null alleles that have a high frequency in a population are more likely to interfere with relationship estimation than null alleles having a low frequency. Therefore, we binned simulated data according to the frequency of null alleles, with a bin width of 0.10 (eg data that had null alleles with a frequency  $\geq 0.15$  and  $< 0.25$  were placed in the '0.2' bin). In some cases, we simulated data for multiple loci having null alleles. Here, data were binned according to the average frequency of null alleles at all loci with null alleles.

We used two statistics to measure how accurately relatedness and relationship could be estimated. The accuracy of estimates of relatedness was measured by the root mean squared error (RMSE) of the estimates. The accuracy of estimates of relationship was measured by the proportion of simulated data that successfully identified the correct relationship from among four possibilities: unrelated, half-siblings, full-siblings, and parent-offspring. Under each set of conditions, at least one thousand simulated data sets were used to estimate these statistics.

We examined the effect of the following variables upon the accuracy of estimates of relatedness and relationship: sample size ( $N_{\text{Samples}}$ ), total number of loci ( $N_{\text{Loci}}$ ), number of loci having null alleles ( $N_{\text{Nulls}}$ ), and the frequency of null alleles ( $p_{\text{null}}$ ) (Tables 6 and 7). In addition, we evaluated six methods for estimating relatedness and relationship – methods that spanned the range of options available to geneticists encountering null alleles. The first method, IGNORE, simply ignored null alleles. The next three methods are variations of the maximum likelihood approach we present above. ML-APRIORI assumes that the user knows *a priori* which loci have null alleles. ML-DETECTED assumes the user does not know which loci have null alleles, and therefore must test for them. We used a Monte-Carlo randomization test (Guo and Thompson, 1992) for excess homozygosity and the  $U$ -statistic (Rousset and Raymond, 1995) to detect null alleles. Loci that had a one-tailed  $P$ -value of  $< 0.05$  divided by the number of loci in the data were classified as having null alleles. ML-ALL assumed null alleles were present at all loci. This strategy may appear unreasonable, but if a locus did not have a null allele, the estimated frequency of a null allele at the locus was usually small. Last, we tested two variants of removing loci with null alleles. REMOVE-APRIORI assumed that loci having null alleles were identified *a priori*. REMOVE-DETECTED used the randomization test described above to detect loci having null alleles. In each case, such loci were removed from the data.

Our method of correcting for the presence of null alleles (ML-DETECTED, ML-APRIORI, or ML-ALL) improves the accuracy of relatedness identification for full-siblings (Table 6), but the differences can appear subtle in this context. However, the ML-DETECTED method improves RMSE by up to 6.2% over IGNORE (average improvement = 2.2%) and represents up to a 14.0% improvement in relatedness estimation over REMOVE-DETECTED (average = 7.6%). Across simulated conditions, increasing the number of loci considered has the greatest impact on reducing error in relatedness

estimation and it is only when more than six loci are considered that REMOVE-DETECTED approaches the accuracy of the IGNORE or ML approaches. Our method also performs the best in relationship estimation, improving the ability to correctly identify parent-offspring relation-

ships (Table 7). The biggest improvements over IGNORE occur when null allele frequencies are high or when many loci are available. The largest improvements in accuracy of our method relative to REMOVE methods occur when null alleles are present at multiple loci.

**Table 6** The root mean square error of estimates of relatedness between full-siblings under simulated conditions varying the sample size ( $N_{\text{Samples}}$ ), total number of loci ( $N_{\text{Loci}}$ ), number of loci having null alleles ( $N_{\text{Nulls}}$ ), and the frequency of null alleles ( $p_{\text{null}}$ ), as indicated (by bold type) in the first four columns. Lower values indicate greater accuracy in relatedness estimation

$N_{\text{Loci}}$	$N_{\text{Nulls}}$	$N_{\text{Samples}}$	$p_{\text{null}}$	Statistical method					
				Ignore	ML detected	ML apriori	ML all	Remove detected	Remove apriori
<i>No null alleles</i>									
6	None	96	—	0.210	0.210	0.210	0.210	0.210	—
<i>Vary number of loci having null alleles</i>									
6	<b>1</b>	96	0.2	0.216	0.214	0.214	0.214	0.226	0.224
6	<b>2</b>	96	0.2	0.221	0.216	0.216	0.216	0.236	0.243
6	<b>3</b>	96	0.2	0.226	0.220	0.220	0.220	0.253	0.268
<i>Vary frequency of null allele</i>									
6	2	96	<b>0.1</b>	0.217	0.215	0.214	0.214	0.228	0.243
6	2	96	<b>0.2</b>	0.221	0.216	0.216	0.216	0.236	0.243
6	2	96	<b>0.3</b>	0.225	0.218	0.218	0.218	0.239	0.243
6	2	96	<b>0.4</b>	0.226	0.219	0.219	0.219	0.242	0.243
<i>Vary sample size</i>									
6	2	<b>48</b>	0.2	0.225	0.221	0.220	0.220	0.238	0.243
6	2	<b>96</b>	0.2	0.221	0.216	0.216	0.216	0.236	0.243
6	2	<b>192</b>	0.2	0.218	0.212	0.212	0.212	0.235	0.243
<i>Vary total number of loci</i>									
<b>6</b>	1	96	0.2	0.216	0.214	0.214	0.214	0.226	0.224
<b>12</b>	1	96	0.2	0.160	0.159	0.158	0.159	0.164	0.164
<b>24</b>	1	96	0.2	0.116	0.116	0.116	0.116	0.118	0.118

Bold characters indicate changes that were made between the simulations and correspond to the labels of the x-axis in Figures 1 and 2.

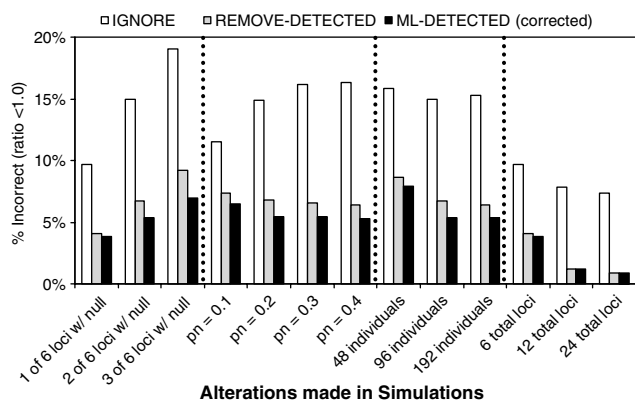
**Table 7** Proportion of simulated data sets successfully able to identify the relationship between a parent-offspring pair when differing characteristics of the data set are varied: sample size ( $N_{\text{Samples}}$ ), total number of loci ( $N_{\text{Loci}}$ ), number of loci having null alleles ( $N_{\text{Nulls}}$ ), and the frequency of null alleles ( $p_{\text{null}}$ ). Higher values indicate greater accuracy in relationship estimation

$N_{\text{Loci}}$	$N_{\text{Nulls}}$	$N_{\text{Samples}}$	$p_{\text{null}}$	Statistical method					
				Ignore	ML detected	ML apriori	ML all	Remove detected	Remove apriori
<i>No null alleles</i>									
6	<b>None</b>	96	—	0.781	0.780	0.780	0.779	0.780	—
<i>Vary number of loci having null alleles</i>									
6	<b>1</b>	96	0.2	0.718	0.757	0.758	0.759	0.740	0.751
6	<b>2</b>	96	0.2	0.663	0.728	0.730	0.731	0.706	0.705
6	<b>3</b>	96	0.2	0.624	0.711	0.713	0.716	0.676	0.653
<i>Vary frequency of null allele</i>									
6	2	96	<b>0.1</b>	0.698	0.731	0.739	0.739	0.715	0.705
6	2	96	<b>0.2</b>	0.663	0.728	0.730	0.731	0.706	0.705
6	2	96	<b>0.3</b>	0.644	0.724	0.725	0.725	0.714	0.705
6	2	96	<b>0.4</b>	0.636	0.718	0.719	0.719	0.707	0.705
<i>Vary sample size</i>									
6	2	<b>48</b>	0.2	0.660	0.714	0.721	0.723	0.703	0.705
6	2	<b>96</b>	0.2	0.663	0.728	0.730	0.731	0.706	0.705
6	2	<b>192</b>	0.2	0.663	0.734	0.735	0.736	0.711	0.705
<i>Vary total number of loci</i>									
<b>6</b>	1	96	0.2	0.718	0.757	0.758	0.759	0.740	0.751
<b>12</b>	1	96	0.2	0.827	0.883	0.889	0.888	0.877	0.884
<b>24</b>	1	96	0.2	0.898	0.965	0.974	0.974	0.963	0.965

Bold characters indicate changes that were made between the simulations and correspond to the labels of the x-axis in Figures 1 and 2.

We also considered the probabilities of drawing false conclusions using two simple ways of evaluating population genotype data. For the same three methods applied in Table 5 (IGNORE, ML-DETECTED, and REMOVE-DETECTED), we determined the percentage of simulated parent-offspring pairs for which the likelihood of the true parent-offspring relationship was less than the likelihood of being unrelated (likelihood ratio less than one), as a measure of the probability of reaching a false conclusion in relationship estimation (Figure 1). In every case, the probability of drawing false conclusions is highest when null alleles are ignored: IGNORE leads to false conclusions 7.4–19.1% of the time. Applying our correction for null alleles also reduces the probability of drawing a false conclusion relative to REMOVE methods, except when a large number of loci are considered: in this case the accuracy of the two approaches is equivalent. We also used the percentage of calculated  $r$ -values for all parent-offspring pairs that deviated by more than  $\pm 20\%$  ( $\pm 0.1$ ) from the true value of 0.5 as a further test of these three competing methods (Figure 2). For relatedness, up to 7% (IGNORE) or 4% (REMOVE-DETECTED) more of the calculations over or underestimated  $r$  by  $>20\%$  than when the correction for null alleles is applied.

Finally, we evaluated the effect of the number of loci ( $N_{\text{Loci}}$ ) on the accuracy of relationship and relatedness estimates from our simulated genetic data having no null alleles (Table 8). For a given number of loci, Table 8 indicates the levels of accuracy that could be expected if the data are error-free and shows that accuracy improves when the number of loci increases. Comparing the results in Table 8 to Tables 6 and 7 also allows for evaluation of the performance of our method to two methods that would eliminate null alleles from the data. First, researchers might consider replacing loci containing null alleles with loci having no null alleles. However, if additional primers are developed, loci with null alleles



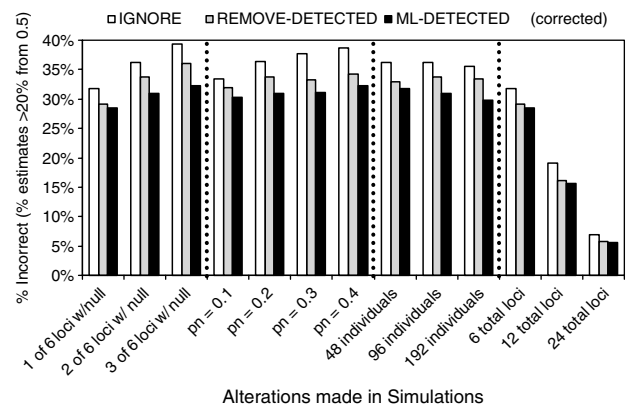
**Figure 1** Probability of falsely concluding that the likelihood of unrelated is greater than the likelihood of the true parent-offspring relationship from the simulated data using the three competing approaches: ignoring problem loci and including loci with null alleles without applying a correction (IGNORE), removing loci where null alleles were detected from data analysis (REMOVE-DETECTED), and applying our correction for null alleles at loci where they were detected (ML-DETECTED). Vertical dotted lines separate sub-sets of the simulated data within which one characteristic of the data set was varied (as described in Table 7).  $y$ -axis indicates the percentage of parent-offspring pairs for which  $L(PO)/L(UR)$  was incorrectly determined to be  $<1.0$ .

should be retained and our method applied: accuracy would be improved by retaining all loci and applying our correction rather than replacing loci because the latter would reduce the total number of loci considered (compare ML results for  $N_{\text{Loci}}$  given  $N_{\text{Nulls}}$  in Tables 6 and 7, to results for  $N_{\text{Loci}} - N_{\text{Nulls}}$  in Table 8). Second, if researchers consider redesigning existing primers, instead of replacing loci, comparing the results in Table 8 to the ML results in Tables 6 and 7 shows that, for the

**Table 8** The accuracy of simulated genetic data in estimating relationship or relatedness when no null alleles are present

$N_{\text{Loci}}$	Proportion $R_{PO}$ correct	RMSE $r_{FS}$
1	0.44	0.42
2	0.59	0.31
3	0.65	0.27
4	0.71	0.24
5	0.75	0.22
6	0.78	0.21
7	0.81	0.20
8	0.83	0.19
9	0.85	0.18
10	0.87	0.17
11	0.88	0.16
12	0.90	0.16
13	0.91	0.15
14	0.92	0.15
15	0.92	0.14
16	0.93	0.14
17	0.94	0.14
18	0.94	0.13
19	0.95	0.13
20	0.95	0.13
21	0.96	0.12
22	0.96	0.12
23	0.96	0.12
24	0.97	0.12

The first column shows the number of loci ( $N_{\text{Loci}}$ ) in the simulated data, the second the proportion of simulated data sets that successfully identified parent offspring pairs ( $R_{PO}$ ), the third the root mean square error (RMSE) of estimates of relatedness between full sibs ( $r_{FS}$ ).



**Figure 2** Probability of over or underestimating relatedness for parents and offspring by more than 20% from the simulated data using three competing approaches: IGNORE, REMOVE-DETECTED, and ML-DETECTED. Vertical dotted lines separate sub-sets of the simulated data within which one characteristic of the data set was varied (Table 6).  $y$ -axis indicates the percentage of parent-offspring pairs for which  $r$  estimates deviated by more than  $\pm 20\%$  from the true value.



same  $N_{\text{Loci}}$ , the differences between the accuracy achieved using error-free data or applying our corrected method are small and diminish as the number of loci increase. When few loci are available, the costs of redesigning primers should then be evaluated relative to the costs (slightly reduced accuracy) of applying our corrected method to the imperfect data at hand. When many loci are available, researchers would receive little benefit by redesigning primers rather than applying our correction.

## Conclusion

Failure to correct for the presence of null alleles in microsatellite data can produce badly biased estimates of relatedness and incorrect assessments of relationships. Even at low frequencies, null alleles can have a large impact. Dropping data from problem loci altogether can significantly alter the likelihoods of competing relationships and this solution needlessly discards valuable information. Even under relatively simple scenarios, dropping loci does not perform as well as, and never performs better than, correcting for null alleles, so it cannot be considered the 'conservative' approach. Inclusion of loci at which null alleles are present, without correcting for them, is the approach most likely to lead to false conclusions in relatedness and relationship estimation.

Our method for including null alleles in calculations of relationship probabilities and relatedness values is easy to apply to co-dominant genotype data. Once null alleles are detected and their frequency estimated, all of the information required for these adjusted calculations is present in the original genotyping data, so application of our method bears no additional costs. Little modification is required to the methods already in place for evaluating relationships and relatedness: all that is required is to use Table 3 rather than Table 2 when applying Equation (1). This new approach provides a means by which a previously recognized and widespread problem, predominantly discussed as a theoretical or conceptual issue, can be corrected in practice.

## Acknowledgements

We thank Mark L Taper and Steve Cherry for their help with the manuscript. Grants from the Living Desert Museum and Gardens, People's Trust for Endangered Species, the Cleveland Metroparks Zoo/Cleveland Zoological Society, National Geographic Society, the Brookfield Zoo/Chicago Zoological Society, British Airways, and National Science Foundation (IBN-0238169) supported field and lab work on striped hyaenas. We thank Kenya Wildlife Service and Ministry of Education, Science, and Technology for permission to conduct field research and collect tissue samples and Laurence G. Frank for his assistance in the field.

## References

Blouin MS (2003). DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. *Trends Ecol Evolut* 18: 503–511.

- Blouin MS, Parsons M, Lacaille V, Lotz S (1996). Use of microsatellite loci to classify individuals by relatedness. *Mol Ecol* 5: 393–401.
- Boehnke M, Cox NJ (1997). Accurate inference of relationships in sib-pair linkage studies. *Am J Hum Genet* 61: 423–429.
- Brookfield JFY (1996). A simple new method for estimating null allele frequency from heterozygote deficiency. *Mol Ecol* 5: 453–455.
- Chakraborty R, De Andrade M, Daiger SP, Budowle B (1992). Apparent heterozygote deficiencies observed in DNA typing and their implications in forensic applications. *Ann Hum Genet* 56: 45–57.
- Cotterman CW (1941). Relatives and human genetic analysis. *Sci Monthly* 53: 227–234.
- Dakin EE, Avise JC (2004). Microsatellite null alleles in parentage analysis. *Heredity* 93: 504–509.
- Devroye L (1986). *Non-Uniform Random Variate Generation*. Springer-Verlag: New York.
- Guo SW, Thompson EA (1992). Performing the exact test of Hardy-Weinberg proportions for multiple alleles. *Biometrics* 48: 361–372.
- Kalinowski ST, Wagner AP, Taper ML (2006). ML-Relate: a computer program for maximum likelihood estimation of relatedness and relationship. *Mol Ecol Notes* 6: 576–579.
- Kalinowski ST, Taper ML (in press). Maximum likelihood estimation of the frequency of null alleles at microsatellite loci. *Conserv Genet*.
- Marshall TC, Slate J, Kruuk LEB, Pemberton JM (1998). Statistical confidence for likelihood-based paternity inference in natural populations. *Mol Ecol* 7: 639–655.
- Milligan BG (2003). Maximum-likelihood estimation of relatedness. *Genetics* 163: 1153–1167.
- Paetkau D, Strobeck C (1995). The molecular-basis and evolutionary history of a microsatellite null allele in bears. *Mol Ecol* 4: 519–520.
- Queller DC, Strassman JE, Hughes CR (1993). Microsatellites and Kinship. *Trends Ecol Evolut* 8: 285–288.
- Raymond M, Rousset F (1995). Genepop (Version-1.2) – population-genetics software for exact tests and ecumenism. *J Heredity* 86: 248–249.
- Rousset F, Raymond M (1995). Testing heterozygote excess and deficiency. *Genetics* 140: 1413–1419.
- Sancristobal M, Chevalet C (1997). Error tolerant parent identification from a finite set of individuals. *Genet Res* 70: 53–62.
- Sobel E, Papp JC, Lange K (2002). Detection and integration of genotyping errors in statistical genetics. *Am J Hum Genet* 70: 496–508.
- Summers K, Amos W (1997). Behavioral, ecological, and molecular genetic analyses of reproductive strategies in the Amazonian dart-poison frog, *Dendrobates ventrimaculatus*. *Behav Ecol* 8: 260–267.
- Thompson EA (1975). Estimation of pairwise relationships. *Ann Hum Genet* 39: 173–188.
- Thompson EA (1991). Estimation of relationships from genetic data. In: Rao CR, Chakraborty R (eds) *Handbook of Statistics*. Elsevier Science: Amsterdam, Vol 8, pp 255–269.
- Van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P (2004). MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. *Mol Ecol* 4: 535–538.
- Wagner AP (2006). Behavioral ecology of the striped hyaena (*Hyaena hyaena*). PhD Dissertation. Montana State University, Bozeman, MT.

## Appendix

We have shown how having a null allele at a locus affects the probabilities of observing each possible pair of genotypes. However, it is possible to have multiple, distinct (non-IBD) null alleles at a single locus. If

multiple null alleles are present at a locus, this expands the number of true genotypes that may have produced an observed genotype and it reasonable to question how this might affect the probabilities of observing each genotype. If there is no effect, the sum of the probabilities of all possible true genotypes when two distinct null alleles are considered should equal the sum of the probabilities in Table 3 (for the same observed genotypes).

For example, if there are two null alleles at a locus ( $n_1$  and  $n_2$ ) and  $ii/jk$  are the genotypes observed, true genotypes may be  $ii/jk$  or  $in_1/jk$  or  $in_2/jk$ . The sum of the probabilities under  $m=0$  for  $ii/jk$  with two distinct null alleles is:

$$\begin{aligned} P(G_1 = ij/G_2 = jk|m = 0) &= P(\text{Observe } ij/ik|m = 0) \\ &= P(ij/ik|m = 0) + P(in_1/jk|m = 0) + P(in_2/jk|m = 0) \\ &= 2p_i^2 p_j p_k + 4p_i p_j p_k p_{n_1} + 4p_i p_j p_k p_{n_2} \\ &= 2p_i^2 p_j p_k + 4p_i p_j p_k (p_{n_1} + p_{n_2}) \end{aligned} \quad (\text{A-1})$$

Recall that we defined  $p_n$  as the total frequency of all null alleles at a locus, so  $p_n = p_{n_1} + p_{n_2}$ , in this example. The sum of the final quantities in the above equation is then identical to the sum of the probabilities of the true genotypes under  $m=0$  for  $ii/jk$  given in Table 3.

A more complex situation occurs when two homozygotes for the same allele are observed ( $ii/ii$ ). For this set of observed genotypes, there are now nine possible true genotypes that can produce that observation (see Table A1) and we break up the possible true underlying genotypes by the expanded set that corresponds with the true genotypes in Table 3. For example,  $ii/in$  expands to  $ii/in_1$  and  $ii/in_2$  and, for  $m=0$ , the probabilities of

those true genotypes are  $2p_i^3 p_{n_1}$  and  $2p_i^3 p_{n_2}$ , respectively. Summing these quantities gives:

$$\begin{aligned} P(ii/in_1|m = 0) + P(ii/in_2|m = 0) \\ = 2p_i^3 p_{n_1} + 2p_i^3 p_{n_2} = 2p_i^3 (p_{n_1} + p_{n_2}) = 2p_i^3 p_n \end{aligned} \quad (\text{A-2a})$$

Similarly,  $in/in$  now expands to  $in_1/in_2$ ,  $in_2/in_1$ ,  $in_1/in_1$ , and  $in_2/in_2$ . For  $m=0$ , these probabilities are  $4p_i^2 p_{n_1} p_{n_2}$ ,  $4p_i^2 p_{n_1} p_{n_2}$ ,  $4p_i^2 p_{n_1}^2$ , and  $4p_i^2 p_{n_2}^2$ , respectively. Summing these quantities gives:

$$\begin{aligned} P(in_1/in_2|m = 0) + P(in_2/in_1|m = 0) \\ + P(in_1/in_1|m = 0) + P(in_2/in_2|m = 0) \\ = 4p_i^2 p_{n_1} p_{n_2} + 4p_i^2 p_{n_1} p_{n_2} + 4p_i^2 p_{n_1}^2 + 4p_i^2 p_{n_2}^2 \\ = 4p_i^2 (2p_{n_1} p_{n_2} + p_{n_1}^2 + p_{n_2}^2) \\ = 4p_i^2 p_n^2 \end{aligned} \quad (\text{A-2b})$$

For  $in/in$ , when  $m=1$ :

$$\begin{aligned} P(in_1/in_2|m = 1) + P(in_2/in_1|m = 1) \\ + P(in_1/in_1|m = 1) + P(in_2/in_2|m = 1) \\ = p_i p_{n_1} p_{n_2} + p_i p_{n_1} p_{n_2} + p_i p_{n_1} (p_i + p_{n_1}) + p_i p_{n_2} (p_i + p_{n_2}) \\ = 2p_i p_{n_1} p_{n_2} + (p_i^2 p_{n_1} + p_i p_{n_1}^2) + (p_i^2 p_{n_2} + p_i p_{n_2}^2) \\ = p_i (2p_{n_1} p_{n_2} + p_i p_{n_1} + p_{n_1}^2 + p_i p_{n_2} + p_{n_2}^2) \\ = p_i (p_n^2 + p_i (p_{n_1} + p_{n_2})) \\ = p_i (p_n^2 + p_i p_n) = p_i p_n (p_i + p_n) \end{aligned} \quad (\text{A-2c})$$

The probabilities given in each cell of Table 3 can similarly be shown not to be affected by the number of distinct null alleles actually present at a locus.

**Table A1** A list of all possible true genotypes for a pair of individuals, where  $ii/ii$  was the observed genotype, given the possible number of null alleles ( $n$ ) when null alleles are considered IBD (column 1) or distinct (column 2), and the probability of each underlying genotype pair given the number of alleles identical by descent ( $m$ ) when two distinct null alleles are present.  $p_x$  represents the frequency of allele  $x$  in population corrected for the presence of null alleles (that is,  $p_n$  or  $p_{n_1} + p_{n_2}$  is considered when summing allele frequencies to 1)

True genotypes from Table 3	True genotypes with two null alleles	n	Probability of true genotypes, given $m$ genes IBD, when two distinct null alleles are at a locus		
			m=0	m=1	m=2
$ii/ii$	$ii/ii$	0	$p_i^4$	$p_i^3$	$p_i^2$
$ii/in$	$ii/in_1$	1	$2p_i^3 p_{n_1}$	$p_i^2 p_{n_1}$	0
	$ii/in_2$	1	$2p_i^3 p_{n_2}$	$p_i^2 p_{n_2}$	0
$in/ii$	$in_1/ii$	1	$2p_i^3 p_{n_1}$	$p_i^2 p_{n_1}$	0
	$in_2/ii$	1	$2p_i^3 p_{n_2}$	$p_i^2 p_{n_2}$	0
$in/in$	$in_1/in_2$	2	$4p_i^2 p_{n_1} p_{n_2}$	$p_i p_{n_1} p_{n_2}$	0
	$in_2/in_1$	2	$4p_i^2 p_{n_1} p_{n_2}$	$p_i p_{n_1} p_{n_2}$	0
	$in_1/in_1$	2	$4p_i^2 p_{n_1}^2$	$p_i p_{n_1} (p_i + p_{n_1})$	$2p_i p_{n_1}$
	$in_2/in_2$	2	$4p_i^2 p_{n_2}^2$	$p_i p_{n_2} (p_i + p_{n_2})$	$2p_i p_{n_2}$

The probability of the observed genotype pair is obtained by summing the probabilities of the possible underlying true genotypes. Note that the sum of the quantities in this table is the same as the sum of the true genotype probabilities for  $ii/ii$  in Table 3.