doi: 10.1111/j.1365-294X.2007.03089.x

Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment

STEVEN T. KALINOWSKI,* MARK L. TAPER* and TRISTAN C. MARSHALL†

**Department of Ecology, Montana State University, Bozeman, MT 59717, USA, †Field Genetics Limited, 23a Oaklands Grove, London W12 0JD, UK*

Abstract

Genotypes are frequently used to identify parentage. Such analysis is notoriously vulnerable to genotyping error, and there is ongoing debate regarding how to solve this problem. Many scientists have used the computer program CERVUS to estimate parentage, and have taken advantage of its option to allow for genotyping error. In this study, we show that the likelihood equations used by versions 1.0 and 2.0 of CERVUS to accommodate genotyping error miscalculate the probability of observing an erroneous genotype. Computer simulation and reanalysis of paternity in Rum red deer show that correcting this error increases success in paternity assignment, and that there is a clear benefit to accommodating genotyping errors when errors are present. A new version of CERVUS (3.0) implementing the corrected likelihood equations is available at www.fieldgenetics.com.

Keywords: CERVUS, estimation, genotype error, likelihood, parentage, paternity

Received 24 April 2006; revision accepted 23 June 2006

Introduction

Genetic data are frequently used to estimate genealogical relationships among individuals (see Blouin 2003 for review), and such analyses have provided remarkable insight to the reproductive lives of plants and animals. Parentage tests, for example, have revealed extra-pair copulation among birds (Charmantier & Blondel 2003; Castro *et al*. 2004), estimated reproductive success in primate hierarchies (Constable *et al*. 2001), and documented kin selection among ground squirrels (Cordero *et al*. 1999; Shelley & Blumstein 2005).

Relationship estimation, however, is notoriously vulnerable to genotyping error. Consider a mother and father with genotypes *ii* and *jj*. Offspring from these parents will have genotype *ij*. But, if the genotype of the offspring is mistakenly scored as *ii*, the actual father will be excluded from paternity. Such false exclusion can be caused by contamination, allelic dropout, microsatellite stutter, null alleles, or human error. The problem is particularly insidious because a single genotyping error can exclude the true father no matter how many loci are scored. In fact, if genotyping errors are not accommodated during data analysis, increasing the number of loci scored will probably increase the probability of a false exclusion.

This problem has long been recognized (see Pompanon *et al*. 2005 for a review), and several authors have suggested statistical methods for accommodating genotyping errors while estimating relationship (e.g. SanCristobal & Chevalet 1997; Marshall *et al*. 1998; Wagner *et al*. 2006). The method of Marshall *et al*. (1998) has been particularly influential, for it is implemented by the popular computer program CERVUS (Marshall *et al*. 1998). Most researchers using CERVUS 1.0 and 2.0 have taken advantage of the option for allowing for genotyping errors (Morrissey & Wilson 2005).

Recently, Morrissey & Wilson (2005) used computer simulations to show that in some circumstances (e.g. studies with small numbers of loci) paternity could be assigned more often by ignoring genotyping errors than by allowing for error — even when genotyping errors were present (Morrissey & Wilson 2005). This result is surprising for it contradicts previous simulations which showed a clear benefit to accommodating genotyping error (SanCristobal & Chevalet 1997; Wang 2004; Vandeputte *et al*. 2006).

Correspondence: Steven Kalinowski, Fax: 406 994–3190; E-mail: skalinowski@montana.edu

Morrissey & Wilson (2005) also found that the best strategy for estimating paternity was to assign a nonzero probability of genotyping error, but to deliberately assign this rate a value smaller than the actual rate.

The surprising results of Morrissey & Wilson (2005) can be explained by examining the likelihood equations used in their computer simulations. Morrissey & Wilson (2005) used the likelihood equations of Marshall *et al*. (1998) that were implemented by versions 1 and 2 of CERVUS. In this study, we show that these equations artificially inflate the rate at which erroneous genotypes are expected to be observed. We use computer simulation to show that this explains the results of Morrissey & Wilson (2005). Using corrected likelihoods, we show a consistent benefit to allowing for genotyping error while estimating paternity. In addition, we show that these corrected likelihood equations make it easier to establish paternity with a high degree of statistical confidence. Lastly, we use a new version of CERVUS (CERVUS 3.0) to re-analyse paternity in Rum red deer (Marshall *et al*. 1998).

Methods

Likelihood equations

We use the random genotype replacement model of Marshall *et al*. (1998) to model genotyping error. This model assumes that when a genotyping error is made, the probability of observing any specific erroneous genotype is equal to the frequency of that genotype in the population. This model is mathematically convenient without being implausible. For example, it is a reasonable model for pipetting, labelling, or data entry errors. Following Marshall *et al*. (1998), we assume genotyping error rates are independent and constant across individuals and loci. Let g represent an observed genotype, let ε represent the genotyping error rate in a laboratory, and let $P(g)$ represent the Hardy–Weinberg frequency of genotype g in a population. Given this model of genotyping error, the probability of observing g is equal to $(1 - \varepsilon)P(g) + \varepsilon P(g)$. The first term in this sum is the probability that a locus is genotyped without error and has genotype g. The second term in the sum is the probability that an error has occurred and genotype g is observed.

Let g_m, g_a, g_o, represent genotypes observed in a mother, alleged father, and offspring (respectively). Our goal is to formulate the likelihood of these genotypes for alternative relationships that might exist between the three individuals. Let H_1 represent the hypothesis that the alleged father is indeed the father of the offspring. Let H_2 represent the hypothesis that the alleged father is unrelated to the mother and offspring. By definition, the likelihood of each hypothesis is the probability of observing g_m, g_a, g_o, if the hypothesis was true. If the mother's genotype is known, the likelihood of H_1 is equal to

$$\begin{aligned}
L(H_1) = {} & (1-\varepsilon)^3[T(g_o \mid g_m, g_a)P(g_m)P(g_a)] \\
& + \varepsilon(1-\varepsilon)^2[T(g_o \mid g_m)P(g_m)\underline{P(g_a)} \\
& + T(g_o \mid g_a)P(g_a)\underline{P(g_m)} + P(g_m)P(g_a)\underline{P(g_o)}] \\
& + \varepsilon^2(1-\varepsilon)[P(g_m)\underline{P(g_a)P(g_o)} \\
& + \underline{P(g_m)}P(g_a)\underline{P(g_o)} + \underline{P(g_m)P(g_a)}P(g_o)] \\
& + \varepsilon^3[\underline{P(g_m)P(g_a)P(g_o)}]
\end{aligned} \qquad \text{(eqn 1)}$$

where $T(\cdot)$ are standard Mendelian transition probabilities (e.g. Marshall *et al*. 1998). The reader may note that equation 1 can be simplified (see the Appendix for simplified likelihood equations for three different patterns of relatedness). We have chosen the above form to facilitate comparison to the corresponding equations developed by Marshall *et al*. (1998) (equation 6 and its predecessor) and used in versions 1 and 2 of CERVUS. The difference between equation 1 (above) and equation 6 of Marshall *et al*. (1998) is that the latter assigned a value of 1 to all terms that we have underlined above.

Deconstructing equation 1 reveals the difference between it and the corresponding equations of Marshall *et al*. (1998). Equation 1 can be understood as follows. If the three genotypes g_m, g_a, g_o are observed, one of four possibilities must be true: (i) all of the genotypes are correct, (ii) two of the genotypes are correct and one is incorrect, (iii) one of the genotypes is correct and two are incorrect, or (iv) all three of the genotypes are incorrect. The four lines on the right side of equation 1 correspond to these cases, where underlined terms refer to erroneous genotypes. The probability of each case is calculated in two steps. For example, in the first line, the probability of all three genotypes being correct is $(1 - \varepsilon)^3$. If all genotypes are correct, the probability of observing g_m, g_a, g_o is equal to $T(g_o \mid g_m, g_a)P(g_m)P(g_a)$. The last line shows the probability of erroneously genotyping the mother, alleged father, and offspring, and observing g_m, g_a, and g_o. The probability that all three genotypes are incorrect is equal to ε^3, and if all genotypes are incorrect, the probability of observing g_m, g_a, g_o is equal to $P(g_m)P(g_a)P(g_o)$. Equation 1 is written for one locus. Likelihoods for multiple unlinked loci are calculated by multiplying across loci.

The corresponding equation for the alternative hypothesis, H_2, that three individuals are mother, offspring, unrelated male is

$$\begin{aligned}
L(H_2) = {} & (1-\varepsilon)^3[T(g_o \mid g_m)P(g_m)P(g_a)] \\
& + \varepsilon(1-\varepsilon)^2[T(g_o \mid g_m)P(g_m)\underline{P(g_a)} \\
& + P(g_o)P(g_a)\underline{P(g_m)} + P(g_m)P(g_a)\underline{P(g_o)}] \\
& + \varepsilon^2(1-\varepsilon)[P(g_m)\underline{P(g_a)P(g_o)} \\
& + \underline{P(g_m)}P(g_a)\underline{P(g_o)} + \underline{P(g_m)P(g_a)}P(g_o)] \\
& + \varepsilon^3[\underline{P(g_m)P(g_a)P(g_o)}]
\end{aligned} \qquad \text{(eqn 2)}$$

Again, we have underlined terms that Marshall *et al*. (1998) assigned a value of 1.

Essentially, Marshall *et al*. (1998) mistakenly indicated that the probability of making a genotyping error and observing g is equal to ε, instead of $\varepsilon P(g)$. This inflates the probability of observing genotyping errors in the likelihood calculations. The most important consequence of this error is that it reduces the impact of mismatched genotypes between an offspring and alleged father. As discussed above, if the genotyping error rate is assumed to be zero, one mismatched genotype between an offspring and an alleged father will exclude that male from paternity. If we allow for genotyping errors, unrelated males cannot be excluded from paternity with certainty. However, if sufficient data are collected, their genotypes can be dismissed as unlikely. If the probability of observing erroneous genotypes is high (either because the error rates is high, or because the likelihood equations unintentionally inflate it), mismatched genotypes appear more likely. This makes it harder to dismiss an unrelated male's genotype as unlikely.

Computer simulations

We used computer simulations to assess the importance of replacing the likelihood equations of Marshall *et al*. (1998) with the new ones described above. In particular, we used these simulations to ask if our reformulation produces better assignment of paternity, and if there is likely to be a cost to including genotyping error in likelihood calculations.

We designed our simulations to be comparable to those of Marshall *et al*. (1998) and Morrissey & Wilson (2005). These studies include several different scenarios; for simplicity, we chose the problem of identifying a father from a list of possible fathers when the mother's identity and genotype are known and all potential fathers in the population have been genotyped. Each iteration of the simulation began by randomly selecting actual genotypes for 100 unrelated adult males and 100 unrelated adult females. While doing this, we assumed that allele frequencies in the population were [0.25, 0.25, 0.2, 0.15, 0.05, 0.05, 0.02, 0.01, 0.01, 0.005, 0.005]. Loci were assumed to be unlinked, and each locus was assumed to have these same allele frequencies. This set of allele frequencies was chosen because they are representative of polymorphic microsatellite loci commonly used to identify parentage and because Morrissey & Wilson (2005) found that these specific allele frequencies resulted in a high cost to allowing for genotyping error. Once the genotypes of the adults were generated, we randomly selected a male and female to be the parents of a single offspring. The actual multilocus genotype of this offspring was obtained by simulating Mendelian inheritance. Genotyping errors were then simulated using the random genotype replacement model of Marshall *et al*. (1998) described above, and a genotyping error rate of 0.01.

Once the observed genotypes were constructed, we used the statistical framework of Marshall *et al*. (1998) to identify the most likely father and to assess the statistical confidence of that identification. This entails calculating the natural logarithm of the likelihood-odds ratio, LOD, for each adult male. The LOD score for an alleged male is equal to

$$\mathrm{LOD}_a = \ln\left[\frac{L(H_1)}{L(H_2)}\right] \qquad \text{(eqn 3)}$$

where likelihoods in the numerator and denominator are calculated from equations 1 and 2. A positive LOD indicates that a male is more likely to be the father than is a male randomly drawn from the population. A negative LOD indicates the male is less likely to be the father than is a male randomly drawn from the population. Once LOD scores are calculated for all males, the male with the highest score is the putative father. The statistical confidence of this estimate is measured by the difference between LOD scores of the male with the highest score and the male with the second highest score. If this difference, denoted Δ, is large, we can be confident that the male with the highest LOD score is actually the father. Like in Marshall *et al*. (1998), we used a LOD score of zero as a threshold while calculating Δ (see Marshall *et al*. 1998; p. 642 for details), and used computer simulation to estimate a critical value of Δ to use to establish statistical confidence. This critical value was chosen so that 99% of the Δ values above Δ critical ($\Delta_{0.99}$) were correct assignments. These calculations were done using both the original likelihood equations of Marshall *et al*. (1998) and also the corrected equations presented above.

In each iteration of the simulations, we recorded whether the male with the highest LOD score was actually the father and the value of Δ for this male. One hundred thousand simulations were run for each set of parameters.

An empirical example: paternity in Rum deer

We also used a new version of CERVUS that incorporates the corrected likelihoods, CERVUS 3.0, to assess the impact of the correction on assignment of paternity in red deer on the Isle of Rum, Scotland. Analysis of this data set using the original likelihoods, based on data from nine microsatellite and three allozyme loci, was previously described in Marshall *et al*. (1998). The analysis included 875 calves born in the study population between 1982 and 1996 that were typed at six or more loci. Of these offspring, 655 had mothers that were also typed at six or more loci. Offspring born in a given year were tested against stags typed at six or more loci that were observed in the study area during the preceding mating season. Simulations carried out to establish critical values of Δ used the corrected likelihoods but otherwise used identical allele frequencies and parameters as the original analysis. Note that for consistency with earlier

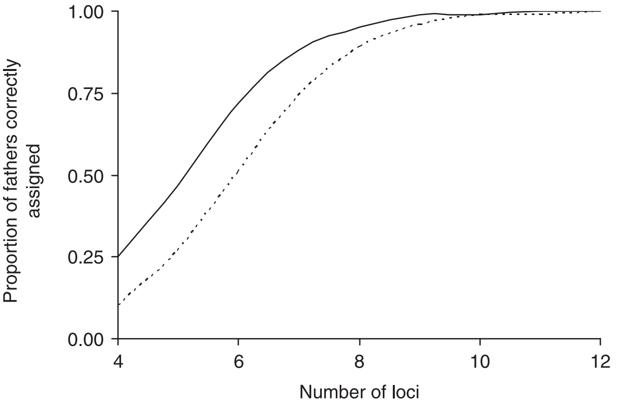

Fig. 1 The proportion of times that the correct father could be identified confidently ($\Delta > \Delta_{0.99}$) in computer simulations for the original likelihood equations of Marshall *et al*. (1998) (dashed line) and for the corrected equations presented in this study (solid line). Simulated data had allele frequencies [0.25, 0.25, 0.2, 0.15, 0.05, 0.05, 0.02, 0.01, 0.01, 0.005, 0.005]. The *x* axis shows the number of loci used for each simulation. A genotyping error rate of 0.01 was used while simulating genotypes and while analysing the data.

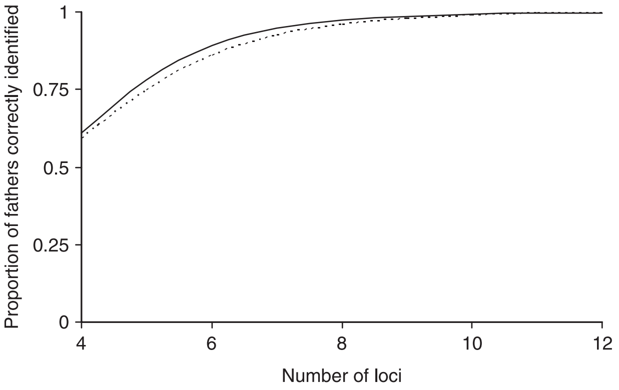

Fig. 2 The proportion of times that the male with the highest LOD score was actually the father in computer simulations for the original likelihood equations of Marshall *et al*. (1998) (dashed line) and for the corrected equations presented in this study (solid line). Simulated data had with allele frequencies [0.25, 0.25, 0.2, 0.15, 0.05, 0.05, 0.02, 0.01, 0.01, 0.005, 0.005]. The *x* axis shows the number of loci used for each simulation. A genotyping error rate of 0.01 was used while simulating genotypes and while calculating likelihoods.

studies, the confidence levels used in this analysis, 80% and 95%, are lower than the 99% used for the simulations described elsewhere in this study.

Results

The new likelihood equations presented above modestly increased the rate at which paternity could be correctly assigned with statistical confidence (i.e. $\Delta > \Delta_{0.99}$) (Fig. 1). For example, when six loci were genotyped, and both the actual and assumed genotyping error rates were 0.01, the original likelihood equations of Marshall *et al*. (1998) were able to correctly assign paternity (with a $\Delta_{0.99}$ level of confidence) to 51% of the offspring. The corresponding result for the corrected equations was 73%. The amount of improvement depended on the number of loci genotyped (Fig. 1). If sufficient loci were genotyped (e.g. 10+), both methods worked well. However, when fewer than 10 loci were genotyped, the corrected equations worked better.

The simulations also showed that the male with the highest likelihood of being the father was in fact usually the actual father — and that this was true using the original likelihoods of Marshall *et al*. (1998) as well as the corrected likelihoods (Fig. 2). In other words, the errors in the equations of Marshall *et al*. (1998) resulted in decreased confidence in a male being the father, but did not have a substantial effect on identification of the correct father.

The simulations also showed a clear benefit to including genotyping error in likelihood equations (Fig. 3). Morrissey & Wilson (2005) reported that allowing for genotyping error produced lower rates of paternity assignment than assuming there were no errors — even when data contained

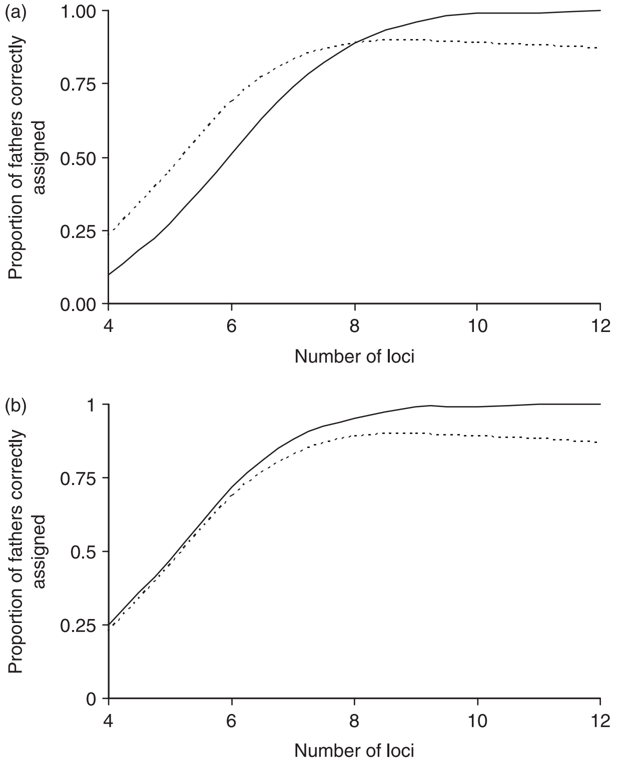

Fig. 3 The proportion of times that the correct father could be identified confidently ($\Delta > \Delta_{0.99}$) in computer simulations for analyses that assumed the error rate was 0.01 (solid line) or 0.0 (dashed line) in the likelihood calculations — when the actual genotyping error rate was 0.01. Simulated data had allele frequencies [0.25, 0.25, 0.2, 0.15, 0.05, 0.05, 0.02, 0.01, 0.01, 0.005, 0.005]. Panel 3a shows results using the original equations of Marshall *et al*. (1998). Panel 3b shows results using the corrected equations described in this study. The *x* axis shows the number of loci used for each simulation.

errors. We obtained the same result using the original likelihoods of Marshall *et al*. (1998) with between three and seven loci (Fig. 3a). However, when we used corrected equations, the apparent cost to including genotyping errors either disappeared (when between four and seven loci were used) or turned into a clear benefit (eight or more loci) (Fig. 3b). Failure to allow for error actually led the number of paternities assigned to fall if more than eight loci were used.

The corrected likelihoods also improved the success of paternity analysis in Rum red deer. More paternities were assigned using the corrected likelihoods than using the original likelihoods of Marshall *et al*. (1998) at 80% and 95% confidence both when mothers were sampled (Fig. 4a) and when mothers were unsampled (Fig. 4b). Increases were large, around 10% of offspring tested, for analyses where most paternities were previously unassigned (i.e. at the higher 95% confidence level when mothers were sampled and at both confidence levels when mothers were unsampled).

However, not all the individual paternities assigned using the original likelihoods were assigned to the same males when using the corrected likelihoods (Fig. 4a, b, columns 1 and 3). Around 17% of paternities assigned at 80% confidence using the original likelihoods were not assigned using the corrected likelihoods, both when mothers were sampled and when mothers were unsampled. For approximately two-thirds of these cases, the male assigned using the original likelihoods was still the most likely male but could no longer be assigned with the same level of confidence. For the remaining third, a different male was most likely. Ten paternities (3%) assigned at 80% confidence using the original likelihoods were actually assigned to a different male at 80% confidence using the corrected likelihoods when mothers were sampled; there were no such cases when mothers were unsampled. At the 95% confidence level, 8% of paternities assigned using the original likelihoods were not assigned using the corrected likelihoods when mothers were sampled. In almost all of these cases, the male assigned using the original likelihoods was still the most likely male but could no longer be assigned with the same level of confidence, and in no cases was another male actually assigned paternity. When mothers were unsampled, all paternities assigned at 95% confidence using the original likelihoods were also assigned using the corrected likelihoods.

The corrected likelihood equations are expected to decrease the likelihood of paternities that involve father–offspring mismatches at one or more loci. Using original and corrected likelihoods, we examined the number of paternities assigned at 80% confidence with zero, one or two or more mismatching loci among the 12 tested (Fig. 5). The number of paternities assigned with one or more mismatches decreased using the corrected likelihoods even as the total number of assigned paternities increased. Furthermore, the 25 paternities assigned using the original likelihoods despite two or more mismatches were reduced to just three when the corrected likelihoods were used.

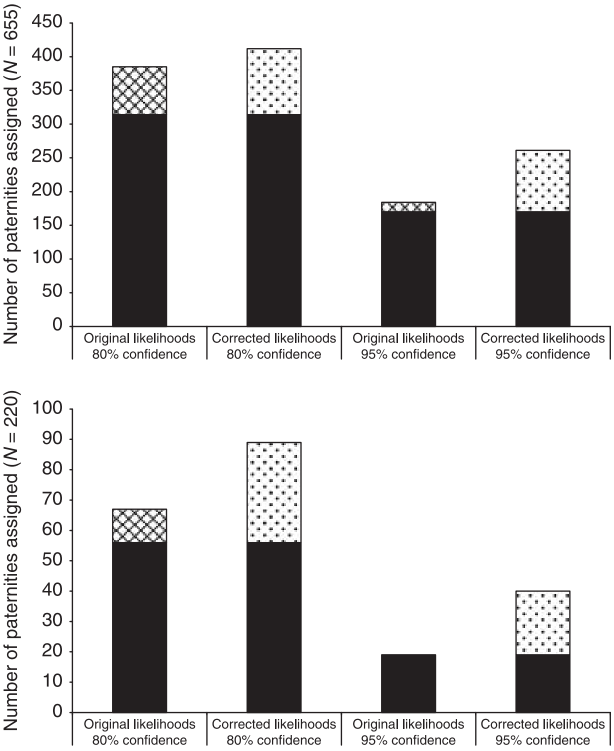

Fig. 4 The results of paternity inference for 875 Rum red deer calves born between 1982 and 1996, comparing the number of paternities assigned using the original likelihood equations of Marshall *et al*. (1998) with the number of paternities assigned using the corrected likelihood equations presented in this study. Cases where the mother was sampled (Fig. 4a) were analysed separately from those where the mother was unsampled (Fig. 4b) and in each case paternities were assigned at two nested confidence levels, 80% and 95%. Columns show paternities divided into two categories: (i) assigned by both methods (dark fill), and (ii) paternities assigned only using the original likelihoods (crosshatching) or paternities assigned only using the corrected likelihoods (light fill). Analysis was based on nine microsatellite and three allozyme loci (Marshall *et al*. 1998); to be included, individuals had to be typed at a minimum of six loci. In each case confidence was determined by simulations using parameter values shown in Marshall *et al*. (1998). Because on average only 65% of candidate males were sampled, it is unlikely that the number of assigned paternities would exceed 426 (Fig. 4a) or 143 (Fig. 4b) even with an unlimited number of loci.

Finally, we repeated this analysis using the corrected likelihoods but with a lower error rate in the likelihood calculations than in the simulated genetic data as suggested by Morrissey & Wilson (2005). Specifically, we used an error rate of 0.01 in the genotypes simulated to calculate critical values of Δ and an error rate of 0.001 to calculate likelihoods in both the simulation and the analysis of the

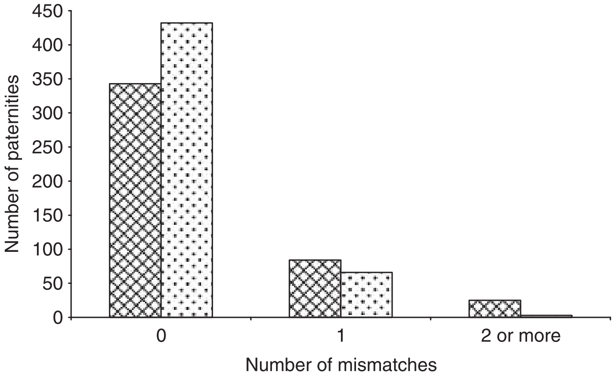

Fig. 5 The number of paternities assigned at 80% confidence with zero, one or two or more mismatches between assigned father and offspring, comparing the results using the original likelihood equations of Marshall *et al*. (1998) (crosshatching, $N = 452$) with the results using the corrected likelihood equations presented in this study (light fill, $N = 501$). The data are as for Fig. 4, combining cases where the mother was sampled with cases where the mother was unsampled.

real data. We found that when mothers were sampled, the number of paternities assigned decreased by 4% at the 80% confidence level and 3% at the 95% confidence level. When mothers were unsampled, the number of paternities assigned increased by 1% at the 80% confidence level and 3% at the 95% confidence level.

Discussion

We have described equations for likelihood of paternity that correct an error in the likelihood equations of Marshall *et al*. (1998). The corrected equations have two important consequences, demonstrated both by simulation and using the reanalysis of Rum red deer. First, they significantly increase the number of paternities that can be assigned at a given level of confidence. Second, they eliminate the cost of allowing for genotyping errors described by Morrissey & Wilson (2005).

Users of versions 1 and 2 of the parentage analysis software CERVUS, which implement the original likelihood equations of Marshall *et al*. (1998), may wonder how the error in the original likelihood equations impact their existing analyses. The main impact is likely to be that fewer paternities were assigned than could be assigned using the corrected likelihood equations. However, the great majority of paternities that were assigned would still be assigned to the same male with at least that level of confidence using the corrected likelihood equations — in the case of Rum red deer, 82% of all paternities assigned at 80% confidence using the original likelihood equations were also assigned at 80% confidence or better using the corrected likelihoods; 93% of paternities assigned at 95% confidence using the original likelihood equations were also assigned at 95% confidence using the corrected likelihoods. The number of cases where corrected likelihoods assign paternity with confidence to a different male is likely to be small — in the case of Rum red deer this happened for just 2% of all paternities assigned using the original likelihoods at 80% confidence and never at 95% confidence. Using corrected likelihoods is also likely to decrease the number of paternities assigned with one or more mismatches between father and offspring, and especially the number of paternities assigned with multiple father–offspring mismatches. In Rum red deer, 6% of paternities assigned at 80% confidence with the original likelihood equations involved multiple father–offspring mismatches. Using the corrected likelihood equations, just 0.4% of paternities involved multiple father–offspring mismatches.

In general, the main benefit of using the corrected likelihood equations, implemented in version 3 of CERVUS, will be to increase the number of paternities that can be assigned at a given level of confidence. However, in studies where paternity can already be assigned for most offspring using the original likelihood equations of Marshall *et al*. (1998), the corrected equations will allow some of these paternities to be assigned with increased confidence. In Rum red deer, 85 paternities assigned at 80% confidence using the original likelihoods (19%) were upgraded to 95% confidence using the corrected likelihoods.

We also assessed Morrissey & Wilson's (2005) suggestion that the error rate used in likelihood calculations could be set to a low (but nonzero) value. The increased success they found when analysing paternity in Rum red deer using the original likelihood equations did not occur with the revised likelihood equations. Indeed there was a slight overall decrease in the number of paternities assigned using this strategy. We believe that our revised likelihood equations eliminate the need to use an artificially low error rate in the likelihood calculations in order to optimize the success of paternity analysis.

Although we have shown a consistent benefit to allowing for genotyping error in paternity tests, this result must be tempered by the acknowledgement that the model of genotyping error employed in this study, and in Marshall *et al*. (1998), is quite simplistic. This model assumed that all loci had the same error rate, and that errors produced 'random' genotypes. We know both assumptions are not likely to be true. Genotyping error rates vary across loci (e.g. Slate *et al*. 2000; Creel *et al*. 2003; Bonin *et al*. 2004), and some types of errors are more common than others (e.g. Bonin *et al*. 2004). Simple errors, such as those caused by null alleles can easily be accounted for in paternity tests (Kalinowski & Taper 2006; Kalinowski *et al*. 2006c; Wagner *et al*. 2006), but other errors present a greater challenge. Progress is being made. Likelihood equations have been developed for increasingly realistic models of genotyping error (e.g. Gill *et al*. 2000; Sobel *et al*. 2002; Wang 2004; Kalinowski *et al*.

2006a), but these models have not yet been tested against simpler models (e.g. SanCristobal & Chevalet 1997; Marshall *et al*. 1998) to see whether they produce better estimates of paternity. Perhaps more importantly, these models have not been validated with empirical data. Other authors have avoided this potential quagmire by focusing on matches or mismatches between individuals (e.g. Slate *et al*. 2000; Vandeputte *et al*. 2006; Kalinowski *et al*. 2006b). This approach has the benefit of simplicity, but may prove less efficient. All of these topics deserve additional research.

Acknowledgements

This research has been supported by an NSF grant to M.L.T. (DEB 0415932). We thank Alastair Wilson, Jon Slate, Loeske Kruuk, Josephine Pemberton and three anonymous reviewers for comments that improved this manuscript, and Josephine Pemberton for letting us use the Rum deer data.

References

Blouin MS (2003) DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. *Trends in Ecology & Evolution*, **18**, 503–511.

Bonin A, Bellemain E, Bronken Eidesen P, Pompanon F, Brochmann C, Taberlet P (2004) How to track and assess genotyping errors in population genetics studies. *Molecular Ecology*, **13**, 3261–3273.

Castro I, Mason KM, Armstrong DP, Lambert DM (2004) Effect of extra-pair paternity on effective population size in a reintroduced population of the endangered hihi, and potential for behavioural management. *Conservation Genetics*, **5**, 381–393.

Charmantier A, Blondel J (2003) A contrast in extra-pair paternity levels on mainland and island populations of Mediterranean blue tits. *Ethology*, **109**, 351–363.

Constable JL, Ashley MV, Goodall J, Pusey AE (2001) Noninvasive paternity assignment in Gombe chimpanzees. *Molecular Ecology*, **10**, 1279–1300.

Cordero PJ, Wetton JH, Parkin DT (1999) Within-clutch patterns of egg viability and paternity in the house sparrow. *Journal of Avian Biology*, **30**, 103–107.

Creel S, Spong G, Sands JL *et al.* (2003) Population size estimation in Yellowstone wolves with error-prone noninvasive microsatellite genotypes. *Molecular Ecology*, **12**, 2003–2009.

Gill P, Whitaker J, Flaxman C, Brown N, Buckleton J (2000) An investigation of the rigor of interpretation rules for STRs derived from less than 100 pg of DNA. *Forensic Science International*, **112**, 17–40.

Kalinowski ST, Taper ML (2006) Maximum likelihood estimation of the frequency of null alleles at microsatellite loci. *Conservation Genetics*, **7**, 991–995.

Kalinowski ST, Taper ML, Creel S (2006a) Using DNA from non-invasive samples to identify individuals and census populations: an evidential approach tolerant of genotyping errors. *Conservation Genetics*, **7**, 319–329.

Kalinowski ST, Sawaya M, Taper ML (2006b) Individual identification and distributions of genotypic differences. *Journal of Wildlife Management*, **70**, 148–150.

Kalinowski ST, Wagner AP, Taper ML (2006c) ML-RELATE: software for estimating relatedness and relationship from multilocus genotypes. *Molecular Ecology Notes*, **6**, 576–579.

Marshall TC, Slate J, Kruuk LEB, Pemberton JM (1998) Statistical confidence for likelihood-based paternity inference in natural populations. *Molecular Ecology*, **7**, 639–655.

Morrissey MB, Wilson AJ (2005) The potential costs of accounting for genotypic errors in molecular parentage analysis. *Molecular Ecology*, **14**, 4111–4121.

Pompanon F, Bonin A, Bellemain E, Taberlet P (2005) Genotyping errors: causes consequences, and solutions. *Nature Reviews Genetics*, **6**, 847–859.

SanCristobal M, Chevalet C (1997) Error tolerant parent identification from a finite set of individuals. *Genetical Research*, **70**, 53–62.

Shelley EL, Blumstein DT (2005) The evolution of vocal alarm communication in rodents. *Behavioral Ecology*, **16**, 169–177.

Slate J, Marshall T, Pemberton J (2000) A retrospective assessment of the accuracy of the paternity inference program CERVUS. *Molecular Ecology*, **9**, 801–208.

Sobel E, Papp JC, Lange K (2002) Detection and integration of genotyping errors in statistical genetics. *American Journal of Human Genetics*, **70**, 496–508.

Vandeputte M, Mauger S, Dupont-Nivet M (2006) An evaluation of allowing for mismatches as a way to manage genotyping errors in parentage assignment by exclusion. *Molecular Ecology Notes*, **6**, 265–267.

Wagner AP, Creel S, Kalinowski ST (2006) Maximum likelihood estimation of relatedness and relationship using microsatellite loci with null alleles. *Heredity*, **97**, 336–345.

Wang J (2004) Sibship reconstruction from genetic data with typing errors. *Genetics*, **166**, 1963–1979.

Appendix

Here we describe corrected likelihood equations for all three 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. Likelihoods for (i) and (ii) may equally be applied to analysis of maternity. These supersede the equations given by Marshall *et al*. (1998) and Morrissey & Wilson (2005).

In each case, hypothesis H_1 that the alleged parent is the true parent is evaluated relative to an alternative hypothesis H_2 that the alleged parent is unrelated, using the observed genotypes of alleged father and offspring (g_a, g_o), and when relevant genotypes of mother or alleged mother (g_m, g_{am}). The likelihood L is calculated for each hypothesis using standard Mendelian transition probabilities T (these are given for autosomal codominant markers in Marshall *et al*. (1998)), genotype probabilities P (calculated from allele frequencies assuming Hardy–Weinberg equilibrium) and the estimated rate of genotyping error, ε. When a genotyping error occurs, the true genotype is replaced by a genotype selected at random under Hardy–Weinberg assumptions. The likelihood ratio is obtained by dividing $L(H_1)$ by $L(H_2)$. In all cases, the probabilities of the parental genotypes cancel out, so the likelihood ratios depend on the transition probabilities, the offspring's genotype frequency, and the rate of genotyping error. The likelihoods for paternity when the mother is unknown are:

$$L(H_1) = P(g_a)\{(1-\varepsilon)^2 T(g_o \mid g_a) + \varepsilon(1-\varepsilon)^2 P(g_o) + \varepsilon^2 P(g_o)\}$$

and

$$L(H_2) = P(g_a)\{(1-\varepsilon)^2 P(g_o) + \varepsilon(1-\varepsilon)^2 P(g_o) + \varepsilon^2 P(g_o)\}.$$

The likelihoods for paternity when the mother is known are

$$L(H_1) = P(g_m)P(g_a)\{(1-\varepsilon)^3 T(g_o \mid g_m, g_a) + \varepsilon(1-\varepsilon)^2[T(g_o \mid g_m) + T(g_o \mid g_a) + P(g_o)] + \varepsilon^2(1-\varepsilon)^3 P(g_o) + \varepsilon^3 P(g_o)\}$$

and

$$L(H_2) = P(g_m)P(g_a)\{(1-\varepsilon)^3 T(g_o \mid g_m) + \varepsilon(1-\varepsilon)^2[T(g_o \mid g_m) + P(g_o)] + \varepsilon^2(1-\varepsilon)^3 P(g_o) + \varepsilon^3 P(g_o)\}.$$

The likelihoods for paternity and maternity jointly are

$$L(H_1) = P(g_{am})P(g_a)\{(1-\varepsilon)^3 T(g_o \mid g_{am}, g_a) + \varepsilon(1-\varepsilon)^2[T(g_o \mid g_{am}) + T(g_o \mid g_a) + P(g_o)] + \varepsilon^2(1-\varepsilon)^3 P(g_o) + \varepsilon^3 P(g_o)\}$$

and

$$L(H_2) = P(g_{am})P(g_a)\{(1-\varepsilon)^3 P(g_o) + \varepsilon(1-\varepsilon)^2 P(g_o) + \varepsilon^2(1-\varepsilon)^3 P(g_o) + \varepsilon^3 P(g_o)\}.$$