Assessing Allelic Dropout and Genotype Reliability Using Maximum Likelihood

Craig R. Miller,*,1 Paul Joyce† and Lisette P. Waits*

*Department of Fish and Wildlife, College of Natural Resources, University of Idaho, Moscow, Idaho 83844 and
†Department of Mathematics, Division of Statistics, University of Idaho, Moscow, Idaho 83844

ABSTRACT

A growing number of population genetic studies utilize nuclear DNA microsatellite data from museum specimens and noninvasive sources. Genotyping errors are elevated in these low quantity DNA sources, potentially compromising the power and accuracy of the data. The most conservative method for addressing this problem is effective, but requires extensive replication of individual genotypes. In search of a more efficient method, we developed a maximum-likelihood approach that minimizes errors by estimating genotype reliability and strategically directing replication at loci most likely to harbor errors. The model assumes that false and contaminant alleles can be removed from the dataset and that the allelic dropout rate is even across loci. Simulations demonstrate that the proposed method marks a vast improvement in efficiency while maintaining accuracy. When allelic dropout rates are low (0–30%), the reduction in the number of PCR replicates is typically 40–50%. The model is robust to moderate violations of the even dropout rate assumption. For datasets that contain false and contaminant alleles, a replication strategy is proposed. Our current model addresses only allelic dropout, the most prevalent source of genotyping error. However, the developed likelihood framework can incorporate additional error-generating processes as they become more clearly understood.

T HE extraction and amplification of DNA from museum, noninvasive, and forensic sources has great potential for studying and managing wild populations (Kohn and Wayne 1997; Taberlet et al. 1999). The power of these approaches lies in their circuity: free-ranging animals can be "tagged" and sexed without ever being handled or observed (Morin et al. 1994; Palsbøll et al. 1997; Reed et al. 1997; Taberlet et al. 1997; Kohn et al. 1999; Woods et al. 1999; Ernest et al. 2000) and generations long dead can reveal the effective size of a population (Miller and Kapuscinski 1997) and changes in patterns and levels of variability over time (Roy et al. 1996; Mundy et al. 1997; Bouzat et al. 1998; Tessier and Bernatchez 1999; Leonard et al. 2000). Unfortunately, the DNA extracted from aged material such as bone or from noninvasive sources such as hair and feces is often at low concentrations and/or highly fragmented. Under these conditions the probability of a genotyping error is severely elevated (Taberlet et al. 1996). At nuclear microsatellite loci there are three types of errors: failure to amplify one of an individual's two alleles (allelic dropout), polymerase error rendering a "false" allele, and amplification of contaminant DNA. Of the three errors, allelic dropout appears to be the most serious problem (Gagneux et al. 1997). Contamination can be controlled by stringent laboratory protocols and the inclusion of multiple negative controls. False alleles are considerably less frequent (Gagneux et al. 1997; Goossens et al. 1998), often show an unusual spectral pattern, or produce three alleles that can be tagged as suspicious and replicated. In contrast, an allele that has dropped out leaves no trace of itself in the genotype data.

The cause of allelic dropout is believed to be stochastic sampling error (Navidi et al. 1992; Taberlet et al. 1996). Sampling occurs when the extract is pipetted into the PCR mix and again when the primers and polymerase bind the template DNA in the PCR. If template DNA is at very low concentrations, then one copy may, by chance, be amplified more than the other. [In this article the term "copies" refers to the one maternally inherited allele and the one paternally inherited allele (without regard to state) so that the term "alleles" can be used to specify the variants at a locus.] If the two copies represent different alleles (a heterozygote) then dropout yields a "false homozygote."

Taberlet et al. (1996) used a stochastic sampling simulation and replicate PCRs in the laboratory to show that at low DNA concentrations the rate of allelic dropout can be >50%. As a worst-case scenario, they reasoned that in a heterozygote where every positive PCR reflects the sampling of only one allele, the probability of sampling the same allele i times consecutively is $\frac{1}{2}^{i-1}$ (assuming that each allele is equally likely to drop out). Treating this as a null hypothesis, a locus where only one allele is observed is accepted as a genuine homozygote

¹Corresponding author: Department of Fish and Wildlife, College of Natural Resources, University of Idaho, Moscow, ID 83844.
E-mail: mill8560@uidaho.edu

only when the heterozygote hypothesis has been made untenable by the data. The probability of false acceptance can be made arbitrarily small (α) by making the integer i arbitrarily large: $i \geq 1 - (\ln \alpha / \ln 2)$. If $\alpha = 0.05$, $i \geq 6$, and if $\alpha = 0.01$, $i \geq 8$. Navidi *et al.* (1992) used a similar hypothesis-based approach to evaluate genotype reliability. Although their model is more complex in that contamination is considered, the homozygote case reduces to the same rejection rule.

If multiple loci are considered simultaneously, then an acceptance error at any heterozygous locus renders the genotype erroneous. The worst-case rationale can be extended to multiple loci by casting it as a decision rule—that is, a procedure specified before collecting any data that will yield a correct genotype with probability $\geq 1 - \alpha$. The procedure accomplishing this is one that renders a correct genotype with probability $= 1 - \alpha$ under the worst possible scenario (a dropout rate of 1 and all loci heterozygous). Under these circumstances, the probability of obtaining a correct multilocus genotype is

$$(1 - (\tfrac{1}{2})^{i-1})^L, \tag{1}$$

where L is the number of loci and i is the number of replicates (see APPENDIX for proof; the word "replicates" is used to specify the per locus number of reactions). This probability can be made arbitrarily large by making i arbitrarily large: $i \geq 1 - [\ln(1 - (1 - \alpha)^{1/L})/2]$ (see APPENDIX). If $L = 8$ then 9 replicates are required to meet the $\alpha = 0.05$ criteria and 11 replicates are required to meet the $\alpha = 0.01$ criteria. For the duration of this article, this is called the "worst-case rule" (WCR).

There are both practical and statistical shortcomings to a worst-case approach. Pragmatically, it leads to the need to perform large numbers of replicates. Acquiring accurate genetic information on a population will often involve typing hundreds or thousand of samples. As is shown, a study using eight loci and the WCR could easily require an average of 35 reactions per sample. This equates to 35,000 single reactions to accurately genotype 1000 samples. The financial costs associated with this number would prohibit many noninvasive and historical genetic studies. Furthermore, the limited amount of DNA extract may be consumed before 30 or 40 reactions can be performed. The statistical problem with a worst-case approach is that it makes essentially no use of the data in hand.

In contrast to ignoring the available data, Gagneux *et al.* (1997) estimated the dropout rate from shed hairs (rather than assuming it is 1) by replicating genotypes on 18 chimpanzees up to six times. Maintaining the heterozygous null, the p value is given by $(p) \times (p/2)^{i-1}$, where p is the dropout rate (Gagneux *et al.* 1997). The problem with this approach is that obtaining an accurate estimate of p at every locus for every individual requires many replicates and thus defeats the promise of improved efficiency. The problem might be circumvented by pooling results across loci and samples, but this requires the assumption that dropout rates are equal across loci

and samples. Gagneux *et al.* (1997) observed significantly different dropout rates between individuals (ranging from 0 to 75%) but not between loci. As with the WCR, this approach makes the assumption that every locus is heterozygous regardless of how likely this is to be true.

What are the implications of committing genotyping errors? Clearly, a high genotyping error rate could bias most current applications of microsatellites, including genetic mark-recapture studies, forensic identification of individuals, parentage analysis, population assignment, and estimates of population substructure. D. Roon, L. Waits and K. Kendall (unpublished data) and Waits and Leberg (2000) have shown that genotyping errors in noninvasive mark-recapture studies can result in severe overestimates of population size. Compounding this problem is the fact that resolution between individuals is improved by adding loci (Mills *et al.* 2000). Because the probability of obtaining a correct multilocus genotype declines as the product across individual loci, adding loci while using a single-locus statistical procedure (or worse yet, ignoring statistical issues of reliability altogether) results in an increasing per sample error rate. The potentially severe implications of committing errors, along with the costly and inefficient methods currently employed for avoiding them, provide strong motivation for developing an approach that uses the available data more efficiently. In this article we develop the theoretical basis of a more efficient method for acquiring reliable genotype data, using the method of maximum likelihood. We then evaluate its performance and compare it with the worst-case approach under a range of simplified scenarios.

METHODS

General approach: Before developing the proposed approach, it is helpful to overview the rationale behind it. Suppose that an individual is genotyped at each of a number of diploid loci i times. Assuming that contamination and false alleles do not occur or can be removed from the data (see CONCLUSIONS AND OUTSTANDING ISSUES), the observation of two different alleles at a locus implies that the individual is a heterozygote. If only one allele is observed, however, then the individual may either be a true homozygote or it may be a heterozygote at which i dropouts of the same allele have occurred. The probability of the latter event can be estimated as it is a function of the probability that the two copies differ (*i.e.*, the heterozygosity) and the probability of a dropout. If the allele frequencies are known and Hardy-Weinberg equilibrium is assumed, then the heterozygosity (conditional on the allele observed) is readily obtained. The dropout probability for the sample at hand can be estimated by finding that dropout rate that makes the observed data most likely. The dropout rate must be estimated for each sample because samples differ in age, environmental exposure, etc., and therefore in quality and quantity of DNA. The (un)reliability of the observed

multilocus genotype can then be estimated by weighting the probability that sequential dropout errors have occurred by the probability that the locus is heterozygous multiplied across the observed homozygous loci. Samples that are not reliable must be replicated until they are.

The model: Consider a population where all loci under study are independent, at Hardy-Weinberg equilibrium, and have known allele frequencies. The model involves two sampling events. First, during reproduction alleles are sampled from the gamete pool and fixed into individuals. Second, during the PCR alleles are sampled from individuals across loci. Let i_j denote the number of times sampling occurs at locus j. Each time one copy drops out with probability p_j. If this error occurs, we assume that each copy is equally likely to be the dropout. Otherwise, both copies are observed. We do not consider the event that both copies drop out because PCR failure may be due to more than stochastic sampling error such as PCR reagents, thermocycler problems, etc. Hence, p_j is actually the conditional probability of detecting one allele given that at least one allele amplifies. (It can be shown that, for realistic error rates, neglecting double dropouts has a negligible effect.) If the two copies are labeled a and b (where a could be the same or different from b), the results can be summarized as the number of times in which a, b, and ab are sampled: $r_{a,j}$, $r_{b,j}$, and $r_{ab,j}$, respectively. Let the vector of these counts be denoted by $\tilde{n}_j$. When the individual's true genotype, g, is known, the likelihood of the data given the genotype and given the dropout rates is trinomial multiplied across the T heterozygous loci.

$$L(\tilde{n}_1, \tilde{n}_2, \ldots \tilde{n}_T | g, p_1, p_2, \ldots p_T) = \prod_{j=1}^{T} \left(\frac{i_j!}{(r_{a,j})!(r_{b,j})!(r_{ab,j})!} \right)$$
$$\times (p_j/2)^{(r_{a,j}+r_{b,j})}(1-p_j)^{r_{ab,j}}. \quad (2)$$

Truly homozygous loci can be ignored because the probability of the data at them is 1.

Of course, the true genotype is not known. This is addressed by writing the likelihood as the sum of genotype-specific probabilities weighted by the unconditional probability of the genotype (*i.e.*, its expected frequency) over all possible genotypes. In addition, the model is greatly simplified by assuming that across all samples the dropout rates at different loci are related to one another by a collection of constants such that $p_j = c_j p$. Then

$$L(\tilde{n}_1, \ldots \tilde{n}_T | p) = \sum_G \left[\left(\prod_{j=1}^{T} \frac{i_j!}{(r_{a,j})!(r_{b,j})!(r_{ab,j})!} \right) (c_j p/2)^{(r_{a,j}+r_{b,j})} \right.$$
$$\left. \times (1 - c_j p)^{r_{ab,j}} \right) P(G = g) \right]. \quad (3)$$

If the dropout rates are equal across loci then $c_j = 1$ for all j and the likelihood reduces to

$$L(\tilde{n}_1, \ldots \tilde{n}_T | p) = \sum_G \left[\left(\prod_{j=1}^{T} \frac{i_j!}{(r_{a,j})!(r_{b,j})!(r_{ab,j})!} \right) (p/2)^{(r_{a,j}+r_{b,j})} \right.$$
$$\left. \times (1 - p)^{r_{ab,j}} \right) P(G = g) \right]. \quad (4)$$

Although all subsequent theory and analyses in this article are based on the assumption that the dropout rate is even across loci, this assumption can be relaxed for any of the derivations that follow by substituting $c_j p$ in for p.

It is the reliability of the genotype, not the dropout rate, that is of interest to the investigator. Let E_j be the event that the observed genotype is correct at locus j and let E_g be the event that it is correct across all loci. Let f_j denote the frequency of the observed allele at locus j. Note that it is at the M loci observed as homozygous where errors may be hidden. The reliability of a genotype, $P(E_g)$, is given by

$$\text{Reliability} = P(E_g | \tilde{n}_1, \tilde{n}_2, \ldots \tilde{n}_M, p) = \prod_{j=1}^{M} P(E_j | \tilde{n}_j, p)$$

$$= \prod_{j=1}^{M} \left[1 - \frac{2 f_j (1 - f_j)(p/2)^{i_j}}{2 f_j (1 - f_j)(p/2)^{i_j} + f_j^2} \right]$$

$$= \prod_{j=1}^{M} \left[\frac{f_j}{f_j + 2(1 - f_j)(p/2)^{i_j}} \right]. \quad (5)$$

For purposes of study design, the unconditional probability that a genotype will be correctly identified is useful. Let Z_l be the event that the lth locus is heterozygous, Z_l^C the event that the lth locus is homozygous, H_l the heterozygosity at locus l, and L the number of loci typed. Then

$$P(E_g | p) = \prod_{l=1}^{L} [P(Z_l^C) + P(E_j | Z_l, p) P(Z_l)]$$

$$= \prod_{l=1}^{L} [1 - 2(p/2)^{i_l} H_l]. \quad (6)$$

By comparing Equations 5 and 6 we see how (5) is dependent on the observed data while (6) is not. The dependency on the data occurs in two ways. Since it cannot be determined before viewing the data which loci will be observed as homozygous, the observed allele with frequency f_j at homozygous locus j is itself data dependent. Also, note that the total number of observed homozygous loci, M, is a random variable that is determined only after viewing the data.

Estimating the dropout rate, genotype reliability, and number of additional replicates: The value of p that maximizes Equation 4 is the maximum-likelihood estimate (MLE), $\hat{p}$. If $\hat{p}$ is substituted for p into Equation 5 and $\hat{p} \leq p$, then accepting only those genotypes that exceed a reliability criteria $(1 - \kappa)$ of say, 95%, would limit the long run frequency of accepting false genotypes to $\leq 5\%$. But when $p > \hat{p}$, substituting $\hat{p}$ into Equation 5 overstates the reliability, a nonconservative error. This error can be guarded against by using an upper confidence bound on p in Equation 5, $\hat{p}(up)$. Formally, we define the estimated reliability,

$$\text{Estimated reliability} = \prod_{j=1}^{M} \left(\frac{f_j}{f_j + 2(1 - f_j)(\hat{p}(up)/2)^{i_j}} \right), \quad (7)$$

where $\hat{p}(up)$ is chosen so that $P(\text{Reliability} \geq \text{Estimated}$

reliability) $= \kappa$. A method for finding $\hat{p}(up)$ is described in the simulation section below. This general approach of estimating the reliability on the basis of a MLE of the dropout rate is referred to as the MLR method.

Suppose that between one and three PCR replicates are conducted initially at each locus and the reliability is estimated. If the estimated reliability is $< 1 - \kappa$, then further replication is necessary, but how much and at which loci? The answer lies in Equation 7. Because the estimated reliability is a product across individual loci, it follows that the largest per reaction increase in estimated reliability will occur by adding a replicate to the most unreliable locus. In theory, the most efficient procedure is to add one replicate to the most unreliable locus, reestimate the reliability, and continue in this manner until the estimated reliability $\geq 1 - \kappa$. Adding replicates one at a time and reevaluating the data between each is called the "single addition method" (SAM). When the (re)evaluation entails the MLR method as just described, the abbreviation MLR$_{SAM}$ is used.

In the laboratory, however, SAM will usually be impractical as it entails a PCR, gel, and analysis between every additional reaction. A more practical approach is to add additional replicates in a block. The following algorithm provides an efficient way to choose the size of the block: mathematically add one to the number of replicates at the most unreliable locus, assume that the same homozygote is observed, and reestimate the reliability using the original $\hat{p}(up)$. Repeat this process until the estimated reliability $\geq 1 - \kappa$. Perform this pattern of replication in the laboratory. Unless $\hat{p}(up)$ increases substantially with the new data, this method will yield a genotype estimated as reliable. If $\hat{p}(up)$ does jump and the genotype is still estimated as unreliable, perform another block of replicates. Adding replicates *en masse* is called the "blockwise addition method" (BAM). When the block size and acceptability criteria are based on the MLR method, the abbreviation MLR$_{BAM}$ is used. (The MLR prefix is necessary here because SAM and BAM are also used in conjunction with the WCR evaluation criteria.)

Simulations: Simulations are used here for two basic purposes: (i) to find the upper confidence bound on p, $\hat{p}(up)$, to be used in Equation 7 and (ii) to evaluate and compare the performances of the various approaches. Specific methods to these ends are detailed in the subsections below, but the same basic simulation algorithm is used throughout. First, model parameters are fixed. These include the number of loci, number of initial replicates, allele frequencies, and the dropout rate. Unless otherwise noted, the dropout rate is equal across loci. In each run a multilocus genotype is created and then sampled with errors as described in *The model* above, to yield the data. A computer algorithm is used to search the potential dropout rates between zero and one at 1% increments until the MLE from Equation 4 is found. Simulations are carried out using Visual Basic 6.0.

Performance of the MLE: The performance of Equation 4 in estimating the dropout rate is evaluated by conducting 1000 runs for a given set of parameters and then calculating the (estimated) bias and standard error. Simulations are run to assess the effects of the number of loci (three to six), number of replicates (one to three), heterozygosity (18–67%), and the parametric dropout rate (0–1) on the performance.

Upper confidence bound on the dropout rate: There are two steps in determining $\hat{p}(up)$. First we need to know how large p could be and still yield estimates as small or smaller than $\hat{p}$ only a specified proportion of the time (α). Thus the upper $1 - \alpha$ confidence bound on p [denoted $\hat{p}(up_{1-\alpha})$] is defined as the value of p that would yield MLEs $\leq \hat{p}$, $\alpha(100)\%$ of the time. Because the data are discrete, there may be a range of p's that meet this criteria; $\hat{p}(up_{1-\alpha})$ is the largest of these. Finding $\hat{p}(up_{1-\alpha})$ requires knowing the sampling distribution of $\hat{p}$ for any given p, number of replicates, number of loci, and allele frequencies. When repeated many times, the algorithm described in the first paragraph of this subsection on simulations generates just this. Candidate values of p are examined individually until the largest one having α of its probability $\leq \hat{p}$ is found. This is achieved by a computer algorithm that searches the candidate value of p at 1% increments.

This establishes how $\hat{p}(up_{1-\alpha})$ can be determined for any value of α; the second task is to determine what α renders a rate of false inclusions $\approx \kappa$. If the reliability was a deterministic function of p, then a $1 - \alpha$ upper bound on p would render a $1 - \alpha$ lower bound on the reliability. However, the reliability is a stochastic function depending on both p and the data. As is shown, a $1 - \kappa$ of 95% corresponds to a $1 - \alpha$ of only 70%. There does not appear to be an analytic method for finding α for a given κ. Instead, simulations are employed to test how different upper bounds limit the rate of false inclusions across a range of reasonable conditions. In these simulations, the upper bound $(1 - \alpha)$, number of loci, number of initial replicates, allele frequencies, and the reliability criteria $(1 - \kappa)$ are fixed. In each run, data are generated by performing the initial replicates, determining $\hat{p}(up_{1-\alpha})$, and estimating the reliability according to Equation 7. If the estimated reliability is $< 1 - \kappa$, then SAM is used to add replicates until the estimated reliability is $\geq 1 - \kappa$. Upon acceptance, the true and observed genotypes are compared to yield a binary result of either correct or incorrect. This process is repeated 1000 times and the observed incidence of false inclusions is calculated. The appropriate upper bound is defined as the smallest value of $1 - \alpha$ that limits the observed incidence of false inclusions to $\leq \kappa$ for all values of p. Upper bounds are searched at 5% increments.

Efficiency of the different approaches: Four different methods—MLR$_{SAM}$, MLR$_{BAM}$, WCR$_{SAM}$, and WCR$_{BAM}$—are evaluated and compared by running simulations under common parametric conditions (*i.e.*, fixed dropout rate, number of loci, allele frequencies, and number of initial

replicates). In all simulations, the data are evaluated after the initial reactions have been performed. In the MLR simulations, a sample estimated to be unreliable is replicated at observed homozygous loci either in a single (SAM) or a blockwise (BAM) fashion as described above until it is estimated as reliable. By comparing it to the true genotype, the accepted genotype is scored as either correct or incorrect and the number of reactions invested in it is recorded. This is repeated 1000 times to yield an observed incidence of false inclusions and a mean number of reactions per sample.

The WCR simulations are analogous except in the rules governing acceptance and additional replication. Under the WCR$_{SAM}$, all observed homozygous loci are replicated once and reevaluated, replicated once and reevaluated, and so on. Loci that do not turn up heterozygous by this process are replicated until Equation 1 is satisfied. With the WCR$_{BAM}$, one block of reactions is added to all observed homozygous loci so that Equation 1 is satisfied. Because the BAM is more practical in the lab, further simulations focus on the MLR$_{BAM}$ and WCR$_{BAM}$ approaches. Specifically, the effects of heterozygosity (50–80%), number of loci (four to eight), number of initial replicates (one to three), and reliability criteria (95–99.9%) on the number of reactions per sample are explored.

Interlocus dropout heterogeneity: Simulations are conducted to investigate how well the MLR$_{BAM}$ approach performs when it is assumed that the dropout rates are even across loci, but they are not. This is accomplished by running successive simulations where the dropout rates across loci are made increasingly uneven but other parameters are held constant. Reliability is estimated under the assumption of dropout rate homogeneity (*i.e.*, using Equations 4 and 7), and replicates are added using BAM. Each simulation consists of 1000 runs from which the incidence of false inclusions is calculated.

RESULTS AND DISCUSSION

Performance of the MLE: When the number of loci is less than four, genotypes are unreplicated, or the heterozygosity is <50%, there is little information in the data regarding the dropout rate. In this case, the MLE from Equation 4 tends to be biased high for small values of p (data not shown). While no bias is desirable, overestimating the dropout rate is a conservative error. Above these values (or for $p > 0.5$ below these values), the estimator becomes approximately unbiased. As expected in a binomial model, the largest standard errors are observed when p is near 50%.

Upper confidence bound on the dropout rate: To make a conservative estimate of the reliability, we need a sufficiently conservative estimate of the dropout rate. Preliminary simulations showed that, irrespective of what upper bound is used, the incidence of false inclusions is highest when p is between 0.5 and 0.8 (data not shown).

TABLE 1

Upper bound as a function of κ, number of loci, and H

No. loci	κ = 5%		κ = 1%	
	$H = 50\%$	$H = 67\%$	$H = 50\%$	$H = 67\%$
4	65	65	70	75
6	70	70	75	75
8	75	70	75	75

The upper bound was defined as the smallest value of $1 - \alpha$ that limited the observed incidence of false inclusions to $\leq \kappa$ in 1000 runs at $p = 0.5, 0.6, 0.7,$ and 0.8. In all cases two initial replicates were performed.

We therefore concentrated on finding the appropriate upper bound for p in this range as it will be sufficiently large for other values of p. When a 95% reliability criteria is required for acceptance (*i.e.*, $1 - \kappa = 0.95$) and two initial replicates are performed, the approximate upper bound on p for four, six, or eight loci with $H = 50$ or 67% is between 65 and 75% (Table 1). Increasing the reliability criteria to 99% elevates the upper bounds slightly to between 70 and 75%. In the case of three initial replicates, the appropriate upper bounds are in the 60–70% range, while with initially unreplicated data upper bounds are between 75 and 85% (data not shown). These results are used to set the upper bound in subsequent simulations.

Efficiency of the different approaches: This study is motivated largely by the apparent inefficiency of the worst-case approach. The central question is, therefore, how efficient is the proposed MLR method by comparison? Consider a case where two initial replicates are performed at six loci, all with 67% heterozygosity. The reliability criteria are set at $1 - \kappa = 95\%$ for MLR simulations and the probability of a correct genotype is likewise set at $1 - \alpha = 95\%$ for the WCR simulations. In comparing the mean number of total reactions required to achieve acceptable genotypes under each of the four methods (Figure 1A), several important trends emerge. First, the MLR methods are virtually always more efficient than the WCR method of the same replication strategy. Second, differences between the MLR and the WCR methods are largest when $p = 0$ and they disappear as p approaches 1. For $p \leq 0.2$, the MLR methods require 10–12 reactions fewer than the WCR methods (a 40–50% reduction). We do note, however, that the total number of reactions in this and all other simulations in this article does not include any failed PCR reactions. While PCR failures should be rare for samples with low values of p, failures will increase the total reaction counts and reduce the proportional difference between approaches. Third, the SAM and BAM approaches are nearly equivalent for small values of p, but SAM is increasingly superior as p approaches 1.

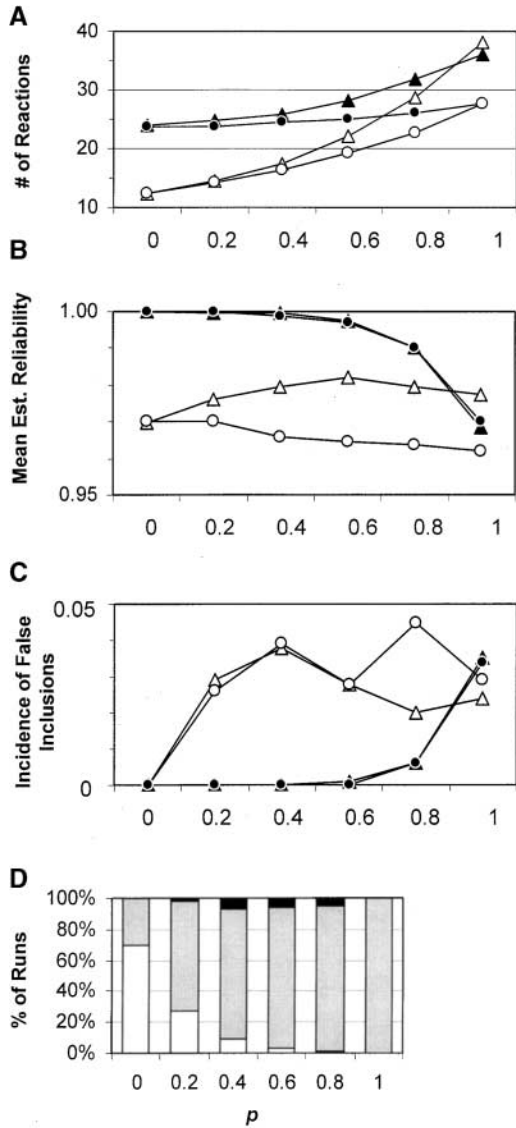This third trend is important because BAM is far more practical than SAM in the laboratory and because drop-

FIGURE 1.—Comparative performance of the WCR$_{SAM}$ (●), WCR$_{BAM}$ (▲), MLR$_{SAM}$ (○), MLR$_{BAM}$ (△) across dropout rate, p. (A) Mean total number of reactions to achieve acceptable genotypes. (B) Mean estimated reliability of accepted genotypes. For WCR$_{SAM}$ and WCR$_{BAM}$, reliability estimated after genotype acceptance using Equation 7 and a 70% upper bound on p. (C) Incidence of false inclusion over the 1000 runs. (D) The number of rounds of reactions needed under MLR$_{BAM}$ to achieve genotype acceptance. (■) Three rounds, (▨) two rounds, (□) one round. Based on 1000 runs/simulation, six loci, two initial replicates/locus, $H = 67\%$ [allele frequencies: {0.33, 0.33, 0.34}], reliability criteria $1 - \kappa = 95\%$, upper bound on p ($1 - \alpha$) = 70%.

out rates should be <0.5 for most samples. For example, GAGNEUX *et al.* (1997) and GOOSSENS *et al.* (1998) observed mean dropout rates of 31 and 14%, respectively, on single hair extracts and ERNEST *et al.* (2000) observed an 8% dropout rate from felid fecal DNA. The practicality of the MLR$_{BAM}$ approach is apparent when one considers the number of rounds of replication (*i.e.*, the number of blocks including the initial block of 12 reactions)
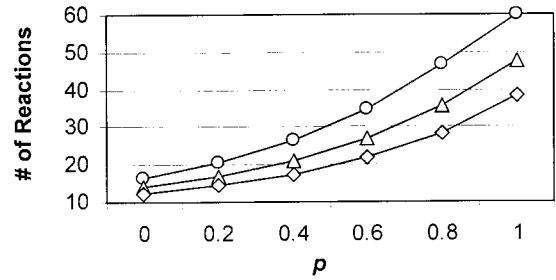

FIGURE 2.—Effect of reliability criteria $(1 - \kappa)$ on the number of reactions needed to achieve acceptable genotypes under the MLR$_{BAM}$ method. Based on 1000 runs/simulation, six loci, two initial replicates/locus, $H = 67\%$. Upper bounds on p $(1 - \alpha)$: 70% for $\kappa = 0.05$ (◇), 75% for $\kappa = 0.01$ (△), 80% for $\kappa = 0.001$ (○). In all simulations, observed incidence of false inclusions $\leq 1 - \kappa$.

that were necessary in the MLR$_{BAM}$ simulations of Figure 1A (Figure 1D). The sample is nearly always accepted within two rounds and when p is small it is not uncommon to accept it without further replication.

The WCR is generally inefficient because it is designed to guard against a worst possible scenario, $p = 1$. When p is not large, the payoff of (over)replicating is, of course, that it renders genotypes with high estimated reliabilities and very few errors (Figure 1, B and C). In contrast, the MLR methods yield moderate estimated reliabilities and an incidence of false inclusion below—but generally not far below—κ. This is true of all the MLR simulations in this article except those involving interlocus dropout heterogeneity (see below). To put these false inclusion rates in perspective, if genotypes in these simulations were unconditionally accepted after initial replication (*i.e.*, without being subject to a reliability criteria) the incidence of erroneous genotypes would be ~0, 7, 29, 54, 75, and 93% for $p = 0, 0.2, 0.4, 0.6, 0.8$, and 1, respectively.

It might be argued that the lower false inclusion rate observed with WCR approaches over most of the range of p suggests it actually outperforms the MLR approaches in one respect. Recall, however, that the investigator is willing to tolerate up to 5% errors to reduce the number of reactions. If the investigator wishes to reduce the incidence of false inclusions below 5%, the reliability criteria in the MLR methods can simply be raised. Figure 2 shows the number of replicates required to achieve estimated reliabilities of 95, 99, and 99.9% with the MLR$_{BAM}$ approach for six loci, two initial replicates, and $H = 67\%$. When p is low, only a few more reactions are required to achieve the higher estimated reliabilities, and even at $p = 0.4$, 99% estimated reliability is just four reactions >95% estimated reliability. As p gets large the cost of higher estimated reliability grows considerably.

In addition to assuming that $p = 1$, the WCR assumes that $H = 1$. As these assumptions are approached, we expect the relative performance of the WCR methods to improve. Examining the comparative effect of heterozygosity on efficiency in the WCR$_{BAM}$ and MLR$_{BAM}$ ap-
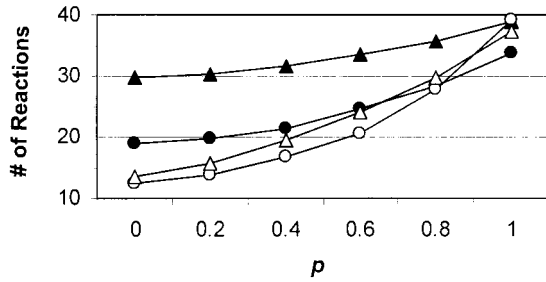
FIGURE 3.—Effect of heterozygosity on the number of reactions needed to achieve acceptable genotypes under WCR_{BAM} and MLR_{BAM} approaches. (●) $H = 80\%$, WCR_{BAM}; (○) $H = 80\%$, MLR_{BAM}; (▲) $H = 50\%$, WCR_{BAM}; (△) $H = 50\%$, MLR_{BAM}. Based on 1000 runs/simulation, six loci, two initial replicates/locus, 95% reliability criteria, and 70% upper bound on p. Allele frequencies: $H = 50\%$, {0.5, 0.5}; $H = 80\%$, {0.2, 0.2, 0.2, 0.2, 0.2}. In all simulations, observed incidence of false inclusions $\leq 1 - \kappa$.

proaches shows this to be true (Figure 3). But even at $H = 80\%$, MLR_{BAM} outperforms WCR_{BAM} across the range where most samples will realistically fall (for $p < 70\%$). When $H = 50\%$ and p is small, MLR_{BAM} renders acceptable genotypes in approximately one-half as many reactions. Interestingly, the efficiency of the MLR_{BAM} approach is only slightly improved by increasing heterozygosity.

Because an error anywhere in a genotype renders it erroneous, adding loci elevates the estimated reliability required of each locus and thereby the number of replicates per locus. The impact of this effect was investigated by running simulations at four and eight loci while holding all other parameters constant (two initial replicates, $H = 67\%$; Figure 4). Surprisingly, when p is near zero, doubling the number of loci approximately doubles the total number of reactions in the MLR_{BAM} approach. This near linear increase reflects a near constancy in per locus replication. As p increases, however, the cost of adding loci escalates in a nonlinear manner (as indicated by the divergent MLR_{BAM} lines in the figure). Figure 4 also shows that unless p is near one, adding loci increases the disparity between the MLR_{BAM} and WCR_{BAM} approaches.
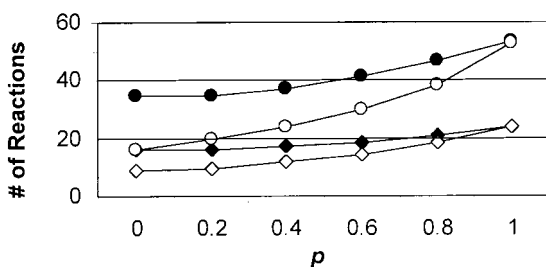

FIGURE 4.—Effect of number of loci on mean number of reactions needed to reach acceptable genotypes under WCR_{BAM} and MLR_{BAM} approaches. (●) Eight loci, WCR_{BAM}; (○) eight loci, MLR_{BAM}; (◆) four loci, WCR_{BAM}; (◇) four loci, MLR_{BAM}. Based on 1000 runs/simulations, two initial replicates/locus, $H = 67\%$, 95% reliability criteria, and 70% upper bound on p. In all simulations, observed incidence of false inclusions $\leq 1 - \kappa$.
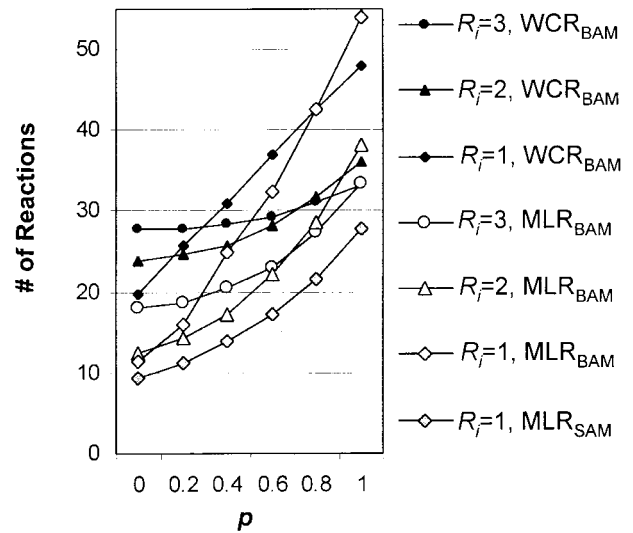

FIGURE 5.—Effect of number of initial replicates, R_i, on mean number of reactions to achieve acceptable genotypes under WCR_{BAM} and MLR_{BAM} approaches. The results of $R_i = 1$ under MLR_{SAM} (the theoretical optimum) is also shown for comparison. Based on six loci, $H = 67\%$, 95% reliability criteria. Upper bounds on $p(1 - \alpha)$: 75% for $R_i = 1$, 70% for $R_i = 2$, and 70% for $R_i = 3$. In all simulations, observed incidence of false inclusions $\leq 1 - \kappa$.

One parameter that affects efficiency and is easily manipulated by the investigator is the number of initial replicates per locus. TABERLET *et al.* (1996) proposed performing three initial replicates—though this choice was influenced by considerations of false and contaminant alleles as well as allelic dropout. We consider the cases of one, two, and three initial replicates while the other parameters are held constant (six loci, $H = 67\%$). The MLR_{BAM} approach on unreplicated data is very inefficient except when p is near 0 (Figure 5). This is especially apparent when it is compared to the theoretical optimum: MLR_{SAM} on initially unreplicated data. Note how much poorer the performance of BAM is relative to SAM with unreplicated (Figure 5) compared to two-replicate data (Figure 1A). This is because unreplicated data yield a higher upper bound and more observed homozygote loci and these drive the size of the reaction block up. In practice, the most efficient replication strategy is two initial replicates under the MLR_{BAM} method. Three initial replicates under MLR_{BAM} is more efficient than two only when $p \geq 0.8$. The two- and three-replicate results under the WCR_{BAM} approach are even less efficient except when p is near 1.

Interlocus dropout heterogeneity: All simulations to this point have assumed that dropout rates are equal across loci. Here we address a simple question: How well does MLR_{BAM} perform when the dropout rates are assumed to be even but they are not? Figure 6A shows the observed incidence of false inclusions under MLR_{BAM} for increasing degrees of dropout rate heterogeneity when the other parameters are fixed (six loci, $H = 67\%$, two initial
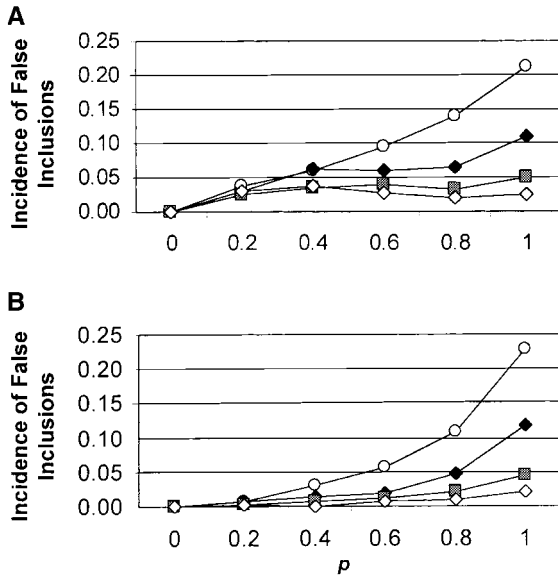
A

B

FIGURE 6.—Effect of interlocus dropout rate heterogeneity on incidence of false inclusions under MLR$_{BAM}$ when rates are assumed to be even in the analysis. Using (A) a 70% upper bound on p and (B) a 95% upper bound on p. The constants dictating the relative dropout rates between the six loci: (○) 1 1 .7 .7 .4 .4, (■) 1 1 .9 .9 .8 .8, (◆) 1 1 .8 .8 .6 .6, (◇) 1 1 1 1 1 1. For example, when $p = 0.6$, the dropout constants {1 1 .7 .7 .4 .4} render dropout rates of {.6 .6 .42 .42 .24 .24}. Based on 1000 runs/simulation, six loci, two replicates/locus, and 95% reliability criteria.

replicates). An upper bound of 70% is used as would be appropriate when the error rates genuinely are homogenous (Table 1). Even in the moderately uneven case where two of the loci are at 60% of the maximum rate, two are at 80%, and two are at the full error rate (coded "1 1 .8 .8 .6 .6" in Figure 6), the incidence of false inclusions remains near 5% so long as $p \leq 0.8$. When the unevenness is more severe with loci at 40 and 70% of the maximum error rate (1 1 .7 .7 .4 .4), the incidence of false inclusion becomes unacceptably large for $p > 0.4$.

One remedy might be to use a larger upper bound on p to estimate reliability. Figure 6B shows the incidence of false inclusions across the same set of uneven dropout rates when a $1 - \alpha = 95\%$ upper bound is used. This reduces the incidence of false inclusions at low and moderate p values, but the effect diminishes as p gets large. For the most uneven case considered, the false inclusions rate is acceptable so long as $p < 0.6$. Surprisingly, this increase in the upper bound on p from 70 to 95% elevates the number of reactions only slightly (Figure 7). These results suggest that, if the dropout rates across loci are not highly uneven and/or if the base rate is not large, analyzing data under the even dropout rate assumption still yields reliable results. Using a higher upper bound on p increases the range of violations over which the model remains robust while not increasing the number of reactions appreciably.

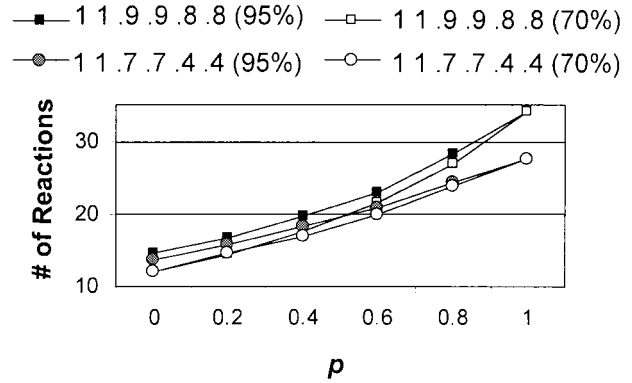**Study design:** When designing and budgeting a study,

FIGURE 7.—Effect of upper bound on number of reactions needed to achieve acceptable genotypes under MLR$_{BAM}$ for two levels of interlocus dropout rate heterogeneity. Upper bounds on p are indicated within parentheses in legend. Based on 1000 runs/simulation, six loci, two replicates/locus, and 95% reliability criteria. See Figure 6 for incidence of false inclusions associated with these simulations.

it is often valuable to have an estimate of the number of reactions that will be required. Equation 6, the unconditional probability of obtaining a correct genotype, can provide such an estimate. This requires three things of the investigator: (1) confidence that the model is appropriate; (2) knowledge of the heterozygosity per locus, or a willingness to make an educated guess; and (3) knowledge of the dropout rate, or an educated guess. To avoid underbudgeting, the investigator can use conservatively low heterozygosities and conservatively high dropout rates. It should be noted that the number of replicates need not be even across loci; the SAM algorithm can be used to forecast how replication will proceed on average after the initial replicates are performed. We also note that the number of reactions will additionally include failures.

CONCLUSIONS AND OUTSTANDING ISSUES

The most important result of this article is that under the model assumptions the MLR$_{BAM}$ represents an efficient method for obtaining reliable genotypes, especially in comparison to the WCR$_{BAM}$ approach. Although a number of variables are shown to affect this efficiency, two valuable points emerge. First, the MLR$_{BAM}$ method is especially efficient when p is small—and published data suggest it generally will be 0–40% (GERLOFF et al. 1995; GAGNEUX et al. 1997; GOOSSENS et al. 1998; ERNEST et al. 2000). This result is robust to changes in the number of loci, heterozygosity, reliability criteria, and, to a lesser extent, the number of initial replicates. But there will be samples, if only occasionally, where allelic dropout occurs far more often. The second important point is that these problem samples may require more reactions, but they do not compromise accuracy under the MLR$_{BAM}$ approach.

Tantamount to the performance of MLR$_{BAM}$ in simulations is the issue of its applicability to real data. Several of the assumptions upon which the model is based warrant closer scrutiny. One such assumption made here is that

the dropout rates are even across loci. The findings that the model is robust to mild departures from evenness and the failure of GAGNEUX *et al.* (1997) to reject the even error rate null hypothesis across 11 microsatellite loci suggest that this assumption may often be sufficiently approximated. Furthermore, the model can accommodate heterogeneity by assuming that the dropout rates across loci are related by a set of constants (see Equation 3). This would require a preliminary experiment to estimate the constants for the loci being used.

A second assumption made in the MLR model is that the two alleles in a heterozygote are equally likely to drop out. It has been suggested that the longer allele may drop out more often than the shorter (GERLOFF *et al.* 1995; GOOSSENS *et al.* 1998), but to our knowledge the empirical data have failed to detect significant departures from an even dropout rate (GERLOFF *et al.* 1995; GAGNEUX *et al.* 1997). If future data reveal how length, or some other allele attribute such as repeat motif, affects the dropout rate, the likelihood framework will readily allow the incorporation of this knowledge. Likewise, it has been assumed here that allele frequencies are known. If the sample size is large and there is no tendency for certain alleles to drop out more than others, then the observed allele frequencies should be relatively accurate. When these conditions are not met, the current model may need to be modified to address this source of uncertainty.

The most serious assumption made in our model is that there are no false or contaminant alleles in the analyzed data. We do not assume that such alleles never occur, but rather that they can be flagged and removed from the data. Although several studies have reported occurrences of false and contaminant alleles that are nonnegligible (TABERLET *et al.* 1996; GAGNEUX *et al.* 1997; GOOSSENS *et al.* 1998), three reasons suggest that these estimates may overstate the problem in the current context: (1) Some counts include the cases where three or more alleles (one or more false) appear in one reaction as well as the more insidious cases where one or two alleles are observed but one is false; (2) counts may include false alleles that have never otherwise been observed in the population or species and would therefore be viewed with great suspicion; and (3) the current practice of scoring microsatellites using electropherograms provides much greater resolution for identifying anomalous patterns compared to the former technique of scoring alleles by autoradiography.

Nevertheless, cryptic false and contaminant alleles do occur and when they are undetected, they will cause genotype errors. It may be possible to explicitly incorporate these errors into the likelihood model, for example, by assigning each allele a conditional probability of being true and a probability of being false. Certainly, there is information regarding how likely an allele is to be false *vs.* true such as its frequency and its length relative to the other allele (most false alleles are one repeat shorter or longer than a true allele). Likewise, the contamination probability can be estimated by using numerous blanks during DNA extraction and PCR. In practice, one current option is to follow TABERLET *et al.* (1996) in assuming that the probability that the same false or contaminant allele will occur twice at the same locus is remote. This leads to the requirement that every allele be observed twice before accepting a genotype.

A dilemma arises with this approach, however, when a series of reactions at a locus yield one heterozygote result and the same homozygote for all the rest (*e.g.*, *ab*, *a*, *a*, *a*). If in replicating we continue to observe the same homozygote, at what point should we begin to have serious doubts about it being a genuine heterozygote? Suppose that the locus is truly a heterozygote. The probability of observing one heterozygous result and $i - 1$ homozygotes of the same allele given the model is

$$2i(p/2)^{i-1}(1 - p). \qquad (8)$$

Replacing p with $\hat{p}$, or more conservatively with $\hat{p}(up)$, estimated from data at the other loci, a p value can be calculated. Even in the worst case, where p is near 80–85%, it is very unusual not to observe both alleles twice in a true heterozygote within five or six replicates. Once replication drives this p value below, say, 5%, the sample should be abandoned at this locus. The sample cannot be declared a homozygote because casting doubt on the heterozygote hypothesis and accepting the homozygote-false/contaminant allele hypothesis are not equivalent. The competing hypotheses cannot be resolved (as with a likelihood-ratio test) without placing a probability on the event of a false/contaminant allele (see also NAVIDI *et al.* 1992). An alternative to discarding the locus completely would be to count it as one-half a locus occupied by the allele observed in the homozygote.

Though false and contaminant alleles are generally rare, homozygotes are not. Each time a homozygote is observed, there is the possibility that it represents a true heterozygote where consecutive dropout errors have occurred. In this article we developed a general mathematical framework for dealing with this source of uncertainty. While incomplete, the model shows promise for vastly improving the efficiency in acquiring reliable genetic data—a critical step toward realizing the potential of noninvasive, historic, and forensic genetic sampling.

LITERATURE CITED

BOUZAT, J. L., H. A. LEWIN and K. N. PAIGE, 1998 The ghost of genetic diversity past: historical DNA analysis of the greater prairie chicken. Am. Nat. **152:** 1–6.

ERNEST, H. B., M. C. T. PENDO, B. P. MAY, S. M. SYVANEN and W. M. BOYCE, 2000 Molecular tracking of mountain lions in the Yosemite Valley region in California: genetic analysis using microsatellite and faecal DNA. Mol. Ecol. **9:** 433–441.

GAGNEUX, P., C. BOESCH and D. S. WOODRUFF, 1997 Microsatellite scoring errors associated with noninvasive genotyping based on nuclear DNA amplified from shed hairs. Mol. Ecol. **6:** 861–868.

Gerloff, U., C. Schlötterer, K. Rassmann, I. Rambold, G. Hoh-mann et al., 1995 Amplification of hypervariable simple sequence repeats (microsatellites) from excremental DNA of wild living bonobos (*Pan paniscus*). Mol. Ecol. **4:** 515–518.

Goossens, B., L. P. Waits and P. Taberlet, 1998 Plucked hair samples as a source of DNA: reliability of dinucleotide microsatellite genotyping. Mol. Ecol. **7:** 1237–1241.

Kohn, M. H., and R. K. Wayne, 1997 Facts from feces revisited. Trends Ecol. Evol. **12:** 223–227.

Kohn, M. H., E. C. York, D. A. Kamradt, G. Haught, R. M. Sauvajot et al., 1999 Estimating population size by genotyping faeces. Proc. R. Soc. Lond. Ser. B **266:** 1–7.

Leonard, J. A., R. K. Wayne and A. Cooper, 2000 Population genetics of ice age brown bears. Proc. Natl. Acad. Sci. USA **97:** 1651–1654.

Miller, L. M., and A. R. Kapuscinski, 1997 Historical analysis of genetic variation reveals low effective population size in northern pike (*Esox lucius*) population. Genetics **147:** 1249–1258.

Mills, L. S., J. J. Citta, K. P. Lair, M. K. Schwartz and D. A. Tallmon, 2000 Estimating animal abundance using noninvasive DNA sampling: promise and pitfalls. Ecol. Appl. **10:** 283–294.

Morin, P. A., J. Wallis, J. J. Moore and D. S. Woodruff, 1994 Paternity exclusion in a community of wild chimpanzees using hypervariable simple sequence repeats. Mol. Ecol. **3:** 469–478.

Mundy, N. I., C. S. Winchell, T. Burr and D. W. Woodruff, 1997 Microsatellite variation and microevolution in the critically endangered San Clemente Island loggerhead shrike (*lanius ludovicianus mearnsi*). Proc. R. Soc. Lond. Ser. B **264:** 869–875.

Navidi, W., N. Arnheim and M. S. Waterman, 1992 A multiple-tubes approach for accurate genotyping of very small DNA samples by using PCR: statistical considerations. Am. J. Hum. Genet. **50:** 347–359.

Palsbøll, P. J., J. Allen, M. Bérubé, P. J. Clapham, T. P. Feddersen et al., 1997 Genetic tagging of humpback whales. Nature **388:** 767–769.

Reed, J. Z., D. J. Tollit, P. M. Thompson and W. Amos, 1997 Molecular scatology: the use of molecular genetic analysis to assign species, sex and individual identity to seal faeces. Mol. Ecol. **6:** 225–234.

Roy, M. S., E. Geffen, D. Smith and R. K. Wayne, 1996 Molecular genetics of pre-1940 red wolves. Conserv. Biol. **10:** 1413–1424.

Taberlet, P., S. Friffin, B. Goossens, S. Questiau, V. Manceau et al., 1996 Reliable genotyping of samples with very low DNA quantities using PCR. Nucleic Acids Res. **24:** 3189–3194.

Taberlet, P., J. J. Camarra, S. Griffin, E. Uhrès, O. Hanotte et al., 1997 Noninvasive genetic tracking of the endangered Pyrenean brown bear population. Mol. Ecol. **6:** 869–876.

Taberlet, P., L. P. Waits and G. Luikart, 1999 Noninvasive genetic sampling: look before you leap. Trends Ecol. Evol. **14:** 323–327.

Tessier, N., and L. Bernatchez, 1999 Stability of population structure and genetic diversity across generations assessed by microsatellites among sympatric populations of landlocked Atlantic salmon (*Salmo salar L.*). Mol. Ecol. **8:** 169–179.

Waits, J. L., and P. L. Leberg, 2000 Biases associated with population estimation using molecular tagging. Anim. Conserv. **3:** 191–200.

Woods, J. G., D. Paetkau, D. Lewis, B. N. McLellan, M. Proctor et al., 1999 Genetic tagging of free-ranging black and brown bears. Wildl. Soc. Bull. **27:** 616–627.

APPENDIX: PROOF OF EQUATION 1

Assumptions:

1. All L loci are true heterozygotes.

2. Allelic dropout occurs in every reaction. Denoting the two alleles a and b, the true genotype is revealed only if a drops out during one replication and b drops out during another.

3. The probability of observing a particular allele is one-half per replicate per locus.

4. Loci are independent.

Probability experiment: Each locus is initially replicated k times. The observed homozygotes are then replicated an additional $i - k$ times for a total of i replicates.

Recording the results of the experiment: If at the end of the experiment all replicates produce the same allele, the locus is typed as a homozygote; otherwise it is typed as a heterozygote. The genotype is correctly identified if at the end of the experiment each locus is typed as a heterozygote.

Mathematics: Let C be the event that the genotype is typed correctly. Let M_k equal the number of observed homozygotes after k replicates. Let $X_j = 1$ if, after k replicates, locus j is an observed homozygote and $X_j = 0$ otherwise. It follows that $M_k = X_1 + X_2 + \ldots + X_L$. Now let s be any real number. It follows that

$$E(s^{M_k}) = E(s^{X_1 + X_2 + \cdots + X_L})$$

$$= [E(s^{X_1})]^L$$

$$= [s(\tfrac{1}{2})^{k-1} + (1 - (\tfrac{1}{2})^{k-1})]^L. \quad \text{(A1)}$$

Therefore,

$$P(C) = E[P(C|M_k)] = E([1 - (\tfrac{1}{2})^{i-k}]^{M_k}).$$

Now replace $s = 1 - (\tfrac{1}{2})^{i-k}$ in (A1) to get

$$P(C) = [(1 - (\tfrac{1}{2})^{i-k})(\tfrac{1}{2})^{k-1} + (1 - (\tfrac{1}{2})^{k-1})]^L$$

$$= (1 - (\tfrac{1}{2})^{i-1})^L,$$

which is Equation 1. Note that $P(C)$ does not depend on k. The initial number of replicates does not affect the probability of a correct genotype.

If a genotype error probability is set to α, then $P(C) = 1 - \alpha$. Therefore,

$$[1 - (\tfrac{1}{2})^{i-1}]^L = 1 - \alpha.$$

Therefore,

$$(\tfrac{1}{2})^{i-1} = 1 - (1 - \alpha)^{1/L},$$

implying

$$i = 1 - \frac{\ln(1 - (1 - \alpha)^{1/L})}{\ln 2}.$$