2004

# Genetic Errors Associated with Population Estimation Using Non-Invasive Molecular Tagging: Problems and New Solutions

Kevin S. McKelvey
*U.S. Forest Service,* kmckelvey@fs.fed.us

Michael K. Schwartz
*U.S. Forest Service*

# GENETIC ERRORS ASSOCIATED WITH POPULATION ESTIMATION USING NON-INVASIVE MOLECULAR TAGGING: PROBLEMS AND NEW SOLUTIONS

KEVIN S. McKELVEY,[1] U.S. Forest Service, Rocky Mountain Research Station, 800 East Beckwith, Missoula, MT 59801, USA
MICHAEL K. SCHWARTZ, U.S. Forest Service, Rocky Mountain Research Station, 800 East Beckwith, Missoula, MT 59801, USA

*Abstract:* The use of non-invasive DNA-based sampling is becoming increasingly popular. However, the misidentification of individuals due to genotyping error (primarily allelic dropout) is a critical problem, especially when using individuals in the capture–mark–recapture (CMR) approach to estimate population size. We propose 2 simple and cost-effective tests, Examining Bimodality (EB) and Difference in Capture History (DCH), to determine whether a sample contains genotyping errors and the relative magnitude of the problem. These tests formalize currently used approaches for identifying genotyping errors. We evaluate the efficacy of these tests using simulated CMR data. Results show that both tests are effective at detecting genotyping errors and provide a strong indication of whether the data are error free. The EB and DCH tests apply to data in which multiple samples are associated with individuals, such as those generated by CMR sampling. Managers need to be able to identify and eliminate genotyping errors to produce population estimates that are both unbiased and scientifically defensible.

As DNA techniques become more commonplace, non-invasive, DNA-based monitoring of vertebrate species has become increasingly popular (e.g., Taberlet et al. 1997, Woods et al. 1999, Mowat and Paetkau 2002). Initial monetary, logistical, and technological constraints have been alleviated by recent laboratory advancements, providing researchers with relatively inexpensive methods for developing molecular markers appropriate for a wide range of species. Thus, non-invasive sampling makes generation of count data for difficult-to-study species, such as rare carnivores, potentially feasible. For example, Kohn et al. (1999) extracted DNA from scat collected on paths and roads to estimate the size of a coyote (*Canis latrans*) population in California, USA, and Woods et al. (1999) used hair collection devices to estimate grizzly bear (*Ursus arctos*) population size in the Columbia River basin of British Columbia, Canada. Both studies used non-invasive genetic sampling in a CMR framework to estimate population size. Others have used non-invasive sampling of hair and scat to provide an index of population size when small sample sizes precluded population estimation (e.g., Taberlet et al. 1997).

We suspect, as molecular markers become more plentiful and DNA analyses faster and less expensive, that non-invasive DNA sampling will become the dominant tool for monitoring many species. Currently, the molecular markers of choice for non-invasive genetic CMR analyses are microsatellites. Microsatellites are appropriate markers to use in CMR analyses for many reasons: (1) microsatellites are variable in many species; (2) microsatellites generally are believed to be selectively neutral genetic markers with largely independent evolution; (3) many markers exist for many species; (4) cross-species microsatellite amplification allows the use of many microsatellite markers for unstudied species (Primmer et al. 1996); (5) protocols to develop new or additional microsatellites are getting simpler and less time consuming; and (6) combining many, highly variable, independent microsatellites can provide a unique "molecular tag" or "fingerprint" for every individual in a population. For example, Mowat and Paetkau (2002) used 6 microsatellite markers with an average heterozgosity of 0.69 and had sufficient power to individually identify 139 of 141 marten (*Martes americana*) based on hair samples collected in British Columbia.

Despite the power of non-invasive sampling for enumerating populations, this technique is still in its infancy and is fraught with many difficulties, especially when tied to CMR methods. Capture–mark–recapture methods assume no error in individual identification, and estimates can become severely biased if tags are lost or read improperly. The use of genetic tags derived through non-invasive sampling complicates CMR
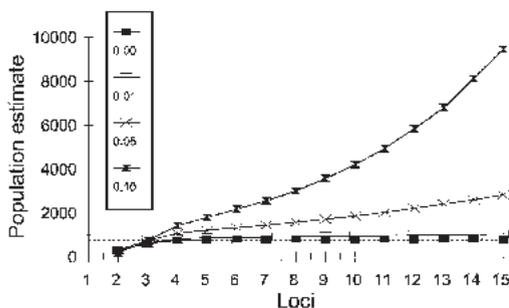
[1] E-mail: kmckelvey@fs.fed.us

Fig. 1. Following Waits and Leberg (2000), population estimates are based on sampling a theoretical population of 800 individuals and applying genetic markers using 2–15 loci and per-locus allelic dropout rates of 0.0–0.10. Estimates are averages associated with 5 sampling periods, a 20% probability of capture (*p*) per period, and using the null model in program CAPTURE. Each line is the average of 500 simulations.

estimation by introducing identification errors that lead to either positive or negative biases in abundance estimates. Underestimation of abundance is due to the "shadow effect" (Mills et al. 2000), a phenomenon of multiple individuals having identical tags caused by using too few loci or loci with low heterozygosity. Fortunately, shadow effects can be removed by using sufficiently large, heterozygous tags. Under these circumstances, probability of identity (PI), defined as the probability that 2 randomly chosen individuals in a population have identical genotypes (Paetkau and Strobeck 1994), becomes minuscule, and the shadow effect largely disappears.

Abundance overestimation due to genotyping errors is much more difficult to control (Taberlet et al. 1996, 1999; Goossens et al. 1998; Taberlet and Luikart 1999; Waits and Leberg 2000). Several sources of genotyping error exist: allelic dropout, false alleles, misreading banding patterns (laboratory error), and transcription errors (e.g., typing errors associated with transferring genetic data to a data file; Taberlet et al. 1996, Gagneux et al. 1997, Paetkau 2003). Most genotyping error rates are highly correlated with the amount of DNA collected (Taberlet et al. 1996, Morin et al. 2001). Allelic dropout, the amplification of only 1 of 2 alleles in a heterozygote pair producing a false homozygote, and amplification of false alleles occur in non-invasively collected samples because DNA is often in the picogram range (Morin et al. 2001), 3 orders of magnitude less than typically collected from tissue samples (Foucault et al. 1996, Gagneux et al. 1997, Taberlet and Luikart 1999). Furthermore, this low-quantity DNA may

be damaged due to exposure to moisture, heat, and ultraviolet radiation encountered during non-invasive sample collection.

Allelic dropout is possibly the most severe problem with non-invasive DNA sampling. In a study on chimpanzees (*Pan troglodytes*), Gagneux et al. (1997) showed that 31% of all single-hair amplifications produced allelic dropout. Goossens et al. (1998) found that error rates fell from 14.0 to 4.9 to 0.3% as numbers of alpine marmot (*Marmota marmota*) hairs increased in a polymerase chain reaction (PCR) from 1 to 3 to 10 hairs. While error rates can be reduced when multiple hairs are used to collect DNA, errors likely will not be eliminated given the quality of DNA generally associated with non-invasive genetic sampling (Goossens et al. 1998).

Genotyping errors usually produce "new" individuals because these errors are largely random and the probability of duplication of identical random patterns across multiple loci is extremely unlikely (Waits and Leberg 2000). Therefore, these samples are unique and are identified as new captures in a CMR experiment. The presence of false individuals created by genotyping error both elevates the number of unique individuals and decreases the recapture rate. Thus, the effects of these errors on population estimators are multiplicative. For instance, Creel et al. (2003) found in a study of wolves (*Canis lupus*) in Yellowstone National Park that genotyping errors could cause up to a 5.5-fold increase in estimated population size.

What makes these errors pernicious is not only their strong impact on abundance estimates, but, even more importantly, the rate of misidentification increases with both the number of loci analyzed (Waits and Leberg 2000; Fig. 1) and the total number of samples collected. Genotyping errors occur at the level of the locus. Therefore, the total number of errors will be the per-locus error rate × the number of loci × the number of samples. To properly evaluate CMR data, the tags (in this case the multilocus genotypes) should be unique, and the design should generate multiple recaptures such that deviations from null expectations can be evaluated (Otis et al. 1978). However, with genotyping error, every sample representing a recapture could potentially produce a new individual. With a 10% misidentification rate per sample, an individual that produced 4 samples would have a 34% chance of producing at least 1 false new individual. Therefore, a well-designed study based on CMR theory that seeks to obtain many samples from each individual

across multiple sampling periods will have higher rates of genotyping error—and may produce worse estimates—than more casual studies that collect relatively few samples. This problem is compounded by the fact that many non-invasive sampling designs produce many redundant samples per individual. For instance, a single individual may leave many separate hair samples on a barbed-wire corral in a single sampling session.

Current methodologies for dealing with allelic dropout suggest a "multi-tube" approach (Taberlet et al. 1996). The multi-tube approach involves multiple assays performed for each locus followed by the application of a dichotomous set of rules. The first rule is that a homozygous genotype is not accepted until it has been repeated 7 times; the second rule is that no allele is accepted until it has been identified at least twice (Taberlet et al. 1996, but see Paetkau 2003). This approach is expensive and time consuming. Furthermore, even after thorough repetition, no independent test exists to determine whether the allelic dropout errors were removed (Creel et al. 2003); thus, a 7-fold increase in the number of assays to remove dropout errors potentially increases laboratory errors.

We propose 2 tests that can be used to screen datasets to determine whether allelic dropout is a problem and to test whether the problem has been removed by multi-tube or other approaches (Mowat and Paetkau 2002). We present our tests in a context similar in design to a large, multi-million dollar effort to count grizzly bears in western Montana, USA (K. Kendal, U.S. Geologic Survey, personal communication), and we relate our results to other situations. Our objectives were to develop tests that both identify and demonstrate the removal of genotyping errors from non-invasively collected genetic samples used for CMR estimation.

## METHODS

To understand patterns associated with genotyping errors, we generated a simulated dataset using program EASYPOP (Balloux 2001). We designed simulated data to be similar to samples obtained through non-invasive grizzly bear sampling containing 15 loci with 5 alleles per locus and an average heterozygosity of 0.78. To produce a realistic population at equilibrium, a large population ($n = 800$) was allowed to "evolve" over 1,000 generations with a low mutation rate ($\mu = 0.001$) consistent with mutation rates for microsatellites (Valdes et al. 1993, Di Rienzo et al. 1994, Schug et al. 1997).
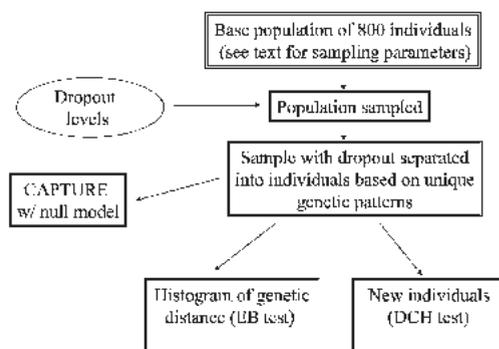


Fig. 2. A flow diagram showing the design of simulations used to demonstrate the Examining Bimodality (EB) and Difference in Capture History (DCH) tests, which allow evaluation of genetic errors introduced through allelic dropout.

We created 5 sampling periods by drawing from the population 5 times, with each individual having a 20% probability of being sampled on each occasion, consistent with null model expectations (Otis et al. 1978). All error was due to allelic dropout, and dropout rates were equally likely for each locus. For each locus, we generated a uniform random deviate ranging between zero and 1. If the random number was less than the specified dropout rate, 1 of the 2 alleles was randomly discarded, producing a homozygous locus. This process had no effect on loci that were already homozygous. We used these data for all subsequent analyses (Fig. 2).

Following Waits and Leberg (2000), we evaluated the impacts of these errors on CMR estimators by entering sampling data into program CAPTURE (White et al. 1978). During the sampling process, all new samples were compared to the samples previously acquired and were either classified as recaptures or new captures based on whether they matched any of the existing multi-locus genotypes. We subsequently used these data to generate a capture history file that could be read by CAPTURE, and we performed analyses using the null model ($M_0$; Otis et al. 1978). CAPTURE then produced an output file containing the population estimate. We repeated the process 500 times for each combination of allelic dropout and genetic tag size. After determining the bias caused by each level of allelic dropout, we ran the sample through our 2 tests to determine whether error rates could be detected.

*Test 1: Examining Bimodality.*—Genetic transfer to offspring produces a bell-shaped relatedness distribution in an ideal population (Queller and
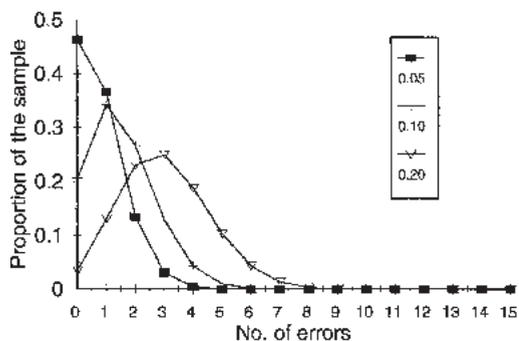
Fig. 3. Binomial error rates associated with per-locus allelic dropout rates of 0.05–0.20, and assuming 15 heterozygous loci in simulations used to demonstrate the Examining Bimodality (EB) and Difference in Capture History (DCH) tests, which allow evaluation of the genetic errors introduced through allelic dropout. Because genotyping errors occur randomly, error rates will follow binomial expectations. At low error rates, most samples will contain no more than a few errors. Samples identified as new captures rather than recaptures due to genotyping error will generally differ only at a few loci.
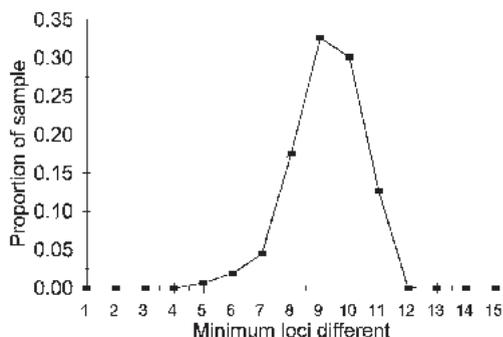


Fig. 4. The minimum number of loci at which 1 individual differs from all other individuals in a simulated population with no genotyping errors based on a sample of 15 heterozygous loci. Individuals in a population differ at multiple loci, and a capture–mark–recapture sample evaluated at 15 loci and without error will, therefore, be composed of identical recaptures and new captures that differ at several loci.

Goodnight 1989, Ruzzante et al. 2001, Rousset 2002). In a population sampled in a manner that generates recaptures, some of the samples (i.e., recaptures) will be genetically identical, whereas the rest will differ. Assuming, for simplicity, that all loci have the same number of alleles in the same proportions, the number of loci at which unique individuals differ will follow binomial expectations. Given $n$ total amplified loci, the proportion of the population that is identical at $k$ loci will be:

$$S_k = \binom{n}{k} P^k (1 - P)^{n-k}, \qquad (1)$$

where $S$ is the proportion of the population that is the same at $k$ loci, and $P$ is the probability of being different at each locus. If $n$ is sufficiently large and $P$ sufficiently small, then most samples will be either unique individuals or identical recaptures.

For recaptures, genotyping errors generally lead to unique genotypes at each locus where an error occurred. Therefore, the number of loci at which a recapture sample differs from other recaptures due to errors will follow similar binomial expectations, substituting the error rate ($E$) for $P$ in Equation 1 (Fig. 3). If $E$ is close to 0.0 and $P$ is close to 1.0, $n$ is sufficiently large, and the sample contains a reasonable proportion of recaptures, computing $S_k$ for all samples and ignoring the recaptures will produce a bimodal distribution. The popula-

tion of unique individuals will form a mode well away from $S_0$, and recaptures containing errors will form a mode close to $S_0$. The samples in the lower mode likely contain error, and the location of the lower mode indicates the approximate error rate. This approach is a formalization of the ad hoc methods presented in Mowat and Paetkau (2002) and Paetkau (2003), who reexamine samples that differed by only 1 or 2 alleles.

In our simulated population with 15 loci analyzed, the distribution of minimum number of loci at which individuals differ is unimodal with the center being far from zero: very few individuals differ at all loci, and few differ at zero loci (Fig. 4). Therefore, unless per-locus allelic dropout rates are extreme, most of the recaptures identified as new individuals due to allelic dropout will only be separated from other individuals at 1, 2, or 3 loci and will be clearly separated from the population distribution. Using these approaches, genotyping errors can only be found within the recapture data; however, for the limited purpose of CMR estimation, only these errors bias the estimates.

*Test 2: Difference in Capture History.*—An additional test is suggested by Fig. 1. We can ask the question: as additional loci are added to the genetic tag, does the population estimate or number of new individuals in the sample increase faster than would be expected through the removal of the shadow effect? Assuming a population with high heterozygosity and multiple allelic states at each locus, the probability of 2 individuals having identical alleles decreases

multiplicatively with increasing loci. At some intermediate number of loci (with high heterozygosity, generally about 6–7), with no genotyping error, the probability of identifying new individuals through expanding the genetic tag becomes infinitesimal (Mills et al. 2000). However, genotyping errors occur with each locus added to the tag, generating genetic differences that are interpreted as new individuals. Assuming that all new individuals are the product of dropout, then the probability of detecting at least 1 error—and hence at least 1 new individual—in a sample is:

$$P_{err\_indiv} = 1 - \prod_{i=1}^{l}(1 - P_{h,i}P_{a,i});$$

$$P_{err\_samp} = 1 - \prod_{j=1}^{n}(1 - P_{err\_indiv,j}), \qquad (2)$$

where $P_{err\_indiv}$ is the probability that an individual recapture is mistyped, $l$ is the number of loci, $P_{h,i}$ is the probability that an individual is heterozygous at locus $i$, $P_{a,i}$ is the probability that a genotyping error occurs at locus $i$, $n$ is the number of recaptures, and $P_{err\_samp}$ is the probability that a sample contains errors.

Therefore, the second test proceeds as follows: (1) Amplify all available loci (15 for our simulations). (2) Screen out recaptures (all organisms with the same multilocus genotype). (3) Compute the probability of 2 individuals being genetically identical by computing PI (Paetkau and Strobeck 1994) or the more conservative probability of identity for siblings (PI$_{sib}$; Evett and Weir 1998, Waits et al. 2001). Compute PI and/or PI$_{sib}$ for 2 through $n$ (in our case 15) loci and choose a number of loci where PI and PI$_{sib}$ are small. We suggest that these choices be very conservative; we recommended PI < 1 × 10$^{-6}$ and/or PI$_{sib}$ < 1 × 10$^{-3}$. We will refer to the number of loci necessary to achieve this as $L\_base$, the base number of loci necessary to reduce the shadow effect to insignificant levels (Mills et al. 2000). (4) Generate a list of unique individuals when the tag is $L\_base$ in size ($N_{L\_base}$), and compare this to the number of unique individuals generated through adding additional loci to the genetic tag. If the sample is error free, no new individuals likely will be generated. An efficient way to evaluate the contribution of each locus to the genetic tag is to compare $N_{L\_base}$ and the number of individuals identified at $L\_base + 1$ loci ($N_{L\_base+1}$), rotate the order of the loci, repeat the process until all loci have been moved through the $L\_base + 1$ position, and sum the number of new individuals produced.

Those samples that change from being viewed as recaptures to unique individuals when reevaluated at $N_{L\_base+1}$ loci are likely to contain genotyping errors. Therefore, by extension, if specific combinations of loci produce more new individuals, these groups of loci contain higher error rates. Specifically, if many new individuals are produced when a specific locus is rotated through the $L\_base + 1$ position, that locus is likely problematic. A straightforward method to assess this possibility is to keep track of the number of individuals produced at each rotation. A chi-square homogeneity test can be applied to these data and, if significant, simultaneous Bonferroni confidence intervals constructed to determine which loci are contributing more new individuals than expected. To test the ability to identify individual loci with high dropout rates, we ran an additional simulation (500 repetitions) with an allelic dropout rate of 0.05 per locus, except for 1 locus that had a 0.20 dropout rate.

Based on Equation 2, our second test should be exquisitely sensitive for situations in which genotyping error is low and can be applied with only $L\_base + 1$ loci amplified (Table 2). However, problems occur with this approach when error rates are high. At high error rates (i.e., >0.25/locus), many recaptures will appear unique with $L\_base$ or fewer loci, and these errors will be hidden from the test. In extreme cases, all recaptures may be identified as unique individuals. In this case, $N_{L\_base}$ = number of samples, and the test will fail.

## RESULTS

For the simulated population, a 7-locus tag produced PI = 3.6E$^{-8}$ and PI$_{sib}$ = 1.0E$^{-3}$ (Table 1) and was chosen as $N_{L\_base}$. The first test clearly demonstrated bimodality for per-locus allelic dropout rates of up to 20% (Fig. 5). The test population, without errors, had no individuals that differed at <5 loci, meaning that at lower rates of allelic dropout, the entire lower mode was due to error (compare Figs. 4 and 5).

The second test, comparing the number of individuals identified at $N_{L\_base}$ and $N_{L\_total}$ was indeed very sensitive to the presence of genotyping error. For the simulated sample sizes ($\bar{x}$ = 538) without dropout, the test never produced additional individuals. Furthermore, at per-locus error rates ≥0.002, we never failed to demonstrate the production of at least 1 individual (1,000 repetitions; Fig. 6). In 497 out of 500 repetitions, a chi-square homogeneity test indicated that the number of

Table 1. Probabilities of identity (PI; Paetkau and Strobeck 1994) and probability of identity for siblings ($PI_{sib}$; Evett and Weir 1998) derived from sampling a theoretical population with overall heterozygosity of 0.78. Statistics associated with each locus and the product obtained by combining 2–15 loci are shown. Product calculation assumes that allele frequencies for all loci are independent.

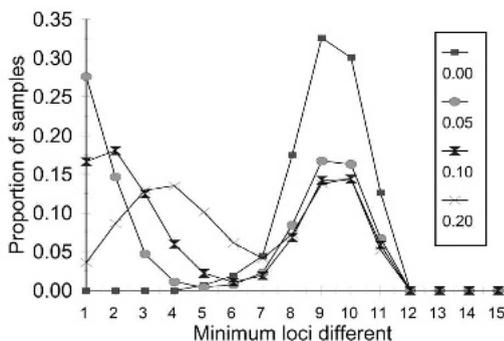| Locus | PI | $PI_{sib}$ | $\Pi PI$ | $\Pi PI_{sib}$ |
|---|---|---|---|---|
| 1 | 0.081 | 0.378 | 0.0808 | 0.3779 |
| 2 | 0.084 | 0.381 | 0.0068 | 0.1441 |
| 3 | 0.094 | 0.395 | 0.0006 | 0.0569 |
| 4 | 0.081 | 0.379 | $5.192E^{-05}$ | 0.0216 |
| 5 | 0.101 | 0.404 | $5.249E^{-06}$ | 0.0087 |
| 6 | 0.081 | 0.378 | $4.249E^{-07}$ | 0.0033 |
| 7 | 0.084 | 0.382 | $3.573E^{-08}$ | 0.0013 |
| 8 | 0.096 | 0.398 | $3.439E^{-09}$ | 0.0005 |
| 9 | 0.082 | 0.380 | $2.832E^{-10}$ | 0.0002 |
| 10 | 0.081 | 0.378 | $2.304E^{-11}$ | $7.174E^{-05}$ |
| 11 | 0.081 | 0.378 | $1.873E^{-12}$ | $2.715E^{-05}$ |
| 12 | 0.086 | 0.383 | $1.605E^{-13}$ | $1.040E^{-05}$ |
| 13 | 0.081 | 0.378 | $1.305E^{-14}$ | $3.938E^{-06}$ |
| 14 | 0.088 | 0.386 | $1.154E^{-15}$ | $1.519E^{-06}$ |
| 15 | 0.083 | 0.381 | $9.605E^{-17}$ | $5.781E^{-07}$ |



Fig. 5. Minimum number of loci different between each sample and all other samples, given a theoretical population with average heterozygosity of 0.78. Lines represent the proportion of the derived sample (5 sampling events, 20% capture probability per event) given per-locus allelic dropout rates of 0.0–0.20. Lines are the average of 30 simulations.

new individuals produced per rotation was not homogeneous ($P \leq 0.05$). In all 497 cases in which the chi-square test was significant, subsequent Bonferroni confidence intervals successfully detected the locus with a 0.20 allelic dropout rate when the locus was in the $L\_base + 1$ position.

## DISCUSSION

The ability to conduct a genetic CMR study has been plagued by problems associated with genotyping errors. This problem appears ubiquitous when using non-invasive samples from some species, and almost nonexistent for other species—probably due to the nature of the material collected (Goossens et al. 1998). However, virtually

no methods are available to determine whether a genetic dataset generated from non-invasively collected samples contains genotyping errors, except by repeating analyses many times (Taberlet et al. 1996). While a multi-tube approach provides data on error rates, it can make non-invasive DNA studies as expensive as traditional CMR studies and, in the end, does not provide a test of whether genetic errors have been successfully removed. In contrast, the DCH test can provide strong evidence that a sample containing recaptures is error free.

Although not detailed here, we ran additional simulations with smaller sample sizes ($n = 200$), lower multilocus heterozygosity ($H = 0.45$), and substantial population substructure (e.g., Fst = 0.22). Decreasing sample size had little effect; the bimodal distributions were still well defined and

Table 2. Probabilities of producing at least 1 false individual given a genetic sample with 8 amplified loci[a]. Heterozygosity is 0.8, $L\_base$[b] = 7, and allelic dropout is assumed to be equally probable for all loci.

| Allelic dropout | Number of recaptures in sample | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 |
| 0.005 | 0.274 | 0.473 | 0.618 | 0.723 | 0.799 | 0.854 | 0.894 | 0.923 | 0.944 | 0.959 |
| 0.010 | 0.474 | 0.723 | 0.855 | 0.923 | 0.960 | 0.979 | 0.989 | 0.994 | 0.997 | 0.998 |
| 0.015 | 0.619 | 0.855 | 0.945 | 0.979 | 0.992 | 0.997 | 0.999 | 1.000 | 1.000 | 1.000 |
| 0.020 | 0.725 | 0.924 | 0.979 | 0.992 | 0.998 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 0.025 | 0.801 | 0.961 | 0.992 | 0.998 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 0.030 | 0.857 | 0.979 | 0.997 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 0.035 | 0.897 | 0.989 | 0.999 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 0.040 | 0.926 | 0.995 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 0.045 | 0.947 | 0.997 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 0.050 | 0.962 | 0.999 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

[a] Based on Equation 2.
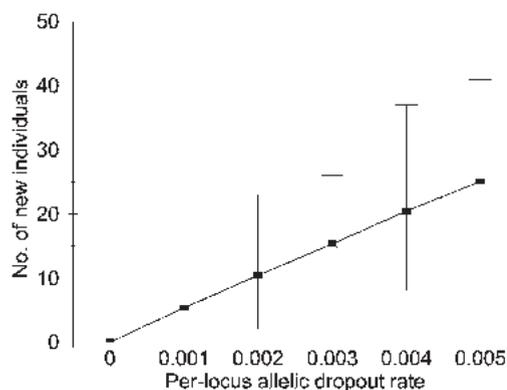[b] Number of loci needed to reduce the shadow effect to insignificant levels.

Fig. 6. Average number of individuals (1,000 simulations) produced with a base genetic tag of 7 loci and a total of 15 loci based on our Difference in Capture History test given a population of 800, 5 sampling periods, a 20% probability of capture in each period, and differing rates of allelic dropout. Average sample size was 538. Error bars are the maximum and minimum numbers of new individuals produced.

the DCH test was still sensitive, as expected based on Table 2. Reducing heterozygosity required us to use more loci to achieve a reasonable *L_base* in the DCH test, but the test performed well. Populations with undetected, significant substructure can produce confusing patterns. If a subset of the population has much lower heterozygosity (e.g., a small, island population enclosed within the main population), then a second mode will be produced. Similarly, because the shadow effect has not been sufficiently reduced in the proportion of the sample with low heterozygosity, new individuals will arise in the DCH test. Thus, the presence of a bimodal structure does not necessarily mean that the samples contain errors; it could indicate that the sample contains individuals from >1 population. However, the lack of any bimodal structure in the EB test strongly indicates a lack of error. Similarly, identifying new individuals when the genetic tag size is increased either indicates error or insufficient removal of the shadow effect, but the lack of new individuals strongly indicates that both the tag size is sufficient and the sample is error free. Note that, even lacking error, samples with these characteristics (i.e., shadow effect) will produce unreliable CMR estimates unless these patterns are recognized and the genetic tag size increased to remove shadow effects.

To avoid the multi-tube approach, some researchers have scrutinized the signals inherent in the datasets (Poole et al. 2001, Mowat and Paetkau 2002, Paetkau 2003, Schwartz et al.

2004). For example, Mowat and Paetkau (2002) screened their samples, discarding any sample that amplified at <50% of their loci (3 of 6 loci) and re-amplifying any samples that only produced scoreable products at 50–83% of loci (3–5 loci). Mowat and Paetkau (2002) also reexamined any pairs of genotypes in their dataset differing at only 1 locus. This approach may work well if error rates are low (typical of the species studied by Mowat and Paetkau [2002] and Paetkau [2003]), but may fail if high error rates are observed (Fig. 5). Further, like the multi-tube methods, Mowat and Paetkau (2002) and Paetkau (2003) provided no demonstrable evidence that allelic dropout had indeed been removed. In fact, we have seen situations in our laboratory in which all loci amplify, yet the genotype is inconsistent across multiple runs.

We believe that a practical approach therefore begins with amplifying a tag sufficiently large to run the proposed tests. For our simulated population, the minimum tag size, based on our criteria for computing *L_base* was 8 loci. However, our sample was more heterozygous than many, and therefore a larger tag often will be required. If bimodality is apparent after applying the EB test to the sample, then error is likely. Because the error is likely concentrated in the samples that compose the lower mode, the degree to which error is the cause of the bimodality can be determined by data checking and performing multitube analyses on the suspect samples (see Woods et al. 1999, Mowat and Paetkau 2002, Paetkau 2003). If the lower mode is >1 locus and is determined to be generated by error, then error rates are likely >10%, and few of the samples will be error free (Fig. 5). In this case, the response depends on the uses of the dataset. The only errors exposed through these methods are those on samples representing recaptured individuals that will affect CMR estimation. Therefore, if the sample is to be used for any purposes other than CMR estimation, we recommend using a multitube approach for all samples.

After multi-tube analysis of the suspect samples and error removal, the bimodality test should be repeated; hopefully, bimodality will be eliminated. As a final test, once allelic dropout has been reduced, we suggest applying our DCH test. Because of the sensitivity of this test, failure to produce new individuals would indicate that genotyping errors were reduced to insignificant levels.

If error rates are low, the cost of amplifying 8–20 loci 1 time will be significantly less than amplify-

ing 6 loci 6–7 times, and multi-tube analysis can be limited to samples likely containing errors. Furthermore, a dataset with larger numbers of loci will allow better estimation of additional population genetic parameters, such as substructure, genetic bottlenecks, or relatedness (Ritland 1996, Luikart and Cornuet 1998, Manel et al. 2003).

We feel that formal population estimates should not be based on genetic tagging in which only 6–7 loci are amplified, unless samples are re-run with a high enough frequency to detect all errors (Taberlet et al. 1996). Although a 6–7 locus tag, in theory, may be sufficient to identify most of the individuals if the loci are highly heterozygous, it is not sufficient to check the validity of those identities using the methods that we propose. Even without error, with a small tag, some individuals will be expected to differ at only 1–2 loci and, changing the tag size and composition likely will affect the capture history. Therefore, with few loci, the tests we propose will produce mixtures of real data and genotyping error with results that are difficult to interpret.

For our tests to be effective, a sample must have several properties. First, enough loci must be amplified to reduce the shadow effect to insignificant levels; this requires the amplification of highly heterozygous loci. The EB test is critically dependent on having a high per-locus difference between PI and (1 – error rate). As these 2 probabilities converge, the error mode and population mode also will converge and the first test will cease to be diagnostic. Similarly, if heterozygosity is low, $L\_base$ will increase and the second test will become both less tractable and less sensitive. Second, the number of loci amplified must be sufficient to reduce the shadow effect to insignificant levels. Last, the sample must contain a substantial number of recaptures (see Table 2).

With these restrictions, many samples collected using non-invasive methods will not be of sufficient quality for the meaningful application of our tests. If only a few loci amplify and the amplified loci differ from sample to sample, we believe that genotyping error cannot be properly evaluated, and its impact on CMR estimates will be largely unknown. Additionally, the very heavy screening necessary when using low-quality samples may lead to biases in abundance estimates (Creel et al. 2003). Given the sensitivity of CMR estimators to genotyping errors, we do not advise the use of CMR estimators on extremely poor-quality genetic samples.

Our tests apply only when multiple samples are available from the same individual; the more samples per individual, the stronger the tests and the more coherent the output. For instance, if an individual is recaptured several times and a genotyping error occurs in 1 of the samples, this same sample will differ from several others at only 1 locus. If the difference is due to a heterozygote becoming a homozygote, then the error is likely allelic dropout and the sample containing the error becomes obvious. Multiple samples from individuals are common in non-invasive sampling and are particularly prevalent if the study is designed to generate CMR statistics. However, genotyping errors occur in all large analyses of samples, even if the samples are associated with high-quality DNA sources such as fresh tissue. For sample sets derived from unique individuals, researchers may take advantage of our tests by inserting repeat samples as a blind test and running the tests on the DNA results. We recommend developing these types of blind tests whenever subsequent DNA analyses are sensitive to genotyping error (e.g., paternity estimates; Marshall et al. 1998).

The problems we discussed are similar to problems found using conventional bands or tags. For example, bird bands can be difficult to read after several years. Field personnel commonly report that a bird tag was probably 1 number, but the last digit of the band was obscured or worn. For instance, an "8" may be significantly worn on some tags such that it may appear like a "3" or "0." These errors have largely been ignored in many studies. However, the consequences of these types of misidentification are similar to those we report here. The advantage of genetic tagging is that methods are available to "recover" the initial tag or at least determine whether the reading is in fact an error.

## MANAGEMENT IMPLICATIONS

Using genetic tags in CMR studies has huge potential and already has been instrumental in estimating population sizes of rare, difficult-to-study species (Palsboll et al. 1997, Poole et al. 2001, Paetkau 2003). Initial excitement has been tempered by realization that these techniques have some pitfalls (Taberlet et al. 1999). The ability to locate genotyping errors, remove the errors, and demonstrate that the resulting dataset is error-free is essential for the generation of accurate and defensible population estimates. Error-checking methods, such as those we presented, allow managers to use genetically based CMR methods with greater confidence, and

genetically based methods offer many benefits other than cost and feasibility.

Population estimates from CMR studies—genetically based or otherwise—will always be sensitive to a variety of sampling assumptions, which, if improperly modeled, lead to biased estimates (McKelvey and Pearson 2001). In many cases, non-invasive sampling allows greater freedom in sampling design, both because obtaining samples is easier and—more importantly—because the organisms are already "marked." Samples representing separate sampling periods often can be collected simultaneously rather than sequentially, greatly improving population closure. Furthermore, passive collection methods such as the collection of scats are not prone to trap-related problems. If samples are passively collected, trap-happy and trap-shy behaviors, latency to first capture due to trap aversion, and differential capture rates between individuals are not issues. Once tag-identification errors associated with non-invasive genetic sampling are resolved, managers and biologists will have a powerful tool for estimating animal abundance.

## ACKNOWLEDGMENTS

## LITERATURE CITED

BALLOUX, F. 2001. EASYPOP (version 1.7): a computer program for population genetics simulations. Journal of Heredity 92:301–302.

CREEL, S., G. SPONG, J. L. SANDS, J. ROTELLA, J. ZEIGLE, L. JOE, K. M. MURPHY, AND D. SMITH. 2003. Population size estimation in Yellowstone wolves with error-prone noninvasive microsatellite genotypes. Molecular Ecology 12:2003–2009.

DI RIENZO, A., A. C. PETERSON, J. C. GARZA, A. M. VALDES, M. SLATKIN, AND N. B. FREIMER. 1994. Mutational processes of simple sequence repeat loci in human populations. Proceedings of the National Academy of Sciences 91:3166–3170.

EVETT, I. W., AND B. S. WEIR. 1998. Interpreting DNA evidence: statistical genetics for forensic scientists. Sinauer, Sunderland, Massachusetts, USA.

FOUCAULT, F., F. PRAZ, C. JAULIN, AND M. AMOR-GUERET. 1996. Experimental limits of PCR analysis of (CA)n repeat alterations. Trends in Genetics 12:450–452.

GAGNEUX, P., C. BOESCH, AND D. S. WOODRUFF. 1997. Microsatellite scoring errors associated with noninvasive genotyping based on nuclear DNA amplified from shed hair. Molecular Ecology 6:861–868.

GOOSSENS, B., L. P. WAITS, AND P. TABERLET. 1998. Plucked hair samples as a source of DNA: reliability of dinucleotide microsatellite genotyping. Molecular Ecology 7:1237–1241.

KOHN, M. H., E. C. YORK, D. A. KAMRADT, G. HAUGHT, R. M. SAUVAJOT, AND R. K. WAYNE. 1999. Estimating population size by genotyping faeces. Proceedings of the Royal Society of London, Series B 266:657–663.

LUIKART, G., AND J-M. CORNUET. 1998. Empirical evaluation of a test for identifying recently bottlenecked populations from allele frequency data. Conservation Biology 12:228–237.

MANEL, S., M. K. SCHWARTZ, G. LUIKART, AND P. TABERLET. 2003. Landscape genetics: combining landscape ecology and population genetics. Trends in Ecology and Evolution 18:189–197.

MARSHALL, T. C., J. SLATE, L. E. B. KRUUK, AND J. M PEMBERTON. 1998. Statistical confidence for likelihood-based paternity inference in natural populations. Molecular Ecology 7:639–655.

MCKELVEY, K. S., AND D. E. PEARSON. 2001. Population estimation with sparse data: the role of estimators versus indices revisited. Canadian Journal of Zoology 79:1754-1765.

MILLS, L. S., J. J. CITTA, K. P. LAIR, M. K. SCHWARTZ, AND D. A. TALMON. 2000. Estimating animal abundance using noninvasive DNA sampling: promise and pitfalls. Ecological Applications 10:283–294.

MORIN, P. A., K. E. CHAMBERS, C. BOESCH, AND L. VIGILANT. 2001. Quantitative polymerase chain reaction analysis of DNA from noninvasive samples for accurate microsatellite genotyping of wild chimpanzees (*Pan troglodytes verus*). Molecular Ecology 10:1835–1844.

MOWAT, G., AND D. PAETKAU. 2002. Estimating marten *Martes americana* population size using hair capture and genetic tagging. Wildlife Biology 8:201–209.

OTIS, D. L., K. P. BURNHAM, G. C. WHITE, AND D. R. ANDERSON. 1978. Statistical inference from capture data on closed animal populations. Wildlife Monographs 62.

PAETKAU, D. 2003. An empirical exploration of data quality in DNA-based population inventories. Molecular Ecology 12:1375–1387.

———, AND C. STROBECK. 1994. Microsatellite analysis of genetic variation in black bear populations. Molecular Ecology 3:489–495.

PALSBOLL, P. J., J. ALLEN, M. BERUBE, P. J. CLAPHAM, T. P. FEDDERSEN, P. S. HAMMOND, R. R. HUDSON, H. JORGENSEN, S. KATONA, A. H. LARSEN, F. LARSEN, J. LIEN, D. K. MATTILA, J. SIGURJONSSON, R. SEARS, T. SMITH, R. SPONER, P. STEVICK, AND N. OIEN. 1997. Genetic tagging of humpback whales. Nature 388:767–769.

POOLE, K. G., G. MOWAT, AND D. A. FEAR. 2001. DNA-based population estimate for grizzly bears in northeastern British Columbia, Canada. Wildlife Biology 7:105–115.

PRIMMER, C., A. MOLLER, AND H. ELLENGREN. 1996. A wide-range survey of cross-species microsatellite amplification in birds. Molecular Ecology 5:365–378.

QUELLER, D. C., AND K. F. GOODNIGHT. 1989. Estimating relatedness using genetic markers. Evolution 43:258–275.

RITLAND, K. 1996. Estimators for pairwise relatedness and individual inbreeding coefficients. Genetical Research 67:175–185.

ROUSSET, F. 2002. Inbreeding and relatedness coefficients: what do they measure? Heredity 88:371–380.

RUZZANTE, D. E., M. M. HANSEN, AND D. MELDRUP. 2001. Distribution of individual inbreeding coefficients, relatedness and influence of stocking on native anadromous brown trout (*Salmo trutta*) population structure. Molecular Ecology 10:2107–2128.

SCHWARTZ, M. K., K. L. PILGRIM, K. S. MCKELVEY, E. L. LINDQUIST, J. J. CLAAR, S. LOCH, AND L. F. RUGGIERO. 2004. Hybridization between Canada lynx and bobcats: genetic results and management implications. Conservation Genetics 5:349–355.

SCHUG, M. D., T. F. C. MACKAY, AND C. F. AQUADRO. 1997. Low mutation rates of microsatellite loci in Drosophila melanogaster. Nature Genetics 15:99–102.

TABERLET, P., J. J. CAMARRA, S. GRIFFIN, E. UHRÈS, O. HANOTTE, L. P. WAITS, C. DUBOIS-PAGANON, T. BURKE, AND J. BOUVET. 1997. Noninvasive genetic tracking of the endangered Pyrenean brown bear population. Molecular Ecology 6:869–876.

———, S. GRIFFIN, B. GOOSSENS, S. QUESTIAU, V. MANCEAU, N. ESCARAVAGE, L. P. WAITS, AND J. BOUVET. 1996. Reliable genotyping of samples with very low DNA quantities using PCR. Nucleic Acids Research 24:3189–3194.

———, AND G. LUIKART. 1999. Non-invasive genetic sampling and individual identification. Biological Journal of the Linnean Society 68:41–55.

———, L. P. WAITS, AND G. LUIKART. 1999. Noninvasive genetic sampling: look before you leap. Trends in Ecology and Evolution 14:323–327.

VALDES, A. M., M. SLATKIN, AND N. B. FREIMER. 1993. Allele frequencies at microsatellite loci: the stepwise mutation model revisited. Genetics 133:737–749.

WAITS, J. L., AND P. L. LEBERG. 2000. Biases associated with population estimation using molecular tagging. Animal Conservation 3:191–199.

WAITS, L. P., G. LUIKART, AND P. TABERLET. 2001. Estimating the probability of identity among genotypes in natural populations: cautions and guidelines. Molecular Ecology 10:249–256.

WHITE, G. C., K. P. BURNHAM, D. L. OTIS, AND D. R. ANDERSON. 1978. Users manual for program CAPTURE. Utah State University Press, Logan, Utah, USA.

WOODS, J. G., D. PAETKAU, D. LEWIS, B. N. MCLELLAN, M. PROCTOR, AND C. STROBECK. 1999. Genetic tagging free ranging black and brown bears. Wildlife Society Bulletin 27:616–627.