THE OPTIMAL NUMBER OF MARKERS IN GENETIC CAPTURE–MARK–RECAPTURE STUDIES

DAVID PAETKAU,[1] Wildlife Genetics International, Box 274, Nelson, BC V1L 5P9, Canada

Abstract: McKelvey and Schwartz (2004, this issue) propose that the number of markers used to assign individual identity in DNA-based population inventories should be doubled or tripled relative to established practice, primarily to facilitate indirect statistical tests for genotyping errors. If applied to studies that use plucked hair samples, this suggestion would cause a proportional increase in the effort required to generate results, and an even greater increase in the number of errors initially present in those results. Since no empirical or deductive evidence was presented to show how established methods of selective reanalysis can fail to detect errors, I conclude that this proposal would dramatically increase costs without improving data quality. While the optimal number of markers will vary between study populations, I present 1 example in which identical results would have been achieved with 3 markers or with the 15 suggested by McKelvey and Schwartz (2004).

JOURNAL OF WILDLIFE MANAGEMENT 68(3):449–452

Key words: bias, DNA, efficiency, error, hair, non-invasive, reliability, capture–mark–recapture, population size.

Any study that uses genetic analysis to establish individual identity, whether in the context of human forensics or capture–mark–recapture (CMR) abundance estimation, is vulnerable to a type of error that occurs when separate samples from the same individual are recorded as having different genotypes. Since each unique genotype is assumed to correspond to a different individual, this error causes an excess of individuals to be recognized (Taberlet et al. 1996). For genetic CMR capture studies to be practical in large-scale applications, one must identify highly efficient methods that can achieve an uncompromising standard of data quality.

Reasoning that inconsistent genotyping would create very similar pairs of genotypes, Woods et al. (1999) selectively reanalyzed the mismatched markers in pairs of genotypes that matched at 5 of the 6 markers used in their study (1MM-pairs). More recently, I extended this selective reanalysis protocol to include certain 2MM-pairs (pairs of genotypes that match at all but 2 of the markers being analyzed; Paetkau 2003). I recommended reanalyzing 3MM-pairs in rare cases in which a particular genotype is unreplicated (observed in just 1 sample) and differs from a second genotype in a manner consistent with "allelic dropout" (a phenomenon in which just 1 allele is detected for a heterozygous gentoype; Taberlet et al. 1996, Gagneux et al. 1997). In situations where allelic dropout is suspected but where the reanalysis confirms the original (suspect) data, I now repeat the reanalysis up to 7 times, although clear evidence of both alleles usually is obtained by the second or third analysis.

This protocol is most efficient if stringent thresholds are used when scoring the initial genotypes and is therefore well suited to working with relatively high-quality DNA samples like plucked hair. Repeated wholesale reanalysis of every genotype (the "multiple tubes" approach; Taberlet et al. 1996) may be more appropriate when working with scat samples, in which concentrations of DNA are relatively low but much sample material is available. I will restrict this discussion to the plucked-hair samples that are most commonly used in non-invasive CMR studies.

I used the expanded version of selective reanalysis (Paetkau 2003) to scrutinize data from 68 studies that used plucked hairs and 5–7 genetic markers to identify individuals from 9 species of carnivores (lynx [*Lynx canadensis*], ocelot [*Leopardus pardalis*], pine marten [*Martes americana*; Mowat and Paetkau 2002], wolverine [*Gulo gulo*], fisher [*M. pennanti*], badger [*Taxidea taxus*], brown bear [*Ursus arctos*; Poole et al. 2001, Boulanger et al. 2003], black bear [*U. americanus*; Mowat et al. 2004], and sun bear [*U. malayanus*]). In 17 studies where detailed records were published, this protocol identified 210 samples that had 1 error in their initial genotype, and 16 samples with 2-locus errors (Paetkau 2003). The trend in these data suggests that reanalysis of >3MM-pairs is not necessary to detect every error that is initially present in most studies.

This logic is independent of the number of markers used. To take extreme examples, a study that made use of a single marker would have to be reanalyzed in its entirety under this protocol

[1] E-mail: dpaetkau@wildlifegenetics.ca

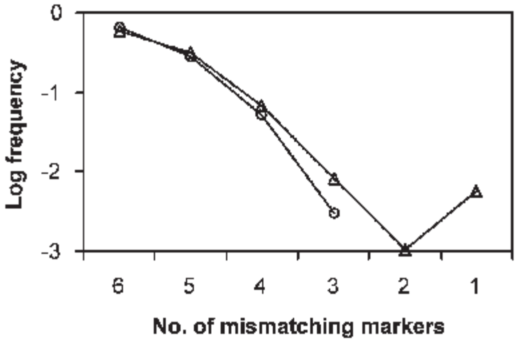

Fig. 1. Mismatch distributions for a 6-locus Yukon brown bear file before (△) and after (○) selective reanalysis (frequency = number of pairs of genotypes with a given number of mismatching markers [MM] divided by the total number of pairs). The distribution was bimodal until selective reanalysis showed that all 16 1MM-pairs and 3 2MM-pairs were caused by errors. One genotype was even shown to have been affected by allelic dropout at 3 markers. The final distribution, in which all pairs of genotypes differed at ≥3 markers, conforms to the steeply sloped unimodal shape observed when analyzing tissue samples from known individuals (Paetkau 2003).

(each distinct genotype would be a 1MM relative to every other genotype), whereas reanalysis would be extremely efficient in a study that used 20 markers (the only 1MM-, 2MM-, or 3MM-pairs would be those caused by errors; McKelvey and Schwartz 2004). Ignoring the fact that the single-locus project would have absurdly high match probabilities and would grossly underestimate the number of individuals sampled, the difference between the single-locus project and the 20-locus project is in the amount of effort that gets "wasted" confirming similar pairs of genotypes that represent different individuals rather than errors. The capacity of the selective reanalysis to detect and correct every error is the same in both studies because all of the inaccurate genotypes would be reanalyzed. In short, the efficiency and not the efficacy of the selective reanalysis approach is affected by altering the number of markers. This focuses the discussion on the number of markers required to maximize efficiency.

The total amount of work required to complete a project using my protocol (Paetkau 2003) is the sum of the first pass (a relatively efficient process in which every sample is analyzed at every marker) and the error checking (a less efficient process in which selected single-locus genotypes are reanalyzed to confirm the differences between similar pairs of genotypes). The number of pairs of genotypes that will fall subject to selective reanalysis is the number of initial errors in genotypes plus the number of pairs of individuals that have similar genotypes by chance. While the work required to complete the first pass increases linearly with the number of markers, the 2 factors that make selective reanalysis necessary respond in opposite directions to changes in the number of markers.

Two reasons explain why the number of initial errors increases with the number of markers. First, each data point that is recorded has an associated probability of error, so increasing the number of markers that are analyzed will increase the total number of errors (Waits and Leberg 2000). Second, the probability that 1 (allelic dropout) or both (outright failure) chromosomes will fail to amplify from a sample is dependent on the amount of DNA that goes into each polymerase chain reaction (PCR; Taberlet et al. 1996). When using finite samples, minimizing the total number of reactions (i.e., markers) allows more DNA to be used in each individual reaction, increasing the probability of observing a result and lowering the probability that this result will be inaccurate (Paetkau 2003).

The frequency of pairs of individuals with similar genotypes is determined by the number and variability of the markers that are analyzed. This frequency cannot be predicted using match statistics or simulations because it depends on the distribution of degrees of relatedness among the sampled individuals; a distribution which is not only unknown, but which varies considerably with the size and degree of isolation of the study population. Knowledge of the number and variability of markers can provide preliminary insight into the frequency of similar genotypes (Paetkau 2003: Fig. 1), but demographic isolation in small populations can lead to unexpectedly high frequencies of similar genotypes (D. Paetkau, unpublished observation).

I will use a recent brown bear study from the Yukon Territory, Canada (R. Maraj, University of Calgary, unpublished data), to illustrate the influence of these factors on efficiency and data quality. This study used 6 markers to analyze 370 hair samples, of which 335 produced sufficient data (solid genotypes for at least 5 markers, as judged by 2 experienced technicians) to allow assignment to 58 unique genotypes (i.e., individuals). A convenient feature of this study is that we have great confidence in the number of individuals identified; since 2MM-pairs are expected to outnumber 0MM-pairs (errors in which 2 individuals have identical multilocus genotypes) by a factor of approximately 100 (Paetkau 2003), the absence of 2MM-pairs in this study (Fig. 1) assures us that the number of individuals represented by the 335 samples was no greater than 58 (i.e., no 0MM-pairs).

I reanalyzed the Yukon file to see how much work would have been required at each phase of analysis if fewer markers had been used. I started with the most variable marker and added markers in order of descending variability (Fig. 2). The number of individuals identified would have been the same whether 3, 6, or (presumably) 15 markers were used. The total number of single-locus analyses required to achieve an accurate, finalized results file ranged from 1,173 (3 markers × 370 samples + 63 reanalyses) to over 5,600 (15 markers). The number of reanalyses would have been minimized (at 23 reactions) by the use of 4–6 markers, going up if the number of markers was either decreased (causing more chance similarities between individuals) or increased (causing more errors in initial genotypes).

The mismatch distribution for the Yukon study was strongly bimodal prior to selective reanalysis (Fig. 1). This is a convenient feature for illustration, or for statistical testing (McKelvey and Schwartz 2004), but it is not a precondition to detecting and correcting errors. Indeed, retrospective comparisons of what the results would have looked like using 4, 5, or 6 markers (Fig. 2) clearly illustrates that the use of 6 markers was overly conservative. One could even argue that the quality of the final estimate of abundance would have been improved had extra effort been directed toward increasing sample size—and thus precision—rather than toward the collection of genetic data from more markers than necessary. Such criticism must be tempered for 3 reasons: (1) the actual performance of a marker system can be evaluated only in retrospect, (2) erring on the side of excess power is preferred to the alternative, and (3) the use of too few markers removes the margin of comfort that allows samples to be retained when they are missing data for 1 marker. I would feel comfortable repeating this study with 5 markers but not with 4. The suggestion that 15 markers should have been used in this study (McKelvey and Schwartz 2004) is preposterous.

While I consider direct testing of data reproducibility through reanalysis of specified pairs of genotypes (Paetkau 2003) to be a more productive activity than indirect statistical scrutiny of preliminary mismatch distributions (McKelvey and Schwartz 2004), the mismatch distribution for a finalized dataset is an invaluable tool for defending a project. As long as 1MM-pairs are rare or absent, and the similar pairs of genotypes that are likely to include errors have been replicated according to formal guidelines, critics can find no logical basis to suspect either source of error to which DNA-based CMR studies are uniquely vulnerable; both insufficiently variable marker systems and genotyping error are expected to produce 1MM-pairs.

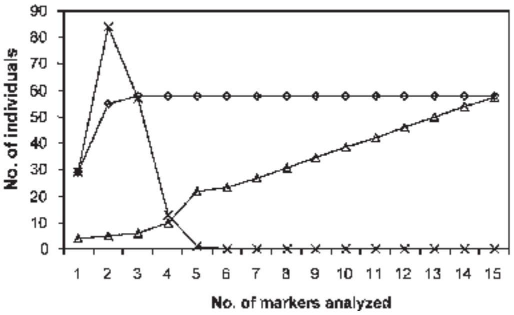

Fig. 2. Effect of changing the number of markers in a Yukon brown bear study that was originally analyzed using 6 markers. The number of unique genotypes (△) is lower than the presumed number of individuals (58) when using 1 or 2 markers, but stabilizes thereafter. The minimum number of reanalyses—under the conservative assumption that each error is detected and corrected with a single polymerase chain reaction—is the number of errors detected (△) plus the number of reanalyses required as a result of different individuals having genotypes that are similar enough to meet criteria for reanalysis (5; Paetkau 2003). The number of errors detected is taken from actual records for the first 6 markers, and then increased proportionally thereafter. This ignores the increase in error that would result from using less DNA per reaction as the number of markers increases.

The other support that is required to defend results is data images showing clear replications of each genotype underlying each 1MM- and 2MM-pair remaining in the final dataset. This replication can be through reanalysis or through the observation of identical multilocus genotypes in independent samples. For example, 312 of the 335 samples that were assigned to individuals in the Yukon project had multilocus genotypes that were replicated in at least 1 other sample. Field biologists should not view a lack of training in genetics as a reason to avoid raw data; good data images will be self-explanatory to any audience, and nothing will ensure conservative lab work like the knowledge that one's raw data may be scrutinized by others.

A remarkable chasm separates the positive outcomes that have been experienced in scores of applied DNA-based population studies (e.g., Sloane et al. 2000), most of which can no longer be published due to a lack of novelty, and the everexpanding cautionary commentary that dominates the literature, replete with dramatic bylines like "promise and pitfalls" (Mills et al. 2000), "cautions and guidelines" (Waits et al. 2001), and "look before you leap" (Taberlet et al. 1999). Unfortunately, the suggestions of McKelvey and Schwartz (2004) only serve to widen this gap, offering no demonstrable benefit over established methods at a cost that would threaten the practical utility of a well-established technique. In most studies of outbred populations, existing, scientifically rigorous data-replication protocols should achieve perfect accuracy of individual assignments using between 5 and 7 markers.

ACKNOWLEDGMENTS

I thank R. Maraj for allowing me to use her Yukon brown bear data and C. Strobeck for feedback on this manuscript.

LITERATURE CITED

BOULANGER, J., G. C. WHITE, B. N. MCLELLAN, J. WOODS, M. PROCTOR, AND S. HIMMER. 2003. A meta-analysis of grizzly bear DNA mark–recapture projects in British Columbia, Canada. Ursus 13:137–152.

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.

MCKELVEY, K. S., AND M. K. SCHWARTZ. 2004. Genetic errors associated with population estimation using non-invasive molecular tagging: problems and new solutions. Journal of Wildlife Management 68:439–448.

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. Ecological Applications 10:283–294.

MOWAT, G., D. C. HEARD, D. R. SEIP, K. G. POOLE, G. STENHOUSE, AND D. PAETKAU. 2004. Grizzly and black bear densities in the interior mountains of North America. Wildlife Biology 10: in press.

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

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

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

SLOANE, M. A., P. SUNNUCKS, D. ALPERS, B. BEHEREGARAY, AND A. C. TAYLOR. 2000. Highly reliable genetic identification of individual northern hairy-nosed wombats from single remotely collected hairs: a feasible censusing method. Molecular Ecology 9:1233–1240.

TABERLET, P., 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.

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

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.

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