Molecular Ecology (2003) **12**, 1375–1387 doi: 10.1046/j.1365-294X.2003.01820.x

An empirical exploration of data quality in DNA-based population inventories

D. PAETKAU
Wildlife Genetics International Inc., Box 274, Nelson, BC, Canada V1L 5P9

Abstract

I present data from 21 population inventory studies — 20 of them on bears — that relied on the noninvasive collection of hair, and review the methods that were used to prevent genetic errors in these studies. These methods were designed to simultaneously minimize errors (which can bias estimates of abundance) and per-sample analysis effort (which can reduce the precision of estimates by limiting sample size). A variety of approaches were used to probe the reliability of the empirical data, producing a mean, per-study estimate of no more than one undetected error in either direction (too few or too many individuals identified in the laboratory). For the type of samples considered here (plucked hair samples), the gain or loss of individuals in the laboratory can be reduced to a level that is inconsequential relative to the more universal sources of bias and imprecision that can affect mark–recapture studies, assuming that marker systems are selected according to stated guidelines, marginal samples are excluded at an early stage, similar pairs of genotypes are scrutinized, and laboratory work is performed with skill and care.

Keywords: abundance, bias, error, mark–recapture, microsatellite, noninvasive

Received 29 August 2002; revision received 18 December 2002; accepted 24 January 2003

Introduction

Noninvasive, DNA-based techniques have become a routine approach to population inventory for black (*Ursus americanus*) and brown bears (*U. arctos* including grizzlies; Taberlet *et al*. 1997; Woods *et al*. 1999; Mowat & Strobeck 2000; Poole *et al*. 2001; Boulanger *et al*. 2003) and have shown a great deal of promise, if not outright success, in many other species (Palsbøll *et al*. 1997; Reed *et al*. 1997; Kohn *et al*. 1999; Ernest *et al*. 2000; Sloane *et al*. 2000; Lucchini *et al*. 2002; Mowat & Paetkau 2002). These techniques have been used to produce dozens of formal, mark–recapture estimates of abundance, to the extent that routine studies, which lack the novelty required by the primary scientific literature, are generally published at the level of internal agency reports.

DNA-based inventory methods offer solutions to some sources of error associated with traditional methods (e.g. loss of 'marks' in mark–recapture studies; see discussion in Palsbøll 1999), and have no impact on other sources of error (e.g. violation of assumptions in mark–recapture models; Seber 1982), but they also introduce two new sources of potential error. First, too few individuals may be identified if the genetic markers being used lack the variability (power) necessary to produce unique genotypes for each individual that is sampled (Woods *et al*. 1999; Mills *et al*. 2000; Waits *et al*. 2001; the 'shadow effect'). Second, inconsistencies in the genotypes recorded for different samples taken from the same individual can result in the genetic identification of an excess of individuals (Taberlet *et al*. 1996; Gagneux *et al*. 1997; Taberlet *et al*. 1997; Goossens *et al*. 1998; Taberlet *et al*. 1999; Woods *et al*. 1999; Miller *et al*. 2002).

When Woods *et al*. (1999) performed the first large-scale, mark–recapture inventory of a brown bear population, they recognized these two sources of error and developed methods to control them. Specifically, they ensured that their marker system had adequate power by selecting the most variable markers from a larger set that had been tested on the study population, and they developed a match statistic that allowed for the fact that many of the individuals sampled were likely to be close relatives. They dealt with genotyping error by scrutinizing and selectively reanalysing pairs of samples whose genotypes were highly similar, but not identical, reasoning that such

Correspondence: D. Paetkau. Fax: 1250 3523567; E-mail: dpaetkau@wildlifegenetics.ca

near-matches might result from inconsistencies in the genotypes recorded for different samples taken from the same individual.

The error-prevention methods employed by Woods *et al*. (1999) were not codified, could not be validated by reference to empirical data, and were not described in detail. The lack of a published description of a refined, validated and detailed error-prevention protocol has facilitated the growth of a remarkably rich literature on the two sources of error that are unique to DNA-based inventories (above). Most recently, some noninvasive genetic surveys have fallen under such intense scrutiny (e.g. Stokstad 2002) that wildlife managers may become unwilling or unable to apply these methods unless more evidence is presented to support the capacity of such surveys to produce reliable data on a routine basis.

The cautionary literature that has grown up around DNA-based population inventory makes little reference to empirical data from the types of large-scale applications whose value is under question. However, at the same time that the academic community has been critically examining these methods for their potential, the wildlife management community — faced with urgent needs for information, and often lacking viable alternatives for collecting that information — has moved on to the widespread application of genetic methods, generating a large quantity of empirical data in the process.

Without denying the value of theoretical arguments and simulated data, there are good reasons to examine empirical evidence when assessing the prevalence of errors in DNA-based inventories. For example, the degree of relatedness has a large impact on the probability that a pair of individuals will have the same genotype (Woods *et al*. 1999), but one does not know the distribution of degrees of relatedness in one's study populations. A straightforward solution to this problem is to examine data from collections of known individuals drawn from similar populations, and to ask how often similar or identical genotypes are seen in practice (Waits *et al*. 2001). Similarly, while some sources of genotyping error have been clearly identified (Taberlet *et al*. 1996; Gagneux *et al*. 1997; Gossens *et al*. 1998), it remains unclear what the relative and absolute frequencies of different types of errors are with the type and quality of samples that are collected in inventory projects. Examining the errors that have been found in applied projects is an informative approach to clarifying these issues.

The purpose of this study is to describe a refined and codified version of Woods *et al*.'s (1999) error-prevention methods, and to use a large body of empirical data to estimate the capacity of the described methods to yield reliable results. The empirical data will be drawn from 21 recent hair-based inventories of black bear, brown bear and pine marten (*Martes americanus*) populations to which the described methods were applied under my supervision (Table 1). This work deals specifically with preventing errors, but the field and laboratory methods used in these studies are in keeping with published reports (Woods *et al*. 1999; Mowat & Strobeck 2000; Poole *et al*. 2001; Mowat & Paetkau 2002).

The similarity of pairs of genotypes (meaning multilocus, microsatellite genotypes unless otherwise specified) comes up repeatedly in the following discussion, so a convention will be adopted of referring to samples with identical genotypes for the relevant markers as zero-mismatch-pairs (0MM-pairs), while pairs of samples with genotypes that match at all but one marker will be called 1MM-pairs, and so on out to 3MM-pairs, which differ at three of the markers for which the samples hold data in common.

Methods

Selection of markers

It is only through retrospective analysis that specific guidelines, such as those suggested later in this study, can be developed. In order to generate data that could form the basis for such an analysis, we developed an *ad hoc* marker-selection guideline based on informal explorations of data. This guideline was that a suite of six microsatellite markers would be used when the mean expected heterozygosity (H_E) for those markers was between 0.7 and 0.8, and that a larger or smaller number of markers should be considered outside this range. Power requirements depend on the number of individuals for whom unique genotypes must be created, so some drift from these basic guidelines was allowed based on the expected number of individuals that were likely to be sampled in a given project (n). In some projects, n exceeded expectation, especially when data were added from new field seasons. In such cases, an additional microsatellite or gender marker was analysed retroactively for all samples.

In most projects, marker-selection was facilitated by the availability of approximately 30 samples from live-captured individuals. These samples were normally analysed using 12–15 microsatellite markers, which were then ranked by H_E. If the best five markers had $H_E > 0.8$ they were selected, but otherwise larger suites of markers were tested until a satisfactory set was identified. In the majority of projects, a six-locus suite of markers was selected (Table 1).

Initial genetic analysis

The first pass at genotyping used standard laboratory methods [polymerase chain reaction (PCR), electrophoresis] that have been described many times. In some studies, a species-specific marker (microsatellite or mtDNA) was

Table 1 Summary of 21 noninvasive, population inventory projects that employed genetic analysis of hair samples, sorted by the number and variability of the microsatellite markers that were used. The number of samples and number of genetically defined individuals are summarized by whether their genotypes were complete for all markers (*l*) or missing data for one (*l*-1) or two (*l*-2) markers; 'similar genotypes' refers to the number of pairs of genotypes in the final data set (after error-checking) that matched at all but one (1MM), two (2MM) or three (3MM) of the markers for which data were available

Project	Species	Contact for further information	Markers (number: names, most variable first)	H_E	No. of samples			No. of individuals			Similar genotypes		
					l	*l*-1	*l*-2	*l*	*l*-1	*l*-2	1MM	2MM	3MM
Owikeno*	*U. arctos*	Stefan Himmer, Arctos Wildlife, Hagensborg, BC	7: *G10B, G10X, G1A, G10L, G1D, G10C, G10J*	0.67	233	27	14	96	9	2	3	13	116
Slocan	*M. ameri.*	Garth Mowat, Aurora Wildlife, Crescent Valley, BC	6: *Ma1, Ma2, Ma8, Ma10, Ma18, Ma1*	0.67	113	21	1	80	8	0	1	38	238
Big Cypress	*U. ameri.*	Thomas Eason, Florida Fish & Wildlife, Tallahassee, FL	6: *G10L, G10B, G1D, G10J, G10P, G10H*	0.71	215	8	1	42	0	0	0	3	32
Ocala†	*U. ameri.*	Thomas Eason, Florida Fish & Wildlife, Tallahassee, FL	6: *G1A, G1D, G10H, MU59, G10B, MU50*	0.71	1695	21	0	223	0	0	6	82	766
Eglin†	*U. ameri.*	Thomas Eason, Florida Fish & Wildlife, Tallahassee, FL	6: *G10M, G1D, G10L, G10X, G10B, G1A*	0.72	266	17	2	63	0	0	1	9	72
Bowron	*U. arctos*	Garth Mowat, Aurora Wildlife, Crescent Valley, BC	6: *G1D, G10B, G10J, G1A, G10L, G10M*	0.72	188	5	0	53	0	0	1	2	23
St. John's	*U. ameri.*	Thomas Eason, Florida Fish & Wildlife, Tallahassee, FL	6: *MU59, G1A, G1D, G10H, MU50, G10B*	0.73	282	9	1	33	0	0	0	2	14
Osceola	*U. ameri.*	Thomas Eason, Florida Fish & Wildlife, Tallahassee, FL	6: *G10H, G10B, G1D, G1A, MU50, MU59*	0.74	224	5	1	64	2	0	0	3	47
Parsnip G	*U. arctos*	Garth Mowat, Aurora Wildlife, Crescent Valley, BC	6: *G10B, G1D, G1A, G10J, G10M, G10L*	0.75	864	26	11	267	4	0	4	59	628
Purcell	*U. arctos*	Michael Proctor, Birchdale Ecological, Kaslo, BC	6: *G10B, G1A, G1D, G10M, G10P, G10J*	0.76	164	3	1	29	0	0	0	1	6
New Jersey†	*U. ameri.*	Kelsey Burgess, Division of Fish & Game, Hampton, NJ	6: *G10P, G10L, MU50, G10J, MU59, G1D*	0.76	551	7	0	350	3	0	11	119	868
Apalachicola	*U. ameri.*	Thomas Eason, Florida Fish & Wildlife, Tallahassee, FL	6: *G10X, G1A, G10L, G1D, G10M, G10B*	0.76	244	3	0	40	0	0	0	2	22
SW MT	*U. ameri.*	Rick Mace, Department Fish Wildlife & Parks, Kalispell, MT	6: *G10J, G10L, G10H, MU59, G10P, G1D*	0.76	310	5	3	158	2	1	3	20	205
Taku G	*U. arctos*	Kimberley Heinemeyer, Round River Cons. Stu., SLC, UT	6: *MU59, G10B, G1D, G1A, G10X, G10J*	0.77	180	24	4	100	0	0	2	11	57
Taku B	*U. ameri.*	Kimberley Heinemeyer, Round River Cons. Stu., SLC, UT	6: *G10X, G10J, G1A, G10B, G1D, MU59*	0.78	63	6	0	45	0	0	0	2	6
Hoopa	*U. ameri.*	Mark Higley, Hoopa Tribal Forestry, Hoopa, CA	6: *G10J, MU59, G10H, G10L, G10M, G10P*	0.78	156	56	1	113	11	0	0	8	102
West Slope B*	*U. ameri.*	John Woods, West Slopes Bear Project, Revelstoke, BC	6: *G10X, G10L, G10C, G10B, G1D, G1A*	0.80	747	34	0	367	9	0	0	31	470
Minnesota	*U. ameri.*	Kristina Timmerman, U. of Minnesota, St Paul, MN	5: *G10M, G10L, G10C, G10B, G1A*	0.79	238	22	0	91	0	0	2	18	174
Oregon	*U. ameri.*	Dave Immell, Department Fish and Wildlife, Portland, OR	5: *G10P, G10J, G10M, G10H, G10*	0.81	83	1	0	39	0	0	0	2	29
Swan	*U. ameri.*	Rick Mace, Department Fish Wildlife & Parks, Kalispell, MT	5: *G10H, G10J, G10X, MU59, G10M*	0.83	353	8	0	199	2	0	1	29	447
Parsnip B	*U. ameri.*	Garth Mowat, Aurora Wildlife, Crescent Valley, BC	5: *C10L, G10J, G10H, G10X, G10C*	0.86	563	8	0	274	1	0	1	17	452
Mean or total				0.76	7732	316	40	2726	51	3	36	471	4774

*These projects include data from more than one field season, of which the first season's data were analysed under the supervision of Curtis Strobeck at the University of Alberta. Samples from all seasons were made available during the error-checking phase of the combined, multiseason database.

†Gender data were included in the actual analysis of individuals in these projects, reducing the number of similar genotypes relative to the numbers reported here, which are based on microsatellite data alone.

used to exclude samples from nontarget species (e.g. Woods *et al*. 1999), or a very robust marker (i.e. the microsatellite marker that amplified in the greatest proportion of samples in earlier projects) was run first to exclude poor quality samples with little prospect of producing multilocus genotypes. However, the usual procedure was to analyse each sample at all the chosen markers at the same time, and in the same lane of an automated sequencer. Single-locus genotypes that could not be scored with high confidence (below) were attempted a second time, using more genomic DNA in the PCR reaction (5 μL instead of 3 μL in the initial reaction).

In order to prevent contamination of genomic DNA samples with amplified DNA, we established an isolated facility for amplified DNA, and enforced strict rules governing the movement of people and materials between facilities. Routine monitoring for contamination was based on the use of extraction and PCR blanks, and records were maintained for these control samples to provide quality assurance.

In most projects, we were able to avoid completely the typing or writing of sample numbers (and thus typographical errors) by building the laboratory database around the field database. This allowed the direct printing of permanent, information-rich sample labels. Sample sheets for automated sequencers were created from files that were exported directly from the database. Similarly, the sample information present in sample sheets was used to allow direct import of genetic results back into the database, while confirming that the correct data were associated with the correct record.

We used a convention of recording results in which we had a high degree of confidence with 3-digit numbers (e.g. allele 182), but removed the first digit when we had lower confidence in the results (i.e. 82). All downstream computer analyses were designed to treat 2-digit numbers as missing data. Each worker became expert at knowing when 3-digit numbers were merited by starting with very conservative practices (i.e. assigning many 2-digit numbers) and keeping track of every inconsistency that was detected by reanalysis. This record of inconsistencies provided a feedback loop that allowed technicians to refine their skills. While the lack of discrete rules for assigning 3-digit allele scores is frustrating, it is difficult to codify the subtle visual cues, such as the relative intensity of dinucleotide shadow peaks, that an expert technician uses to assess data quality. Certainly signal intensity alone cannot form an adequate basis for decision making.

In the current projects (Table 1), alleles were scored automatically using category definitions in a Genotyper (Applied Biosystems) file, and then scrutinized separately by two people, at least one of whom was highly experienced, before being imported (not manually entered) into a database file.

Quality control

One approach to confirming genotypes is to repeat the analysis of all data many times (Taberlet *et al*. 1996; Gagneux *et al*. 1997). This 'multiple tubes' approach can be argued against on two grounds. First, increasing laboratory costs by any multiple would have a devastating impact on the practical utility of the methods. Second, there is a common misconception that the multiple tubes approach is equivalent to multiple 'standard' analyses. This is not the case, because the multiple tubes approach necessitates that the available DNA be diluted across a greater number of tubes, thus increasing the per-tube probability of encountering the errors and amplification failures that are associated with an inadequate quantity of DNA (Taberlet *et al*. 1996; Gagneux *et al*. 1997; Morin *et al*. 2001). Therefore, measuring the gain in quality that comes from using multiple tubes, compared to a single tube with more DNA, is more complex than it first appears.

We used an alternative approach, as follows: (1) DNA was extracted from up to 10 guard hair roots as available; (2) a large proportion of the sample (~1/3) was used in the first pass at genotyping, leaving a comfortable amount for selective reanalysis or downstream analyses (e.g. gender); (3) 2-digit allele scores were used liberally; (4) samples that performed poorly were culled (below), along with samples that showed evidence of three or more alleles, as expected when hairs from two different animals are combined; (5) similar genotypes were reanalysed selectively; and (6) all errors (i.e. any change to a recorded 3-digit number) were documented, and used to adjust practices at step 3. This protocol involves the aggressive consumption of minute, irreplaceable samples, so it requires the type of systemization and control of laboratory procedures that eliminates trivial errors, such as forgetting to add a reagent to a reaction, or adding the wrong sample to a tube.

The critical parameter that determined the amount of initial genotyping error allowed by our approach was the point at which poor samples are culled (step 4). We used a series of increasingly stringent thresholds for culling samples. During the first pass, any sample that did not produce high-confidence (3-digit) genotypes for three or more markers (four or more when using seven markers) was culled before even making an attempt to improve 2-digit or missing data. Following a round of reanalysis directed at improving data for mid-quality samples, the threshold was raised to a minimum of four (5) markers with high-confidence genotypes. It was our experience that most samples either produced complete data for all markers, or were culled; only a small number of samples were missing data for one or two markers (Table 1).

Because incomplete genotypes contain less information, they have a higher degree of similarity to other samples, especially other incomplete samples that are missing data

for different markers. This meant that they invariably needed to be reproduced, in part or in whole, under our error-checking regime (below). Following error-checking, samples that were missing data were considered on a case-by-case basis — with visual reference to all of the raw data available for the sample, and in discussion between two people — and any sign of inconsistency was taken as grounds for culling. Table 1 shows that 96% of samples which escaped culling had complete genotypes, while only 0.5% of such samples were missing data from more than a single marker. Note that, in practice, the sibling match statistic (Woods *et al*. 1999) did not inform decisions to retain samples, because the combination of marker variability and minimum requirements for complete data removed samples with excessively high ($P > 0.05$) values of this statistic.

The error-checking involved scrutiny of all 1MM- and 2MM-pairs of genotypes, as identified by an exhaustive, computer-based search. The first step of this scrutiny was a confirmation that the raw data were accurately reflected in the database, and any inconsistencies detected at this stage were corrected and recorded, and were not pursued further. If the genotypes were not reduced to 0MM-pairs by visual scrutiny, then the markers causing the mismatch were reamplified and reanalysed. When one or both genotypes involved in a 2MM-pair were observed in multiple samples, the reproduced genotypes (could be one or both members of the pair) were not reanalysed. This means that the protocol will not detect cases where identical, multilocus errors are replicated in multiple samples from the same individual.

This error-checking protocol is not designed to detect errors when only one sample has been collected from an individual. This is not a concern in the context of a population inventory project, because such a sample needs only to have a unique genotype, not an accurate genotype, to identify n accurately. If the same genetic data were being used for applications where the accuracy of genotypes was critical, such as studies of parentage, data quality could be maximized by restricting the analysis to genotypes (individuals) that had been observed in multiple samples, or allowance could be made for the presence of undetected errors.

In some cases data sets were combined, as when samples from different field seasons had been analysed in different laboratories (Table 1). This raised a special problem because, although each data set had been scrutinized for error, there was still a possibility that single samples from a given individual would be scored differently between the data sets. Therefore, it was necessary to repeat the error-checking prototcol, scrutinizing all pairs of similar genotypes between data sets. This approach mandates that laboratories share raw data and samples to facilitate data scrutiny and selective reanalysis. It also means that samples which have passed careful scrutiny in previous years may fall under fresh scrutiny as new samples, with potentially similar genotypes, are added to a data set.

Table 2 Numbers of errors detected by scrutiny or reanalysis of pairs of similar genotypes. Only those projects for which exhaustive records were kept are listed. Of 109 amplification errors (Amp.), 17 involved amplification of a false allele and 99 involved allelic-dropout (seven samples suffered both amplification problems in a single event)

		Number of errors		% Error	
Project	Samples	Scoring	Amp.	Scoring	Amp.
Apalachicola	247	1	11	0.4	4.5
Big Cypress	224	1	13	0.4	5.8
Bowron	193	2	6	1.0	3.1
Eglin	285	5	8	1.8	2.8
Hoopa	213	4	2	1.9	0.9
New Jersey	558	4	18	0.7	3.2
Ocala	1716	26	7	1.5	0.4
Oregon	84	10	4	11.9	4.8
Osceola	230	7	3	3.0	1.3
Parsnip B	571	27	3	4.7	0.5
Parsnip G	901	12	7	1.3	0.8
Purcell	168	3	5	1.8	3.0
St John's	292	4	13	1.4	4.5
SW MT	318	3	1	0.9	0.3
Swan	361	1	5	0.3	1.4
Taku B	69	0	0	0.0	0.0
Taku G	208	7	3	3.4	1.4
Total or mean	6638	117	109	1.8	1.6

Detailed records of the nature of every error detected were kept for the 17 most-recent projects (Table 2). Errors were broadly classified as scoring errors — in which the numbers recorded in the database were not consistent with the appearance of the raw data — and amplification errors — where the raw data had an appearance that would reasonably cause an experienced worker to record a genotype that was incorrect. Amplification errors (Taberlet *et al*. 1996; Gagneux *et al*. 1997) were broken down into allelic dropout (the amplification of only one allele in a heterozygote) and false amplification, although the former was far more common in our experience. In the 17 projects where we kept records, there were 222 1MM-pairs prior to, and 30 1MM-pairs following, error-checking. This means that the identification of 192 spurious individuals was circumvented. The number of 2MM-pairs decreased from 387 to 371.

In addition to the strict protocol outlined above, reanalysis was extended subjectively to other classes of samples, or repeated multiple times, whenever any level of discomfort remained about the reproducibility of genotypes. This final level of scrutiny was not codified, because the circumstances that can raise concerns vary greatly. The most common cause for concern was when the absence of a single

allele at one or more loci was the only difference between a pair of genotypes; a pattern that suggests allelic dropout. If convincingly strong and clear results were not obtained after several attempts at confirmation, the samples were culled from the data set. Altogether, we culled 1279 samples (13%) on the basis of quality, and 397 samples (4%) because they appeared mixed (three or more alleles reproduced at one or more loci), while we retained 8088 samples (83%).

Lack of power (too few individuals)

Errors happen

It is appropriate to state at the outset that errors do happen. The first documented error that I heard about affected the study described by Woods *et al*. (1999), where two individuals that were live-captured after completion of the study were found to have identical genotypes for the markers that were used in the inventory project. These two individuals, which differed at other microsatellite markers and had different genders, were subsequently tied back to two, widely separated geographical clusters of non-invasively collected hair samples (John Woods pers. comm.), demonstrating that the laboratory analysis in the original study had detected at least one too few individuals.

Distributions of similar genotypes

While anecdotal accounts can demonstrate that errors do occur, they provide little insight into the frequency of such errors. One cannot know the actual number of individuals with identical genotypes (0MM-pairs) in noninvasive inventory projects, but an examination of the distribution of similar genotypes in those projects can narrow the probable frequency of 0MM-pairs. Looking across all of the projects listed in Table 1, there were 4774 3MM-pairs, 471 2MM-pairs and 36 1MM-pairs. These figures include comparisons between individuals that were defined by incomplete genotypes, and are inflated by the exclusion of the gender data that were actually used in three projects (Table 1). The slope of this distribution, with order of magnitude differences at each step, suggests that no more than a handful of 0MM-pairs are expected to be present in these projects. Certainly it suggests that the mean number of such errors per project was less than one.

Retrospective probing for errors

Of the 21 inventory projects (Table 1), the Ocala study is arguably the most vulnerable to a lack of power, involving at least 223 individuals that must be differentiated using one of the least variable marker systems employed to date. In this study, 76 of the 223 genetically defined individuals were defined based on a genotype that was only observed in a single sample, and were therefore not candidates for errors caused by lack of power (such errors require the same genotype to be associated with samples from two individuals, and thus a minimum of two samples). We searched the remaining 147 individuals for errors by analysing gender for 418 of the 1640 samples that were tied to these individuals, biasing sample selection towards samples that were collected at different times and places. Because even the closest relatives have an ~50% chance of having different genders, this gender analysis should detect ~50% of cases where samples from two real individuals were lumped together in the same genetically defined individual. This test failed to identify any instances where samples with the same multilocus microsatellite genotype were scored as having different genders. While not an exhaustive test, these results demonstrated that errors caused by lack of power were not causing a dramatic underestimate of the number of individuals sampled, even in a project whose power was at the lower end of our allowable range (they also confirmed the reproducibility of the gender analysis).

Tests with known individuals

Unlike data from genetically defined individuals, genetic data from physically captured individuals can be examined directly for matching genotypes (e.g. Waits *et al*. 2001). I had access to five large collections of known individual bears (Table 3). In two cases, results were available for eight markers (Paetkau *et al*. 1998), so the analysis was performed by using data from the six, least-variable markers. For the other three collections, the only data that were available were from the markers that were used in the associated inventory projects (Table 1). Five-locus data were also selected from each of these collections, choosing the most variable markers available in an attempt to produce systems with similar variability to those used in the applied five-locus projects (Table 1).

Searches for matching genotypes in these five collections identified two cases where two known individuals had matching six-locus genotypes, and five cases where five-locus genotypes were held in common between individuals (Table 3). The single case (New Jersey, five loci) where more than one would-be-error (0MM-pair) was found in a single collection of individuals had more 1MM-pairs (48) than all 21 inventory projects combined (36: Table 1), demonstrating that the power present in this test system was much lower than the power of the marker systems that had been applied in practice (and illustrating how rapidly power falls off with variability). These collections of known individuals are similar in size to some of the larger inventory projects, indicating that even large projects can be expected to encounter very few 0MM-pairs when using our methods. Because power requirements are proportional

to the square of the number of individuals that need to be resolved [number of pairs = $n/2*(n-1)$, where n is the number of individuals sampled], these data provide strong evidence that the majority of the small ($n < 100$) projects listed in Table 1 are expected to include zero 0MM-pairs (errors).

Matches between known relatives

Reasoning that siblings are typically more difficult to differentiate than parent–offspring pairs, and making a rough guess at the number of first-order relatives in their study, Woods *et al*. (1999) decided that a conservative approach to match declarations was to use a sibling match probability, and to require that this match probability be below 5% for the data held in common between two samples before those samples could be declared to come from the same individual (note that larger studies, with their exponentially higher power requirements, would require more stringent critical values to maintain similar rates of error, although no formal method was proposed for setting such values).

An interesting feature of the Western Brooks Range study (Table 3) is that many first-order relatives were known (Craighead *et al*. 1995). When the 30 pairs of similar (0MM- + 1MM- + 2MM-pairs; Table 3) six-locus genotypes that were found in this group of individuals were sorted by known relationship, six of 26 known sibling pairs (23%) were found to have similar genotypes, while six of 90 known parent–offspring pairs (7%) had similar genotypes, and 18 of 10 615 pairs of unknown relationship (0.16%) were similar (this last value is inflated because there were undoubtedly unidentified pairs of first-order relatives among the pairs of unknown relationship).

These data provide an empirical demonstration that, while first-order relatives constitute the extreme minority of pairs of individuals, they may represent the majority of the challenge in terms of genetically distinguishing all individuals; a validation of Woods *et al*.'s (1999) approach of focusing on close relatives. To put it differently, a marker system that has sufficient power to resolve a small number of close relatives will have sufficient power to resolve a very large number of nonrelatives (Waits *et al*. 2001).

How variable must markers be?

The marker systems that we used were selected based on fairly subjective guidelines. However, it is now possible to examine the power of these marker systems, and re-evaluate these guidelines objectively. The frequencies of 0MM-pairs and 1MM-pairs are poor measures of power (the former because it is unknown, the latter because it is too small to be measured accurately), so the proportion of pairs of individuals that matched at all but two markers (number of 2MM-pairs/total number of pairs) was selected as a measure of power. In the interest of consistency, comparisons were limited to genetically defined individuals for which complete data were available for the selected markers, excluding the < 2% of individuals that were defined on incomplete genotypes (Table 1).

The comparison of marker variability (H_E) and power demonstrated that slight changes in variability have a large affect on power (Fig. 1). The question is, where is the threshold beyond which more markers should be employed? One would expect that the arguments for the location of this threshold would focus on match probabilities. However, while the preceding sections have demonstrated that we had sufficient power to differentiate the overwhelming majority of individuals, inadequate variability expressed itself in several projects through the error-checking aspect of the protocol, which dictates that 2MM-pairs be scrutinized for possible error. When too many 2MM-pairs

Table 3 The number of similar or identical (0MM) pairs of five- or six-locus genotypes in five collections of known (physically captured) individual bears. If these were DNA-based inventory projects, the number of 0MM-pairs would be unknown, and the number of individuals sampled would be underestimated by this number

				Pairs of genotypes		
	Markers (number: names)	H_E	n	0MM	1MM	2MM
W. Brooks Range	6: *G10B, G10X, G10C, G1A, G10M, G10L*	0.73	148	1	2	27
Richardson Mts	6: *G10M, G10B, G1A, G10C, G10X, G10L*	0.72	119	0	5	29
Ocala	6: *G1A, G1D, G10H, MU59, G10B, MU50*	0.71	126	0	2	36
New Jersey	6: *G10P, G10L, MU50, G10J, MU59, G1D*	0.76	292	1	10	85
Hoopa	6: *G10J, MU59, G10H, G10L, G10M, G10P*	0.78	98	0	0	1
W. Brooks Range	5: *G1D, G10P, G10B, G10X, G10C*	0.78	148	1	9	71
Richardson Mts	5: *G1D, G10P, G10M, G10B, G1A*	0.81	119	1	6	23
Ocala	5: *G1A, G1D, G10H, MU59, G10B*	0.73	126	0	14	98
New Jersey	5: *G10P, G10L, MU50, G10J, MU59*	0.77	292	3	48	225
Hoopa	5: G10J, MU59, G10H, G10L, G10M	0.80	98	0	0	12

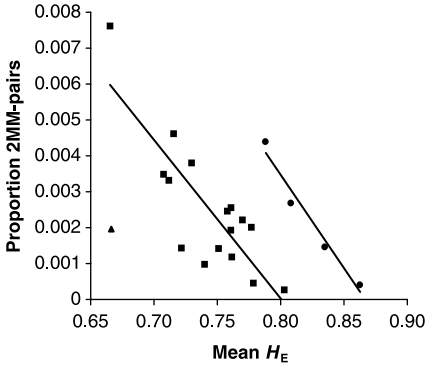

Fig. 1 Variability of the chosen marker system vs. power to resolve individuals, as quantified through the proportion of pairs of individuals with genotypes that match at all but two markers. The 21 projects (Table 1) are separated by whether they employed five (circles), six (squares) or seven (triangle) microsatellite markers.

are present in a project, the required scrutiny of 2MM-pairs can be a more onerous task than analysing an extra marker from the outset. In other words, a set of markers that keeps 2MM-pairs at a low enough frequency for our error-checking protocol to be practical will also keep 0MM-pairs (errors) to an acceptable level.

Based on this practical experience with error-checking, I suggest that small projects ($n < 100$) are feasible with a minimum H_E of 0.69 or 0.78, for six- or five-locus systems, respectively, but that these minimums should rise to an H_E of 0.75 or 0.83, respectively, for large projects ($200 < n < 400$). These thresholds correspond to expected 2MM-pair frequencies of approximately 0.005 for small projects, or 0.002 for large projects (Fig. 1). To illustrate the logic underlying these choices, consider a project that produces genotypes for samples from 200 individuals. In such a project, there would be 19 900 (200*199/2) pairs of individuals to compare and, assuming a marker system with a 2MM-pair frequency of 0.005, the estimated number of 2MM-pairs would be 99.5 (0.005*19 900). While scrutinizing 100, 2MM-pairs is not unmanageable, it is approaching the point where the work required to analyse an extra marker for all samples might be offset completely by the resulting reduction in the number of 2MM-pairs that required scrutiny.

The single seven-locus project (Fig. 1) had much higher power than would have been expected if we had used six markers in this population, but the rapid decline in power that is seen with falling variability suggests that many markers may be required to produce adequate power in populations with low variability. This can be illustrated by another empirical example. Eight-locus data are available for a small number ($n = 34$) of known, individual brown bears from Kodiak Island (Paetkau *et al*. 1998). While power requirements are comparatively trivial with a data set of this size, this insular population is characterized by low variability ($H_E = 0.27$). Comparisons of these 34 genotypes revealed eight 0MM-pairs and 127 1MM- or 2MM-pairs. In such a data set, the number of individuals sampled would be underestimated — a problem that would worsen exponentially with increased n — and the approach of scrutinizing pairs of similar genotypes on a case-by-case basis would be less efficient than simply reproducing the entire data set (a multiple tubes approach). There will clearly be cases where DNA-based population inventories are rendered impractical by low genetic variability.

An interesting implication of these recommendations for marker variability is that, because logistical concerns impose greater constraints on variability than match probabilities, the sibling match statistic (Woods *et al*. 1999) has been rendered functionally obsolete; it effectively never produces values above 5% when incomplete genotypes are culled according to our guidelines.

Genotyping errors (too many individuals)

Assumptions in the protocol

Our error-checking protocol relies on two assumptions: that errors are effectively never present at more than two markers in a single sample, and that the chance of making the same errors at multiple markers in multiple samples from the same individual is negligible. This amounts to an assumption that errors occur reasonably independently between markers — such that the rate of two-locus errors is the square of the rate of single-locus errors — and reasonably independently between samples — such that the probability of the same error(s) being made in two samples is the square of the probability of the error(s) being made in a single sample. Using the data that we recorded on errors, we can look at the independence of errors and, if we are convinced that the error-checking protocol has been effective in detecting most errors, learn something about the frequency and nature of the errors that are made.

This discussion ignores another source of error that could be caused by inaccurate genotyping, which is when samples are taken from individuals with very similar genotypes (1MM or 2MM), and when genotyping errors cause those genotypes to be recorded as identical (0MM). For such an error to occur, a pair of genotypes must be nearly identical to begin with (Table 1 shows that this is unlikely), the errors must affect the markers that differ between the two genotypes, and no other markers, and the errors must specifically convert the genotype of one individual to the genotype of the other individual, and not to one of the dozens of possible genotypes that typically exist. I consider this combination of events to be vanishingly unlikely, and ignore this type of error.

Table 4 Number of errors detected in the 17 inventory projects listed in Table 2, categorized by the number of markers affected, and whether the errors were unique, or were exactly replicated in another sample with the same ultimate genotype. Errors in the cells marked 'unk.' cannot be detected by our protocol, and go uncorrected

No. of markers	Unique	Replicated	Total
1	178	32 (14 events)	210
2	16	unk.	≥ 16
> 2	unk.	unk.	unk.
Total	≥ 194	≥ 32	≥ 226

Independence of errors between samples

We detected 210 cases where a single error was present in the initial genotype that was recorded for a sample, corresponding to about 3% of samples. If errors occur independently between samples, and we consider a case where an error has been made in one sample from a particular individual, then we would expect the probability of an identical error in a second sample from the same individual to be: the probability that the second sample contains an error (0.03) ÷ the number of markers where that error could take place (usually six) ÷ the number of different possible errors that could be made at the affected locus (a great many, although some errors, such as failure to amplify an allele, would be relatively common). A reasonable guess might be that an error would be exactly reproduced in 1/1000th of samples from the same individual as was affected by the initial error.

There were 178 unreplicated, single-locus errors in the studies that we documented (Table 4), and we had a mean of 3.1 samples per individual across these projects. With these numbers, one might not expect to see any cases where the same error was repeated in multiple samples from the same individual. Contrary to this expectation, we found that 32 of the samples with single-locus errors were involved in 14 cases where a particular erroneous genotype was replicated exactly in multiple samples from the same genetically defined individual. In other words, errors are not independent between multiple samples from the same individual.

Looking at the source of the replicated error, we see that 13 of 14 cases involved scoring errors, and that in 12 of these cases the errors affected two or more samples that were analysed on the same run. In these cases, a single technical problem went undetected, or a single oversight was made, that affected a series of adjacent samples in the same way. To illustrate, the two replicated errors that occurred in the Oregon project were caused by a malfunctioning climate control system that cooled the room to the point where the relative mobility of *G10H* was affected, causing peaks to fall outside the categories that we had defined for this marker using Genotyper software. When the technician scored this run manually, she shifted over by one allele [2 base pairs (bp)] on the first sample, and maintained that shift through a whole series of adjacent samples. As the relative values of adjacent genotypes appeared correct, the person who checked the scoring failed to notice the 2 bp shift in the absolute value of allele scores. While the circumstances that give rise to errors are many and varied, a perfect knowledge and execution of established laboratory protocols would have prevented these scoring errors.

Independence of errors between markers

Turning to the independence of errors between markers (within a single sample), we can once again use the observed rate of single-locus errors (0.03) to estimate that the chance of a sample being affected by two errors would be about 1 in 1000 (0.03^2). We had 6638 samples that were good enough to be assigned to individuals (Table 2), so we might have expected to see about six cases where data from two markers were incorrect for a single sample. The observed number of such cases was 16, suggesting less than perfect independence of errors between markers.

Once again, a look at the nature of the observed double errors provides a quick explanation for the apparent lack of independence. In most cases the double errors involved allelic dropout, and involved samples that failed to produce full genotypes on the first pass (i.e. marginal samples). This suggests that the rate of amplification error is a function of sample quality (Taberlet *et al*. 1996; Gagneux *et al*. 1997; Morin *et al*. 2001), and that the excess of double errors can be explained by a relatively higher rate of errors in low quality samples.

The observed rate of amplification error differed considerably between projects (Table 2). It is normally difficult to compare sample quality between projects, because study designs, climate and storage conditions vary between projects. However, these variables were relatively constant for a series of studies of neighbouring black bear populations in Florida. When the observed rate of amplification error was compared to a measure of sample quality (mean number of guard hair roots available for extraction), a clear trend was observed (Fig. 2); low sample quality increases the rate of error.

If sample quality affects rate of error, then an easy way to control error is to exclude low quality samples. Unfortunately, the relationship between the quantity of material that is available (e.g. number of guard hair roots) and the quality of genetic results is quite loose in our experience, presumably because of variation in sample condition. For example, a sample with abundant material could be of

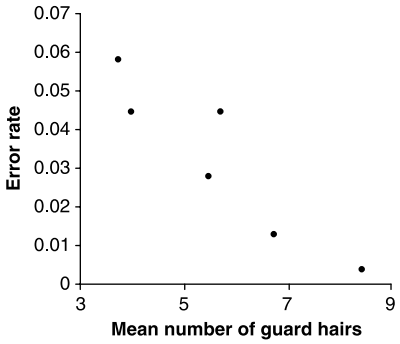

Fig. 2 Proportion of samples in which amplification errors were detected as a function of sample quality (quantity of material available for DNA extraction) in six projects run by the Florida Fish and Wildlife Conservation Commission (Table 1).

poor quality if it spent 2 weeks outdoors in wet conditions. Our experience is that the best way to exclude error-prone samples is to be very conservative when assigning high-confidence genotypes during the initial analysis, and to cull marginal samples according to our guidelines (see Methods). Note that prescreening methods, whether based on use of a single, robust microsatellite marker (see Methods), or whether based on quantitative PCR of another marker (Morin *et al*. 2001), do not affect the application of our protocol, although they do reduce the number of samples that are expected to be culled during multilocus genotyping, thereby increasing efficiency when a large proportion samples are of low quality.

How are we doing?

While the two previous sections have argued that the main assumptions underlying the error-checking protocol are violated, for practical purposes we are interested only in whether the degree of these violations is sufficient to result in data sets that contain a significant number of undetected errors. Our error-checking protocol is blind to certain classes of errors, and the conclusion that one draws about the reliability of our results comes down to one's estimate of the values in the cells marked 'unk' in Table 4. As noted previously, confirmation of 1MM-pairs averted the definition of 192 spurious individuals, compared to the 16 such errors that were avoided by checking a much larger number of 2MM-pairs. A nearly identical ratio was observed between unique errors and replicated errors. A direct extrapolation from these trends would produce an estimate of three undetected errors in total. However, recognizing the risk of extrapolation, and wishing to err on the side of caution, a safe estimate is that undetected errors averaged less than one per project (i.e. the sum of cells marked 'unk' in Table 4 is < 17). This level of residual genotyping error is probably trivial compared to more mundane sources of error, such as mislabelling sample envelopes in the field, and certainly won't have a large impact on mark–recapture estimates of abundance.

The main limitation with the conclusion of low genotyping error rates is that it is specific to the type of samples collected in the projects under consideration (Table 1). Unlike the conclusions concerning resolving power, which are based on match probabilities, genotyping error rates are expected to vary with the quality of DNA that one manages to collect, such that the reliability of the described methods cannot be considered proven for samples such as scat or shed hair.

Redundancy

The preceding sections have focused on data from observed errors, but a different line of reasoning that provides a general sense of error rates takes advantage of sampling redundancy. Given the variability of the marker systems that we use, and the low rate at which replicated errors are observed, a match in genotypes between two samples can be taken as strong evidence that those two samples are from the same individual, and that neither genotype contains errors. This reasoning is of little value in projects where many samples have unique genotypes, but in some projects there are dozens of samples with the same genotype, giving tremendous confidence that the observed genotype is correct (or at least reproducible).

The best example of this is the Ocala project, in which 99% (1573 of 1590) of noninvasively collected hair samples (that were not culled) produced genotypes that were observed in one or more other samples from that population. When such a large proportion of the data have been replicated in this way, it becomes hard to imagine a scenario in which these data are riddled with genotyping error. While this logic cannot be applied in projects with less redundant sampling, the methods that are applied are consistent across projects, so the reassurance provided though this reasoning can be extrapolated to other projects, at least to the extent that those projects share the relatively high sample quality (low rate of amplification error; Table 2) observed in the Ocala study.

Signals of error: heterozygote deficit

While I have argued that it is possible to keep genotyping errors to a low frequency, it is useful to identify signs that may be indicative of undetected genotyping errors. Because allelic dropout is the most common single class of error that we encounter, differences between H_O (observed heterozygosity) and H_E can be used to identify a widespread failure to amplify both alleles in heterozygous genotypes. Looking at the 21 inventory projects (Table 1), mean

H_E was 0.7560; effectively indistinguishable from the H_O of 0.7565. As allelic dropout was detected in 1.6% of samples (Table 2), there would have been a larger difference between H_O and H_E prior to error-checking.

Another example comes from examining the mean H_O of the single-locus genotypes that were scored with high confidence (3-digit numbers) for samples that were ultimately culled. H_O in these rejected samples was 0.51; fully 25% below expectation. This large deficit of heterozygotes, apparently affecting one in three heterozygous genotypes, indicates that the overall rate of genotyping error is vastly higher than the values reported in Table 2, but that most of the errors occur in samples that are culled before error-checking takes place. This supports the earlier argument that error rates depend heavily on sample quality, and emphasizes the value of removing low-quality samples from a project at the earliest possible stage.

An obvious limitation of heterozygote deficit is that the deficit becomes large, and thus statistically detectable, only when the error rate is high. Another difficulty with using heterozygote deficit to detect allelic dropout is that the same indicator results from the presence of nonamplifying (null) alleles (albeit, only at the affected loci; Callen *et al*. 1993), and from Wahlund effects (Wahlund 1928). Wahlund effects may be seen when study areas are much larger than the dispersal capacity of the study animal, such that the assumption of random mating, which underlies the calculation of H_E, is violated. Similarly, there is little reason for inventory projects to eschew markers with a few null alleles, because the accuracy of genotypes is far less important to inventory projects than the reproducibility of genotypes. Notwithstanding these limitations, checking for heterozygote deficit is merited because it requires little effort, and can provide reassurance that allelic dropout is not rampant.

Other signals of genotyping error

Another indirect indicator of genotyping problems is the ratio of 2MM- to 1MM-pairs, which we found to be > 10 after error-checking. Because most errors occur in isolation, they create 1MM-pairs. In many of our studies, 1MM-pairs outnumbered 2MM-pairs prior to error-checking, providing a clear and routinely observable indicator of a data set with quality problems.

A final resource for confirming data quality is for the participants in the project who have an expert knowledge of the ecology of the study species, or of the field data or the mark–recapture analysis, to ensure that the genetic data are consistent with this knowledge. For example, do the 'capture' locations and times for a given genetically defined individual exceed the reasonable movement capacity of the study species? Or do the gender results obtained in the lab correspond to known genders from live-captured animals whenever the microsatellite genotype indicates a match with such an animal?

Another example that has been observed on several occasions (John Boulanger pers. comm.) is that data sets which have not been heavily scrutinized (i.e. contain a number of errors) are flagged for closure violation by goodness-of-fit testing of the mark–recapture model. Closure violation normally refers to individuals who are not present in the study area during all capture sessions, and thus have a capture probability of zero in some sessions (e.g. Boulanger & McLellan 2001), but this phenomenon is mimicked when a spurious individual is defined based on a genotyping error, and thus has no chance of recapture. When these data sets were subsequently scrutinized and errors were removed, the apparent closure violation disappeared.

Training and supervision

An examination of Table 2 highlights an area of major concern; that rates of scoring error (human error) can fluctuate wildly between projects. In several cases, the frequency of scoring errors rose above 2% of samples, and when this happened it could always be traced to insufficient training and supervision of new technical staff. It is my opinion that the potential for DNA-based inventory projects to go badly wrong does not lie in the technical details that form the bulk of this manuscript, but in underestimating the amount of skill, knowledge and diligence that supervisors must transfer to their staff to enable those workers to generate data files containing thousands of error-free allele scores.

Effective use of resources

I have argued that the number of errors present in the projects listed in Table 1 is very small, but probably not zero. An important point to remember is that many estimates of abundance are compromised by very large confidence intervals (e.g. Boulanger *et al*. in press); sometimes of a magnitude that is similar to that of the estimate itself. Faced with this reality, and with finite budgets, there is a limit to the amount of energy which can responsibly be invested in searching for errors. At some point, the reductions in bias that can be brought about through reduced error are trivial in comparison to the increase in precision that could be achieved through diverting energy into increasing the sample size. If residual errors are present at the frequency that I am suggesting, then our ultimate understanding of the number of animals in the study populations will be improved more through the analysis of additional samples (i.e. increasing precision) than through searching for that last undetected error that may or may not be lurking in a given data set.

Conclusion

Noninvasive, DNA-based population inventories have a proven record of providing high quality information to wildlife managers. With the development of sample collection techniques for a greater range of species I expect the use of these methods to increase substantially, particularly in situations where physical capture is difficult or undesirable. With DNA samples of the quality that were available to the studies listed in Table 1, and with adherence to a few basic principles, there is little reason to fear the gain or loss of individuals in the laboratory. The relevant principles include thorough training and dedicated supervision of laboratory workers, selection of powerful marker systems, systemization and automation of laboratory methods, early culling of low quality samples and confirmation of similar pairs of genotypes. I believe that the approach described here strikes an appropriate balance between maximizing the quantity and the quality of data produced.

Acknowledgements

I thank the individuals and agencies listed in Table 1 for involving me in their projects, and for allowing me to use their data. I am grateful for the skill and care that Monique Paradon, Kristopher Sabourin, Jennifer Bonneville, Kelly Stalker, Leni Neumeier, Nicole Thomas, Jennifer Weldon, Sarah Waterhouse and July Lenek brought to bear in performing the laboratory work for the projects listed in Table 1, and for assistance in compiling data for this manuscript.

Addendum

Subsequent to the preparation of this manuscript, a sample with allelic-dropout at three markers was discovered in a project that is not described here. As a result, we have added a category of samples to those scrutinized for errors: members of 3MM-pairs with unique genotypes (found in just one sample) in which at least half of single-locus genotypes are homozygous.

References

Boulanger J, McLellan B (2001) Closure violation in DNA-based mark–recapture estimation of grizzly bear populations. *Canadian Journal of Zoology*, **79**, 642–651.

Boulanger J, White GC, McLellan BN *et al.* (2003) A meta-analysis of grizzly bear DNA mark–recapture projects in British Columbia, Canada. *Ursus*, **13**, 137–152.

Callen DF, Thompson AD, Shen Y *et al.* (1993) Incidence and origin of 'null' alleles in the $(AC)_n$ microsatellite markers. *American Journal of Human Genetics*, **52**, 922–927.

Craighead L, Paetkau D, Reynolds HV, Vyse ER, Strobeck C (1995) Microsatellite analysis of paternity and reproduction in Arctic grizzly bears. *Journal of Heredity*, **86**, 255–261.

Ernest HB, Penedo MCT, May B, Syvanen M, Boyce WM (2000) Molecular tracking of mountain lions in the Yosemite Valley region in California: genetic analysis using microsatellites and faecal DNA. *Molecular Ecology*, **9**, 433–441.

Gagneux P, Boesch C, Woodruff DS (1997) Microsatellite scoring errors associated with noninvasive genotyping based on nuclear DNA amplified from shed hair. *Molecular Ecology*, **6**, 861–868.

Goossens B, Waits LP, Taberlet P (1998) Plucked hair samples as a source of DNA: reliability of dinucleotide microsatellite genotyping. *Molecular Ecology*, **7**, 1237–1241.

Kohn MH, York EC, Kamradt DA, Haught G, Sauvajot RM, Wayne RK (1999) Estimating population size by genotyping faeces. *Proceedings of the Royal Society of London, Series B*, **266**, 657–663.

Lucchini V, Fabbri E, Marucco F, Ricci S, Biotani L, Randi E (2002) Noninvasive molecular tracking of colonizing wolf (*Canis lupus*) packs in the western Italian Alps. *Molecular Ecology*, **11**, 857–868.

Miller C, Joyce P, Waits LP (2002) Assessing allelic dropout and genotype reliability using maximum likelihood. *Genetics*, **160**, 357–366.

Mills LS, Citta JJ, Lair KP, Schwartz MK, Tallmon DA (2000) Estimating animal abundance using noninvasive DNA sampling: promise and pitfalls. *Ecological Applications*, **10**, 283–294.

Morin PA, Chambers KE, Boesch C, Vigilant L (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, Paetkau D (2002) Estimating marten *Martes americana* population size using hair capture and genetic tagging. *Wildlife Biology*, **8**, 201–209.

Mowat G, Strobeck C (2000) Estimating population size of grizzly bears using hair capture, DNA profiling, and mark–recapture analysis. *Journal of Wildlife Management*, **64**, 183–193.

Paetkau D, Waits LP, Clarkson PL *et al.* (1998) Variation in genetic diversity across the range of North American brown bears. *Conservation Biology*, **12**, 418–429.

Palsbøll PJ (1999) Genetic tagging: contemporary molecular ecology. *Biological Journal of the Linnean Society*, **68**, 3–22.

Palsbøll PJ, Allen J, Bérubé M *et al.* (1997) Genetic tagging of humpback whales. *Nature*, **388**, 767–769.

Poole KG, Mowat G, Fear DA (2001) DNA-based population estimate for grizzly bears *Ursus arctos* in northeastern British Columbia, Canada. *Wildlife Biology*, **7**, 105–115.

Reed JZ, Tollit DJ, Thompson PM, Amos W (1997) Molecular scatology: the use of molecular genetic analysis to assign species, sex and individual identity to seal faeces. *Molecular Ecology*, **6**, 225–234.

Seber GAF (1982) *The Estimation of Animal Abundance*. Charles Griffin, London.

Sloane MA, Sunnucks P, Alpers D, Beheregaray B, Taylor AC (2000) Highly reliable genetic identification of individual northern hairy-nosed wombats from single remotely collected hairs: a feasible censusing methods. *Molecular Ecology*, **9**, 1233–1240.

Stokstad E (2002) Fur flies over charges of misconduct. *Science*, **295**, 250–251.

Taberlet P, Camarra J-J, Griffin S *et al.* (1997) Noninvasive genetic tracking of the endangered Pyrenean brown bear population. *Molecular Ecology*, **6**, 869–876.

Taberlet P, Griffin S, Goossens B *et al.* (1996) Reliable genotyping of samples with very low DNA quantities using PCR. *Nucleic Acids Research*, **24**, 3189–3194.

Taberlet P, Waits LP, Luikart G (1999) Noninvasive genetic sampling: look before you leap. *Trends in Ecology and Evolution*, **14**, 323–327.

Wahlund S (1928) Zusammensetzung von populationen und korrelationserscheinungen vom standpunkt der vererbungslehre aus betrachtet. *Hereditas*, **11**, 65–106.

Waits LP, Luikart G, Taberlet P (2001) Estimating the probability of identity among genotypes in natural populations: cautions and guidelines. *Molecular Ecology*, **10**, 249–256.

Woods JC, Paetkau D, Lewis D *et al*. (1999) Genetic tagging of free-ranging black and brown bears. *Wildlife Society Bulletin*, **27**, 616–627.

The author is President (and Molecular Artificer) at a private genetics laboratory that sits at the interface between academic research and routine, postacademic application. His customers include wildlife researchers, managers and breeders, and analysis of noninvasive population inventories is one of the main services provided by his group.