A simulation test of the effectiveness of several methods for error-checking non-invasive genetic data

David A. Roon
University of Idaho, roon8505@uidaho.edu

Lisette P. Waits
University of Idaho, lwaits@uidaho.edu

Katherine C. Kendall
USGS-NRMSC Science Center, kkendall@usgs.gov

A simulation test of the effectiveness of several methods for error-checking non-invasive genetic data

David A. Roon[1,*], Lisette P. Waits[1] and Katherine C. Kendall[2]

[1] Department of Fish and Wildlife, University of Idaho, Moscow, Idaho 83844-1136, USA
[2] USGS-NRMSC Science Center, c/o Glacier National Park, West Glacier, Montana, 59936-0128, USA

Abstract

Non-invasive genetic sampling (NGS) is becoming a popular tool for population estimation. However, multiple NGS studies have demonstrated that polymerase chain reaction (PCR) genotyping errors can bias demographic estimates. These errors can be detected by comprehensive data filters such as the multiple-tubes approach, but this approach is expensive and time consuming as it requires three to eight PCR replicates per locus. Thus, researchers have attempted to correct PCR errors in NGS datasets using non-comprehensive error checking methods, but these approaches have not been evaluated for reliability. We simulated NGS studies with and without PCR error and 'filtered' datasets using non-comprehensive approaches derived from published studies and calculated mark–recapture estimates using CAPTURE. In the absence of data-filtering, simulated error resulted in serious inflations in CAPTURE estimates; some estimates exceeded N by $\geq 200\%$. When data filters were used, CAPTURE estimate reliability varied with per-locus error ($E\mu$). At $E\mu = 0.01$, CAPTURE estimates from filtered data displayed $< 5\%$ deviance from error-free estimates. When $E\mu$ was 0.05 or 0.09, some CAPTURE estimates from filtered data displayed biases in excess of 10%. Biases were positive at high sampling intensities; negative biases were observed at low sampling intensities. We caution researchers against using non-comprehensive data filters in NGS studies, unless they can achieve baseline per-locus error rates below 0.05 and, ideally, near 0.01. However, we suggest that data filters can be combined with careful technique and thoughtful NGS study design to yield accurate demographic information.

INTRODUCTION

Accurate demographic information is essential to the effective conservation of wild populations (Lande, 1988; Chao, 1989). For some species and populations, robust estimates of abundance, density and population trend can be obtained through the use of traditional sampling methods, such as mark–recapture (White *et al.*, 1982), line-transect sampling (Burnham, Anderson & Laake, 1980) and visual surveys (Eberhardt, Knight & Blanchard, 1986; Eberhardt & Knight, 1996; Anderson *et al.*, 1998). However, for many populations of conservation concern, these approaches have serious limitations (Seber, 1982; Lancia, Nichols & Pollock, 1994). Threatened populations are often low in density, with individuals that are difficult to capture (Shaffer, 1981; Ratti & Garton, 1994). Also, handling poses serious risks to many organisms (Blanchard & Knight, 1995).

Over the past decade, non-invasive genetic sampling (NGS) has emerged as an important tool in demographic estimation (Taberlet, Waits & Luikart, 1999; Piggott & Taylor, 2003). Studies that combine NGS with micro-satellite genotyping can circumvent factors that limit traditional census methods (Morin & Woodruff, 1996; Paxinos *et al.*, 1997; Taberlet *et al.*, 1999; Woods *et al.*, 1999). However, demographic studies assume that 'marks' can be identified repeatedly and without ambiguity (White *et al.*, 1982). Mark–recapture analysis is vulnerable to violations of this assumption; loss of a mark or creation of a false mark can skew apparent recapture rates, significantly biasing population estimates. In NGS mark–recapture studies, every unique microsatellite genotype is treated as a distinct mark, representing one capture of an individual and repeat observations of unique genotypes are treated as repeat captures of the same individual. Both assumptions can be violated in NGS studies. Two distinct individuals in a population may carry the same genotype by chance. The likelihood of this occurrence, or probability of identity (P_{ID}), is contingent on allelic diversity, number of loci analysed, and the percentage and degree of related individuals in a population (Waits, Luikart & Taberlet, 2001). As P_{ID} increases, so does the likelihood of incorrectly obtaining matching genotypes for two distinct individuals. Mills *et al.* (2000) demonstrated that this 'shadow effect' could lead to a negative bias in mark–recapture estimates, but they also demonstrated that P_{ID}-statistics could be used to predict the number of loci

*All correspondence to: David A. Roon. Tel: (208) 885-7323; Fax: (208) 885-9080; E-mail: roon8505@uidaho.edu.

necessary to reduce the occurrence false matches to less than 1%.

In contrast, errors during polymerase chain reaction (PCR) amplification can result in the creation of false genotypes, adding spurious marks to capture histories (Taberlet *et al.*, 1999). Two types of amplification error are observed during microsatellite genotyping: allelic dropout and false amplification (Taberlet *et al.*, 1996). Allelic dropout is the most common error in field studies where error rates have been quantified (Gerloff *et al.*, 1995; Goossens, Waits & Taberlet, 1998; Goossens *et al.*, 2000; Launhardt *et al.*, 2000; Frantz *et al.*, 2003). Waits & Leberg (2000) found that a per-locus error rate of 0.05 could cause an inflation of up to 200% in mark–recapture estimates derived using the program CAPTURE (Otis *et al.*, 1978) if the data are not filtered to remove errors. Goossens *et al.* (1998) and Gagneux, Boesch & Woodruff (1997) observed per-locus error rates of as high as 0.14 and 0.35, respectively, in microsatellites amplified from hair extracts. Estimates for mean per locus error rates from NGS studies using faeces have ranged from 0.05 to 0.29 (Kohn *et al.*, 1999; Bayes *et al.*, 2000; Goossens *et al.*, 2000; Murphy *et al.*, 2002).

In early NGS studies, protocols for detecting genetic error were not cited (Morin & Woodruff, 1992) or were restricted to cross-comparisons between NGS genotypes and genotypes from tissue extractions (Reed *et al.*, 1997). Over time, various methods for detecting errors have been proposed. The multiple-tubes method of Taberlet *et al.* (1996) is statistically well supported, but may be cost-prohibitive for large-scale studies. In response to the higher error rates observed while amplifying materials such as shed hair and scat, more efficient variations of the muliple-tubes approach have been proposed (Miller, Joyce & Waits, 2002; Frantz *et al.*, 2003), but these methods require a minimum of two to three PCRs per sample per locus. Less comprehensive error-checking mechanisms, or 'filters', have been utilised in several NGS studies (Woods *et al.*, 1999; Eggert, Eggert & Woodruff, 2003; Paetkau, 2003). These methods generally are predicated on detecting either or both of the following patterns:

- Unique genotypes that appear only once within a data set
- Pairs of unique genotypes that closely resemble one another, such as differing at a single locus.

Criteria for which genotypes are re-analysed, which loci are re-amplified and how many re-amplifications are carried out vary from study to study (Kohn *et al.*, 1999; Woods *et al.*, 1999; Poole, Mowat & Fear, 2001; Eggert *et al.*, 2003). The effectiveness of these non-comprehensive filtering methods for reducing bias in mark–recapture estimates and census counts has not been evaluated.

In the present study, our specific goals were:

(i) To simulate microsatellite-based NGS mark–recapture studies under a range of population sizes, sampling intensities and per-locus error rates.
(ii) To determine the impact of amplification error on population estimates in the absence of any mitigation.
(iii) To evaluate the effectiveness of non-comprehensive data 'filters'.

MATERIALS AND METHODS

Background

Our simulation framework is based on data collected for the Glacier Bear DNA Project (GBP), a large-scale NGS survey of bear populations using hair-snagging methods within the Northern Continental Divide Ecosystem (NCDE: Kendall, Waits & Schirokauer, 1997). Within our simulations, we used NCDE brown bear allele frequencies for the loci G1A, G10B, G10C, G10L, G10M and G10P (Paetkau & Strobeck, 1994; Paetkau *et al.*, 1995). Global heterozygosity was 0.71, with allele counts ranging from five (G10L) to 12 (G10B).

We used these data because of the recent prevalence of NGS studies in bear research (Woods *et al.*, 1996, 1999; Taberlet *et al.*, 1997; Mowat & Strobeck, 2000; Poole *et al.*, 2001) and to evaluate the potential bias in our GBP dataset using different non-comprehensive data filters. In choosing loci, we sought to maximise the allelic diversity, the ease and accuracy of amplification and the potential for multiplexing during electrophoresis (Taberlet *et al.*, 1999). Our chief objective was to minimise the number of loci used, while setting the population-wide sibling-probability of identity (P_{ID-Sib}), the upper limit of the possible ranges of P_{ID} in a population, to below 0.01 (Waits *et al.*, 2001). Selection of loci for other NGS studies will follow similar guidelines (Waits & Leberg, 2000).

Model assumptions

Our model assumed closed and panmictic populations. Although many NGS study populations will be subject to emigration and immigration, closure assumptions may be reasonable if studies are carried out over a brief temporal scale (White *et al.*, 1982; Lancia *et al.*, 1994). Even if closure is violated, abundance estimates may still be biologically meaningful if the study population is considered to be an extended super-population (White, 1996; Boulanger & McLellan, 2001). Panmixia may be violated in the field; however, many of the organisms studied using NGS display high levels of gene flow and lack highly structured mating systems (Morin & Woodruff, 1996; Reed *et al.*, 1997; Kohn *et al.*, 1999; Woods *et al.*, 1999).

Model framework

For each simulation, populations of individuals with six-locus genotypes were generated. These were:

- A source population. Six-locus genotypes for this population were generated randomly, based on allele frequencies observed in the GBP data set.
- A series of *T* sample populations.

Simulation of sample populations

For some simulations, six-locus genotypes (samples) for these populations were drawn with replacement from the source population, with uniform sampling probability across all source population genotypes. However, in many NGS studies, sample deposition will be clustered. Bears surveyed using hair-snagging studies may leave multiple samples at a single trapping location; other organisms may defaecate in spatially correlated patterns. These patterns may reduce the overall 'coverage' (percentage of the superpopulation sampled: Boulanger et al., 2002) for a given level of S and T (see below). Thus, for a second set of simulations, each randomly chosen genotype was represented by $y(i)$ copies of that genotype, where $y(i)$ was treated as a Poisson random variable with mean value $(\lambda) = 3$. We used $\lambda = 3$ based on results from the GBP.

Simulations were programmed in C++. For each simulation, parameterisation included 'N', the size of the source population. We used $N = 50$, 100 and 500. Other parameters were 'T', the number of sampling occasions, 'S', the number of samples drawn at each sampling occasion and '$E\mu$', the mean per-locus error rate. Simulated values for S ranged from 25–200 in increments of 25. For the number of trapping occasions (T) we used values from 3–9 inclusive.

Simulation of amplification error

As samples were drawn from the source population, errors were generated stochastically. In the GBP data set, error rates varied considerably between samples; variance in error rates between loci was much less. The quality of template sources will vary widely in NGS studies, although quantitative PCR can mitigate this variation to some degree (Morin et al., 2001). Thus, each sample drawn from the source population was assigned a per-locus error rate (E_s) specific to that sample and held constant across loci. Most surveys using NGS have noted little or no variation in error rates between loci (although see Lucchini et al., 2002; Frantz et al., 2003). For each sample, E_s was generated from a beta-binomial distribution with expected value $E\mu$. We modelled four per-locus $E\mu$ values; 0.005 (cited in Paetkau, 2003), 0.01 (estimated for the Greater Glacier Bear DNA Project), 0.05 (observed from faecal DNA in Kohn et al. (1999)) and 0.09 (analogous to rates observed in multiple scat studies and surveys using shed hair: Gagneux et al., 1997; Bayes et al., 2000; Murphy et al., 2002).

For low quality DNA extracts, some loci may fail to amplify. The decision to drop or include 'incomplete' samples from datasets may have ramifications for overall data reliability. In our simulations, each locus within a sample genotype had a probability E_s of failing to amplify (equal to the probability of amplifying with error). Samples were incorporated into capture histories if:

1. At least four loci amplified successfully and
2. The sample's multilocus genotype appeared only once in the dataset, or, given the subset of amplified loci,

matched only one unique genotype pattern in the overall sample population.

For each locus, errors were generated with probability E_s. Based on patterns observed in the Glacier Bear Project, as well as other studies (Gagneux et al., 1997, Goossens et al., 1998), 20% of errors were modelled as false amplifications. For each false amplification event, allele sizes were increased or decreased in two base-pair increments with equal probability. Microsatellite mutational dynamics may diverge from stepwise model assumptions (Ohta & Kimura, 1973; Goldstein et al., 1999). However, at the rates of error simulated in our study, alternative mutational models would have minimal impact on our results. Remaining errors (80%) were modelled as allelic dropout.

We implicitly define 'error rate' as the chance (prior to any error-checking) of having an incorrect result at a given locus. Thus, our per-locus error rate encompasses amplification error, electropherogram scoring error, sample mix-up and other human error (Paetkau, 2003). NGS genetic analysis is likely to occur in labs with a range of experience with microsatellite genotyping; our definition encompasses this variation.

Data filters: 'Singles'

Based on published NGS studies (Woods et al., 1999; Poole et al., 2001), we modelled a 'Singles' filter as follows. Samples with matching genotypes were considered reliable. If a sample's genotype appeared only once in a data set, that sample was 're-amplified'. The number of simulated re-amplifications was determined using a 'three strikes' criterion. If the initial re-amplification failed to match a sample's original genotype, one additional amplification was carried out for all loci. If two out of three of these results matched, the consensus genotype was assigned to the sample. If all three amplifications diverged, the sample was eliminated from the data set. For each re-amplification, amplification failures and errors were generated according to the original sample-specific error rate (E_S).

Data filters: 'Mismatch'

In simulating a mismatch-type filter, pairwise comparisons were carried out between all unique genotypes within the sample population (note – a unique genotype might be carried by multiple samples in a given data set). For each pair of unique genotypes, loci were flagged for re-amplification under the following conditions:

1. Two unique genotypes diverged at only one locus (single-locus mismatch), or
2. The two unique genotypes diverged at two loci (two-locus mismatch), but in a manner that could be attributed to allelic dropout.

A unique genotype might compare closely to more than one alternative genotype under these criteria; thus,
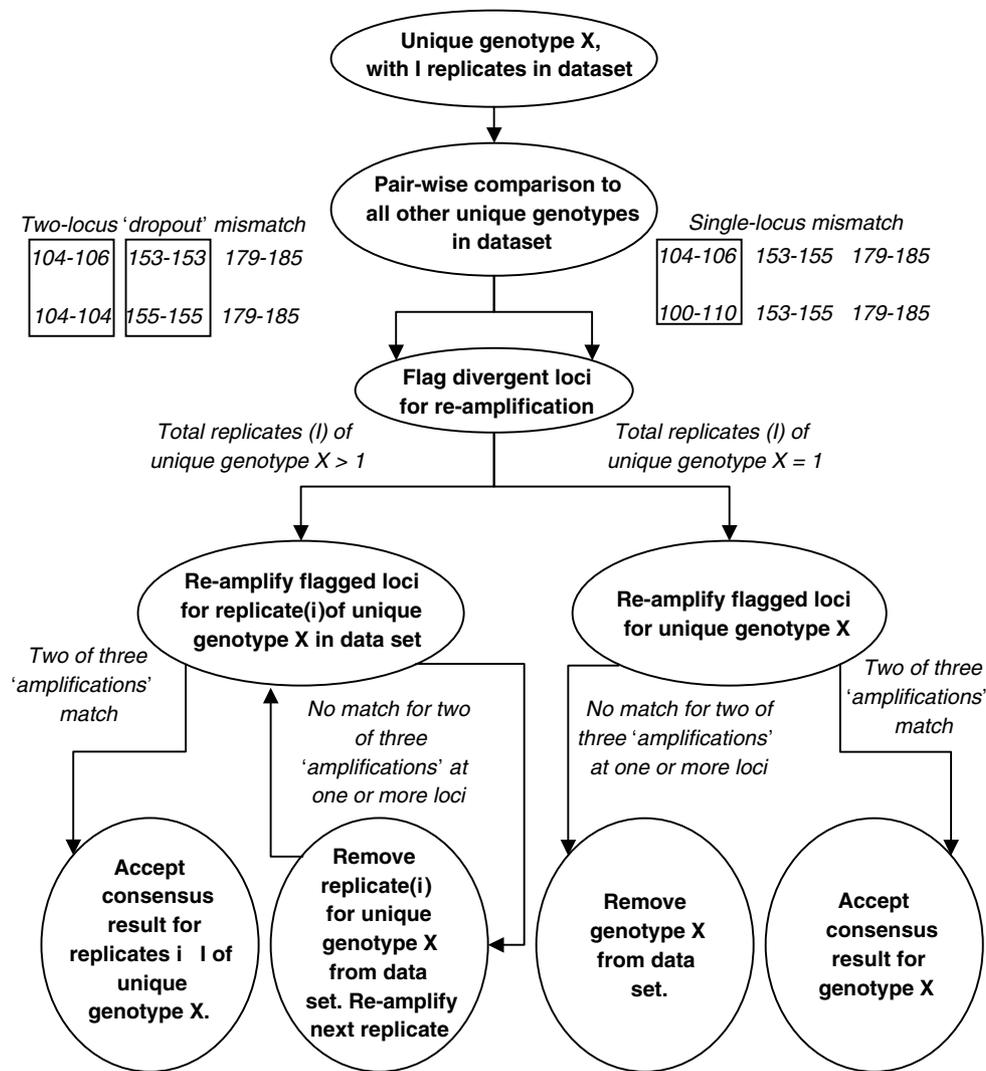
Fig. 1. Flow chart diagram for the 'Mismatch' filter as simulated in this study.

unique genotypes often had multiple loci flagged for 're-amplification'.

These criteria were used in the GBP. The re-amplification of single-locus mismatch genotypes is also a centerpiece of the screening protocols suggested by Paetkau (2003); similar approaches were used in earlier studies (Woods *et al*., 1999; Poole *et al*., 2001).

If a unique genotype appeared more than once in a given data set, flagged loci would be re-amplified for the first replicate of that unique genotype. Amplifications of flagged loci were used to validate the genotype replicate according to the 'three strikes' criteria described for the Singles approach. If two out of three 're-amplifications' matched for flagged loci, we accepted those results for all replicates of that particular unique genotype (Fig. 1). Flagged loci were re-drawn from the original source genotype. Amplification errors and failed amplifications were generated using the sample-specific error rate (E_S).

In cases where a replicate of a unique genotype was rejected as untrustworthy, only that specific replicate was eliminated from analysis; the next replicate of the unique genotype was then re-amplified (Fig. 1).

Data analysis

The Singles and Mismatch filters were tested independently and in combination with Mismatch followed by Singles. Subsequent to error-generation and 'filtering', each genotype in the sample data set was assigned two marks; one based on its source genotype, the other based on its simulated multi-locus genotype (potentially impacted by error). Based on these marks, data sets were condensed into histories for analysis in CAPTURE (Otis *et al*., 1978). Replicated captures of unique genotypes within sampling occasions (T) were collapsed into single capture events. Of the available CAPTURE estimators, the null model (H_0) was used; we also analysed datasets using Mh-Jackknife, a heterogeneity estimator that allows for variation in individual capture probabilities (White *et al*., 1982). We acknowledge that H_0 can be negatively biased in some field situations where the equal capture assumption

is violated (White *et al.*, 1982); however, Waits & Leberg (2000) have observed severe positive biases while using heterogeneity estimators to analyse data sets containing simulated error.

For each combined set of parameters we ran 200 simulations. We recorded the range and variability of CAPTURE estimates, using both the Null and Jackknife estimators, as well as the number of unique genotypes for each simulation, an equivalent to a minimum census count. When comparing estimates derived from populations with different *N* values, we calculated Percentage Relative Bias (PRB), equal to $100 \times (X - N)/N$, where *X* is the estimate mean and *N* is the source population size (White *et al.*, 1982). We recorded the confidence intervals generated within CAPTURE for each simulation. We assessed the frequency at which different estimators were chosen by CAPTURE's model selection function, under different simulation scenarios. To evaluate filter efficiency, we recorded the number of amplification events required by each method. We also estimated the number of amplifications required to apply the multiple tubes method of Taberlet *et al.* (1996), according to the following formula:

$$R = (H_0 \times \mathrm{MCC} \times 3) + ((1 - H_0) \times \mathrm{MCC} \times 8)$$

where *R* is the estimated number of amplifications necessary to validate a given data set, MCC is the minimum census count, or simulated count of unique genotypes for the dataset and H_0 is the observed heterozygosity (0.71). Three and eight are the minimum number of amplifications needed to validate heterozygous and homozygous loci, respectively, using the Taberlet *et al.* (1996) approach.

RESULTS

Population size

Initial results (Fig. 2) suggested that changes in population size had minimal impact on the level of bias in CAPTURE estimates based on error-impacted data. Increases in *N* did result in increased precision for CAPTURE estimates; however, this was true for all data sets, with and without simulated error. Based on these results, we present our remaining data with *N* held at 100.

In presenting our results, we simplify as follows: either *S* or *T* are held constant ($S = 100$, $T = 5$) while the other sampling parameters (*S*, *T*, or $E\mu$) are varied. Observed trends for other iterations are adequately reflected within these parameter limits.

CAPTURE estimates: Poisson versus non-Poisson sampling

In simulations where the sampling of unique genotypes was modelled as a Poisson process ($\lambda = 3$), CAPTURE estimates based on data sets without error were biased below *N* for all simulations with $S < 100$ (see Fig. 4). These negative biases were driven by inherently low capture probabilities for Poisson-based data sets (Fig. 3; White *et al.* 1982). For example, at $S = 25$ and $T = 5$, the mean capture probability (as calculated within CAPTURE) was 0.0873 for Poisson data sets, while the corresponding value for uniform data sets was 0.2258. Poisson sampling also reduced the overall coverage, or percentage of population sampled (Boulanger *et al.* 2002), for all iterations (Fig. 3). Interestingly, at $S > 100$, error impacted
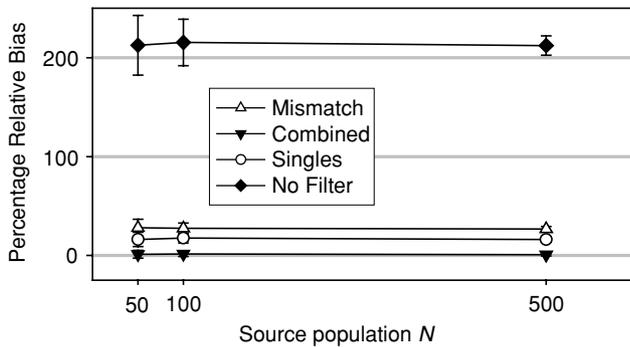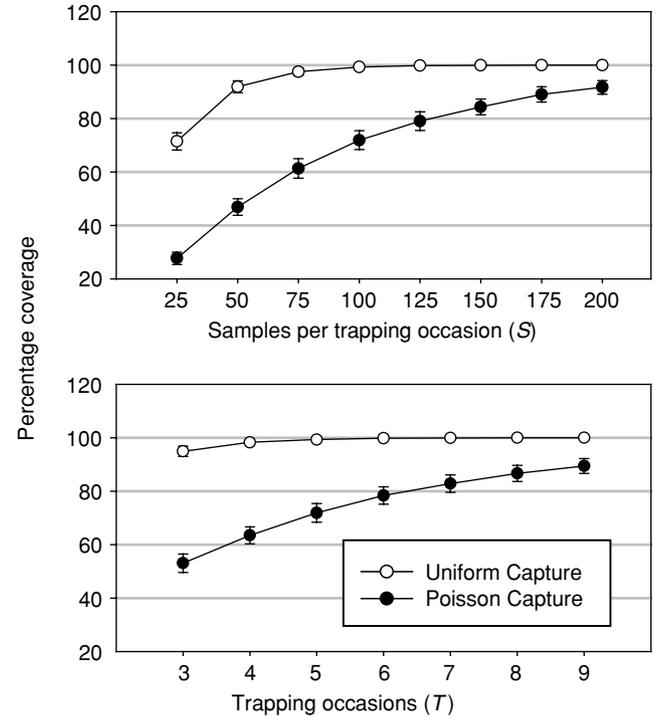

Fig. 2. Mean Percentage Relative Bias (PRB) in CAPTURE population estimates from non-invasive genetic sampling (NGS) data sets as a function of filter method and changes in the source population size (*N*). Simulation parameters: number of samples per occasion ($S = 1N$; number of trapping occasions (T) = 5; per locus error ($E\mu$) = 0.09; 200 iterations per data point. Trends were similar at other levels of *T* and $E\mu$ (data not shown).

Fig. 3. Mean coverage (percentage of *N* sampled) for data sets with no error as a function of increases in the number of samples per trapping occasion (*S*), the number of trapping occasions (*T*) and sampling distribution (Poisson or uniform). Default simulation parameters: trapping occasions (*T*) = 5; source population size (*N*) = 100; samples per trapping occasion (*S*) = 100; 200 iterations per data point.
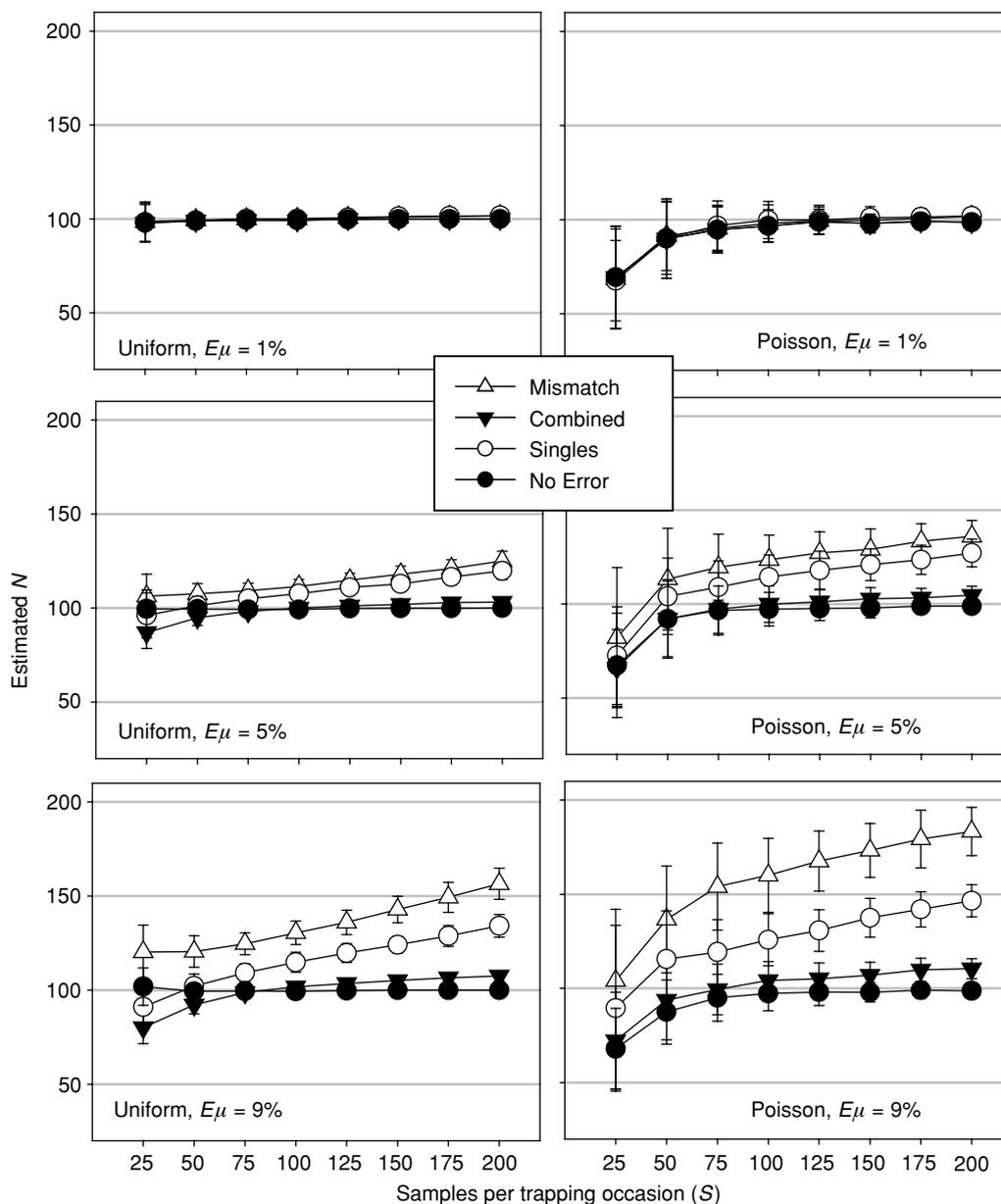
Fig. 4. Mean CAPTURE population estimates for simulated error-impacted non-invasive genetic sampling (NGS) data, as a function of filter method, changes in the number of samples per trapping occasion (S), the mean per-locus error rate ($E\mu$) and the sampling distribution (Poisson or uniform). Simulation parameters: number of trapping occasions (T) = 5; source population size (N) = 100; 200 iterations per data point.

CAPTURE estimates for Poisson data sets displayed positive biases of a greater magnitude than those observed for uniform sampling data sets (Fig. 4). In presenting our results, we will emphasise data from our non-Poisson simulations, but give point comparisons to Poisson-based results.

CAPTURE estimates: heterogeneity estimator

For data sets impacted by error, analyses using Mh-Jackknife resulted in increased levels of estimate bias (Fig. 5). For example, at $S = 200$, $T = 5$, $E\mu = 0.05$ (a comparatively high sampling intensity), mean CAPTURE estimates were inflated above N by 76% (uniform

sampling: Fig. 5) for data sets analysed using Mh-Jackknife, while corresponding estimates using the Null estimator were inflated by 25% (Fig. 5). These differences were evident even at very low $E\mu$ values; at $S = 200$, $T = 5$ and $E\mu = 0.005$, estimates were inflated by 1.5% (Jackknife) and 0% (Fig. 5). Thus, in our subsequent presentation of results, we focus on results using the more conservative Null estimator.

CAPTURE estimates: $E\mu = 0.005$

For filtered data, there was no discernible difference between CAPTURE estimates at $E\mu = 0.005$ and $E\mu = 0.01$ (i.e. estimate means were within one percentage point).
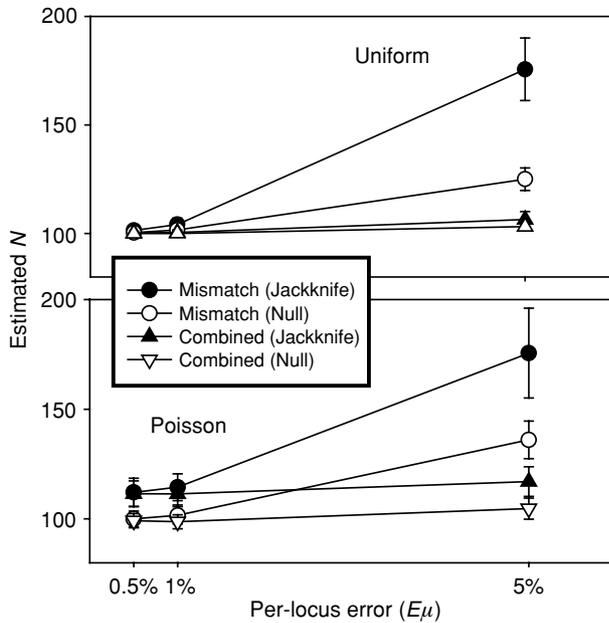
Fig. 5. Mean CAPTURE population estimates for simulated error-impacted non-invasive genetic sampling (NGS) data as a function of filter method (combined approach or mismatch method), per-locus error ($E\mu$), CAPTURE estimator choice (Mh-Jackknife or the Null estimator) and the sampling distribution (Uniform versus Poisson). Simulation parameters: number of trapping occasions (T) = 5; $E\mu = 0.005$, $E\mu = 0.01$ and $E\mu = 0.05$; samples per trapping occasion (S) = 200.
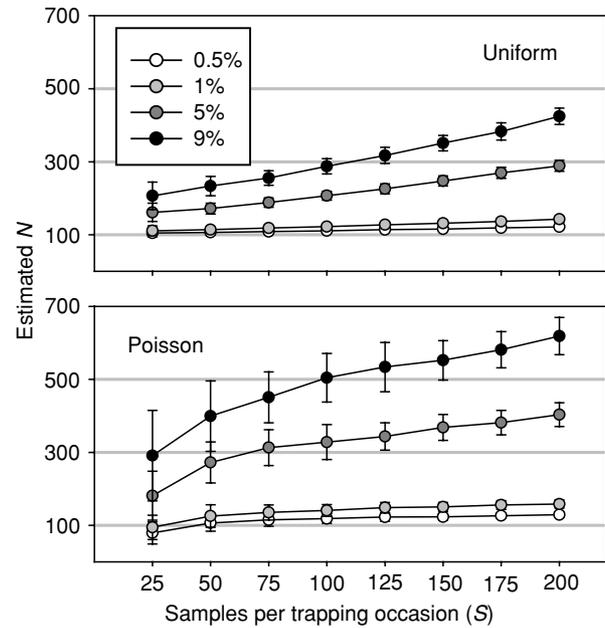
Fig. 6. Mean CAPTURE population estimates for simulated error-impacted non-invasive genetic sampling (NGS) data with no filtering, as a function of changes in the number of samples per trapping occasion (S), the mean per-locus error rate ($E\mu$) and the sampling distribution (Poisson versus uniform). Simulation parameters: number of trapping occasions (T) = 5; source population size (N) = 100; 200 iterations per data point.

In our subsequent figures and discussion, we emphasise data from $E\mu = 0.01$, $E\mu = 0.05$ and $E\mu = 0.09$.

CAPTURE Estimates: unfiltered and error-free data

For data sets with unmitigated error, mean CAPTURE estimates exceeded N by factors ranging from 5% ($S = 25$, $T = 5$, $E\mu = 0.005$) to 324% ($S = 200$, $T = 5$, $E\mu = 0.09$: Fig. 6) for uniform sampling data sets. Under Poisson sampling, biases were even more severe, with a maximum of 518% observed at $S = 200$, $T = 5$, $E\mu = 0.09$ (Fig. 6). These biases increased with sampling pressure, as a function of S and/or T.

In the absence of simulated genetic error, CAPTURE estimates converged to N (100) in tandem with increases in T or S (Figs 4 & 7).

CAPTURE estimates: filtered data

At $E\mu = 0.005$–0.01, mean CAPTURE estimates for error-impacted data sets did not deviate from data sets without error by more than 2% ($T = 5$, $S = 200$, $E\mu = 0.09$, uniform and Poisson sampling, Mismatch filter: Fig. 4). Data sets filtered using Singles or the Combined approach produced CAPTURE estimates that were indistinguishable from error-free estimates.

At $E\mu = 0.05$ and $E\mu = 0.09$, mean CAPTURE estimates based on Mismatch-filtered data were consistently biased for data without error (Figs 4 & 7). For uniform sampling data, these estimates consistently exceeded N (100). Mean biases ranged from 6% ($T = 5$, $S = 25$) to 25% ($T = 5$, $S = 200$) at $E\mu = 0.05$; the corresponding means ranged from 20% to a maximum of 57% at $E\mu = 0.09$ (Figs 4 & 7). For Poisson-based data, we observed a negative bias at $T = 5$, $S = 25$ and $E\mu = 0.05$; all other biases were positive, with a maximum bias above N of 83% ($T = 5$, $S = 200$, $E\mu = 0.09$: Figs 4 & 7).

The implications of using the Singles or Combined filters varied at $E\mu = 0.05$ and $E\mu = 0.09$, according to the levels of S and T and depending on whether sampling was uniform or Poisson. For uniform sampling data, the Singles or Combined filters produced CAPTURE estimates that were both biased above N (at higher levels of S and T) and biased below N (at lower levels of S and T: Figs 4 & 7). When the combined filter approach was used, observed biases at $E\mu = 0.05$ were generally less than 5%, the exception being $S = 25$, $T = 5$, where we observed a negative bias of 13.1% (Fig. 4). Biases were more severe at $E\mu = 0.09$; we observed mean negative biases of -20% below N for the Combined approach at $T = 5$, $S = 25$ and $E\mu = 0.09$ (Fig. 4). A similar result (bias $= -8\%$) was observed for Singles data (Fig. 4). In contrast, we recorded mean biases of 34% above N for Singles data at $T = 5$, $S = 200$ and $E\mu = 0.09$ (Fig. 4). Positive biases for the Combined approach were not as extreme; still, we observed a maximum of 8% above N at $T = 5$, $S = 200$, $E\mu = 0.09$ (Fig. 4).
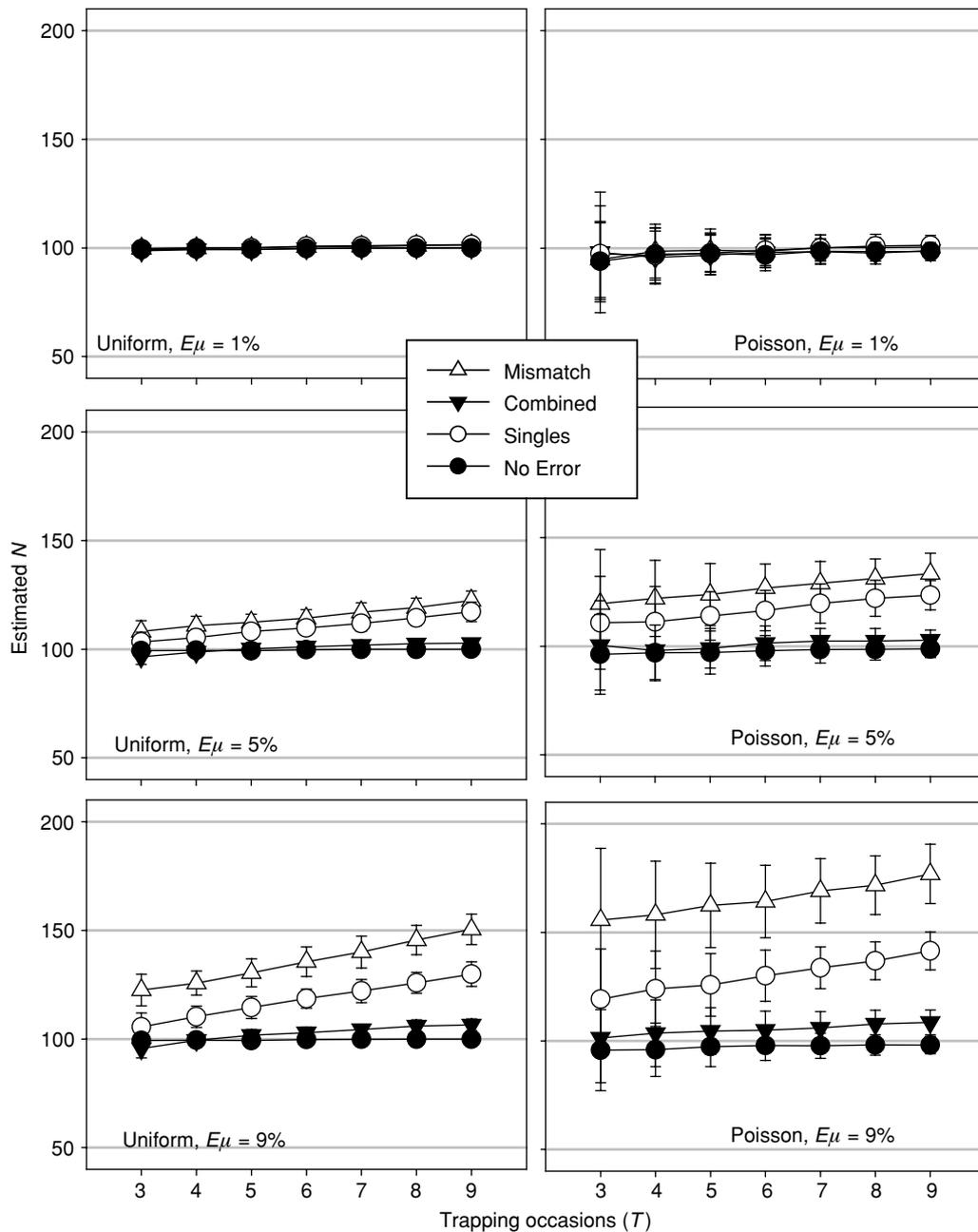
Fig. 7. Mean CAPTURE population estimates for simulated error-impacted non-invasive genetic sampling (NGS) data, as a function of filter method, changes in the number of trapping occasions (T), the mean per-locus error rate ($E\mu$) and the sampling distribution (Poisson versus uniform). Simulation parameters: samples per trapping occasion (S) = 100; source population size (N) = 100; 200 iterations per data point.

For Poisson data sets, error-impacted CAPTURE estimates were never biased below estimates without error. For Poisson data sets filtered with Singles, a maximum bias above N (46.5%) was observed at $T = 5$, $S = 200$ and $E\mu = 0.09$ (Fig. 4). The same parameter iteration resulted in a bias above N of 10.5% when we used the Combined approach (Fig. 4).

Minimum census counts

The total number of unique genotypes observed in an NGS study serves as a minimum census count ('MCC').

The degree to which error-impacted MCCs diverged from non-impacted MCCs varied with changes in S, T and $E\mu$ (data not shown). These patterns of variability were similar to patterns observed for CAPTURE estimates. At $E\mu = 0.005$–0.01, mean MCCs for error-filtered data diverged from MCCs derived from data with no error by less than 2%. In contrast, biases were present at $E\mu = 0.05$ and $E\mu = 0.09$; magnitude varied with changes in $E\mu$, S and T. There were some simulations at $E\mu = 0.05$ and $E\mu = 0.09$ for which MCCs were inflated above N; maximum positive biases of 56% (uniform sampling) and 47% (Poisson sampling) were observed at $T = 5$, $S = 150$, $E\mu = 0.09$ (Mismatch filter).
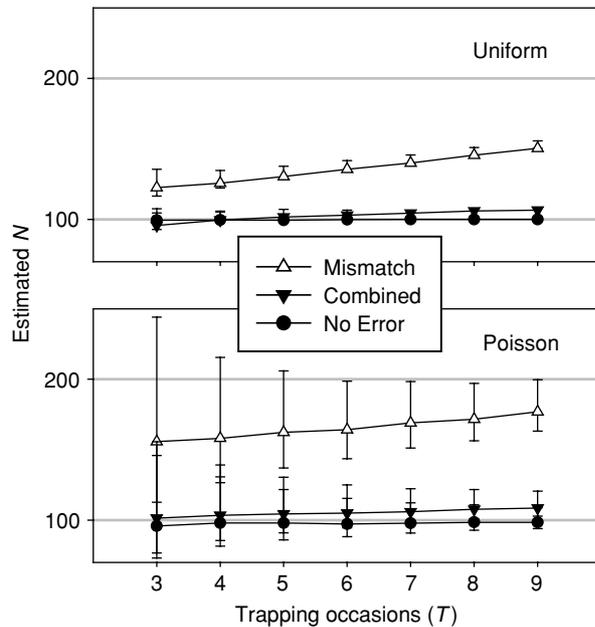
Fig. 8. Changes in mean confidence intervals calculated within CAPTURE for filtered and unfiltered data as a function of the number of trapping occasions (T) and the sampling distribution. Simulation parameters: samples per occasion (S) = 100; source population size (N) = 100; mean per-locus error ($E\mu$) = 0.09; 200 iterations per data point. Trends were similar when T was held constant and S increased from 25 to 200.

Confidence intervals and Model Selection

As calculated within CAPTURE, 95% confidence intervals for error-impacted estimates often did not contain the true N. The percentage of intervals where the true N lay below the -95% boundary generally increased as $E\mu$, S and T increased (Fig. 8.)

For data that were not impacted by simulated error, M(o) (the null model) was chosen most frequently under CAPTURE's model selection criteria (Fig. 9). Alternative selections occurred with increasing frequency as S and T increased, but percentage selection of alternative models never exceeded 23% ($T = 5$, $S = 150$: Fig. 9). When simulated error was present and uncorrected, model selection rapidly shifted to M(h) (heterogeneity) or other alternatives. Model selection patterns for data filtered using the Singles or the Combined approach were generally similar to those for data without error, particularly at $E\mu = 0.01$. However, at $E\mu = 0.05$ and higher, there was a false heterogeneity signal in the Singles data and, at higher levels of S and T, within the combined data (Fig. 9). Data filtered using the Mismatch approach yielded a high percentage selection of M(h) and other 'incorrect' models, even at $E\mu = 0.01$ (Fig. 9). We present these data not as a critique of the model selection function in CAPTURE, but as a means of highlighting error-based changes to our simulated data structure.

Amplification effort

The number of simulated amplifications needed to validate data sets increased with $E\mu$, S and T (Fig. 10).

The Mismatch method required the smallest number of simulated amplifications. Three times as many 're-amplifications' were required for Singles, while only twice as many were required for the Combined approach. For all iterations, the multiple-tubes method (Taberlet *et al.*, 1996) would have required at least twice as many amplifications as the Combined approach, even under 'best case' assumptions (Fig. 10).

Probability of identity issues

In a simulation test where 1000 genotypes were produced ($E\mu = 0.09$) and then filtered using the Combined approach, the average probability of identity (sibs) was 0.0074, with one false match. Thus, false identity was not a confounding factor in our results.

DISCUSSION

Evaluating population size and trend is a key component in the development of informed conservation and management programs (Lande, 1988; Chao, 1989). NGS studies present an exciting alternative approach to population estimation. Genotype artifacts generated through PCR amplification errors can be a significant hazard in the appropriate application of these methods (Taberlet *et al.*, 1999; Waits & Leberg, 2000). Our results demonstrate that non-comprehensive data filtering approaches can measurably improve the accuracy of NGS demographic assessments. However, our results also document the limitations of these methods. Most significantly, none of the tested methods produced consistently unbiased demographic estimates at per-locus $E\mu$ values of 0.05 and 0.09 across all sampling parameters.

Estimate bias

At $E\mu = 0.05$ and 0.09, filtered data often yielded biased CAPTURE estimates. The direction and magnitude of these biases varied with changes in the number of trapping occasions (T), the percentage of the population sampled, S_I, and the error-checking method used. Of these parameters, only T is fixed during study design. S_I is a function of samples per occasion (S), the true population size (N) and capture patterns (such as the Poisson sampling process modelled in this study) and is therefore unknown. The presence of positive biases in CAPTURE estimates at high levels of S, T and $E\mu$ is not surprising. However, negative biases were also observed at low levels of S and T and $E\mu \geq 0.05$. For these iterations, stringent application of the Singles or Combined approaches resulted in a disproportionate rejection of samples that represented valid one-time-only capture events. Researchers struggling with amplification error in NGS studies might be tempted to pursue a 'zero-tolerance' policy towards samples with inconsistent amplification histories. Our results suggest that hyper-conservative error-checking approaches can
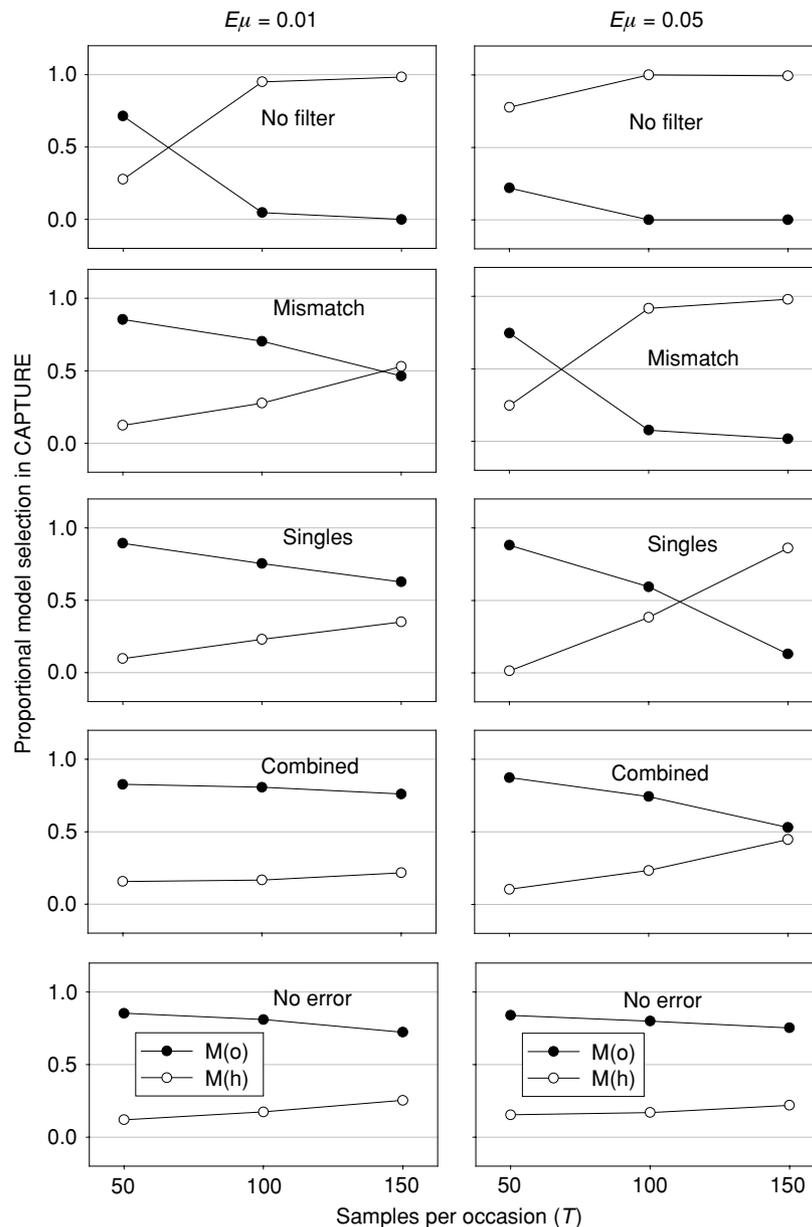
Fig. 9. Percentage CAPTURE model selection for filtered non-invasive genetic sampling (NGS) data as a function of mean per-locus error ($E\mu$) and samples per occasion (S). Simulation parameters: number of trapping occasions (T) = 5; source population size (N) = 100; 500 simulations per data point; uniform sampling distribution. Although alternate models e.g. M(t) or M(tbh), were sometimes selected, M(o) and M(h) accounted for >90% of the selected models at all data points.

negatively bias NGS data when error rates are high; however, this is a conservative error.

Misleading confidence intervals

The confidence intervals calculated within CAPTURE were often misleading for error-impacted data. Estimate precision should increase with increased sampling effort (Otis *et al*., 1978); we observed this pattern in both error-free and error-impacted data (Fig. 8). However, for error-impacted data sets, increased precision in CAPTURE estimates did not reflect a corresponding increase in estimate accuracy (Fig. 8). Genotyping errors violate the

fundamental assumption of mark–recapture theory: marks are known without ambiguity. Researchers who ignore such violations may feel a level of security in their $\hat{N}$ which is not justified.

Poisson versus uniform capture distribution

In our study, datasets with 'clustered' sampling via Poisson process displayed reductions in capture probabilities and overall coverage compared to uniform sampling. These results are congruent with work by Boulanger *et al*. (2002). For these data sets, the addition of false data points (due to simulated genotype error) resulted in
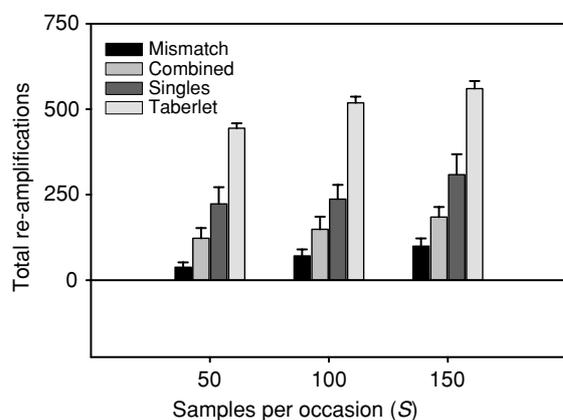
Fig. 10. Total number of simulated 're-amplifications' required to perform four different error-checking protocols; the Mismatch, Combined and Singles approaches and the multiple tubes method of Taberlet *et al.* (1996). Simulation parameters: number of trapping occasions (T) = 5; samples per occasion (S) = 50, 100 and 150; source population size (N) = 50; mean per-locus error ($E\mu$) = 0.01; uniform capture distribution; 200 simulations per data point. Patterns were similar at other levels of S, T, and $E\mu$.

extreme estimate bias at high levels of sampling pressure. Researchers who anticipate clustered deposition during NGS sampling should be particularly careful to minimise per-locus error rates.

Distributions of capture events and model selection

Genetic errors produce false genotypes that persist within data sets as apparent single-capture individuals. Changes in model selection patterns within CAPTURE reflect this process (Fig. 9). The apparent 'catchability' of individuals is skewed, with consequences for the observed probability of recapture (p). Estimators such as Mh-Jackknife or Mh-Chao are highly sensitive to p (Chao, 1989) and thus may yield highly inflated estimates for error-impacted data sets. In our study, the use of the Jackknife estimator (Mh-Jackknife) exacerbated the impact of error within our CAPTURE estimates. Mills *et al.* (2000) noted that the Jackknife estimator might mitigate potential bias due to the 'shadow effect' (false genotype matches). Our results suggest that heterogeneity estimators should be used with caution in NGS studies.

In addition to impacting population estimation procedures, changes in the perceived distribution of genotype captures and creation of false individuals will add a spurious component to any analysis of spatial pattern and kin structure (Queller, Strassman & Hughes, 1993; Pritchard, Stephens & Donnelly, 2000).

Recommendations

Researchers intending to use a non-comprehensive data filter should optimise methods so as to minimise per-locus error rates. As previously noted, most Combined approach estimates displayed biases of less than 5% at $E\mu$ = 0.05,

the exception being S = 25, T = 5 (Fig. 4). However, given the inherent difficulties in quantifying error rates, we suggest that NGS projects strive for mean per-locus error rates closer to 0.01. We have encountered no published scat studies with error rates near this threshold (Bayes *et al.*, 2000; Murphy *et al.*, 2002). Similarly, studies using single hairs or shed hairs have often recorded higher error rates (Gagneux *et al.*, 1997). These types of studies will probably need to use comprehensive data filters (Taberlet *et al.*, 1996, Miller *et al.*, 2002; Frantz *et al.*, 2003) to minimise bias. Plucked hair studies, particularly those utilising multiple hairs (Goossens *et al.*, 1998, Paetkau, 2003), offer greater promise for non-comprehensive error filters.

Multiple strategies for minimising error rates are available to the researcher, ranging from protocol optimisation (Taberlet *et al.*, 1999; Paetkau, 2003) and careful preservation of samples (Murphy *et al.*, 2002; Roon, Kendall & Waits, 2003) to the use of quantitative PCR (Morin *et al.*, 2001).

Our results suggest that heterogeneity estimators (Mh-Jackknife and Mh-Chao) should be used with caution in NGS studies, particularly if per-locus error rates cannot be certified at or below 0.01. The biological traits of the study organism should guide model selection; we also suggest that researchers compare results from heterogeneity-based estimators with results from the more conservative Null model.

Of the three filters tested in this study, both Singles and the Combined approach performed adequately at $E\mu$ = 0.01, within the range of parameters tested. Furthermore the Combined approach required fewer PCR replicates (Fig. 10). CAPTURE estimates for Mismatch data showed some suggestion of positive bias at high levels of S and T, even at $E\mu$ = 0.01. Thus, we recommend that researchers applying a 'Mismatch' type filter use greater rigor than modelled in this study. It may be possible to use the methods of Taberlet *et al.* (1996) or Miller *et al.* (2002) to set a minimum number of amplifications needed for validating specific loci. Alternatively, researchers could choose to apply the Combined filter if willing to double the total number of re-amplifications (Fig. 10).

Our study modelled three common and intuitive data filtering methods, but many alternative approaches exist. When applying alternative filtering methods, we would recommend simulation experiments to evaluate their effectiveness under the sampling conditions and objectives of the study. We also note that our results are based on final genotypes of 4–6 loci and the addition of more loci would add additional genotyping errors. Thus, NGS projects requiring more than 6 loci should aim to keep genotyping errors at or below 1%.

Our results demonstrate that it is possible, under certain conditions, to mitigate the impact of genetic error in NGS studies using less comprehensive methods than multiple-tubes approaches (Taberlet *et al.*, 1996; Miller *et al.*, 2002; Frantz *et al.*, 2003). Data filters cannot correct for all levels of genetic error and are no substitute for a diligent optimisation of methods (Paetkau, 2003). We caution that a traditional means of improving estimate

precision – increasing sampling intensity – may lead to increasingly inaccurate estimates if error filtering is ineffective. However, our results demonstrate that data filtering methods such as the Singles or Combined approach can be effective across a range of parameter values, when per-locus error is ~ 0.01 or less. Researchers who apply NGS methods to demographic research can combine such filters with careful lab technique and thoughtful study design and obtain biologically robust results.

REFERENCES

Anderson, C. R., Moody, D. S., Smith, B. L., Lindzey, F. G. & Lanka, R. P. (1998). Development & evaluation of sightability models for summer elk surveys. *J. Wildl. Mgmt.* **62**: 1055–1066.

Bayes, M. K., Smith, K. L., Alberts, S. C., Altmann, J. & Bruford, M. W. (2000). Testing the reliability of microsatellite typing from faecal DNA in the savannah baboon. *Conserv. Genet.* **1**: 173–176.

Blanchard, B. M. & Knight, R. R. (1995). Biological consequences of relocating grizzly bears in the Yellowstone ecosystem. *J. Wildl. Mgmt.* **59**: 560–565.

Boulanger, J. & McLellan, B. N. (2001). Closure violation in DNA-based mark–recapture estimation of grizzly bear populations. *Can. J. Zool.* **79**: 642–651.

Boulanger, J., White, G. C., McLellan, B. N., Strobeck, C. & Paetkau, D. P. (2002). A meta analysis of grizzly bear DNA mark–recapture projects in British Columbia. *Ursus* **13**: 137–152.

Burnham, K. P., Anderson, D. A. & Laake, J. L. (1980). Estimation of density from line transect sampling of biological populations. *Wildl. Monogr.* **72**: 1–202.

Chao, A. (1989). Estimating population size for sparse data in capture–recapture experiments. *Biometrics* **45**: 427–438.

Eberhardt, L. L. & Knight, R. R. (1996). How many grizzlies in Yellowstone? *J. Wildl. Mgmt.* **60**: 416–421.

Eberhardt, L. L., Knight, R. R. & Blanchard, B. M. (1986). Monitoring grizzly bear population trends. *J. Wildl. Mgmt.* **50**: 613–618.

Eggert, L. S., Eggert, J. A. & Woodruff, D. S. (2003). Estimating population sizes for elusive animals: the forest elephants of Kakum National Park, Ghana. *Mol. Ecol.* **12**: 1389–1402.

Frantz, A. C., Pope, L. C., Carpenter, P. J., Roper, T. J., Wilson, G. J., Delahay, R. J. & Burke, T. (2003). Reliable microsatellite genotyping of the Eurasian badger (*Meles meles*) using faecal DNA. *Mol. Ecol.* **12**: 1649–1661.

Gagneux, P., Boesch, C. & Woodruff, D. (1997). Microsatellite scoring errors associated with noninvasive genotyping based on nuclear DNA amplified from shed hair. *Mol. Ecol.* **6**: 861–868.

Gerloff, U., Schlötterer, C., Rassmann, K., Rambold, I., Hohmann, G., Fruth, B. & Tautz, D. (1995). Amplification of hypervariable simple sequence repeats (microsatellites) from excremental DNA of wild living Bonobos (*Pan Paniscus*). *Mol. Ecol.* **4**: 515–518.

Goldstein, D., Roemer, G., Smith, D., Reich, D., Bergman, A. & Wayne, R. (1999). The use of microsatellite variation to infer population structure and demographic history in a natural model system. *Genetics* **151**: 797–801.

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

Goossens, B., Chikhi, L., Utami, S., Ruiter, J. & Bruford, M. (2000). A multi-sample, multi-extracts approach for microsatellite analysis of faecal samples in an arboreal ape. *Conserv. Genet.* **1**: 157–162.

Kendall, K. C., Waits, L. P. & Schirokauer, D. (1997). *Using DNA to monitor grizzly and black bear populations*. Draft Study Plan through the U.S. Geological Survey. Available at: http://www.nrmsc.usgs.gov/research/glac_beardna.htm

Kohn, M., York, E., Kamradt, D. A., Haught, G., Sauvajot, R. M. & Wayne, R. K. (1999). Estimating population size by genotyping feces. *Proc. R. Soc. Lond. Ser. B* **226**: 657–663.

Lancia, R., Nichols, J. & Pollock, K. (1994). Estimating the number of animals in wildlife populations. In *Research and management techniques for wildlife and habitats*: 419–444. 5th Edition. Bookhout, T. (Ed.). Bethesda, MD: The Wildlife Society.

Lande, R. (1988). Genetics and demography in biological conservation. *Science* **241**: 1455–1460.

Launhardt, K., Epplen, J. T., Epplen, C. & Winkler, P. (2000). Amplification of microsatellites adapted from human systems in faecal DNA of wild Hanuman langurs (*Presbytis entellus*). *Electrophoresis* **19**: 1356–1361.

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

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

Mills, L., Citta, J., Lair, K. & Schwartz, M. (2000). Estimating animal abundance using non-invasive DNA sampling: promises and pitfalls. *Ecol. Applic.* **10**: 283–294.

Morin, P. A. & Woodruff, D. S. (1992). Paternity exclusion using multiple hypervariable microsatellite loci amplified from nuclear DNA of hair cells. In *Paternity in primates: genetic tests and theories*: 63–81. Martin, R. D., Dixson, A. F. & Wickings, E. J. (Eds). Basel, Switzerland: Karger.

Morin, P. A. & Woodruff, D. S. (1996). Noninvasive genotyping for vertebrate conservation. In *Molecular genetic approaches for conservation*: 298–313. Smith, T. B. & Wayne, R. K. (Eds). Oxford: Oxford University Press.

Morin, P. A., Chambers, K. E., 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*). *Mol. Ecol.* **10**: 1835–1844.

Mowat, G. & Strobeck, C. (2000). Estimating population size of grizzly bears using hair capture, DNA profiling, and mark–recapture analysis. *J. Wildl. Mgmt.* **64**: 183–193.

Murphy, M. A., Wait, L. P., Kendall, K. C., Wasser, S. K., Higbee, J. A. & Bogden, R. (2002). An evaluation of long-term preservation methods for brown bear (Ursus arctos) faecal DNA samples. *Conserv. Genet.* **3**: 435–440.

Ohta, T. & Kimura, M. A. (1973). A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genet. Res.* **22**: 201–204.

Otis, D. L., Burnham, K. P., White, G. C. & Anderson, D. P. (1978). Statistical inference from capture data on closed animal populations. *Wildl. Monogr.* **62**: 1–135.

Paetkau, D. (2003). An empirical exploration of data quality in DNA-based population inventories. *Mol. Ecol.* **12**: 1375–1387.

Paetkau, D. P. & Strobeck, C. (1994). Microsatellite analysis of genetic variation in black bear populations. *Mol. Ecol.* **3**: 489–495.

Paetkau, D. P., Calvert, W., Stirling, I. & Strobeck, C. (1995). Microsatellite analysis of population structure in Canadian polar bears. *Mol. Ecol.* **4**: 347–354.

Paxinos, E., McIntosh, C., Ralls, K. & Fleischer, R. (1997). A noninvasive method for distinguishing among canid species: amplification and enzyme restriction of DNA from dung. *Mol. Ecol.* **6**: 483–486.

Piggott, M. P. & Taylor, A. C. (2003). Remote collection of animal DNA and its applications in conservation management and understanding

the population biology of rare and cryptic species. *Wildl. Res.* **30**: 1–13.

Poole, K. G., Mowat, G. & Fear, D. (2001). DNA-based population estimate for grizzly bears *Ursus arctos* in northeastern British Columbia, Canada. *Wildl. Biol.* **7**: 105–115.

Pritchard, J., Stephens, M. & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics.* **155**: 945–959.

Queller, D., Srassman, J. & Hughes, C. (1993). Microsatellites and kinship. *Trends Ecol. Evol.* **8**: 285–288.

Ratti, J. & Garton, E. O. (1994). Estimating the number of animals in wild populations. In *Research and management techniques for wildlife and habitats*: 215–253. 5th Edition. Bookhout, T. A. (Ed.). Bethesda, MD: The Wildlife Society.

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

Roon, D. A., Kendall, K. C. & Waits, L. P. (2003). A quantitative evaluation of two methods for preserving hair samples. *Mol. Ecol. Notes* **3**: 163–166.

Seber, G. A. F. 1982. *The estimation of animal abundance and related parameters*. 2nd Edition. New York: Macmillian Publishing Company inc.

Shaffer, M. (1981). Minimum population sizes for species conservation. *Bioscience* **31**: 131–134.

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

Taberlet, P., Griffen, S., Goosens, B., Questiau, S., Manceau, V., Escaravage, N., Waits, L. P. & Bouvet, J. (1996). Reliable genotyping of samples with very low DNA quantities using PCR. *Nucl. Acids Res.* **24**: 3189–3194.

Taberlet, P., Camarra, J. J., Griffin, S., Uhres, E., Hanotte, O., Waits, L. P., Dubois-Paganon, C., Burke, T. & Bouvet, J. (1997). Noninvasive genetic tracking of the endangered Pyrenean brown bear population. *Mol. Ecol.* **6**: 869–876.

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

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

White, G. C. (1996). *Noremark*: population estimation from mark recapture studies. *Wildl. Soc. Bull.* **24**: 50–52.

White, G. C., Anderson, D. R., Burnham, K. P., & Otis, D. L. (1982). *Capture–recapture and removal methods for sampling closed populations*. Los Alamos, New Mexico: Los Alamos National Laboratory, LA 8787-NERP.

Woods, J. G., McLellan, B., Paetkau, D., Strobeck, C. & Proctor, M. (1996). DNA fingerprinting applied to mark–recapture bear studies. *Int. Bear News* **5**: 9–10.

Woods, J. G., Paetkau, D., Lewis, D., McLellen, B., Proctor, M. & Strobeck, C. (1999). Genetic tagging of free-ranging black and brown bears. *Wildl. Soc. Bull.* **27**: 616–627.