Kansas State University Libraries

New Prairie Press

Conference on Applied Statistics in Agriculture

2007 - 19th Annual Conference Proceedings

ADJUSTING POPULATION ESTIMATES FOR GENOTYPING ERROR IN NON-INVASIVE DNA-BASED MARK-RECAPTURE EXPERIMENTS

Shannon M. Knapp

Bruce A. Craig

ADJUSTING POPULATION ESTIMATES FOR GENOTYPING ERROR IN NON-INVASIVE DNA-BASED MARK-RECAPTURE EXPERIMENTS

Shannon M. Knapp and Bruce A. Craig
Department of Statistics, Purdue University, 250 N. University St., West Lafayette, IN 47906-2066, USA

Abstract

DNA from non-invasive sources is increasingly being used as molecular tags for mark-recapture population estimation. These sources, however, provide small quantities of often contaminated DNA, which can lead to genotyping errors that will bias the population estimate. We describe a novel approach, called Genotyping Uncertainty Added Variance Adjustment (GUAVA), to address this problem. GUAVA incorporates an explicit model of genotyping error to generate a distribution of complete-information capture histories that is used to estimate the population size. This approach both reduces the genotyping-error bias and incorporates the additional uncertainty due to genotyping error into the variance of the estimate. We demonstrate this approach via simulated mark-recapture data with a range of genetic information, population sizes, sample sizes, and genotyping error-rates. The bias, variance, and coverage of the GUAVA estimates are shown to be superior to those of other available methods used to analyze this type of data. Because GUAVA assumes each sample is genotyped only once per locus, it also has the potential to save a great deal of time and money collecting consensus molecular information.

Keywords: DNA markers, genotyping error, mark-release-recapture, microsatellite, non-invasive sampling, population size estimation

1. Introduction

Mark-recapture techniques provide a powerful tool to estimate the number of individuals in wildlife populations. With the increasing accessibility of molecular methods, DNA can now be used as a molecular mark in population estimates. Non-invasive sources of DNA, such as hair or scat, are advantageous for species that are secretive, endangered, sparsely distributed, or trap-shy and, therefore, difficult to study using traditional marks. Non-invasive DNA has been used in population estimates in such varied species as badger, bear, cougar, coyote, elephant, marten, otter, seal, whale, wolf, wolverine, and wombat (Waits and Paetkau 2005).

Unfortunately, DNA marks are subject to pitfalls not found with traditional marks. First, non-invasive sources often provide low-copy and poor-quality DNA, which increase the chances of genotyping errors. As a result, samples from the same individual may have different observed genotypes and be treated as if they came from different individuals. This would result in an overestimation of the population size. Second, when a limited number of loci are examined, different individuals can have the same true genotype, a phenomenon known as the shadow effect (Mills et al. 2000). This results in an underestimation of the population size. These problems can be minimized by repeat genotyping of each sample and using a larger number of loci respectively. These remedies, however, dramatically increase the cost of the study. It would be more cost-effective to better analyze the data available.

Lukacs and Burnham (2005) introduced a mark-recapture-based maximum likelihood method to estimate population size and a genotyping-error rate simultaneously. Their method follows the CAPTURE (White, et al. 1982) paradigm, with an additional term to represent the probability that a genotype is read correctly. They assume all individuals in the population have a unique genotype (i.e., there is no shadow effect), sampling is done without replacement, and that a genotyping error will always lead to a unique observed genotype.

In this paper, we propose an alternative approach, which we refer to as Genotyping Uncertainty Added Variance Adjustment (GUAVA). This approach takes full advantage of the molecular markers used and the information on the genotyping errors. Unlike the method of Lukacs and Burnham (2005), our method is explicitly designed to be used with microsatellites, the genetic markers commonly used in non-invasive DNA-based mark-recapture studies, and accounts for the two types of genotyping errors found in microsatellites: misprinting and allelic dropout. Microsatellite alleles vary in their number of motif-repeats, effectively the length of the allele. Misprinting, also known as false alleles, occurs when an allele appears to have more or fewer motif-repeats than it truly has. Allelic dropout occurs when one allele in a heterozygote does not amplify, giving the appearance that the individual is a homozygote. We also relax the assumptions of Lukacs and Burnham (2005) that there is no shadow effect and that all genotyping errors will lead to a unique genotype. Additionally, we allow for sampling with replacement, which is more realistic for many forms of non-invasive sampling.

In the next section, we detail the steps to generate our pseudo complete-information capture histories based on the observed data. It is this distribution of capture histories that is at the heart of GAUVA method. This is followed by some discussion of concerns with the Lukacs-Burnham approach, the other method that incorporates genotyping error. We then present a 2-capture session simulation study to compare the resulting population estimates using GUAVA with those from Lukacs-Burnham and other commonly-used approaches and conclude with a discussion of the advantages and limitations of the GUAVA approach.

2. The GUAVA Approach

GUAVA generates a distribution of complete-information capture histories, based on the observed data. For a given population estimator, the population size estimate is the mean from this distribution and the uncertainty of this estimate incorporates the variability in the size estimates from this distribution. The underlying driver of this approach is the probability that two samples (observed genotypes) come from the same individual. The derivation of this probability is based on two key principles. First, given the true genotype of an individual, GUAVA's genotyping error model defines a distribution of observed genotypes. Second, GUAVA assumes the population is in Hardy-Weinberg equilibrium so the probability of sampling each genotype in the population is assumed known (i.e., genotype frequencies can be obtained from allele frequencies).

Given the set of probabilities that a specific pair of samples come from the same individual, we can match observed genotypes (i.e., declare they are from the same individual) and generate a complete-information capture history. GUAVA does this by permuting the order of the samples and then testing each sample against all previous ones. Testing for a sample ends when the sample is matched to a previous sample or after it has been tested against all previous samples (i.e., declared a new individual). From this generated capture history, a population estimate can be obtained using any standard population estimator.

This process of generating a complete-information capture history is repeated many times to obtain an approximate distribution of the population estimate. As we show in Section 2.1, the population size, *N*, is a term in the probability of a match. Thus an arbitrary *N* is used to generate the first capture history and subsequent iterations use the population estimate from the previous iteration to generate a new capture history. After a sufficient burn-in period, every *k*th estimate of N is recorded. The variance in the estimates of *N* over the iterations of the Markov Chain approximates the variance due to genotyping errors.

To summarize, the GUAVA population estimate is

$$\hat{N}_{GUAVA} = \frac{1}{r}\sum_{i=1}^{r}\hat{N}_i \ , \tag{1}$$

where *r* is the number of recorded estimates from burn-in to convergence, and the variance of the estimate is

$$\hat{V}\left(\hat{N}_{GUAVA}\right) = \frac{1}{r}\sum_{i=1}^{r}\left[\hat{V}\left(\hat{N}\right)\right]_i + \left\{\frac{1}{r}\sum_{i=1}^{r}\hat{N}_i^2 - \left[\frac{1}{r}\sum_{i=1}^{r}\hat{N}_i\right]^2\right\}. \tag{2}$$

The first variance term can be considered the variance due to sampling error and the second term can be considered variance due to genotyping error.

2.1 Calculation of the Probability a Pair of Samples Came from the Same Individual, $s_{i,j}$

Consider the term "Observed Genotype" (GO) to refer to the observed result of genotyping a sample, and the term "True Genotype" (GT) to refer to the unobservable, actual genotype of a sample. Given a genotyping error model, it is straightforward to find the probability distribution of the observed genotype g_l at locus *l* given the true genotype is t_l. Based on these probabilities, the unconditional probability that the sample will have observed genotype g_l at locus *l* is

$$P\left(GO_l = g_l\right) = \sum_{t_l} P\left(GO_l = g_l \middle| GT_l = t_l\right) PID_{t_l} \ , \tag{3}$$

where $PID_{t_l} = P\left(GT_l = t_l\right)$ is simply the frequency of the genotype t_l, often referred to as the probability of identity. Because errors at one locus are assumed to be independent of errors at another locus, the probability of the observed multilocus genotype is

$$P\left(GO = g\right) = \prod_{l=1}^{L} P\left(GO_l = g_l\right) \tag{4}$$

where *L* is the number of loci.

When considering a pair of samples *i* and *j*, the samples either (1) came from the same individual, (2) came from different individuals that have the same true genotype, or (3) came from different individuals with different true genotypes. In terms of drawing two individuals from the population with replacement, these cases have probabilities N^{-1}, *EPID*(*N*-1)N^{-1}, and (1-

EPID)(*N*-1)N^{-1}, respectively, where *N* is the population size and $EPID = \sum_t PID_t^2$ is the expected value of the multi-locus PID. For each case, we now present the probability that the observed genotypes are g_i and g_j. These probabilities will be combined with the ones above to generate the unconditional probability that two samples, *i* and *j*, will have observed multilocus genotypes g_i and g_j.

Case 1: Samples from the same individual. Because genotyping errors in one sample are assumed to be independent from genotyping errors in another sample, the probability that two samples, *i* and *j*, from the same individual would have observed multilocus genotypes g_i and g_j, respectively, is

$$P\left(GO_i = g_i \cap GO_j = g_j \middle| S\right) = \sum_t P\left(GO_i = g_i \middle| GT_i = t\right) P\left(GO_j = g_j \middle| GT_j = t\right) PID_t\,, \qquad (5)$$

where *S* is the event that the two samples are from the same individual. The summation is taken over all possible true multilocus genotypes *t*. Because the number of multilocus genotypes *t* is prohibitively large in these studies, this term can be more efficiently calculated as the product of the per-locus probabilities

$$P\left(GO_i = g_i \cap GO_j = g_j \middle| S\right) = \prod_{l=1}^{L}\left[\sum_{t_l} P\left(GO_{l,i} = g_{l,i} \middle| GT_{l,i} = t_l\right) P\left(GO_{l,j} = g_{l,j} \middle| GT_{l,j} = t_l\right) PID_{t_l}\right]. \qquad (6)$$

Case 2: Samples from different individuals with the same true genotype. In this case, the formula is similar except for the fact that we are considering two individuals rather than one. That is why we weight each true genotype by $PID_t^2 / EPID$ instead of by PID_t .

$$P\left(GO_i = g_i \cap GO_j = g_j \middle| S^c \cap GT_i = GT_j\right) = \frac{\sum_t P\left(GO_i = g_i \middle| GT_i = t\right) P\left(GO_j = g_j \middle| GT_j = t\right) PID_t^2}{EPID}, \qquad (7)$$

Case 3: Samples from different individuals with different true genotypes. In this case, we sum over all combinations of different genotypes and weight accordingly.

$$P\left(GO_i = g_i \cap GO_j = g_j \middle| S^c \cap GT_i \neq GT_j\right) = \frac{\sum_{t_1}\sum_{t_2 \neq t_1} P\left(GO_i = g_i \middle| GT_i = t_1\right) P\left(GO_j = g_j \middle| GT_j = t_2\right) PID_{t_1} PID_{t_2}}{1 - EPID} \qquad (8)$$

From these results,

$$\begin{aligned} P\left(GO_i = g_i \cap GO_j = g_j\right) = {} & P\left(GO_i = g_i \cap GO_j = g_j \middle| S\right) N^{-1} \\ & + P\left(GO_i = g_i \cap GO_j = g_j \middle| S^c \cap GT_i = GT_j\right) EPID\left(N-1\right) N^{-1} \\ & + P\left(GO_i = g_i \cap GO_j = g_j \middle| S^c \cap GT_i \neq GT_j\right)\left(1 - EPID\right)\left(N-1\right) N^{-1}, \end{aligned} \qquad (9)$$

which reduces to

$$P\left(GO_i = g_i \cap GO_j = g_j \middle| S\right) N^{-1} + P\left(GO_i = g_i\right) P\left(GO_j = g_j\right)\left(N-1\right) N^{-1}. \qquad (10)$$

Finally with this probability, we can find the probability two samples *i* and *j* came from the same individual, given their observed genotypes are g_i and g_j using Bayes Theorem,

$$s_{i,j} = P\left(S \middle| GO_i = g_i \cap GO_j = g_j\right) = \frac{P\left(GO_i = g_i \cap GO_j = g_j \middle| S\right) P(S)}{P\left(GO_i = g_i \cap GO_j = g_j\right)}. \tag{11}$$

This expression can be expanded into the more computationally efficient form

$$s_{i,j} = \frac{P\left(GO_i = g_i \cap GO_j = g_j \middle| S\right)}{P\left(GO_i = g_i \cap GO_j = g_j \middle| S\right) + (N-1) P\left(GO_i = g_i\right) P\left(GO_j = g_j\right)}. \tag{12}$$

It is this probability that is used to match up the samples and create a pseudo-complete information capture history. Please note that a pair of samples with identical observed genotypes will not be matched with probability 1 and that a pair with different observed genotypes may be matched. Most other methods utilize various assumptions to create a single pseudo complete-information capture history. This not only leads to bias, if the assumptions are not true, but also does not account for the increased uncertainty due to genotyping error.

3. The Lukacs-Burnham Approach

In a similar vain, Lukacs and Burnham (2005) specify probabilities for all possible capture histories. There are, however, some concerns with their approach in the non-invasive setting. For illustration, we consider the 2-capture-session case and focus on the potential capture histories of a genotype.

The capture history [10] occurs when a genotype is observed only during the first capture session. This is defined to have probability $p_1\alpha(1 - c) + p_1(1 - \alpha)$, where p_i is the probability that a genotype was first captured during session i; α is the probability that a genotype is identified correctly, given it is observed for the first time; and c is the probability of recapture (Lukacs and Burnham 2005). Essentially, this capture history occurs if a genotype is caught during the first session, genotyped correctly, and then not recaptured, or if any genotype is captured during the first session and genotyped incorrectly.

Note that because Lukacs and Burnham (2005) assume there is no shadow effect, a genotype is synonymous with an individual in the construction of their probabilities. Because Lukacs and Burnham (2005) assume that any genotyping error will lead to a unique genotype, if there is a genotyping error at the initial capture, that genotype will never be observed again. Also, in contrast to GUAVA, Lukacs and Burnham assume that if two samples have the same observed genotype, they come from the same individual with probability 1.

The capture history [01], where the genotype is observed only during the second capture session, has probability $(1 - p_1)[p_2\alpha + p_2(1 - \alpha)] = (1 - p_1)p_2$. According to Lukacs and Burnham (2005), this capture history can occur if the individual is not captured during the first session, but is captured during the second session and is either genotyped correctly or incorrectly at that time.

The capture history [11], where the genotype is observed during both the first and second capture sessions, has probability $p_1\alpha c$. This capture history occurs if the genotype is sampled during the first capture session, genotyped correctly, then recaptured during the second sampling session. Note that under the Lukacs and Burnham (2005) assumptions, recapture implies the genotype was both captured and genotyped correctly.

A final capture history, [00], where the genotype is not observed during either capture session, has probability (1 - p_1)(1 - p_2). These capture histories are, of course, not observable.

Because the parameters p_1, p_2, c, and α are not simultaneously estimable with only two capture sessions, we assumed there was no time effect (*i.e.*, $p_1 = p_2 = p$) and there is no behavior effect, that is, the probability of first capture is equal to the probability of recapture. With this second assumption, the event a genotype is recaptured is equivalent to the individual being captured and genotyped correctly, $c = \alpha p$, (P. Lukacs, pers. comm., 16 February 2007). With these simplifying assumptions, the MLEs of α and p for the two-capture session are

$$\hat{\alpha} = \sqrt{\frac{n_{[11]}}{n_{[10]} + n_{[11]} - n_{[01]}}} \tag{13}$$

and

$$\hat{p} = \frac{n_{[11]} + n_{[10]} - n_{[01]}}{n_{[11]} + n_{[10]}}, \tag{14}$$

where $n_{[11]}$, $n_{[10]}$, $n_{[01]}$, are the number of individuals in the study with capture histories [11], [10], and [01], respectively. The population estimate is

$$\hat{N}_{LB} = \frac{\hat{\alpha}\left(n_{[11]} + n_{[10]} + n_{[01]}\right)}{1-\left(1-\hat{p}\right)^2}. \tag{15}$$

A closed-form solution to the associated variance estimator is also available via the method suggested in Lukacs and Burnham (2005), but it is rather cumbersome and not presented here.

Although the probabilities of the four capture histories, [00], [01], [10], and [11] sum to 1, we believe there is another possible sample outcome that have been neglected. Specifically, we believe that there is a fifth "capture history" which involves two of the previous capture histories. The "capture history" {[10], [01]} occurs when a genotype is sampled during both sessions, but there was a genotyping error in one or both sessions. This fifth capture history has probability $p(1 - \alpha)p\alpha + p(1 - \alpha)\, p(1 - \alpha) + p\alpha\, p(1 - \alpha)$, which reduces to $p^2(1 - \alpha^2)$. The capture history {[10], [01]} is not observable, but will add one count to each capture history [10] and [01].

This additional capture history makes the sum of the probabilities of the possible outcomes sum to more than 1. One could scale the four simple capture history ([00], [01], [10], and [11]) probabilities so they sum to one. The resulting probabilities would then be:

$$P[10] = P[01] = \frac{p(1-p) + p^2\left(1-\alpha^2\right)}{1+p^2\left(1-\alpha^2\right)}$$

$$P[11] = \frac{p^2\alpha^2}{1+p^2\left(1-\alpha^2\right)} \tag{16}$$

$$P[00] = \frac{(1-p)^2}{1+p^2\left(1-\alpha^2\right)}.$$

The MLEs of α and p are not simultaneously estimable with only two-capture sessions under this new formulation. However, a similar formulation for the three-capture session case, does lead to closed form MLEs of both α and p.

$$\hat{\alpha} = \frac{6n_2n_3 + n_2^2 + 9n_3^2}{3n_3\left(n_1 + 2n_2 + 3n_3\right)}, \tag{17}$$

and

$$\hat{p} = \frac{9n_3^2\left(n_1 + 2n_2 + 3n_3\right)}{n_2^3 + 27n_3^3 + 27n_2n_3^2 + 9n_2^2n_3}, \tag{18}$$

where n_1, n_2, and n_3 are the number of capture histories where the genotype was sampled only once ([001], [010], and [100]), twice ([011], [101], and [110]), and three times ([111]), respectively. The population estimate for the three capture session case is then

$$\hat{N}_{LB-\text{modified}} = \frac{\hat{\alpha}_3\left(n_1 + n_2 + n_3\right)}{1-\left(1-\hat{p}_3\right)^3}. \tag{19}$$

We will not pursue this estimator any further but thought some discussion about their specific assumptions and this capture history omission may help shed light on the following simulation results.

4. Simulations

We evaluated the accuracy and precision of GUAVA via simulation studies varying four factors: marker set, population size, sample size, and genotyping error rates. We ran 1000 replications of each factor combination. The levels of each factor are as follows:

- **MARKER SET –** Considered marker sets with Poor, Fair, and Good genetic information. Allele frequencies were taken from population I(BR) in Paetkau *et al.* (1997). The three levels of marker sets used the first 3, the first 5, and all 8 loci, respectively. Each locus contained between 7 and 14 alleles. The EPID for the three marker sets were 4.0×10^{-4}, 7.3×10^{-6}, and 7.3×10^{-9}, with EPID$_{\text{sib}}$ (Waits *et al.* 2001) values of 0.05461, 0.01015, and 0.00067. EPID$_{\text{sib}}$ is the estimated probability that a pair of siblings would share a genotype.
- **POPULATION SIZE –** Used population sizes of 50, 200, and 1000 individuals.
- **SAMPLE SIZE –** Considered 25, 50, 100, 200, and 500 samples per capture session. Because some of these levels are unrealistically large or insufficiently small for some levels of population size, only 3 levels of sample size were used for each level of population size. For population size of 50, we used sample sizes 25, 50, and 100. For population size of 200, we used sample sizes 50, 100, and 200. For population sizes of 1000, we used sample sizes 100, 200, and 500.

- **GENOTYPING ERROR –** We used a set of Low and High error rates. Low had a misprint probability of 0.01 per allele and a dropout rate of 0.05 per locus. High had a misprint and dropout rates of 0.10 and 0.25, respectively. The High genotyping error rate level was only used with a population size of 200.

Individuals were simulated by randomly assigning an allele size to each allele at a locus based on the frequencies of those alleles. Because of this, it is possible for individuals in the population to share the same true genotype (i.e., the shadow effect). This method of simulation assumes Hardy-Weinberg equilibrium, so genotype frequencies could be calculated from allele frequencies. Given the true genotypes of the individuals in the population, samples were obtained by first randomly sampling individuals with replacement for each of two capture sessions and then imposing genotyping errors. The method of simulating samples meant that the simplifying assumptions for the Lukacs-Burnham method (*i.e.*, that $p_1 = p_2 = p$ and $c = \alpha p$) were satisfied. Misprinting errors were randomly imposed on each allele; if misprinting occurred, it was equally likely to increase or decrease the allele size by one repeat. Next dropout was simulated with the two alleles at the locus equally likely to be dropped.

For each replication, capture histories were generated as described above. For each capture history, we used the Bailey's Binomial estimate of population size and the associated estimates of variance (Seber 1982:61). We also examined the Lincoln-Peterson estimator (Seber 1982: 60), but initial results suggested the Lincoln-Peterson estimator did not perform as well as the Bailey's estimator (data not shown). The key difference between the two estimators is that the Lincoln-Peterson assumes the second sample is taken without replacement while Bailey's Binomial assumes the second sample is taken with replacement. Sampling is done with replacement in the simulations, and would be expected to be taken with replacement in field trials, so the Bailey's estimator is more appropriate.

We determined the required burn-in period and number of iterations to skip between recorded values of $\hat{N}$ by examining 10 replicates of each treatment for 10,000 iterations. Using the total number of unique observed genotypes as N_0, the burn-in time appeared negligible and was set at 100 for all factor combinations. Autocorrelation of successive estimates was tested and the largest significant lag over the 10 replicates for a treatment was used as the lag (range 2 to 15 iterations, Table 1). The chain was continued until at least 100 values of $\hat{N}$ were recorded and then ceased when the change in the average $\hat{N}$, was less than 0.01. All replications converged with no more than 13,577 recorded values, although most converged with less than 200 recorded values (Table 1).

For the purpose of comparison, for each replicate, we also calculated population estimates for the following methods

- **TRUE** - the Bailey's Binomial population estimates that would result if true identities of individuals were discernable.
- **GT** - the Bailey's Binomial population estimates that would result had true genotypes been discernable (i.e., if there was no genotyping error).
- **GO** - the Bailey's Binomial population estimates based on the observed genotypes

- **BC** - the Bailey's Binomial population estimates based on a "biologist-corrected" capture history. For the biologist correction we allow two scats to be a match if they have the same observed genotype at all loci or if the observed genotypes are the same at all but one locus and the two scats share one allele at the non-matching locus.
- **LB** - the population estimates obtained using the Lukacs and Burnham (2005) method. These are based on observed genotypes.

In addition to the population estimates, large-sample confidence intervals were constructed for each method.

5. Results and Discussion

Success rate. In some replications, a population estimate for a specific method was not obtainable. This was not due to insufficient sample size as population estimates were obtained in 100% of the replicates using TRUE. Other methods that did not struggle were GT and GUAVA. For other methods, however, as few as 2.8% of replicates produced estimates using observed genotypes (GO) as the data source and as few as 24.3% for the biologist-corrected data (BC) (Table 1). The Lukacs-Burnham method (LB) never produced estimates in more than 60.9% of replicates, and for some factor combinations in as few as 2.8% of replicates (Table 1). Estimates are not obtainable if there are no recaptures. When observed genotypes are used as the data source, genotyping errors reduce the number of apparent recaptures. Recaptures of observed genotypes decrease as the probability of genotyping error increases. As the number of loci used in the marker set increases, the probability that there will be an error at one or more loci increases, subsequently reducing apparent recaptures. With the Low error rates, the probability of at least one genotyping error is 0.19 with 3 loci, and increases to 0.44 with 8 loci. With the High error rates these probabilities increase to 0.78 and 0.98. The BC method reduces this problem by allowing for some genotyping error, but cannot entirely overcome it. The LB method frequently did not produce estimates because the estimates of the parameters α and p can be negative, imaginary, or greater than 1 depending on the relative numbers of each type of capture history. When this occurs, no estimate can be obtained.

Bias. GUAVA estimates had very low bias, ranging from -5.3% to +3.6% (and only -1.0% to +3.4% under Low combinations). These levels are very comparable to the TRUE estimates (Table 2). GT estimates, when different from TRUE estimates, are biased low, estimating the number of genotypes in the population, instead of the number of individuals in the population. The GT estimates tend to underestimate even the number of genotypes in the population. This is because when there are multiple copies of some genotypes in the population, the assumption of equal capture probability inherent in the Bailey's estimate is violated; a multi-session estimator that allows for individual heterogeneity, such as models available in CAPTURE (White *et al.* 1982) would be more appropriate, but would still not result in an estimate of the number of individuals in the population. This result indicates that there are still estimation problems, even when all genotyping error has been eliminated.

The bias of the uncorrected data (GO) is unacceptably high, ranging from -2.7% to +8,919.8% (-2.7 to +219.2% when High error factor combinations are excluded). The bias of

GO estimates increase as the number of loci increase, because the probability of a genotyping error at one or more loci increases.

The BC estimates partially overcome the problems with the GO estimates, but this method has its own shortcomings. When only 3 loci are used, the BC method pairs samples that match at only 2 loci, something unlikely to be done in practice, and greatly underestimating the population size in these cases. The bias of the BC estimates ranged from -81.8% to +4,477.1% (-81.8% to +55.0% when High error factor combinations are excluded).

LB estimates were also highly biased, with percent bias ranging from -17.1% to +14,539.7% (-17.12% to +245.0% when High error factor combinations are excluded). As previously mentioned, the probability that a multi-locus genotype is read correctly decreases with increasing number of loci. With the Low error rates, the probability a genotype is read correctly, α, ranges from 0.81 for 3 loci to 0.56 for 8 loci (with High error, these values reduce to 0.22 and 0.02). Lukacs and Burnham (2005), assumed low levels of genotyping error, testing α from 0.95 to 0.99. Several loci are required for the LB method to avoid the shadow effect, which the LB method assumes is not present.

Variance and Standard Errors. In general, the variance estimate of the Bailey's Binomial estimate is proportional to the estimate, so the standard errors tend to be high when the population is overestimated and low when the population is underestimated; this explains much of the differences between TRUE standard errors and the standard errors of the GO and BC methods (Table 3).

The variance (or standard error) of the GUAVA estimate is expected to be at least as large as that of the TRUE estimate. The average GUAVA standard errors were never more than 32.3% higher than that of the respective TRUE estimate (16.6% when High error factor combinations are excluded). The average GUAVA standard error decreased with increasing sample size and with an increase of the number of loci used in the marker set; however, doubling the number of samples taken lead to a larger reduction in the average standard error than did a doubling of the number of loci used.

As with the GUAVA estimates, the average standard error of the LB estimate decreases with increasing sample size, but, in contrast to GUAVA, the average standard error of the LB estimate increased as the number of loci increased. Again, this is attributable to the increase of genotyping errors (reduction in α) as the number of loci increases. Lukacs and Burnham (2005) realized that their variance term was potentially very large, but only if α was small, which they assumed it was not.

Coverage. For each treatment we calculated the percent of replicates where the true population size was included in the 95% confidence interval (Table 4). Ideally this value should be close to 95%. Coverage values less then 93.6% or higher than 96.4% are significantly different from 95% (at the 5% level). Coverage, in part, evaluates the accuracy of the standard error. However, coverage values must be examined in conjunction with the accuracy of the estimate and the size of its standard error. For several factor combinations, the coverage of the GO and LB estimates reaches 100% due to astronomically high standard errors.

The coverage of the GUAVA estimate ranged from 85.2% to 97.7% (87.4% to 97.7% when High error factor combinations are excluded). Where the GUAVA coverage diverged significantly from 95%, the TRUE coverage typically did as well. This suggests that the normality assumption in the construction of the confidence intervals is more at issue than is the accuracy of the estimate or its standard error. According to Seber (1982:63), the normality assumption is appropriate when the sample size and number of recaptures is large; some of our factor combinations apparently pushed the boundaries for normality.

6. Summary

This simulation study demonstrates the potential benefits of GUAVA. By generating the distribution of likely complete-information capture histories, the population estimate had low bias with standard errors such that the coverage probability of a standard large-sample confidence interval was comparable to estimates based on perfect information (TRUE). Furthermore, the accuracy and precision of the GUAVA estimates were comparable whether 3 or 8 loci were used. This suggests that fewer loci can be used either to reduce cost or as a trade-off to increasing sample size. Work is ongoing to develop the most efficient “matching” approach and expand the algorithm to more than two sessions. The eventual goal is to incorporate this procedure into standard software like CAPTURE (White *et al.* 1982) since any population estimate can be used with the approach.

Literature Cited

Lukacs, P. M. and K. P. Burnham. 2005. Estimating population size from DNA-based closed capture-recapture data incorporating genotyping error. Journal of Wildlife Management 69: 396-403.

Mills, L. S., J. J. Citta, K. P. Lair, M. K. Schwartz, and D. A. Tallmon. 2000. Estimating animal abundance using noninvasive DNA sampling: promise and pitfall. Ecological Applications 10: 283-294.

Paetkau, D., L. P. Waits, P. L. Clarkson, L. Craighead, and C. Strobeck. 1997. An empirical evaluation of genetic distance statistics using microsatellite data from bear (Ursidae) populations. Genetics 147: 1943-1957.

Seber, G. A. F. 1982. The Estimation of Animal Abundance and Related Parameters. Second edition. Macmillan. New York, New York, USA..

Waits, L. P., G. Luikart, and P. Taberlet. 2001. Estimating the probability of identity among genotypes in natural populations: cautions and guidelines. Molecular Ecology 10: 249-256.

Waits, L. P. and D. Paetkau. 2005. Noninvasive genetic sampling tools for wildlife biologists: A review of applications and recommendations for accurate data collection. Journal of Wildlife Management 69: 1419-1433.

White, G. C., D. R. Anderson, K. P. Burnham, and D. L. Otis. 1982. Capture-recapture and removal methods for sampling closed populations. Los Alamos National Laboratory LA-8787-NERP. 235 pp.

Table 1. Lag time, time to convergence, and percent of replications where an estimate was obtained for GO, BC, and LB methods. *N* is the true population size, Error is the genotyping error treatment level, *n* is the sample size, Markers is the marker set treatment level, *k* is the lag between recorded iterations, and *r* is the number of recorded estimates until convergence.

							% reps with estimate		
N	**Error**	***n***	**Markers**	***k***	**mean *r***	**max *r***	**GO**	**BC**	**LB**
50	**Low**	**25**	**Poor**	3	111.8	651	100	100	60.9
			Fair	3	112.8	671	99.6	100	59.1
			Good	3	101.9	419	98.7	100	60.1
		50	**Poor**	3	103.9	133	100	100	56.1
			Fair	3	101.2	119	100	100	53.5
			Good	2	100.1	109	100	100	54.2
		100	**Poor**	3	102.4	118	100	100	55.0
			Fair	3	100.7	112	100	100	52.9
			Good	2	100.0	103	100	100	53.0
200	**Low**	**50**	**Poor**	5	149.6	2,303	100	100	58.8
			Fair	3	217.9	4,216	100	100	57.4
			Good	4	138.1	1,671	98.5	100	59.9
		100	**Poor**	4	116.8	203	100	100	54.4
			Fair	3	110.0	190	100	100	54.6
			Good	2	107.9	303	100	100	55.2
		200	**Poor**	5	110.4	161	100	100	51.0
			Fair	3	105.3	141	100	100	51.5
			Good	4	100.9	127	100	100	53.8
	High	**50**	**Poor**	15	163.0	1,960	82.3	100	53.3
			Fair	6	156.1	2,985	28.0	82.1	24.0
			Good	4	163.6	3,132	2.8	24.3	2.8
		100	**Poor**	10	130.8	342	99.8	100	57.3
			Fair	6	122.8	258	71.8	99.8	49.0
			Good	4	113.8	192	13.9	63.9	12.7
		200	**Poor**	10	117.8	193	100	100	52.5
			Fair	6	113.7	170	99.4	100	59.5
			Good	4	108.8	152	42.3	98.6	34.8
1000	**Low**	**100**	**Poor**	7	673.8	8,856	100	100	57.6
			Fair	4	1,281.4	13,577	99.5	100	56.9
			Good	3	735.3	12,426	97.5	100	59.7
		200	**Poor**	6	184.3	554	100	100	56.9
			Fair	4	158.3	480	100	100	56.3
			Good	3	272.7	2,795	100	100	53.6
		500	**Poor**	6	145.5	306	100	100	50.5
			Fair	3	126.1	235	100	100	52.9
			Good	3	109.1	175	100	100	52.4

Table 2. Average estimate. *N* is the true population size, Error is the genotyping error treatment level, *n* is the sample size, Markers is the marker set treatment level, $\bar{N}_{GT}$ is the average number of unique genotypes in the population over the 1,000 replications of the treatment.

N	**Error**	***n***	**Markers**	$\bar{N}_{GT}$	**GUAVA**	**TRUE**	**GT**	**GO**	**BC**	**LB**
50	**Low**	**25**	**Poor**	**49.5**	50.6	49.6	48.7	72.1	41.2	63.1
			Fair	**50.0**	50.2	50.3	50.3	91.2	56.8	80.8
			Good	**50.0**	49.5	49.5	49.5	125.3	66.4	121.1
		50	**Poor**	**49.5**	50.7	50.0	49.3	75.2	44.0	71.5
			Fair	**50.0**	50.0	49.9	49.9	98.0	57.9	92.6
			Good	**50.0**	50.2	50.3	50.3	144.9	70.9	133.7
		100	**Poor**	**49.5**	51.1	49.9	49.4	81.5	46.7	87.4
			Fair	**50.0**	50.1	50.0	50.0	108.0	61.1	116.2
			Good	**50.0**	50.0	50.0	50.0	159.6	77.5	172.0
200	**Low**	**50**	**Poor**	**192.5**	202.9	197.7	185.4	262.1	100.7	225.9
			Fair	**199.8**	202.4	201.3	201.1	365.1	219.5	320.0
			Good	**200.0**	201.6	201.4	201.4	499.0	268.7	473.4
		100	**Poor**	**192.5**	203.3	200.6	188.7	271.9	113.5	247.5
			Fair	**199.9**	200.6	200.5	200.3	376.1	222.3	341.6
			Good	**200.0**	199.5	199.5	199.5	532.5	268.6	471.6
		200	**Poor**	**192.5**	204.6	200.3	190.3	286.6	128.9	287.4
			Fair	**199.9**	200.2	199.7	199.5	389.8	228.0	392.1
			Good	**200.0**	199.9	199.9	199.9	561.9	280.7	556.2
	High	**50**	**Poor**	**192.4**	207.1	197.2	184.4	876.2	193.2	1,014.3
			Fair	**199.9**	206.3	198.0	197.9	1,176.5	886.5	2,006.0
			Good	**200.0**	201.4	200.7	200.7	1,244.6	1,189.3	2,410.7
		100	**Poor**	**192.5**	199.2	200.1	188.4	1,438.9	228.5	1,271.2
			Fair	**199.8**	198.9	200.6	200.3	3,915.2	1,425.0	5,219.3
			Good	**200.0**	198.0	199.0	199.0	4,918.6	4,086.5	9,009.0
		200	**Poor**	**192.6**	189.4	199.4	189.5	1,499.6	277.2	1,406.2
			Fair	**199.9**	195.1	200.0	199.8	7,878.0	1,547.4	6,805.1
			Good	**200.0**	197.6	200.0	200.0	18,039.6	9,154.2	29,279.4
1000	**Low**	**100**	**Poor**	**853.6**	1,033.8	997.6	732.5	973.3	182.2	828.8
			Fair	**996.4**	1,001.9	991.2	983.8	1,727.2	944.0	1,420.3
			Good	**1,000.0**	1,006.9	1,007.0	1,007.0	2,361.5	1,312.5	2,149.7
		200	**Poor**	**853.8**	1,014.8	999.3	745.9	994.1	222.8	912.9
			Fair	**996.5**	999.5	993.3	987.0	1,783.3	959.5	1,553.3
			Good	**1,000.0**	994.9	995.0	995.0	2,568.7	1,304.7	2,239.4
		500	**Poor**	**854.6**	1,007.9	996.5	777.8	1,065.9	298.3	1,084.7
			Fair	**996.4**	1,003.3	999.2	993.9	1,842.0	1,003.8	1,775.2
			Good	**1,000.0**	1,002.6	1,002.7	1,002.7	2,659.2	1,348.9	2,539.1

Table 3. Average standard error of the estimate. *N* is the true population size, Error is the genotyping error treatment level, *n* is the sample size, and Markers is the marker set treatment level.

N	Error	*n*	Markers	GUAVA	TRUE	GT	GO	BC	LB
50	**Low**	**25**	**Poor**	12.5	11.6	11.3	21.4	8.7	78.0
			Fair	12.1	12.1	12.1	30.9	14.8	142.4
			Good	11.6	11.6	11.6	49.5	18.8	388.1
		50	**Poor**	5.5	5.3	5.2	10.8	4.3	26.8
			Fair	5.3	5.2	5.2	16.8	6.8	48.7
			Good	5.3	5.3	5.3	31.6	9.8	112.9
		100	**Poor**	2.1	1.9	1.9	5.4	1.8	14.6
			Fair	2.0	1.9	1.9	8.9	3.0	27.3
			Good	1.9	1.9	1.9	17.3	4.9	64.0
200	**Low**	**50**	**Poor**	55.6	50.6	45.9	78.7	17.5	354.8
			Fair	53.5	52.4	52.3	129.9	59.9	850.3
			Good	52.4	52.3	52.3	204.5	82.2	2240.6
		100	**Poor**	25.7	24.7	22.3	40.4	9.6	122.1
			Fair	24.8	24.7	24.6	67.5	29.2	250.2
			Good	24.5	24.4	24.4	115.9	39.6	506.1
		200	**Poor**	11.5	10.8	9.8	20.3	5.1	53.5
			Fair	10.8	10.7	10.7	33.7	13.6	105.7
			Good	10.7	10.7	10.7	60.8	19.5	224.8
	High	**50**	**Poor**	66.8	50.5	45.5	444.6	49.6	9,155.9
			Fair	60.3	50.9	50.9	651.0	450.8	26,954.7
			Good	53.9	52.0	52.0	699.5	660.3	34,728.5
		100	**Poor**	26.8	24.6	22.3	523.8	31.0	5,245.4
			Fair	25.4	24.7	24.7	2,081.3	512.7	79,413.8
			Good	24.5	24.3	24.3	2,791.9	2,200.7	181,922.4
		200	**Poor**	10.4	10.6	9.7	279.7	19.5	1,607.2
			Fair	10.7	10.7	10.7	3,317.3	293.1	54,260.0
			Good	10.7	10.7	10.7	10,061.4	4,095.4	768,925.2
1000	**Low**	**100**	**Poor**	350.4	300.5	188.7	291.5	21.5	1,753.3
			Fair	310.8	298.1	294.8	678.9	277.6	6,177.4
			Good	305.2	304.7	304.7	1,056.6	453.8	15,616.6
		200	**Poor**	157.7	149.1	94.7	148.2	13.6	628.9
			Fair	150.1	147.5	146.0	362.5	140.0	1,936.5
			Good	148.2	148.1	148.1	632.7	225.0	4,303.0
		500	**Poor**	56.9	55.1	36.8	61.9	7.5	204.3
			Fair	55.9	55.3	54.8	147.3	55.9	568.6
			Good	55.7	55.7	55.7	261.2	90.0	1,208.9

Table 4. Percent of replications in which the 95% confidence interval included the true population size. *N* is the true population size, Error is the genotyping error treatment level, *n* is the sample size, and Markers is the marker set treatment level.

N	**Error**	***n***	**Markers**	**GUAVA**	**TRUE**	**GT**	**GO**	**BC**	**LB**
50	**Low**	**25**	**Poor**	90.6	86.7	85.1	99.4	62.6	99.7
			Fair	90.2	88.6	88.6	99.9	94.6	100.0
			Good	87.4	86.9	86.9	100.0	98.8	100.0
		50	**Poor**	94.0	91.8	88.7	29.0	55.9	100.0
			Fair	92.9	89.9	89.8	0.9	87.5	100.0
			Good	92.6	92.1	92.1	0.0	40.9	100.0
		100	**Poor**	97.7	93.0	88.4	0.0	49.1	0.5
			Fair	95.2	92.8	92.8	0.0	6.7	3.0
			Good	93.0	93.1	93.1	0.0	0.0	43.6
200	**Low**	**50**	**Poor**	89.2	87.7	81.4	98.6	4.9	100.0
			Fair	89.6	88.0	87.7	100.0	94.4	100.0
			Good	88.8	88.4	88.4	100.0	98.4	100.0
		100	**Poor**	93.8	94.3	84.1	66.7	0.2	100.0
			Fair	95.0	94.2	94.2	5.8	97.2	100.0
			Good	93.6	93.4	93.4	0.0	68.9	100.0
		200	**Poor**	95.9	95.6	76.4	0.0	0.0	99.4
			Fair	94.6	94.1	94.0	0.0	47.9	85.8
			Good	94.2	93.9	93.9	0.0	0.0	98.1
	High	**50**	**Poor**	85.2	88.2	82.0	100.0	82.2	100.0
			Fair	88.6	88.0	88.0	100.0	100.0	100.0
			Good	91.1	90.6	90.6	100.0	100.0	100.0
		100	**Poor**	90.0	93.7	83.4	3.5	89.8	100.0
			Fair	93.6	92.9	92.8	79.4	2.8	100.0
			Good	94.1	93.1	93.1	99.3	85.1	100.0
		200	**Poor**	86.0	94.8	73.7	0.0	1.4	100.0
			Fair	95.7	94.4	94.4	4.2	0.0	100.0
			Good	96.9	94.8	94.8	72.6	8.0	100.0
1000	**Low**	**100**	**Poor**	88.5	89.6	51.4	85.7	0.0	99.1
			Fair	87.6	87.7	87.3	99.7	82.7	100.0
			Good	89.9	89.4	89.4	100.0	98.6	100.0
		200	**Poor**	92.3	94.3	27.8	90.2	0.0	99.8
			Fair	93.9	94.0	93.3	32.1	90.0	100.0
			Good	92.8	93.3	93.3	2.0	89.4	100.0
		500	**Poor**	94.4	95.3	0.1	84.0	0.0	100.0
			Fair	95.9	94.9	94.5	0.0	93.7	100.0
			Good	95.8	95.7	95.7	0.0	0.1	100.0