Conservation Genetics (2006) 7:319–329
DOI 10.1007/s10592-005-9024-6

Using DNA from non-invasive samples to identify individuals and census populations: an evidential approach tolerant of genotyping errors

Steven T. Kalinowski*, Mark L. Taper & Scott Creel
Department of Ecology, Montana State University, 310 Lewis Hall, Bozeman, MT, 59717, USA
*(*Corresponding author: phone: +406-994-3232; fax: +406-994-3190; e-mail: skalinowski@montana.edu)*

Received 22 February 2005; accepted 06 July 2005

Key words: allele dropout, census, DNA, genotyping error, non-invasive, statistical evidence

Abstract

DNA extracted from hair or faeces shows increasing promise for censusing populations whose individuals are difficult to locate. To date, the main problem with this approach has been that genotyping errors are common. If these errors are not identified, counting genotypes is likely to overestimate the number of individuals in a population. Here, we describe an algorithm that uses maximum likelihood estimates of genotyping error rates to calculate the evidence that samples came from the same individual. We test this algorithm with a hypothetical model of genotyping error and show that this algorithm works well with substantial rates of genotyping error and reasonable amounts of data. Additional work is necessary to develop statistical models of error in empirical data.

Introduction

"...there is a critical need for population genetics software... incorporating [genotyping] error" – Bonin et al. (2004)

A census is invaluable for the management of small populations. Capture-mark-recapture methods are currently the standard method for estimating the size of populations, but genetic data offers increasing promise – especially for species whose individuals are difficult to locate. The method is simple in concept (1) Collect a large number of hair or faeces specimens from the field. (2) Genotype DNA extracted from these specimens. (3) Count the number of unique multilocus genotypes observed. This number serves as a minimum number of individuals visiting a watering hole, crossing a road, or living in a population (e.g., Taberlet et al. 1997). More refined estimates of census size can be obtained using genotype accumulation methods (e.g., Kohn et al. 1999) or using capture-mark-recapture analysis of the genotype counts (e.g., Woods et al. 1999).

DNA censuses are vulnerable to genotyping error (e.g., Taberlet et al. 1999; Taberlet and Luikart 1999; Waits and Leberg 2000). This is because, genotyping errors can cause two specimens from the same individual to appear to have different genotypes, and therefore appear to come from two different individuals. Even low error rates can dramatically inflate estimates of census size (Waits and Leberg 2000).

The conventional method for dealing with genotyping errors is to try to reduce their occurrence to a negligible rate. There are several ways to do this (e.g., Taberlet et al. 1999; Morin et al. 2001; Miller et al. 2002; Paetkau 2003). For example, Taberlet et al. (1999) recommended re-genotyping specimens until the correct genotype could be inferred reliably. In contrast, Paetkau (2003) recommended using professional judgment to remove poor quality specimens from analysis. No matter how genotyping errors are prevented or

identified, the protocol must be almost perfect to accurately count individuals.

An alternative to eliminating errors is to accommodate them during data analysis. Many authors have estimated genotyping error rates (e.g., Broquet and Petit 2004), but there has been few suggestions for how to deal with the errors that occur (but see Creel et al. 2003; McKelvey and Schwartz 2004). Incorporating genotyping error into data analysis would represent a paradigm shift for the non-invasive literature. Here, we investigate whether likelihood based methods can be used to sort non-invasive specimens by their identity. The task is not easy; three substantial problems must be solved. First, statistical models of genotyping error must be identified. This is challenging because, to be done well, the correct genotypes of non-invasive specimens must be known. Second, the parameters in such models must be estimated. This is challenging because each specimen is likely to have at least one parameter describing how likely errors will be in that specimen. If there are 100 specimens in a collection, there will be over 100 parameters to estimate – and this is computationally difficult. Third, an algorithm is needed to sort specimens according to their identity. This is challenging because, even small numbers of specimens can be sorted in too many ways to enumerate.

Solving these three problems will require a concerted effort by the non-invasive DNA community. Here, we address the main statistical challenges (the second and third points listed above), and show that even data sets having high genotyping error rates have enough information to identify individuals accurately.

An algorithm for individual identification

A DNA census seeks to estimate the number of individuals in a population. In this paper, we address a more limited question: which specimens in a collection came from the same individuals? Our approach is divided into three steps. First, a model of genotyping error is selected. This may be done on the basis of background knowledge or by model identification from a suite of alternative models (Burnham and Anderson 2002; Johnson and Omland 2004). Second, the parameters of the model are estimated. These will be genotyping error rates and parameters that affect these rates. For example, in the model we present as an example, dropout and misprint rates are estimated for every specimen. Third, and last, specimens are clustered into sets using the estimates of genotyping error rates to evaluate the evidence of identity. We begin by discussing this clustering algorithm, and then discuss the specific genotyping error model that we used to test its effectiveness.

Calculating the evidence that two specimens came from the same individual

When genotyping errors are possible, the term "genotype" can be ambiguous. Where there is the possibility of confusion, we will refer to a true underlying genotype of a specimen as the latent genotype, and a scored or measured genotype as an observed genotype.

The goal of our algorithm is to sort specimens into sets that are each derived from unique individuals. The algorithm begins with each specimen in a set by itself (i.e., a singleton set), and proceeds by calculating the evidence that pairs of sets contain specimens from the same individual (as opposed to different individuals). If this evidence is high, two sets of specimens will then be combined. Essentially, this is an exercise in estimating the relationship between specimens. Let Ω_h represent the hth set of specimens. Let the variable R_{h_1,h_2} represent the relationship between the specimens in sets Ω_{h_1} and Ω_{h_2}

$$R_{h_1,h_2} \in \{\text{SI, U, PO, FS}\} \quad (1)$$

where SI is an abbreviation for "same individual," U for "unrelated individuals," PO for "parent/offspring," and FS for "full sibs". Other relationships between specimens are possible (e.g., half sibs or cousins), but these relationships are intermediate between U and PO or U and FS so we will not consider them.

In order to calculate the likelihood of R_{h_1,h_2}, we need to calculate the probability of the observed genotypes in sets Ω_{h_1} and Ω_{h_2}. Let the vector $\mathbf{g_{ij}}$ represent the genotypes observed at the jth locus of the ith specimen. Let k_j represent a potential latent genotype for the jth locus, and let $P_{\mathbf{g_i}|k_j}$ represent the probability of observing $\mathbf{g_{ij}}$ from k_j. $P_{\mathbf{g_{ij}}|k_j}$ will be estimated from a model of genotyping error that is either assumed from previous experience or identified and fitted with the data of the study of

interest (see below for an example of the latter approach). Let the vector $\mathbf{G_{jh}}$ represent all of the genotypes observed, at the jth locus, for all the specimens in Ω_h. Let $P_{\mathbf{G_{jh}}|k_j}$ represent the probability of observing these genotypes from the latent genotype k_j

$$P_{\mathbf{G_{hj}}|k_j} = \prod_{i \in \Omega_h}^{\text{samples in } \Omega_h} P_{\mathbf{g_{ij}}|k_j}. \tag{2}$$

The likelihood of R_{h_1,h_2} is calculated by summing over all possible latent genotypes for both Ω_{h1} and Ω_{h_2} and multiplying across independent loci

$$L\left(R_{h_1,h_2}\right) = \prod_{j}^{\text{loci}} \left\{ \sum_{k_{j_1}}^{\text{latents for } \Omega_{h_1}} \sum_{k_{j_2}}^{\text{latents for } \Omega_{h_2}} \left[P_{k_{j_1}k_{j_2}|R_{h_1h_2}} P_{\mathbf{G_{h_1j}}|k_{j_1}} P_{\mathbf{G_{h_2j}}|k_{j_2}} \right] \right\}, \tag{3a}$$

where $P_{k_{j_1}k_{j_2}}|R_{h_1,h_2}$ is the probability of observing the latent genotypes k_{j1} and k_{j_2} in two specimens whose relationship is R_{h_1,h_2}. We can estimate $P_{k_{j_1}k_{j_2}|R_{h_1h_2}}$ from the allele frequencies in the population if, we assume random mating (e.g., Thompson 1991). When $R_{h_1,h_2} = \text{SI}$, equation (3a) reduces to

$$L\left(R_{h_1,h_2}=\text{SI}\right) = \prod_{j}^{\text{loci}} \left[\sum_{k_j}^{\text{latents}} \left(P_{k_j} P_{\mathbf{G_{h_1j}}|k_j} P_{\mathbf{G_{h_2j}}|k_j} \right) \right]. \tag{3b}$$

Now we can compare the likelihoods of different relationships between sets of specimens, and use these likelihoods to calculate the evidence that two sets of specimens came from the same individual. Following Royall (1997, 2004), we define the evidence that specimens in Ω_{h_1} and Ω_{h_2} came from the same individual, $\text{EI}(h_1, h_2)$, as the ratio of the likelihood that they came from one individual with the likelihood that they came from two individuals. In our framework, if the sets of specimens came from two individuals, the individuals must be either: unrelated (U), parent/offspring (PO), or full-sibs (FS). The evidence of identity is then

$$\text{EI}(h_1,h_2) \equiv \frac{L\left(R_{h_1,h_2}=\text{SI}\right)}{\text{MAX}\left[L\left(R_{h_1,h_2}=\text{U}\right), L\left(R_{h_1,h_2}=\text{PO}\right), L\left(R_{h_1,h_2}=\text{FS}\right)\right]}. \tag{4}$$

where, the likelihoods are given by equation (3). If $\text{EI}(h_1, h_2)$ is greater than 1, there is evidence that the two sets of specimens came from the same individual (See Mellen and Royall 1997, for a discussion of this definition in forensic identification).

Clustering algorithm

Specimens can be clustered by their individual identity with the following algorithm. (1) Estimate the allele frequencies of the population. (2) Estimate the latent genotype frequencies in the population. (3) Estimate the probability of observed genotypes from latent genotypes $P_{\mathbf{gij}}|k_j$ using an appropriate model of genotyping error. (4) Place each specimen into a singleton set. (5) Calculate $\text{EI}(h_1,h_2)$ for all pairs of sets. (6) Identify the pair of sets for which $\text{EI}(h_1,h_2)$ is highest and call the evidence that these two sets of specimens came from the same individual $\text{EI}_{\max}$. (7) If $\text{EI}_{\max}$ is greater than 1.0, combine these two sets and return to step 5. If $\text{EI}_{\max}$ is less than 1.0, stop. We call this algorithm the Evidence-of-Identity-Clustering Algorithm or EIC algorithm.

A model for genotyping error

The EIC algorithm requires a probabilistic model of genotyping error. More specifically, it requires the probability that a latent genotype k_j is scored as $\mathbf{g_{ij}}$. Recent work on genotyping error in non-invasive samples has emphasized estimating genotyping error rates (e.g., Bonin et al. 2004; Broquet and Petit 2004), but has not developed statistical models of genotyping error. Therefore, we used a reasonably complex heuristic model to test the EIC algorithm. The model we use has two types of genotyping error and assumes that the rates of these errors vary across samples and loci.

Two types of genotyping error are common with non-invasive specimens: dropout and misprinting (e.g., Taberlet et al. 1996; Gagneux et al. 1997). Allele dropout is the failure of one or more alleles in a specimen to amplify because of low concentrations of DNA in the specimen or because of differential amplification of one allele (e.g., the genotype *ab* is scored as either *aa* or *bb*) (Wattier et al. 1998). Misprinting (in the context of this paper) is a PCR artifact that causes a microsatellite

allele to be scored as one repeat motif shorter or longer than the actual allele (e.g., the microsatellite allele 100 is scored as 98 or 102, assuming a dinucleotide repeat motif).

Miller et al. (2002) have presented a statistical model for dropout errors in multilocus genotypes, and have shown how to obtain maximum likelihood estimates of the dropout rate. We extend their model to include single step misprinting. We define the dropout rate, d, as the probability that a latent heterozygote is scored as a homozygote for one of the two alleles in the heterozygote (Note that this assumes that both alleles do not drop out). We assume that error rates vary across specimens and loci. Let d_{ij} represent the dropout rate at the jth locus in the ith specimen. Following Miller et al. (2002), we assume that the dropout rates at different loci are related by $d_{ij} = d_i c_j$ where, d_i is a specimen specific number between zero and one, and c_j is a locus specific number between zero and one. For simplicity, we assume that both alleles in a heterozygote have the same probability, $d_{ij}/2$, of dropping out.

Our model of misprinting is analogous to the single step model of mutation for microsatellite loci (See Jarne and Lagoda 1996 for review). We assume that each allele has a probability of m of being misread by one repeat motif, and that misprinting is equally likely to lead to a smaller allele as to a larger allele. As with dropout rates above, we assume that the misprint rate for each locus is equal to $m_{ij} = m_i c_j$ (where, i indexes specimens and j loci).

Last, we assume that a genotype at one locus may have two errors: for example, a dropout and a misprint or two misprints. With these assumptions, we can formulate the probability of observing any genotype from a latent genotype (Table 1). For example, the probability that the latent genotype 100/106 is scored as a 100/104 (assuming a dinucleotide repeat motif) is equal to the probability that dropout does not occur $(1 - d_{ij})$ times the probability that a misprint does not occur for allele 100 $(1 - m_{ij})$, times the probability that allele 106 is scored as 104 $(\frac{m_{ij}}{2})$.

Maximum likelihood estimation of d, m and c

Next we present a maximum likelihood method for estimating d_{ij} and m_{ij}. We start by calculating the

Table 1. Probabilities of observing all possible genotypes from the latent genotype $a_x\, a_y$ as a function of the locus specific dropout rate (d) and locus specific misprint rate (m)

Observation	Latent genotype:$a_x\, a_y$			
	$x = y$	$y - x = 1$	$y - x = 2$	$y - x > 2$
$a_{x-1}a_{x-1}$	$\left(\frac{m}{2}\right)^2$	$\frac{d}{2}\frac{m}{2}$	$\frac{d}{2}\frac{m}{2}$	$\frac{d}{2}\frac{m}{2}$
$a_{x-1}a_x$	$2\left(\frac{m}{2}\right)(1-m)$	$(1-d)\frac{m}{2}\frac{m}{2}$	0	0
$a_{x-1}a_{x+1}$	$2\left(\frac{m}{2}\right)^2$	–	$(1-d)\frac{m}{2}\frac{m}{2}$	0
$a_{x-1}a_{y-1}$	–	–	–	$(1-d)\left(\frac{m}{2}\right)\left(\frac{m}{2}\right)$
$a_{x-1}a_y$	–	$(1-d)\frac{m}{2}(1-m)$	$(1-d)\frac{m}{2}(1-m)$	$(1-d)\frac{m}{2}(1-m)$
$a_{x-1}a_{y+1}$	–	$(1-d)\frac{m}{2}\frac{m}{2}$	$(1-d)\frac{m}{2}\frac{m}{2}$	$(1-d)\left(\frac{m}{2}\right)\left(\frac{m}{2}\right)$
a_xa_x	$(1 - \mathrm{m})^2$	$\frac{d}{2}(1-m) + \frac{d}{2}\frac{m}{2} + (1-d)(1-m)\frac{m}{2}$	$\frac{d}{2}(1-m)$	$\frac{d}{2}(1-m)$
a_xa_{x+1}	$2\left(\frac{m}{2}\right)(1-m)$	–	$(1-d)(1-m)\frac{m}{2}$	0
a_xa_{y-1}	–	–	–	$(1-d)\frac{m}{2}(1-m)$
a_xa_y	–	$(1-d)(1-m)(1-m) + (1-d)\frac{m}{2}\frac{m}{2}$	$(1 - \mathrm{d})(1 - \mathrm{m})^2$	$(1 - \mathrm{d})(1 - \mathrm{m})^2$
a_xa_{y+1}	–	$(1-d)(1-m)\frac{m}{2}$	$(1-d)(1-m)\frac{m}{2}$	$(1-d)\frac{m}{2}(1-m)$
$a_{x+1}a_{x+1}$	$\left(\frac{m}{2}\right)^2$	–	$\frac{d}{2}\frac{m}{2} + \frac{d}{2}\frac{m}{2} + (1-d)\frac{m}{2}\frac{m}{2}$	$\frac{d}{2}\frac{m}{2}$
$a_{x+1}a_{x-1}$	–	–	–	$(1-d)\left(\frac{m}{2}\right)\left(\frac{m}{2}\right)$
$a_{x+1}a_y$	–	–	$(1-d)(1-m)\frac{m}{2}$	$(1-d)\frac{m}{2}(1-m)$
$a_{x+1}a_{y+1}$	–	–	$(1-d)\frac{m}{2}\frac{m}{2}$	$(1-d)\left(\frac{m}{2}\right)\left(\frac{m}{2}\right)$
$a_{y-1}a_{y-1}$	–	–	–	$\frac{d}{2}\frac{m}{2}$
$a_{y-1}a_y$	–	–	–	0
$a_{y-1}a_{y+1}$	–	–	–	0
a_ya_y	–	$\frac{d}{2}(1-m) + \frac{d}{2}\frac{m}{2} + (1-d)(1-m)\frac{m}{2}$	$\frac{d}{2}(1-m)$	$\frac{d}{2}(1-m)$
a_ya_{y+1}	–	$(1-d)\frac{m}{2}\frac{m}{2}$	0	0
$a_{j+1}a_{j+1}$	–	$\frac{d}{2}\frac{m}{2}$	$\frac{d}{2}\frac{m}{2}$	$\frac{d}{2}\frac{m}{2}$

likelihood of the genotypes observed at the *j*th locus in the *i*th specimen. Let us assume, with no loss of generality, that this locus has been genotyped t_{ij} times. Recall that the genotypes observed at the *j*th locus in the *i*th specimen are represented by the vector $\mathbf{g_{ij}}$. If the t_{ij} genotypes observed at this locus are statistically independent from each other, the probability of observing $\mathbf{g_{ij}}$ from the latent genotype k_j, $P_{\mathbf{gij}}|k_j$, is multinomial with probabilities given by Table 1. Following Miller et al. (2002), we calculate the unconditional probability of observing $\mathbf{g_{ij}}$ by summing over all possible latent genotypes for the locus, and weighting by the probability of each latent occurring in the population:

$$P\left(\mathbf{g_{ij}}|d_{ij}, m_{ij}\right) = \sum_{k_j}^{\text{latents}} P_{k_j} P_{\mathbf{g_{ij}}|k_j} \tag{5}$$

where, P_{k_j} is the probability of observing latent genotype k_j in the population. In practice, P_{k_j} is unknown, but can be estimated from the allele frequencies if we assume Hardy–Weinberg proportions.

Equation (5) shows the marginal probability for one locus in one specimen. The joint probability for all the genotypes observed from a specimen, and for all the specimens observed in a study, is calculated by multiplying across loci and specimens (See Mellen and Royall 1997). Let the vector **G** represent all the data observed in a study. The likelihood of the parameters given **G** is then

$$L(\mathbf{d},\mathbf{m},\mathbf{c}|\mathbf{G}) = \prod_{i}^{\text{samples}} \left[\prod_{j}^{\text{loci}} \left(\sum_{k_j}^{\text{latents}} P_{k_j} P_{g_{\mathbf{ij}}|k_j} \right) \right]. \tag{6}$$

where the vectors **d**, **m**, and **c** specify the dropout and misprint rates for specimens and loci.

Maximum likelihood estimates of **d**, **m**, and **c** are obtained by finding the values of **d**, **m**, and **c** that maximize equation (6). Our experience suggests estimating d_i and for every specimen, and c_j for every locus is difficult. This is because, there are a large number of parameters to estimate, and because the likelihood surface has many peaks. We have found it useful to reduce the dimension of the problem by binning specimens and loci into groups with similar error rates, and assigning all the specimens in a bin a single rate. Specimens and loci are each binned separately. Appendix A describes a simple method to do this, and Appendix B describes how to estimate **d**, **m**, and **c** once the data is binned.

Testing the algorithm

We used computer simulation to examine how the following variables affected the performance of the EIC algorithm: genotyping error rate, number of PCR replicates per specimen, number of loci genotyped, number of alleles at each locus, number of specimens genotyped, and number of individuals sampled (note: *number of individuals* refers to the number of individuals sampled not the number of individuals in the population). For each of these six variables, we tested low, intermediate, and high values (Table 2 lists the specific values used).

The simulation procedure is illustrated with an example. Consider the case that we used as a standard for comparison: 100 specimens from 50 individuals, 4 PCR replicates per specimen, 6 loci genotyped, 6 alleles per locus, "average" data quality. To begin, we simulated multilocus genotypes for the 50 sampled individuals. While doing this, we assumed the 50 individuals represented 10 families of five individuals (dam, sire, and three offspring). We simulated the allele frequencies in the population with broken stick random numbers (Devroye 1986), and then drew alleles from this distribution to create the genotypes of the dam and sire of each family. Then we simulated Mendelian

Table 2. Parameters used to simulate dropout and misprint rates. The dropout rate for each locus was equal to $d_i\, c_j$ where d_i is a specimen specific parameter drawn from a beta distribution, $Beta(\alpha_{\text{sample}}, \beta_{\text{sample}})$, and c_j is a locus specific parameter drawn from a beta distribution, $Beta(\alpha_{\text{loci}}, \beta_{\text{loci}})$. See Figure 1 for graphs of these distributions. The misprint rate, m_i, for each specimen was equal to one half of d_i

	Specimen quality		
	Good	Average	Poor
α_{sample}	1.25	2.5	5
β_{sample}	8.75	7.5	5
α_{loci}	20	5	2
β_{loci}	20	5	2
$E(d_i)$	0.125	0.25	0.5
$E(m_i)$	0.063	0.12	0.25
$E(d_i\, c_j)$	0.032	0.125	0.25
$E(m_i\, c_j)$	0.015	0.063	0.125

segregation to create the genotypes of the three offspring per family. Next, we simulated the origin of each of the 100 specimens. While doing this we assumed that each of the 50 individuals was sampled at least once, and then randomly drew individuals for the remaining 50 specimens (this allowed us to control the number of individuals contributing to a set of specimens).

In the model of genotyping error described above, the dropout rate for the jth locus in the ith individual is equal to $d_i\, c_j$. We obtained values for d_i and c_j by drawing numbers from beta distributions for each specimen and for each locus (Table 2; Figure 1). This product is approximately beta distributed (Fan 1991). We obtained values for m_{ij} by assuming m_i was equal to half of d_i (we assumed that the misprint rate for a specimen was one half of the dropout rates because, dropout rates are usually higher than misprint rates and because the error rates should be correlated). Table 2 lists the parameters of the beta distributions that we used and their expected values. Figure 1 shows their distributions. For example, data of "average" quality had an expected dropout rate of 0.125 and an expected misprint rate of 0.0625. Once genotyping error rates for each specimen and each locus were obtained, the model described above was used to simulate genotyping errors.

Simulated data was analyzed with the EIC algorithm described above. In order to estimate **d**, **m**, and **c**, we sorted specimens into seven bins and loci into 3 bins using the method described in Appendix A. Maximum likelihood estimates were obtained using the maximization technique described in Appendix B.

One hundred simulations were performed for each of the combinations of parameters listed in Table 2 (100 simulations are less than ideal, but the algorithm is computationally intensive). Three statistics were calculated to evaluate the accuracy of the algorithm: average estimate, average proportional error, and percentage of genotypes sorted correctly. The first, average estimate, is the average of the estimated number of individuals contributing to a collection of specimens. The second, average proportional error, was calculated as the average value of

$$\frac{\left|N_{\text{genotypes}} - \hat{N}_{\text{genotypes}}\right|}{N_{\text{genotypes}}} \qquad (7)$$

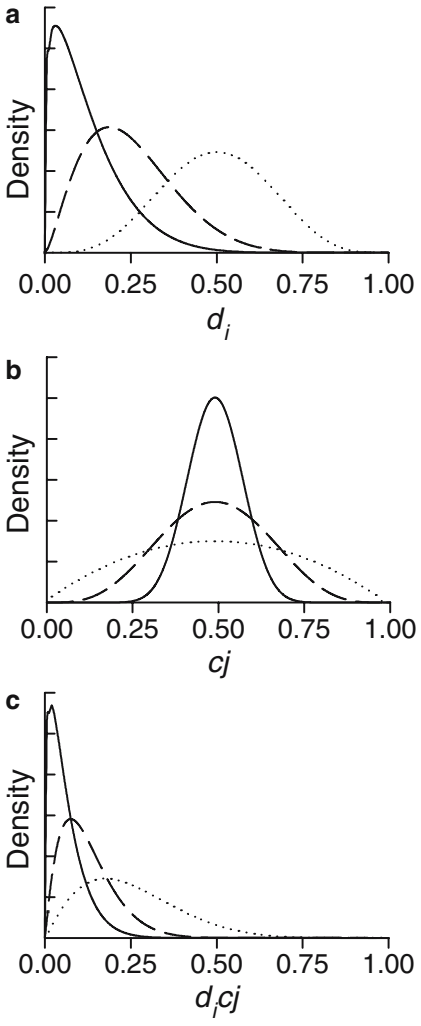

Figure 1. Beta distributions of dropout rates used in simulations. Solid, dashed, and dotted lines show distributions for data having high, average, and poor quality (respectively). The dropout rate for each locus was equal to $d_i\, c_j$ where d_i is a specimen specific parameter drawn from (a) and c_j is a locus specific parameter drawn from (b). Figure 1c shows the approximate distribution of the product $d_i\, c_j$.

observed in the simulated data, where $N_{\text{genotypes}}$ is the number of unique multilocus genotypes among the individuals sampled and $\hat{N}_{\text{genotypes}}$ is the estimate of $N_{\text{genotypes}}$ produced by the EIC algorithm. The third statistic, percentage of genotypes sorted correctly, is equal to the number of genotypes sorted correctly divided by the total number of multilocus genotypes among the individuals. A genotype was considered to be sorted correctly if

all specimens with the same multilocus genotype (and no others) were placed in the same set.

Results

The EIC algorithm did an excellent job sorting specimens: error rates were less than 2% for realistic amounts of data (Table 3). Its performance was positively correlated with the quality of the data, the number of replicates per specimen, the number of loci, the number of alleles per locus, and the number of specimens collected. Note that EIC algorithm has the desirable property of doing better when more data is collected (i.e., more loci, more alleles per locus, or more specimens). This consistency is not shared by genotype counting methods that assume that genotypes are error free – increasing the number of specimens (or loci) is expected to increase the chance of making mistakes (e.g., Waits and Leberg 2000). Note also that the EIC algorithm did extremely well with error free data (the average error was less than 0.1%). Using this method, therefore, with data that has no errors does not appear to sacrifice the quality of the clustering. Last, note that large populations (200 individuals) were just as effectively sorted as were small populations (50 individuals).

The least desirable property of the EIC algorithm is that it requires that each specimen be

Table 3. Performance of the EIC algorithm with simulated data

N[a]	Number of specimens	Number of PCRs[b]	Number of loci	Number of alleles	Data quality[c]	Average estimate	Average |error|	Percent genotypes correct
Experiment i: Data quality varied								
50	100	4	6	6	Poor	48.6	2.4%	95.1%
"	"	"	"	"	Avg.	49.1	1.4%	97.1%
"	"	"	"	"	Good	49.5	0.6%	98.8%
"	"	"	"	"	Perfect	49.8	< 0.1%	> 99.9%
Experiment ii: Number of PCRs varied								
50	100	2	6	6	Avg.	47.6	4.4%	90.0%
"	"	3	"	"	"	48.6	2.5%	95.0%
"	"	4	"	"	"	49.1	1.4%	97.1%
"	"	8	"	"	"	49.7	0.2%	99.6%
Experiment iii: Number of loci varied								
50	100	6	3	6	Avg.	44.9	2.7%	93.4%
"	"	"	6	"	"	49.1	1.4%	97.1%
"	"	"	12	"	"	> 49.9	< 0.1%	99.9%
Experiment iv: Number of alleles varied								
50	100	6	6	3	Avg.	45.4	5.0%	88.4%
"	"	"	"	6	"	49.1	1.4%	97.1%
"	"	"	"	9	"	49.7	0.5%	98.9%
Experiment v: Number of specimens varied								
50	50	6	4	6	Avg.	48.4	2.8%	94.4%
"	100	"	"	"	"	49.1	1.4%	97.1%
"	200	"	"	"	"	49.7	0.2%	99.2%
"	400	"	"	"	"	49.8	< 0.1%	99.3%
Experiment vi: Number of individuals varied								
10	20	6	4	6	Avg.	9.9	0.4%	99.2%
50	100	"	"	"	"	49.1	1.4%	97.1%
100	200	"	"	"	"	98.1	1.5%	97.0%
200	400	"	"	"	"	196.5	1.4%	97.2%

[a]The number of individuals represented in the set of specimens.
[b]The number of times each specimen was genotyped.
[c]See Table 2 and Figure 1 for simulation parameters and expected values. "Perfect" indicates that simulated data had no genotyping errors.

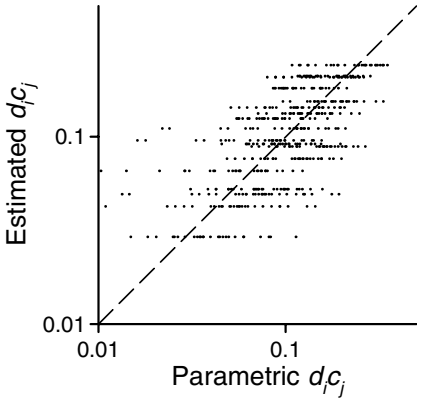

Figure 2. Parametric and estimated dropout rates for each locus in a data set containing 100 specimens, four PCRs per specimen, six loci per specimen, and six alleles per locus. The quality of the data was "Average" (defined in Table 2). Specimens were sorted into seven bins, and loci into three bins, before estimating d_i and c_j.

genotyped at least three- and preferably four-times. However, repeatedly genotyping all specimens to detecting genotyping errors is currently standard practice for non-invasive specimens (See McKelvey and Schwartz 2004 for a brief review), so this necessity is not especially burdensome (but see Paetkau 2003, 2004). If specimen effects were assumed negligible, genotypings per specimen might be reducible. However, because specimen effects are known to be important, we have not pursued development in this direction.

The EIC algorithm requires estimates of **d**, **m**, and **c** to cluster specimens. Therefore, we also informally compared estimates of **d**, **m**, and **c** with the parametric values used in the simulations. Figure 2 shows estimates of the product $d_i\ c_j$ for one set of simulated data. The estimates are slightly biased, but are close enough to the parametric values that the EIC algorithm clustered all specimens correctly for this simulated data set.

Discussion

We have used a hypothetical model of genotyping error to test the EIC algorithm. This is the main drawback of our study, and, as such, deserves comment. There are three points to consider. First, there are no statistical models of genotyping errors available in the literature that we could use to test our algorithm. Second, the EIC algorithm will work with any model of genotyping error, so should be useful once models have been identified. Third, the heuristic model that we used is the most realistic model in the literature to date. For example, Wang (2004) has developed an error tolerant algorithm for partitioning individuals into sibships, but assumed that error rates were constant across individuals and loci – and were known *a priori*.

Most efforts to estimate genotyping error rates have assumed that the latent genotype can be inferred correctly if a specimen is genotyped enough times (e.g., Taberlet et al. 1996; Paetkau 2003). For example, Taberlet et al. (1996) used worst-case scenarios to argue that if a specimen is genotyped three times and {*aa*, *ab*, *bb*} is observed, the correct genotype is almost certainly *ab*. Once the correct genotype is inferred, the number of dropouts and misprints can be counted to calculate error rates (See Broquet and Petit 2004 for a review of 19 studies using methods based on such reasoning). Such estimation is straightforward, but has two drawbacks: it relies on professional judgment to ascertain the correct genotype and it depends heavily on the assumption that the consensus genotype is correct.

Maximum likelihood is logical alternative to professional judgment. The statistical properties of maximum likelihood estimation are extremely well known, and its application can be consistent from study to study. A question arises: which method (professional judgment or maximum likelihood) is best? This answer: we do not know. Maximum likelihood estimation is buttressed by a voluminous statistical literature. Professional judgment takes advantage of subtle visual clues present in the genotyping process that current maximum likelihood models do not use, so might work better than judgment. However, comparing two genotypes and deciding whether they come from the same individual often requires weighing alternative probabilities of errors, and making such decisions is notoriously difficult (e.g. Zeckhauser and Viscusi 1990). Of course, professional judgment and likelihood based approaches are not mutually exclusive, and a combination of methods is likely to work best (Lele 2004).

Once genetic errors are recognized, the next challenge is what to do about them. The conventional approach has been to reduce the frequency of unrecognized errors to a level low enough that the data can be considered error free (e.g., Paetkau

2003, 2004). The main drawback to this approach is that even modest unrecognized error rates can have devastating effects upon a DNA census (Creel et al. 2003). And to make matters worse, demonstrating that a data set is free from errors is difficult (McKelvey and Schwartz 2004). Paetkau's 1 MM checks (2003; 2004) and the tests of McKelvey and Schwartz (2004) will detect some – if not most – errors, but their effectiveness requires further validation.

There are several reasons to believe an error tolerant matching algorithm might produce better results for less cost than conventional methods. First, error tolerant approaches are, by definition, less sensitive to genotyping errors. Second, they may be able to use low quality specimens that would be removed from analysis using stringent genotyping protocols (e.g., Paetkau 2003). Third, an error tolerant approach might save labor costs by eliminating the need to establish consensus genotypes for all samples. Fourth, error tolerant approaches have proven useful in the paternity testing literature (e.g., Marshall et al. 1998; Constable et al. 2001). Fifth, and last, error tolerant algorithms facilitate using large numbers of loci to estimate relatedness accurately.

Conclusions

Our simulations show that error-ridden genotypes can have enough information to accurately sort specimens by individual identity. Our method, therefore, has promise. However, our work here is mostly a proof-of-concept. The dropout/single-step-misprinting model of genotyping error that we used in the simulations seems reasonable and may be useful in practice – nevertheless, its use here has been to demonstrate the utility of the EIC approach. The specific model still requires empirical validation. We recommend that this model and a suite of other genotyping error models be tested (such as the five parameter model of Sobel et al. 2002), and the best model used in the EIC algorithm.

Acknowledgements

This research has been supported by NSF grant DEB-0415932 (MLT). We would like to thank Subhash Lele and three anonymous reviewers for useful comments on an earlier version of this manuscript. We would also like to thank Robert Boik for helpful discussions on optimizing complex constrained problems.

Appendix A. Binning specimens and loci according to number of mismatches observed between replicated genotypes

Specimens potentially could be binned according to many different criteria (e.g., DNA concentration, percentage of missing genotypes, hair vs. faeces). Here we show how genotype inconsistency – measured by allelic mismatches during repeated genotyping – can be used to sort specimens.

Let the function MM(•) indicate the number of allelic mismatches between two genotypes: MM(*aa,aa*) = 0, MM(*aa,ab*) = 1, MM(*aa,bb*) = 2, MM(*aa,bc*) = 2, MM(*ab,ab*) = 0, MM(*ab,ac*) = 1, MM(*ab,cd*) = 2. Let T_{MM} represent the total number of allelic mismatches between one genotype and a set of genotypes.

An example shows how T_{MM} is useful to bin specimens. Consider a locus in a specimen that has been genotyped four times. The genotypes observed are [*aa*, *aa*, *ab*, *ab*]. Let us assume there are three alleles at this locus (*a*, *b*, and *c*). Because, there are three alleles at this locus, there are six possible latent genotypes [*aa*, *ab*, *ac*, *bb*, *bc*, *cc*]. Table A1 shows T_{MM} for the observed genotypes and each possible latent genotype. Let Min(T_{MM}) represent the minimum value of T_{MM}. For example, in Table A1, Min(T_{MM}) = 2.

Values of Min(T_{MM}) can be summed across loci to find the minimum number of allelic mismatches for each specimen in a study. Specimens can then be ranked and divided into bins. The same can be done for loci.

Table A1. Potential latent genotypes and the number of allelic mismatches between them and the set of four observed genotypes [*aa*, *aa*, *ab*, *ab*]

Potential latent genotypes	T_{MM} between latent and [*aa*, *aa*, *ab*, *ab*]
aa	2
ab	2
ac	4
bb	6
bc	6
cc	8

Appendix B. Estimating d, m, and c

The EIC algorithm requires estimates of $d_i\ c_j$ and $m_i\ c_j$ for every locus in each specimen. One obstacle to the estimation of **d**, **m**, and **c** is that these products confound specimen specific and locus specific error rates. For example, $(0.5)(0.3) = (0.3)(0.5)$. Basically, there is only sufficient information in the system to identify the relative error rates of specimens, the relative error rates of loci, and an overall error rate. For clarity of communication, we have chosen to combine overall rate and specimen relative rate into specimen rate and leave the loci effect as a relative rate, but standardized so that the maximum locus effect is 1. This gives us a specimen effect interpretable as the specimen's expected rate at the worst locus.

Algorithmically, we define $\mathbf{c}'$ as a vector of locus specific error rates relative to locus #1.

$$c'_j = \frac{c_j}{c_1} \tag{A1}$$

and find the values of $\mathbf{c}'$ that maximize equation (3a). Before being passed to the likelihood function, each $\mathbf{c}'$ vector is standardized

$$c''_j = \frac{c'_j}{\mathrm{MAX}(c')} \tag{A2}$$

before calculating the likelihood.

Considering the **d**, **m**, and **c** vectors, there are a large number of parameters to be estimated. Maximizing all parameters simultaneously would be cumbersome. We employ the Gauss-Sidell (Kincaid and Cheney 1991) algorithm to break the problem into a large number of maximizations of low dimension. Maximum likelihood values of **d**, **m**, and **c** are found as follows. First, $\mathbf{c}'$ is set to 1.0 for each locus. Then values of d_i and m_i are found that maximize the likelihood of each specimen given $\mathbf{c}'$. We have used the downhill simplex algorithm to do this (Press et al. 1992). Once values for **d** and **m** have been obtained, then the downhill simplex routine is used to find the maximum likelihood values of $\mathbf{c}'$ given **d** and **m**. During this step, the downhill simplex routine explores values of $\mathbf{c}'$, but the likelihood is calculated on $\mathbf{c}''$. When optimum values of $\mathbf{c}'$ have been found, **d** and **m** are again optimized given $\mathbf{c}'$. This continues until estimates converge. Because the object function increases monotonically with each step, and the maximum likelihood is a fixed point for the algorithm, the Gauss-Sidell algorithm will converge to local maxima of the likelihood.

References

Bonin A, Bellemain E, Bronken Eidesen P, Pompanon F, Brochmann C, Taberlet P (2004) How to track and assess genotyping errors in population genetic studies. *Mol. Ecol.*, **13**, 3261–3273.

Broquet T, Petit E (2004) Quantifying genotyping errors in noninvasive population genetics. *Mol. Ecol.*, **13**, 3601–3608.

Burnham KP, Andersen DR (2002) *Model Selection and Inference: A Practical Information-Theoretic Approach*, 2nd edn. Springer-Verlag, New York.

Creel S, Spong G, Sands JL, Rotella R, Zeigle J, Joe L, Murphy KM, Smith D (2003) Population size estimation in Yellowstone wolves with error-prone noninvasive microsatellite genotypes. *Mol. Ecol.*, **12**, 2003–2009.

Constable JL, Ashley MV, Goodall J, Pusey AE (2001) Noninvasive paternity assignment in Gombe chimpanzees. *Mol. Ecol.*, **10**, 1279–1300.

Devroye L (1986) *Non-uniform Random Variate Generation*, Springer-Verlag, New York.

Fan DY (1991) The distribution of the product of independent beta variables *Communications in Statistics – Theory and Methods*, **20**, 4043–4052.

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

Jarne P, Lagoda PJL (1996) Microsatellites, from molecules to populations and back. *TREE*, **11**, 424–429.

Johnson JB, Omland KS (2004) Model selection in ecology and evolution. *TREE*, **19**, 101–108.

Kincaid D, Cheney W (1991) *Numerical Analysis: Mathematics of Scientific Computing*, Brooks/Cole Publishing Company, Pacific Grove, California.

Kohn MH, York EC, Kamradt DA, Haught GH, Sauvajot RM, Wayne RK (1999) Estimating population size by genotyping faeces. *Proc. R. Soc. London. B.*, **266**, 657–663.

Lele SR (2004) Elicit Data, Not Prior: On Using Expert Opinion in Ecological Studies In: The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Considerations (eds. Taper ML, Lele SR), University of Chicago Press, ChicagoChapter 13.

Marshall TC, Slate J, Kruuk LEB, Pemberton JM (1998) Statistical confidence for likelihood-based paternity inference in natural populations. *Mol. Ecol.*, **7**, 639–655.

McKelvey KS, Schwartz MK (2004) Genetic errors associated with population estimation using non-invasive molecular tagging: Problems and new solutions. *J. Wildl. Manage.*, **68**, 439–448.

Mellen BG, Royall RM (1997) Measuring the Strength of Deoxyribonucleic Acid Evidence, and Probabilities of Implicating Evidence. *J. R. Statist. Soc. A.*, **160**, 305–320.

Miller CR, Joyce P, Waits LP (2002) Assessing allelic dropout and genotype reliability using maximum likelihood. *Genetics*, **160**, 357–360.

Morin PA, Chambers KE, Boesch C, Vigilant L (2001) Quantitative polymerase chain reaction analysis of DNA from noninvasive samples for accurate microsatellite genotyping of wild chimpanzees (Pan troglodytes verus). *Mol. Ecol.*, **10**, 1835–1844.

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

Paetkau D (2004) The optimal number of markers in genetic capture-mark-recapture studies *J. Wildl. Manage.*, **68**, 449–452.

Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) *Numerical Recipes in C*, Cambridge University Press, New York.

Royall RM (1997) *Statistical Evidence: a Likelihood Paradigm*, Chapman and Hall, London.

Royall R (2004) The likelihood paradigm for statistical evidence In: The Nature of Scientific Evidence: Empirical, Statistical and Philosophical Considerations (eds. Taper M.L., Lele S.R.), University of Chicago Press, Chicago.

Sobel E, Papp JC, Lange K (2002) Detection and integration of genotyping errors in statistical genetics. *Am. J. Hum. Genet.*, **70**, 496–508.

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

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

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

Taberlet P, Luikart G (1999) Non-invasive sampling and individual identification. *Biol. J. Linn. Soc.*, **68**, 41–55.

Thompson EA (1991) Estimation of relationships from genetic data. In: *Handbook of Statistics, Vol. 8* (eds. Rao CR, Chakraborty R), pp. 255–269. Elsevier Science Publishers.

Wang J (2004) Sibship reconstruction from genetic data with typing errors *Genetics*, **166**, 1963–1979.

Waits JL, Leberg PL (2000) Biases associated with population estimation using molecular tagging. *Anim. Conserv.*, **3**, 191–199.

Wattier R, Engel CR, Saumitou-Laprade P (1998) Short allele dominance as a source of heterozygote deficiency at microsatellite loci: experimental evidence at the dinucleotide locus Gv1CT in Gracilaria gracilis (Rhodophyta). *Mol. Ecol.*, **7**, 1569–1573.

Woods JG, Paetkau D, Lewis D, McLellan BN, Proctor M, Strobeck C (1999) Genetic tagging free ranging black and brown bears. *Wild. Soc. Bull.*, **27**, 616–627.

Zeckhauser RJ, WK Viscusi (1990) Risk within reason. *Science*, **248**, 559–564.