Molecular Ecology (2001) **10**, 249–256

Estimating the probability of identity among genotypes in natural populations: cautions and guidelines

LISETTE P. WAITS,*† GORDON LUIKART† and PIERRE TABERLET†

**Department of Fish and Wildlife Resources, University of Idaho, Moscow, ID 83844–1136, USA, †Laboratoire de Biologie des Populations d'Altitude, CNRS UMR 5553, Université Joseph Fourier, BP 53, F-38041 Grenoble cedex 9, France*

Abstract

Individual identification using DNA fingerprinting methods is emerging as a critical tool in conservation genetics and molecular ecology. Statistical methods that estimate the probability of sampling identical genotypes using theoretical equations generally assume random associations between alleles within and among loci. These calculations are probably inaccurate for many animal and plant populations due to population substructure. We evaluated the accuracy of a probability of identity ($P_{(ID)}$) estimation by comparing the observed and expected $P_{(ID)}$, using large nuclear DNA microsatellite data sets from three endangered species: the grey wolf (*Canis lupus*), the brown bear (*Ursus arctos*), and the Australian northern hairy-nosed wombat (*Lasiorinyus krefftii*). The theoretical estimates of $P_{(ID)}$ were consistently lower than the observed $P_{(ID)}$, and can differ by as much as three orders of magnitude. To help researchers and managers avoid potential problems associated with this bias, we introduce an equation for $P_{(ID)}$ between sibs. This equation provides an estimator that can be used as a conservative upper bound for the probability of observing identical multilocus genotypes between two individuals sampled from a population. We suggest computing the actual observed $P_{(ID)}$ when possible and give general guidelines for the number of codominant and dominant marker loci required to achieve a reasonably low $P_{(ID)}$ (e.g. 0.01–0.0001).

Keywords: DNA fingerprinting, match probability, microsatellites, noninvasive genetic sampling, population estimation, probability of identity

Received 23 May 2000; revision received 12 September 2000; accepted 12 September 2000

Introduction

In conservation genetics and molecular ecology, individual identification or 'DNA fingerprinting' is critical for a growing number of studies that use noninvasive genetic sampling to estimate population size (Taberlet *et al*. 1997; Kohn *et al*. 1999; Woods *et al*. 1999; Ernest *et al*. 2000), to monitor population sizes over time (Kendall *et al*. 1992; Schwartz *et al*. 1998), and to estimate the home range of individuals (Taberlet *et al*. 1997). Individual identification is also critical for molecular studies of clonal plants (Escaravage *et al*. 1998). In addition, the importance of DNA fingerprinting in wildlife forensics and law enforcement is increasing as individual identification is used to match illegally killed animals to samples obtained from poachers and to identify animals that attack livestock or humans (Waits *et al*. 1998). DNA fingerprinting to identify individuals is most frequently accomplished using a group of codominant nuclear DNA microsatellite loci (Bruford & Wayne 1993); however, dominant markers such as amplified fragment length polymorphisms (AFLPs) are also employed (Escaravage *et al*. 1998; Mueller & Wolfenbarger 1999).

For all of these applications, the power of multilocus DNA fingerprinting for identifying individuals can be quantified for individual loci by calculating the match probability on the basis of allele frequency data and Hardy–Weinberg expectation for random union of alleles (National Research Council 1996; Woods *et al*. 1999). The simplified version of this probability, p^2 for homozygotes, 2pq for heterozygotes at a diallelic locus, is the probability that a forensic genotype will match one sample drawn at random from the population and is calculated for multiple loci using the product rule (Li 1976). The forensic applications of DNA fingerprinting in human populations are

Correspondence: L. P. Waits. Fax: 208-885-9080; E-mail: lwaits@uidaho.edu

well established (Chakraborty & Kidd 1991; Jeffreys *et al.* 1991a,b; National Research Council 1992, 1996; Evett & Weir 1998); however, the estimation of match probabilities has been controversial in human populations because of potential biases in data sets due to violations of Hardy–Weinberg equilibrium and linkage disequilibrium (Lander 1989; Lewontin & Hartl 1991; Chakraborty & Jin 1992; Risch & Devlin 1992; Roeder 1994).

Table 1 Theoretical estimated and actual observed probability of identity, $P_{(ID)}$, for microsatellite data sets from brown bear, wolf and wombat populations. The mating system and family size are described because they can influence $P_{(ID)}$ estimates

Species and population name	No. of individuals sampled	No. of loci	Theoretical $P_{(ID)}$	Observed $P_{(ID)}$	Mating system and family sizes
Brown bear					Polygamous mating, multiple paternity litters possible, 1–4 sibs per family, many half-sibs in the population, cubs remain with mother for 1–3 years (Craighead *et al.* 1995; Nowack 1999)
Scandinavia — S	157	4	1.3×10^{-3}	3.1×10^{-3}	
		6	1.0×10^{-4}	7.3×10^{-4}	
		8	1.7×10^{-6}	1.6×10^{-4}	
Scandinavia — M	86	4	1.4×10^{-3}	2.2×10^{-3}	
		6	1.3×10^{-4}	5.5×10^{-4}	
		8	1.3×10^{-6}	2.7×10^{-4}	
Scandinavia — NS	107	4	3.9×10^{-4}	1.4×10^{-3}	
		6	7.1×10^{-6}	7.1×10^{-4}	
		8	1.9×10^{-7}	1.8×10^{-4}	
Wolf					Social monogamy, pack structure, 1–8 pups per pack (Forbes & Boyd 1997; Nowack 1999)
Montana	64	4	3.6×10^{-3}	5.5×10^{-2}	
		6	1.9×10^{-4}	2.5×10^{-2}	
		8	2.5×10^{-5}	8.2×10^{-3}	
Fort Saint Johns	41	4	2.2×10^{-3}	3.7×10^{-2}	
		6	9.3×10^{-5}	3.7×10^{-3}	
		8	1.7×10^{-5}	1.2×10^{-3}	
Hinton	33	4	3.3×10^{-3}	1.5×10^{-2}	
		6	3.3×10^{-4}	7.6×10^{-3}	
		8	3.7×10^{-5}	0.00	
Banff	32	4	2.8×10^{-3}	3.0×10^{-2}	
		6	3.7×10^{-4}	2.0×10^{-2}	
		8	1.6×10^{-5}	2.0×10^{-2}	
Hairy-nosed wombat					Polygamous, 1 offspring per year, burrow clusters contain ~10 animals with few/no full-sibs per cluster, same-sex animals in a burrow cluster are closely related (Taylor *et al.* 1994, 1997)
Epping Forest	28	4	2.3×10^{-3}	1.4×10^{-2}	
		6	4.3×10^{-4}	5.1×10^{-3}	
		8	1.7×10^{-4}	1.1×10^{-3}	
Theoretical data*					
$H = 0.6$ (all loci)		3	6.9×10^{-3}	NA	
		6	4.8×10^{-5}	NA	
		9	3.3×10^{-7}	NA	
$H = 0.6$ (0.4, 0.6, 0.8)		3	4.8×10^{-3}	NA	
		6	2.3×10^{-5}	NA	
		9	1.1×10^{-7}	NA	

Theoretical $P_{(ID)}$ is from eqn 2. Observed $P_{(ID)}$ is computed as the proportion of all possible pairs of individuals with identical multilocus genotypes. The most variable loci are used first, as is the case in most field studies. Data from Waits *et al.* (2000) (bears), Forbes & Boyd (1997) (wolves) and Taylor *et al.* (1994) (wombats).

NA, not available as computations are based on allele frequencies (eqn 2) and not on genotypes; *H*, heterozygosity.

*For $H = 0.6$ (all loci) all loci have same allele frequencies; for $H = 0.6$ (0.4, 0.6, 0.8) one-third of loci have allele frequencies giving $H = 0.4$, 0.6, and 0.8 (see Materials and methods for allele frequencies for each *H* value).

In such circumstances, the product rule, which assumes independence within and among loci, is violated, and the estimator will probably give a value for the probability of a match that is lower than the 'true' probability of a match (Donnelly 1995). In addition, theoretical evaluations of the match probability have demonstrated that the assumption of independence of loci can lead to a substantial underestimate of the match probability when comparing related individuals or populations with substructure even if the population is randomly mating (Nichols & Balding 1991; Donnelly 1995). Because many animal and plant populations are more likely than human populations to contain related individuals and to have substantial substructure, the accurate estimation of match probability is a serious concern for current and future studies in conservation genetics, molecular ecology and wildlife forensics.

Another individual identification estimator is the probability of identity, $P_{(ID)}$, or the probability that two individuals drawn at random from a population will have the same genotype at multiple loci. This theoretical estimator is increasingly used to access the statistical confidence for individual identification (Reed *et al.* 1997; Kohn *et al.* 1999; Mills *et al.* 2000; Waits & Leberg 2000) and to quantify genetic diversity levels in natural populations (Jamieson 1965; Paetkau & Strobeck 1994). This estimator is the square of the match probability for each genotype summed over all possible genotypes as it is comparing two individuals drawn at random from a population. $P_{(ID)}$ can be particularly useful when planning a study that requires individual identification as it can be estimated for differing numbers of loci without having the forensic genotype 'in hand'. To evaluate the usefulness of $P_{(ID)}$ for estimating the probability of identifying individuals, we compared the actual observed and theoretical expected $P_{(ID)}$ using large microsatellite data sets collected for brown bears (*Ursus arctos*) (Waits *et al.* 2000), grey wolves (*Canis lupus*) (Forbes & Boyd 1997), and Australian northern hairy-nosed wombats (*Lasiorinyus krefftii*) (Taylor *et al.* 1994; Taylor 1995). These data sets were chosen because they are among the largest microsatellite data sets for natural populations; samples collected from these species probably contain related individuals due to the social structure; noninvasive genetic sampling is currently being used to generate data sets for these species; and all three species are threatened or endangered (Table 1). The goals of this study were: (i) to evaluate the accuracy of the theoretical $P_{(ID)}$ estimator using data from natural populations; (ii) to provide an upper bound estimator for the number of loci necessary to identify individuals by deriving an equation for the $P_{(ID)}$ of sibs; and (iii) to provide general guidelines for the number of codominant and dominant marker loci required to obtain a reasonably low $P_{(ID)}$ (e.g. 0.01–0.0001).

Materials and methods

The theoretical expected $P_{(ID)}$ was computed for each locus using allele frequencies from a population sample and each of the following two equations:

$$P_{(ID)} = \Sigma p_i^4 + \Sigma\Sigma\,(2p_ip_j)^2 \qquad (1)$$

where p_i and p_j are the frequencies of the ith and jth alleles and $i \neq j$ (Paetkau & Strobeck 1994) and

$$P_{(ID)\text{unbiased}} = \frac{n^3(2a_2^2-a_4)-2n^2(a_3+2a_2)+n(9a_2+2)-6}{(n-1)(n-2)(n-3)} \qquad (2)$$

where n is the sample size, a_i equals $\Sigma p_j^{\,i}$, and p_j is the frequency of the jth allele (Paetkau *et al.* 1998). $P_{(ID)}$ was calculated for each locus and then multiplied across loci to give the overall $P_{(ID)}$. Eqn 1 is the biased formula because it does not correct for sample size differences, and eqn 2 is the unbiased formula with a sample size correction. Eqn 2 was used in all analyses of theoretical $P_{(ID)}$, but the results using eqn 1 are also illustrated in Fig. 1a. Because our sample sizes were fairly large, the results from both equations are very similar.

The observed $P_{(ID)}$ was estimated for each population by computing the proportion of all possible pairs of individuals that have identical genotypes. We chose the grey wolf, brown bear, and hairy-nosed wombat data sets because they are among the largest available from natural populations, and should provide meaningful estimates of $P_{(ID)}$ in natural populations. These data sets also provide a wide range of average expected heterozygosity values (Nei 1978) from approximately 0.40 in the wombats to 0.68 in the brown bears. We computed the theoretical expected and observed $P_{(ID)}$ for one-to-L loci (where L is the total), by sequentially adding one locus, and then re-computing the $P_{(ID)}$. Loci were added in order from the highest to the lowest level of heterozygosity as researchers will generally use the most variable loci first to minimize the number of loci needed to resolve individuals.

A simple equation was derived for estimating $P_{(ID)}$ among sibs ($P_{(ID)\text{sib}}$) for codominant loci, and the equation was later found in Evett & Weir (1998):

$$P_{(ID)\text{sib}} = 0.25 + (0.5\ \Sigma p_i^2) + [0.5(\Sigma p_i^2)^2] - (0.25\ \Sigma p_i^4) \qquad (3)$$

where p_i is the frequency of the ith allele. For related equations of match probability (e.g. sib–sib, parent–offspring), see Woods *et al.* (1999).

Because dominant marker systems such as AFLPs (Vos *et al.* 1995) can also be used to identify individuals based on a particular DNA profile, we included an analysis of dominant loci. The theoretical $P_{(ID)}$ expected for dominant genetic marker systems was computed as follows:

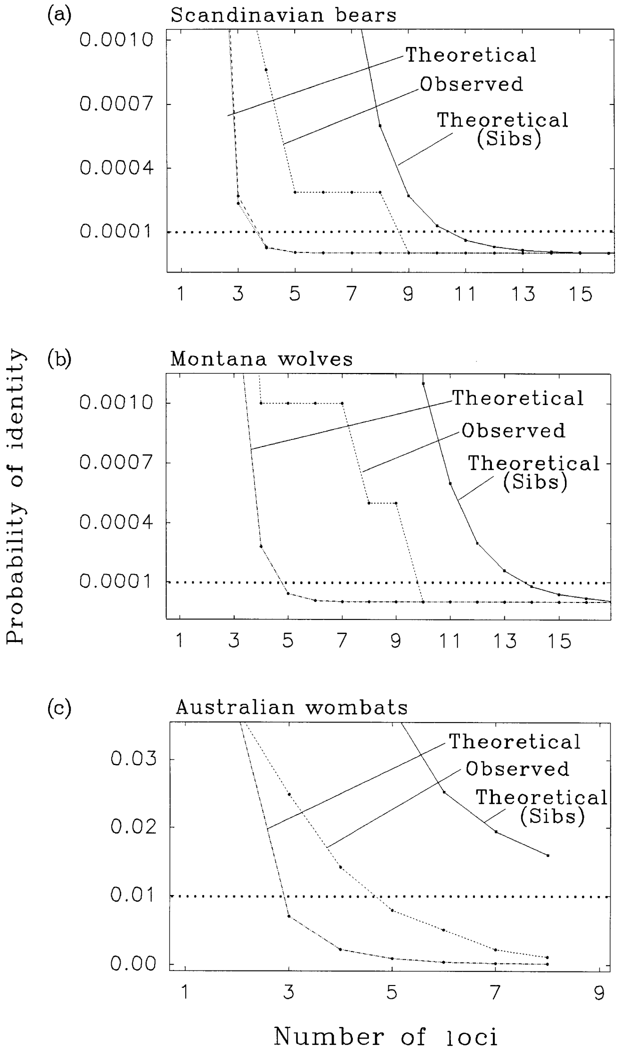

Fig. 1 Relationship between the theoretical, observed and sib probability of identity, $P_{(ID)}$, for nuclear DNA microsatellite loci in: (a) Scandinavian brown bear, (b) Montana wolf, and (c) Australian hairy-nosed wombat populations. Loci on the *x*-axis are arranged from highest (first) to lowest heterozygosity. Theoretical $P_{(ID)}$ is plotted using the unbiased (eqn 2) $P_{(ID)}$ estimator from Paetkau *et al.* (1998). The biased $P_{(ID)}$ estimator (eqn 1) (Paetkau & Strobeck 1994), was also tested and is plotted as a dashed line in (a) only. The observed $P_{(ID)}$ was calculated from a pairwise comparison of genotypes from 84 bears, 64 wolves and 28 wombats. The $P_{(ID)}$ for sibs was derived for this study (see Materials and methods). The horizontal line identifies a $P_{(ID)}$ of 0.0001 (1/10 000) for wolves and bears and a $P_{(ID)}$ of 0.01 (1/100) for wombats.

$$P_{(ID)dom} = p^2 + (2pq)^2 + (q^2)^2 \tag{4}$$

where p is the frequency of a 'present' band (allelic state 1) and q is the absence of a band (allelic state 2). The following equation was derived for estimating $P_{(ID)sib}$ for dominant loci:

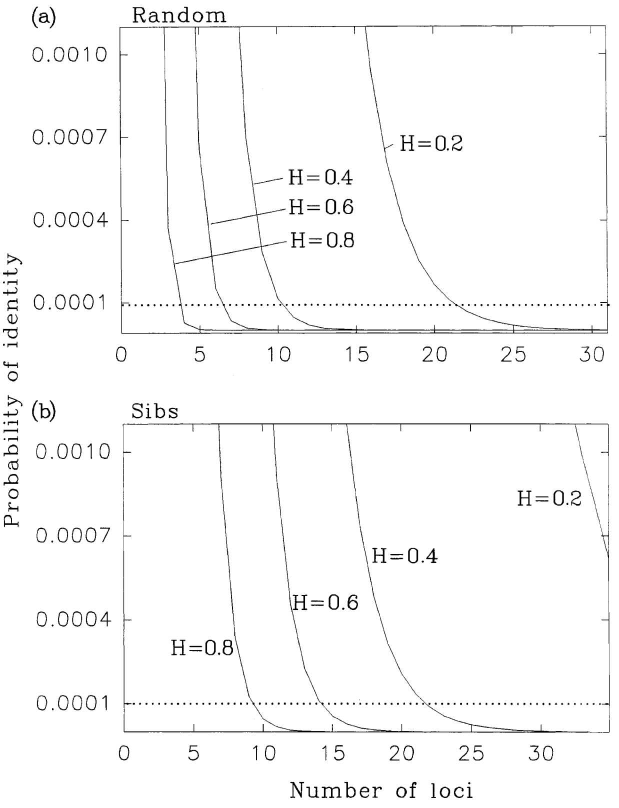

Fig. 2 Relationship between the theoretical probability of identity, $P_{(ID)}$, and the number of codominant loci assayed using four heterozygosity levels for (a) randomly sampled individuals and (b) sibs.

$$P_{(ID)dom-sib} = 1 - \{(3/2p)(q^2)\} \tag{5}$$

We examined microsatellite data sets from one wombat, three bear, and four wolf populations using these methods. For discussion purposes, we chose 0.01 (1 in 100) to 0.0001 (1 in 10 000) as an acceptably low $P_{(ID)}$. A $P_{(ID)}$ of less than approximately 0.01 will be necessary for studies requiring population size estimation (Mills *et al.* 2000) using mark–recapture models, and 0.001–0.0001 should be sufficiently low for most law enforcement forensic applications in natural populations (J. Coffin, S. Fain, personal communication).

To provide approximate guidelines (Figs 2 and 3) for researchers, we computed the theoretical $P_{(ID)}$ values expected for different heterozygosities (Nei 1978) and the numbers of loci that are necessary to achieve a reasonably low $P_{(ID)}$. In Figs 2 and 3, the number of alleles (and allele frequencies) for loci with heterozygosity = 0.2, 0.4, 0.6, and 0.8, were as follows: 2 (0.885, 0.115), 3 (0.76, 0.14, 0.1), 5 (0.59, 0.2, 0.1, 0.07, 0.04) and 10 (0.39, 0.15, 0.11, 0.05, 0.05, 0.05, 0.05, 0.05, 0.05, 0.05). These allele frequencies were necessary to arrive at the chosen heterozygosities. Changing the number of alleles (from three to 10) had

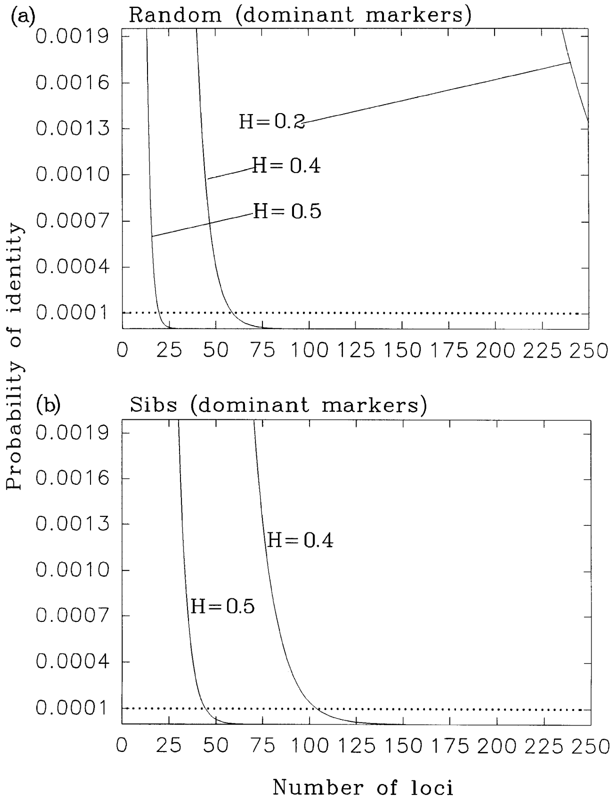

Fig. 3 Relationship between the theoretical probability of identity, $P_{(ID)}$, and the number of dominant loci [e.g. amplified fragment length polymorphisms (AFLP) or random amplified polymorphic DNAs (RAPDS)] assayed using three heterozygosity levels for (a) randomly sampled individuals and (b) sibs.

very little effect on $P_{(ID)}$, as long as the heterozygosity was held constant. We also tested the effect of interlocus variation in heterozygosities on $P_{(ID)}$. For this we computed $P_{(ID)}$ (unbiased eqn 2) for a set of loci all with heterozygosity = 0.6 and compared it with a set of loci with average heterozygosity = 0.6 but with one-third of loci having heterozygosity = 0.4, 0.6 and 0.8. A computer program to compute the theoretical and observed $P_{(ID)}$ is available from the authors.

Results

We evaluated the accuracy of the $P_{(ID)}$ estimator by comparing the theoretical expected and the actual observed $P_{(ID)}$ as summarized for all data sets in Table 1. Substantial differences were detected between the theoretical expected and observed $P_{(ID)}$ using large microsatellite data sets from Montana wolves, Scandinavian brown bears, and the Australian hairy-nosed wombat (Fig. 1, Table 1). The maximum difference between observed and theoretical expected $P_{(ID)}$ was three orders of magnitude. This difference was observed when evaluating seven loci of microsatellite data in the wolf data set (Fig. 1b) and eight loci in the north–south population of brown bears (Table 1). In the bear and wolf data sets, the theoretical and observed $P_{(ID)}$ were <0.0001 when 10 or more loci were analysed. The loci were ranked from highest to lowest based on expected heterozygosity (as would be conducted in most studies). More than 10 loci might have been necessary to reach a $P_{(ID)} < 0.0001$ if each additional locus had been randomly chosen. Also, there was a striking difference between the shapes of the theoretical and observed curves as the observed curve was less smooth due to the presence of a small number of closely related individuals that shared genotypes at a large number of loci.

Figure 1c focuses on the theoretical and observed $P_{(ID)}$ for a highly endangered population of Australian wombats with low levels of heterozygosity and correspondingly higher levels of $P_{(ID)}$. The observed and theoretical $P_{(ID)}$ were 0.014 and 0.0023, respectively, for four loci in the wombats. This represents an underestimation bias of nearly one order of magnitude. The $P_{(ID)}$ values of interest for the wombats are near 0.01 because the main application of $P_{(ID)}$ and individual identification in wombats will be for mark–recapture estimations of population size. A $P_{(ID)}$ of approximately 0.01–0.001 may be appropriate for forensic applications in this population because only approximately 100 wombats remain (Taylor *et al.* 1994).

To address the observed inaccuracies of the theoretical $P_{(ID)}$ and to provide a conservative upper bound on the number of loci necessary for individual identification, a $P_{(ID)}$ equation was derived for sibs. The performance of this equation was compared with the $P_{(ID)}$ equation for random unrelated individuals by examining estimates for 0–30 codominant loci with heterozygosities of 0.2–0.8 (Fig. 2) and 0–250 dominant markers with heterozygosities (i.e. gene diversities) of 0.1 and 0.5 (Fig. 3). These numbers were chosen to represent the range of genetic variation observed when assaying heterozygosity in natural populations. We also tested the importance of the number of alleles (three to 10) at each locus for a given heterozygosity, but the variance in $P_{(ID)}$ due to different allele numbers was extremely small compared with the variance in $P_{(ID)}$ for loci with different heterozygosities (data not shown). Thus, the curves in Figs 2 and 3 and our general guidelines will be approximately applicable to any set of loci with these heterozygosities (see Materials and methods), even loci with differing numbers of alleles or with interlocus variation in heterozygosity.

The $P_{(ID)}$ and $P_{(ID)sib}$ estimators differed by approximately twofold in the number of loci required to achieve a $P_{(ID)} < 0.0001$ using codominant loci. For example, when heterozygosity = 0.4, approximately 11 and 22 loci are required to achieve this $P_{(ID)}$ for random unrelated individuals and sibs, respectively. To achieve a $P_{(ID)} < 0.0001$

for random individuals when using dominant markers, 20–60 loci will be required for heterozygosities of 0.5 and 0.4, respectively. Twice this number of dominant markers will be required to achieve a $P_{(ID)}$ when comparing sibs (Fig. 3). When the standard theoretical $P_{(ID)}$, $P_{(ID)sib}$, and observed $P_{(ID)}$ are plotted on the same graph (Fig. 1), the observed curve always falls between the two theoretical curves.

Discussion

This study clearly demonstrates the potential inaccuracies of using the standard $P_{(ID)}$ estimator to determine the probability of obtaining matching multilocus genotypes when two samples are drawn at random from a population. In some cases, the bias was so large that the observed $P_{(ID)}$ was only 0.001 when the theoretical $P_{(ID)}$ was 0.000001 (Fig. 1b). The theoretical $P_{(ID)}$ curve for wolves and brown bears (Fig. 1) would suggest that five loci are sufficient to identify individuals; however, the observed $P_{(ID)}$ curves indicate that nine to 10 loci would occasionally be necessary to avoid inaccurately identifying a match between two related individuals. The difference between the theoretical and observed $P_{(ID)}$ is probably due to population substructure and the presence of close relatives in the data set which violates the assumptions of the theoretical $P_{(ID)}$ equation (Donnelly 1995). For example, the family pack structure of wolves (Nowack 1999), the tendency of brown bear cubs to remain with their mothers for 2–3 years (Nowack 1999), and an increased level of relatedness between same-sex animals sharing burrows (Taylor *et al.* 1997) may result in sampling a large proportion of closely related individuals (Table 1). In addition, past demographic events that have resulted in a disruption of Hardy–Weinberg equilibrium in a population can also lead to inaccurate estimations using the theoretical $P_{(ID)}$ equation. Thus, the risk of incorrectly assigning a match based on the theoretical $P_{(ID)}$ estimator is potentially higher for populations of social species where many related individuals will probably be sampled and in populations that violate the assumptions of Hardy–Weinberg equilibrium.

To minimize the cost and time involved in studies that require individual identification, researchers generally analyse the minimum number of loci necessary to obtain statistical support for the certainty of individual identification. Thus, when the theoretical $P_{(ID)}$ estimator substantially underestimates the true populational $P_{(ID)}$, it will lead to inaccurate results of individual identification if these biases are not identified and addressed. Before initiating a study that requires individual identification, we suggest that researchers plot the observed $P_{(ID)}$ for multilocus genotypes of known individuals in the study population including known relatives, if possible. As a conservative estimate, researchers can determine the minimum number of loci necessary to identify individuals by choosing the number and combination of loci for which no two individuals share the same genotype. For many projects, it will not be possible to plot the observed $P_{(ID)}$ before choosing the number of loci because of a lack of genotypic data (i.e. >40 genotypes) necessary for accurate estimates. Under these circumstances, it will be important to use both the standard $P_{(ID)}$ and the $P_{(ID)sib}$ to determine the upper and lower bounds for the number of loci required to identify individuals with the desired level of statistical confidence based on heterozygosities estimated from other populations of the same species and/or from Figs 2 and 3. The level of desired statistical confidence will be project specific and should take into account: (i) the severity of the consequences of incorrectly assigning a match to different individuals in the study of interest, and (ii) the proportion of closely related individuals that will probably be sampled. Woods *et al.* (1999) used a criterion of $P_{(ID)sib} < 0.05$ for brown bear population estimation when using multilocus genotypes amplified from hair samples. In wildlife forensic cases, $P_{(ID)} < 0.001$–0.0001 has been used, depending on the size of the source population (J. Coffin, S. Fain, personal communication).

Due to technological advances in molecular biology, individual identification of forensic, noninvasive genetic samples, and clonal plants using hypervariable DNA markers is increasing rapidly. In 1996 and 1997, forensic identification of individuals using microsatellite analysis was used in 140 Canadian and 50 American court cases in order to match a forensic specimen to an illegally killed animal (J. Coffin, S. Fain, personal communication). Forensic identification was recently used to compare genotypes of bears with genotypes of scat and hair samples collected at the scene where bears had consumed a human (L. Waits, unpublished). Each year thousands of animals ranging from parrots to snakes to monkeys are illegally removed from natural populations and wildlife parks and shipped to foreign countries. Individual identification by DNA fingerprinting could provide critical supporting data to prosecute poachers and to detect and help slow the illegal trade of animals and their body parts or products.

A small, but growing, number of studies has demonstrated the use of hair and scat samples collected in the field to identify individuals, estimate population sizes, and home ranges of free-ranging mammals (Gerloff *et al.* 1995; Foran *et al.* 1997; Palsboll *et al.* 1997; Reed *et al.* 1997; Taberlet *et al.* 1997; Kohn *et al.* 1999; Woods *et al.* 1999; Ernest *et al.* 2000). Currently, there is great interest to utilize these methods at a much larger scale to estimate population sizes in large carnivores and endangered species, and to enumerate groups of individuals with unique behaviours. In noninvasive genetic sampling studies, it is important that researchers consider the entire range of potential errors that can lead to inaccurate results such as microsatellite genotyping errors and contamination that can occur when using the low quantity and quality DNA obtained from forensic, hair and scat samples (Gerloff *et al*. 1995; Taberlet *et al*. 1996, 1997, 1999; Gagneux *et al*. 1997; Goossens *et al*. 1998).

Mills *et al*. (2000) have evaluated the impact of incorrectly assigning a match at different levels of $P_{(ID)}$ using mark–recapture population size estimation. They conclude that a $P_{(ID)}$ of less than approximately 0.01 will be necessary for population size estimation and reveal that the bias leads to an underestimate of the population size. Our study comparing theoretical and observed $P_{(ID)}$ has demonstrated that the standard $P_{(ID)}$ estimator can underestimate the true populational $P_{(ID)}$ which could also lead to an underestimate of the minimum population size when too few nuclear DNA microsatellite loci are used to identify individuals in wolf, wombat, and brown bear populations. Potential errors associated with estimating $P_{(ID)}$ can be avoided by choosing the $P_{(ID)sib}$ estimator as a conservative upper bound of the number of loci necessary to distinguish individuals. However, researchers should be aware that the probability of observing a genotyping error in a multilocus genotype will increase as the number of loci analysed increases. Waits & Leberg (2000) have demonstrated that microsatellite genotyping errors can inflate mark–recapture population estimates up to 200% when genotyping seven to 10 loci. Thus, accurate estimates of the number of individuals in a population using noninvasive genetic sampling will require methods to minimize microsatellite genotyping errors and careful evaluation of $P_{(ID)}$ estimators.

Acknowledgements

J.-M. Cornuet kindly derived the equations for $P_{(ID)sib}$ for dominant and codominant markers. Support for G.L. and L.W. was provided by postdoctoral fellowships from NSF-NATO, Université Joseph Fourier, and CNRS, France. L .S. Mills and C. Miller provided helpful critical reviews of this manuscript.

References

Bruford MW, Wayne RK (1993) Microsatellites and their application to population genetic studies. *Current Opinion in Genetics and Development*, **3**, 939–943.

Chakraborty R, Jin L (1992) Heterozygote deficiency, population substructure and their implications in DNA fingerprinting. *Human Genetics*, **88**, 267–272.

Chakraborty R, Kidd K (1991) The utility of DNA typing in forensic work. *Science*, **254**, 1735–1739.

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

Donnelly P (1995) Nonindependence of matches at different loci in DNA profiles: quantifying the effect of close relatives on the match probability. *Heredity*, **75**, 26–34.

Ernest H, Penedo M, May B, Syvanen M, Boyce W (2000) Molecular tracking of mountain lions in the Yosemite Valley region in California: genetic analysis using microsatellites and faecal DNA. *Molecular Ecology*, **9**, 433–442.

Escaravage N, Questiau S, Pornon A, Doche B, Taberlet P (1998) Clonal diversity in a *Rhododendron ferrugineum* L. (Ericaceae) population inferred from AFLP markers. *Molecular Ecology*, **7**, 975–982.

Evett IW, Weir BS (1998) *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists*. Sinauer, Sunderland.

Foran DR, Minta SC, Heinemeyer KS (1997) DNA-based analysis of hair to identify species and individuals for population monitoring. *Wildlife Society Bulletin*, **25**, 840–847.

Forbes SH, Boyd DK (1997) Genetic structure and migration in native and reintroduced Rocky Mountain wolf populations. *Conservation Biology*, **11**, 1226–1234.

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

Gerloff U, Schlötterer C, Rassmann K *et al.* (1995) Amplification of hypervariable simple sequence repeats (microsatellites) from excremental DNA of wild living bonobos (*Pan paniscus*). *Molecular Ecology*, **4**, 515–518.

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

Jamieson A (1965) The genetics of transferrins in cattle. *Heredity*, **20**, 419–440.

Jeffreys AJ, Royle NJ, Patel I *et al.* (1991a) Principles and recent advances in human DNA fingerprinting. In: *DNA Fingerprinting: Approaches and Applications* (eds Burke T, Gaudenz D, Jeffreys AJ, Wolf R), pp. 1–19. Birkhäuser, Basel.

Jeffreys AJ, Turner M, Debenham P (1991b) The efficiency of multilocus DNA fingerprint probes for individualization and establishment of family relationships, determined from extensive casework. *American Journal of Human Genetics*, **48**, 824–840.

Kendall KC, Metzgar LG, Patterson DA, Steele BM (1992) Power of sign surveys to monitor population trends. *Ecological Applications*, **2**, 422–430.

Kohn M, York E, Kamradt D *et al.* (1999) Estimating population size by genotyping faeces. *Proceedings of the Royal Society of London Series B*, **266**, 1–7.

Lander ES (1989) DNA fingerprinting on trial. *Nature*, **339**, 501–505.

Lewontin RC, Hartl DL (1991) Population genetics in forensic DNA typing. *Science*, **254**, 1745–1750.

Li CC (1976) *First Course in Population Genetics*. Boxwood, Pacific Grove, CA.

Mills L, Citta J, Lair K, Schwartz M, Tallmon D (2000) Estimating animal abundance using non-invasive DNA sampling: promise and pitfalls. *Ecological Applications*, **10**, 283–294.

Mueller U, Wolfenbarger L (1999) AFLP genotyping and fingerprinting. *Trends in Ecology and Evolution*, **14**, 389–394.

National Research Council of the USA (1992) *DNA Technology in Forensic Science*. National Academy Press, Washington, DC.

National Research Council of the USA (1996) *The Evaluation of Forensic DNA Evidence*. National Academy Press, Washington, DC.

Nei M (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics*, **89**, 583–590.

Nichols JD, Balding DJ (1991) Effects of population structure of DNA fingerprint analysis in forensic science. *Heredity*, **66**, 297–302.

Nowack RM (1999) *Walker's Mammals of the World*. John Hopkins University Press, Baltimore.

Paetkau D, Strobeck C (1994) Microsatellite analysis of genetic variation in black bear populations. *Molecular Ecology*, **3**, 489–495.

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

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

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

Risch N, Devlin B (1992) On the probability of matching DNA fingerprints. *Science*, **255**, 717–720.

Roeder K (1994) DNA fingerprinting: a review of the controversy. *Statistical Science*, **9**, 222–278.

Schwartz M, Tallmon D, Luikart G (1998) Review of DNA-based census and effective population size estimators. *Animal Conservation*, **1**, 293–299.

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

Taberlet P, Griffin S, Goossens B *et al.* (1996) Reliable genotyping of samples with very low DNA quantities using PCR. *Nucleic Acids Research*, **26**, 3189–3194.

Taberlet P, Waits LP, Luikart G (1999) Noninvasive genetic sampling: look before you leap. *Trends in Ecology and Evolution*, **14**, 321–325.

Taylor AC (1995) *Molecular ecology of the endangered northern hairy-nosed wombat* Lasiorhinus krefftii, *and applications to conservation and management*. PhD Dissertation, University of New South Wales, Sydney.

Taylor AC, Horsup A, Johnson CN, Sunnucks P, Sherwin B (1997) Relatedness structure detected by microsatellite analysis and attempted pedigree reconstruction in an endangered marsupial, the northern hairy-nosed wombat *Lasiohinus krefftii*. *Molecular Ecology*, **6**, 9–19.

Taylor AC, Sherwin W, Wayne R (1994) Genetic variation of microsatellite loci in a bottlenecked species: the northern hairy-nosed wombat *Lasiorhinus krefftii*. *Molecular Ecology*, **3**, 277–290.

Vos P, Hogers R, Bleeker M *et al.* (1995) AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research*, **23**, 4407–4414.

Waits J, Leberg P (2000) Biases associated with population estimation using molecular tagging. *Animal Conservation*, **3**, 191–199.

Waits LP, Paetkau D, Strobeck C (1998) The genetics of the bears of the world. In: *Bear Conservation Act* (ed. Servheen C), pp. 25–32. IUCN, Gland.

Waits L, Taberlet P, Swenson J, Sandegren F, Franzén R (2000) Nuclear DNA microsatellite analysis of genetic diversity and gene flow in the Scandinavian brown bear (*Ursus arctos*). *Molecular Ecology*, **9**, 421–431.

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

Lisette Waits is an assistant professor at the University of Idaho and focuses on the conservation genetics and molecular ecology of carnivores with a current emphasis on noninvasive genetic sampling approaches. Gordon Luikart is a researcher in the Laboratoire de Biologie des Populations d'Altitude (LBPA) currently focusing on the conservation genetics of ungulates and on evaluating new statistical approaches for estimating population parameters and demographic history from molecular data. Pierre Taberlet is the Director of the LBPA, and focuses on the conservation genetics and molecular ecology of many different plant and animal species.