*Research Note*

# Individual Identification and Distribution of Genotypic Differences Between Individuals

**STEVEN T. KALINOWSKI,**[1] *Department of Ecology, Montana State University, Bozeman, MT 59717, USA*

**MICHAEL A. SAWAYA,** *Department of Ecology, Montana State University, Bozeman, MT 59717, USA; Wildlife Conservation Society, Bozeman, MT 59715, USA*

**MARK L. TAPER,** *Department of Ecology, Montana State University, Bozeman, MT 59717, USA*

(JOURNAL OF WILDLIFE MANAGEMENT 70(4):1148–1150; 2006)

## Key words

*census, DNA, estimation, genotype, identification, individual, microsatellite, mismatch, noninvasive, probability, siblings.*

The DNA extracted from forensic evidence has revolutionized law enforcement. The DNA extracted from hair, feces, and other noninvasive samples is having the same effect on ecology and conservation biology. The simplest, and most frequent, use of noninvasive DNA is individual identification. Multilocus genotypes, sometimes referred to as *DNA fingerprints*, can be used to identify individuals and thus count or track animals (e.g., Woods et al. 1999).

Two genetic problems can frustrate individual identification. First, if not enough loci are examined, multilocus genotypes in a population may not be unique (e.g., Taberlet and Luikart 1999). If this happens, individuals with the same genotype will be indistinguishable. Second, genotyping errors can cause samples that came from the same individual to appear to have different genotypes, and therefore appear to have come from different individuals. Genotyping error is a substantial concern in wildlife studies using noninvasive samples because hair and feces contain small amounts of DNA, and this DNA degrades in the field (e.g., Taberlet et al. 1999). The most common problem is that heterozygotes are scored as homozygotes, but other errors are possible (Taberlet et al. 1996). If they are not detected, such errors could dramatically inflate estimates of census size (Waits and Leberg 2000).

The problem of multiple individuals having the same genotype at all loci examined can be solved by increasing the number of loci in a study, and thereby decreasing the probability that 2 individuals have the same multilocus genotype. The probability that 2 individuals have the same multilocus genotype, often called the *probability of identity*, $P_{ID}$, can be estimated from allele frequencies in a population using established formulae (e.g., Waits et al. 2001). This probability is the standard statistic in forensic science to evaluate how well a set of molecular markers discriminates between individuals (e.g., Vazquez et al. 2004). The problem of genotyping error causing samples from the same individual to look different can also (at least in part) be solved by ensuring that a sufficient number of loci are examined. If genotyping error is reasonably small, most errors will cause samples from the same individual to differ (mismatch) at only one locus. If enough loci are scored, distinct individuals will have multilocus genotypes that mismatch at more than one locus. Therefore, if both of these criteria are met, most pairs of samples that mismatch by one locus will differ because of genotyping error, which will make identifying errors much easier (e.g., Paetkau 2003, 2004). For this reason, many researchers will want to design DNA censuses so almost all (if not all) individuals sampled will have genotypes that differ at 2 or more loci. There is, however, no formula available to calculate the probability that 2 individuals have genotypes that differ at 2 or more loci. We provide formulae for this purpose, and introduce a computer program to implement them.

[1] *E-mail: skalinowski@montana.edu*

## Methods

We begin with a few definitions. We define a *genotyping error mismatch* as a mismatch caused by genotyping error between 2 samples from the same individual, and we define a *genotype difference mismatch* as a mismatch caused by underlying genotype differences between 2 individuals. Let $MM_r$ represent an indicator variable that is equal to 1 if 2 individuals have a different genotype at the $r^{\text{th}}$ locus and is equal to 0 if the genotypes are the same

$$MM_r = \begin{cases} 0, & \text{if genotypes identical at the } r^{\text{th}} \text{ locus} \\ 1, & \text{if genotypes different at the } r^{\text{th}} \text{ locus.} \end{cases} \quad (1)$$

Let $k$-$MM$ represent the number of mismatches (out of $L$ loci) between 2 individuals

$$k\text{-}MM = \sum_{r=1}^{L} MM_r. \quad (2)$$

Our goal is to find the probability distribution for $k$-$MM$ for individuals randomly sampled from a population. Relatives will have genotypes that are more similar than unrelated individuals, so we created distributions for different degrees of relatedness. All calculations are for unlinked loci with codominant alleles.

We represented genealogical relationships between individuals by the delta coefficients of Jacquard (1974, Table 6.1; see Lynch and Walsh 1998 for a readable introduction). For noninbred individuals, $\Delta_7$, $\Delta_8$, and $\Delta_9$ represent the probabilities that single locus genotypes for a pair of individuals have 2, 1, or 0 (respectively) alleles identical by descent ($\Delta_7 + \Delta_8 + \Delta_9 = 1$; note $\Delta_1$ through $\Delta_6$ have historically been used to represent patterns of inbreeding and we are assuming individuals are not inbred). These

probabilities specify the genetic relationship between 2 individuals more precisely than the more commonly used relatedness coefficient. We illustrate how this works with a few examples. Let the vector $\mathbf{\Delta}$ represent these 3 delta coefficients, i.e., $\mathbf{\Delta} = \{\Delta_7, \Delta_8, \Delta_9\}$. If 2 individuals are unrelated, then they will not have alleles identical by descent and $\mathbf{\Delta} = \{0, 0, 1\}$. If 2 individuals are parent–offspring, they must have exactly one allele identical by descent and $\mathbf{\Delta} = \{0, 1, 0\}$, and if 2 individuals are full siblings, $\mathbf{\Delta} = \{¼, ½, ¼\}$. Other relationships are possible (e.g., half-siblings, first cousins, and grandparent–grandchild) and can be calculated from pedigrees (see Lynch and Walsh 1998 for a review).

Given $\mathbf{\Delta}$, the probability that 2 individuals have the same genotype at the $r^{\text{th}}$ codominant locus, $P_{ID,r}$, is

$$P_{ID,r} = P(MM_r = 0|\mathbf{\Delta}, \mathbf{f}) = \sum_{i=1}^{m}\sum_{j=i}^{m}\begin{cases}\Delta_9 f_i^4 + \Delta_8 f_i^3 + \Delta_7 f_i^2 & \text{if } i = j \\ \Delta_9 4 f_i^2 f_j^2 + \Delta_8 f_i f_j (f_i + f_j) + \Delta_7 2 f_i f_j & \text{if } i \neq j\end{cases} \quad (3)$$

where the vector $\mathbf{f}$ represents the allele frequencies at the $r^{\text{th}}$ locus, $f_i$ and $f_j$ represent the frequencies of the $i^{\text{th}}$ and $j^{\text{th}}$ alleles, $m$ is the total number of alleles at the locus, and loci are unlinked. Note that we dropped the locus subscript, $r$, from the allele frequencies in Equation 3. The vector $\mathbf{\Delta}$ is the same for all loci, so it doesn't need to be subscripted. Equation 3 is deconstructed as follows. The double summation lists all possible genotypes, $ij$, at the $r^{\text{th}}$ locus, and the terms in the brackets calculate the probability of 2 individuals having that genotype (given $\mathbf{\Delta}$; see Thompson 1991 for a discussion of how to calculate these probabilities). Equation 3 may appear different from previous formulae for calculating probability of identity but only because it accommodates any relationship between 2 noninbred individuals. It is equivalent, for example, to formulae in Waits et al. (2001) and Woods et al. (1999).

The probability that 2 individuals have the same genotype at $L$ loci (i.e., are 0-*MM*) given the relationship ($\mathbf{\Delta}$) between the individuals and the vector of allele frequencies in the population ($\mathbf{f}$), is

$$P(0\text{-}MM|\mathbf{\Delta}, \mathbf{f}) = \prod_{r=1}^{L} P_{ID,r} \quad (4)$$

The probability that 2 individuals have the same genotype at all loci except for one (i.e., are 1-*MM*s) given their relationship ($\mathbf{\Delta}$) and the allele frequencies in the population ($\mathbf{f}$), is

$$P(1\text{-}MM|\mathbf{\Delta}, \mathbf{f}) = \sum_{r=1}^{L}\left[(1 - P_{ID,r})\prod_{\substack{r'=1 \\ r' \neq r}}^{L} P_{ID,r'}\right]. \quad (5)$$

Equation 5 is deconstructed as follows. The term inside the summation is the probability that 2 individuals have the same genotype for all of the $L$ loci except the $r^{\text{th}}$ locus. Summation is over $L$ loci because 2 individuals could differ at each of these loci. In general, the probability that 2 individuals differ at $k$ loci is

$$P(k\text{-}MM|\mathbf{\Delta}, \mathbf{f}) = \sum_{c}^{\binom{L}{k}}\left\{\left[\prod_{r\in c}^{k}(1 - P_{ID,r})\right]\left(\prod_{r'\in\bar{c}}^{L-k} P_{ID,r'}\right)\right\} \quad (6)$$

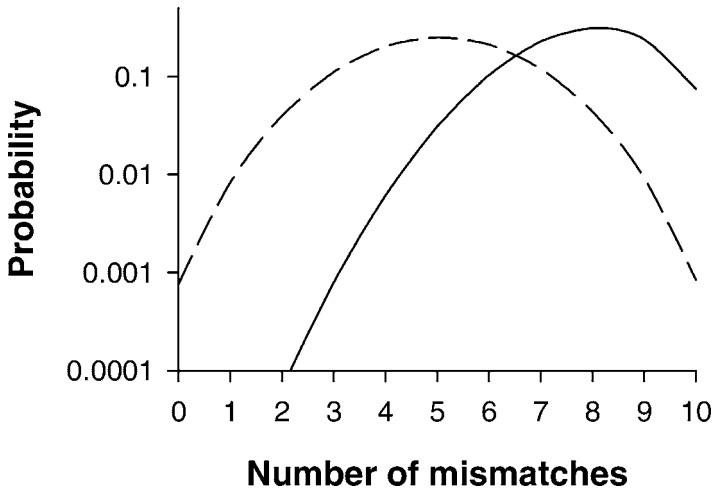


***Figure 1.*** Mismatch probability distributions for desert bighorn sheep in the Eagle Mountains of Southern California, USA. Results are shown for the 10 microsatellite data of Gutiérrez-Espeleta et al. (2000). Solid lines indicate distributions expected for unrelated individuals; dashed lines for full siblings. Note the log scale on the y-axis.

where $c$ is a set of $k$ loci and summation is taken over all sets of $k$ loci that can be drawn from $L$ loci.

We illustrated these formulae with an example. Assume that a DNA census was being planned for the desert bighorn sheep populations of southern California. Ten microsatellite loci were described for these populations (see Gutiérrez-Espeleta et al. 2000 for a description of the loci and their allele frequencies), with an average heterozygosity of approximately 0.50. For simplicity, we assumed that the allele frequencies in the sample from the Eagle Mountains (Gutiérrez-Espeleta et al. 2000) were representative of the region.

## Results

The mismatch distribution for the data of Gutiérrez-Espeleta et al. (2000; Fig. 1) shows that if all 10 loci were genotyped in a DNA census, unrelated individuals are likely to have multilocus genotypes that differ at several loci. For example, the probability of identity for unrelated individuals is less than $10^{-7}$, and there is a high probability ($P \approx 0.9999$) that individuals will differ by 2 or more loci. Siblings, of course, are more genetically similar. The probability of identity for siblings is larger (approx. $10^{-3}$), and the probability of observing 2 or more genotypic mismatches is smaller ($P \approx 0.99$).

Genotyping 10 microsatellite loci is expensive enough to prompt a researcher to ask if fewer loci could provide sufficient discriminatory power. Therefore, we calculated the probability of individuals having 2 or more mismatches for different numbers of loci, always selecting loci having the highest expected heterozygosity (Table 1). This showed that if the 5 most heterozygous loci were used, 99.85% of pairs of unrelated individuals would have 2 or more genotypic mismatches. This percentage drops to approximately 88% for siblings—which again shows how much more genetically similar siblings are than nonrelatives.

## Discussion

Deciding how low mismatch probabilities should be for a DNA census is difficult. It depends on how accurate and cost-effective a

***Table 1.*** Probability of 2 randomly chosen individuals having 2 or more genotypic mismatches (data from Gutiérrez-Espeleta et al. 2000).

| No. of loci[a] | Unrelated | Siblings |
|---|---|---|
| 2 | 0.3509 | 0.1293 |
| 3 | 0.7511 | 0.3675 |
| 4 | 0.9531 | 0.6390 |
| 5 | 0.9985 | 0.8838 |
| 6 | 0.9996 | 0.9352 |
| 7 | 0.9999 | 0.9632 |
| 8 | >0.9999 | 0.9791 |
| 9 | >0.9999 | 0.9881 |
| 10 | >0.9999 | 0.9910 |

[a] We selected loci to maximize expected heterozygosity (e.g., "2" indicates that we used the 2 most heterozygous loci).

DNA census must be. If it is important that no 2 individuals are mistakenly classified as the same individual because they have the same genotype, the probability of 0-*MM* comparisons must be kept low by increasing the number of loci considered. However, increasing the number of loci analyzed will cost more and may result in 2 samples from the same individual being interpreted as distinct individuals due to genotyping errors. These problems are well recognized, and there is no consensus for how to best solve them (see McKelvey and Schwartz 2004*a*,*b*, and Paetkau 2003, 2004 for divergent views). Paetkau (2003, 2004), for example, recommends genotyping samples one time at 6 loci and using stringent quality control protocols to avoid genotyping error. Samples that differ by 1 or 2 loci are then re-genotyped. On the other hand, McKelvey and Schwartz (2004*a*,*b*) recommend using as many as 12–15 loci so that genotyping error mismatch distributions are unlikely to overlap with genotyping difference distributions (see McKelvey and Schwartz 2004*a*; Fig. 5).

No matter which approach is used, however, estimating the mismatch distributions in a population will be useful for geneticists to design a study with genotypic discrimination they desire. A computer program, MM-DIST, is available to compute mismatch distributions for empirical data sets. Program MM-DIST runs on the Windows operating system and reads GENEPOP input files (Raymond and Rousset 1995). Program MM-DIST and documentation are available from www.montana.edu/kalinowski.

## Management Implications

DNA extracted from hair and feces has become an important marker for counting individuals. The principle management implication of our research is that it will assist geneticists in selecting a sufficient number of loci for DNA census so that all individuals sampled are likely to differ at more than one locus. This will help prevent genotyping errors from causing 2 samples from the same individual from being interpreted as 2 individuals. This should improve estimates of population size derived from genetic analysis of noninvasive samples.

## Acknowledgments

Our research was supported by National Science Foundation grant DEB-0415932 to M. Taper.

## Literature Cited

Gutiérrez-Espeleta, G. A., S. T. Kalinowski, W. M. Boyce, and P. W. Hedrick. 2000. Genetic variation and population structure in desert bighorn sheep: implications for conservation. Conservation Genetics 1:3–15.

Jacquard, A. 1974. The genetic structure of populations. Springer-Verlag, New York, New York, USA.

Lynch, M., and B. Walsh. 1998. Genetics and analysis of quantitative traits. Sinauer Associates, Sunderland, Massachusetts, USA.

McKelvey, K. S., and M. K. Schwartz. 2004*a*. Genetic errors associated with population estimation using noninvasive molecular tagging: problems and new solutions. Journal of Wildlife Management 68:439–448.

McKelvey, K. S., and M. K. Schwartz. 2004*b*. Providing reliable and accurate genetic capture–mark–recapture estimates in a cost-effective way. Journal of Wildlife Management 68:453–456.

Paetkau, D. 2003. An empirical exploration of data quality in DNA-based population inventories. Molecular Ecology 12:1375–1387.

Paetkau, D. 2004. The optimal number of markers in genetic capture–mark–recapture studies. Journal of Wildlife Management 68:449–452.

Raymond, M., and F. Rousset. 1995. GENEPOP. Version 1.2. Population genetics software for exact tests and ecumenicism. Journal of Heredity 86: 248–249.

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

Taberlet, P., and G. Luikart. 1999. Noninvasive genetic sampling and individual identification. Biological Journal of the Linnean Society 68:41–55.

Taberlet, P., L. Waits, and G. Luikhart. 1999. Noninvasive genetic sampling: look before you leap. Trends in Ecology and Evolution 14:323–327.

Thompson, E. A. 1991. Estimation of relationships from genetic data. Pages 255–269 *in* C. R. Rao and R. Chakraborty, editors. Handbook of statistics. Volume 8. Elsevier Science, North Holland, New York, USA.

Vazquez, J. F., T. Perez, F. Urena, E. Gudin, J. Albornoz, and A. Dominguez. 2004. Practical application of DNA fingerprinting to trace beef. Journal of Food Protection 67:972–979.

Waits, J. L., and P. L. Leberg. 2000. Biases associated with population estimation using molecular tagging. Animal Conservation 3:191–199.

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.

Woods, J. G., D. Paetkau, D. Lewis, B. N. McLellan, M. Proctor, and C. Strobeck. 1999. Genetic tagging free-ranging black and brown bears. Wildlife Society Bulletin 27:616–627.

*Associate Editor: DeWoody*.