Identification of Source Populations of Mixture Individuals from their Genotypes

Michele Masuda and **Jerome Pella**
U.S. Department of Commerce, NOAA, NMFS,
Alaska Fisheries Science Center, Auke Bay Laboratory,
11305 Glacier Hwy., Juneau, AK 99801-8626, USA

Keywords: Microsatellite, individual assignment, stock mixture analysis

The source populations of individuals of unknown origin can be surmised from their genotypes. In many applications, the sources for more than a single individual are desired and a list of the c (say) potential source populations is available. When such is the case, misidentifications are minimized by assigning each individual with the maximum *a posteriori* probability (MAP) rule. The MAP rule assigns an individual with genotype X to the population for which the posterior source probability,

$$(1) \quad p(s \mid X) = \frac{p_s g_s(X)}{\sum_{i=1}^{c} p_i g_i(X)} \quad s = 1,\ldots,c,$$

is greatest. Here p_i is the prior (before seeing its genotype) probability that the individual comes from population i, and $g_i(X)$ is the relative frequency of the genotype X in population i. Intuitively, $p(s \mid X)$ is the fraction of individuals having genotype X that is contributed by population s to a mixture composed of c populations with proportions, $\boldsymbol{p} = (p_1,\ldots,p_c)$.

If only a single individual is to be identified to source, the equi-probable prior is an obvious choice, i.e., $p_1 = \ldots = p_c = 1/c$. The MAP rule with equi-probable prior produces the same assignments as WHICHRUN[1] (Banks and Eichert 2000), which assigns individuals to the population k in which the genotype is most frequent (MFG rule), i.e., $g_k(X) = \max_i\{g_i(X), i = 1,\ldots,c\}$. However, if sources of several individuals are to be identified, they are better viewed as a sample from a mixture whose unknown source proportions are the prior probabilities in eq. 1. These prior probabilities can be estimated by conditional maximum likelihood with program SPAM[2] (Debevec et al. 2000; Alaska Department of Fish and Game 2003), or by Bayesian methods with program BAYES[3] (Pella and Masuda 2000). Currently, SPAM simply evaluates eq. 1 once using the conditional maximum likelihood point estimates, but BAYES evaluates eq. 1 from draws of all unknowns at each of many cycles.

Paetkau et al. (1995) obtained data for eight microsatellite loci from four Canadian polar bear populations: northern Beaufort Sea, southern Beaufort Sea, western Hudson Bay, and Davis Strait-Labrador Sea (Table 1). Average heterozygosity was near 60% for each population and allele frequency distributions were significantly different between all pairs of populations (Paetkau et al. 1995). Paetkau et al. (1999) in a more extensive genetic study of circumpolar populations of polar bears expanded the number of loci to 16 and the number of populations to 16 (Table 1). A simulation experiment was performed to compare methods of individual assignments using the polar bear data[4]. Only the original four populations were included in our study. The variables controlled in the experiment were the number of loci (either the initial eight loci, or the later 16 loci), the proportions of contribution

Table 1. Number of microsatellite loci and population sample sizes for the two polar bear studies (Paetkau et al. 1995, 1999).

		Sample size Population			
Study	Number of loci	Southern Beaufort Sea	Northern Beaufort Sea	Western Hudson Bay	Davis Strait-Labrador Sea
Paetkau et al. (1995)	8	22	30	30	26
Paetkau et al. (1999)	16	30	30	33	30

[1] WHICHRUN can be obtained from http://www.bml.ucdavis.edu/whichrun.htm.
[2] SPAM can be obtained from http://www.cf.adfg.state.ak.us/geninfo/research/genetics/Software/SpamPage.htm.
[3] BAYES can be obtained from ftp://ftp.afsc.noaa.gov/sida/mixture-analysis/bayes/.
[4] Dr. David Paetkau generously made the data available for this study.

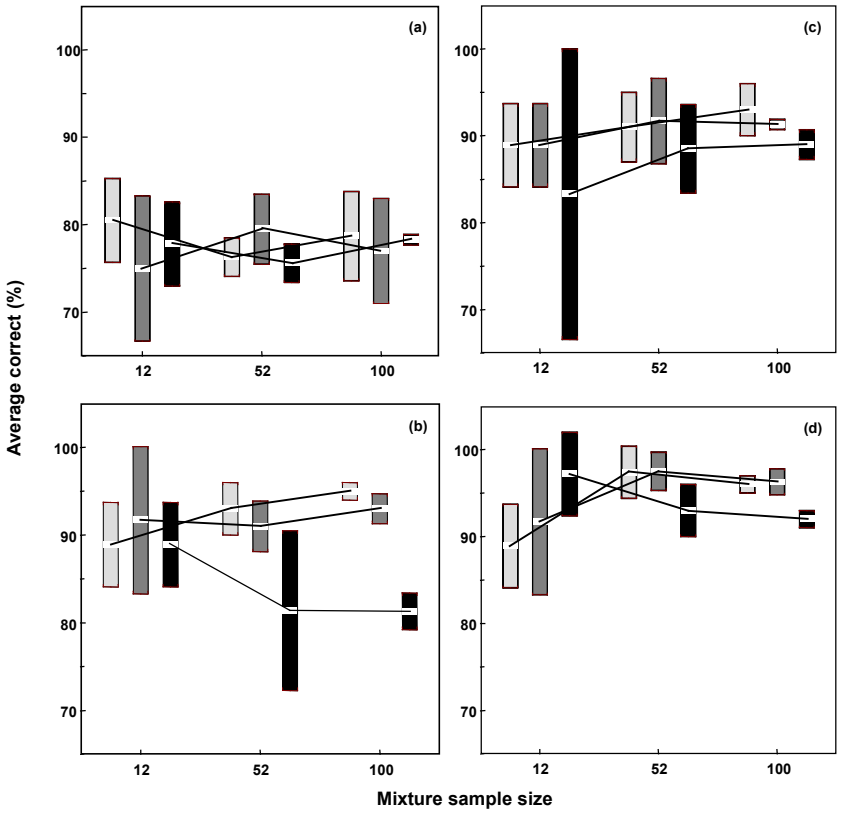
All correspondence should be addressed to M. Masuda.
e-mail: Michele.Masuda@noaa.gov

from the four populations (either equal proportions of 1/4, or an uneven set with western Hudson Bay comprising 75–88.5% of the population mixture and the other stocks comprising equal thirds of the remaining mixture), and mixture sample size (12, 52, or 100 bears). The original baseline samples were independently resampled for each of the 12 cells of the design to provide simulated sets of baseline and population-mixture genotype samples of polar bears, and each cell was independently replicated three times. For each simulated mixture sample, source populations of mixture individuals were identified with the MAP rule applied to the posterior source probabilities using SPAM, the MAP rule applied to average posterior source probabilities using BAYES, and the MFG rule using WHICHRUN. SPAM was run with a Bayesian model of baseline allele frequency distributions, specifically, the mean of the Rannala and Mountain (1997) baseline posterior. BAYES generated a single fixed sequence of 5,000 samples of the unknowns (first 2,500 was discarded as burn-in) for each cell and replicate. The experiment was summarized with the percentage of mixture individuals correctly assigned to their source population.

Although the average correct (%) is imprecisely determined with only three replications per experimental cell, the following generalizations are discernible (Fig. 1). First, with fewer loci and unequal mixture proportions (Fig. 1b), the method used becomes increasingly important with increase in number of individuals to be identified. SPAM and BAYES perform better with increase in mixture sample size, whereas WHICHRUN performance falls progressively below that of the others. The same is true with more loci and unequal mixture proportions (Fig. 1d), but performance differences are smaller. Second, under the contrived equal-proportions mixture for which WHICHRUN is expected to perform best because it is effectively given the correct prior, it performed on average only comparably to SPAM and BAYES for fewer loci (Fig. 1a) and apparently slightly worse for more loci (Fig. 1c).

As the number of individuals to be identified to their source populations increases, the maximum *a posteriori* (MAP) rule with posterior source probabilities computed by likelihood or Bayesian methods performs better than the most frequent genotype (MFG) rule, especially if genetic information is limited. At small sample sizes, little, if any, loss in performance occurs by use of the MAP rule with estimated prior probabilities as compared to the MFG rule. At present, the two programs—SPAM and BAYES—that compute and output the posterior source probabilities of individuals are not designed to perform the assignments by the MAP rule, and the researcher must make the assignments manually (hence the limited number of replications in this study).

Fig 1. Average (n = 3) percentage of mixture individuals (white break in vertical bar) correctly assigned to the population for varying mixture sample sizes, 8-locus (a and b) or 16-locus data set (c and d), equal (a and c) or unequal mixture proportions (b and d), and assignment method: BAYES ( ), SPAM ( ), and WHICHRUN ( ). Extreme ends of bars indicate ±1 standard deviation.



REFERENCES

Alaska Department of Fish and Game. 2003. SPAM Version 3.7: Addendum II to User's Guide for Version 3.2. 16p. Alaska Department of Fish and Game, Commercial Fisheries Division, Gene Conservation Lab, Anchorage, AK, USA.

Banks, M.A., and W. Eichert. 2000. WHICHRUN (Version 3.2) a computer program for population assignment of individuals based on multilocus genotype data. J. Hered. 91: 87–89.

Debevec, E.M., R.B. Gates, M. Masuda, J. Pella, J. Reynolds, L.W. Seeb. 2000. SPAM (Version 3.2): Statistics Program for Analyzing Mixtures. J. Hered. 91: 509–510.

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

Paetkau, D., S.C. Amstrup, E.W. Born, W. Calvert, A.E. Derocher, G.W. Garner, F. Messier, I. Stirling, M.K. Taylor, Ø. Wiig, and C. Strobeck. 1999. Genetic structure of the world's polar bear populations. Mol. Ecol. 8: 1571–1584.

Pella, J., and M. Masuda. 2000. Bayesian methods for analysis of stock mixtures from genetic characters. Fish. Bull. 99: 151–167.

Rannala, B., and J.L. Mountain. 1997. Detecting immigration by using multilocus genotypes. Genetics. 94: 9197–9201.