

PROGRAM NOTE

DROPOUT: a program to identify problem loci and samples for noninvasive genetic samples in a capture-mark-recapture framework

K. S. McKELVEY and M. K. SCHWARTZ

USDA Forest Service, Rocky Mountain Research Station, 800 E. Beckwith, Missoula, MT 59801, USA

Abstract

Genotyping error, often associated with low-quantity/quality DNA samples, is an important issue when using genetic tags to estimate abundance using capture-mark-recapture (CMR). DROPOUT, an MS-Windows program, identifies both loci and samples that likely contain errors affecting CMR estimates. DROPOUT uses a 'bimodal test', that enumerates the number of loci different between each pair of samples, and a 'difference in capture history test' (DCH) to determine those loci producing the most errors. Importantly, the DCH test allows one to determine that a data set is error-free. DROPOUT has been evaluated in McKelvey & Schwartz (2004) and is now available online.

Keywords: allelic dropout, DROPOUT, genotyping error, mark recapture, molecular tagging, noninvasive

Received 09 March 2005; revision accepted 14 April 2005

The fields of ecology, wildlife management and conservation biology have embraced identification of species and individuals using noninvasively collected genetic samples. Projects using noninvasive genetic samples can provide a wealth of data on rare and elusive species once thought impossible to accurately count. Unfortunately, using DNA from noninvasive samples can also lead to genotyping errors which can bias estimates unless carefully controlled (Waits & Leberg 2000; Creel *et al.* 2003; McKelvey & Schwartz 2004).

Genotyping errors often occur when DNA is collected from low quality samples (hair, feathers, faeces, etc.; Taberlet *et al.* 1996; Morin *et al.* 2001). The most common genotyping errors are allelic dropout, the preferential amplification of one of two alleles, false alleles, amplification products that mimic true alleles, and various laboratory errors such as misreading bands or transcription errors. Molecular ecologists have recognized the importance of genotyping error and arrived at multiple solutions. Programs GIMLET (Valiere 2002) and RELIOTYPE (Miller *et al.* 2002) are useful for evaluating errors when samples have been multitubed (Taberlet *et al.* 1996). Program PEDMANAGER (Ewen *et al.* 2000) is useful when pedigree information is available.

Additionally, MICRO-CHECKER (Van Oosterhout *et al.* 2004) compares randomly constructed genotypes to observed genotypes, locating errors due to stutter and short allele dominance. Other approaches include quantifying the amount of extracted DNA and avoiding analysis of samples with low yield (Morin *et al.* 2001).

Here we provide details on a program, DROPOUT, used to identify samples and loci that contain critical errors when using genetic tags (multilocus genotypes) for capture-mark-recapture (CMR) population estimation (McKelvey & Schwartz 2004). For CMR estimation, samples are assigned either to new individuals or recaptures; the only important errors are those that lead to recaptures being classed as new individuals, or vice versa. Many errors, therefore, will not affect estimates, and complete error removal is unnecessary if the data are exclusively used for abundance estimation. Unfortunately, CMR estimates are extremely sensitive to ratios of new individuals to recaptures, and even a 1% per-locus error rate will lead to significant estimation bias (Waits & Leberg 2000; Creel *et al.* 2003; McKelvey & Schwartz 2004). Thus, obtaining accurate population estimates from noninvasive samples requires both efficient ways to locate these errors, and the ability to demonstrate their reduction to trivial levels.

DROPOUT performs two analyses to evaluate errors regardless of source. The first identifies samples that likely

Correspondence: K. S. McKelvey, USFS RMRS, PO Box 8089, Missoula, MT 59807. Fax: (406) 543-2663; E-mail: kmckelvey@fs.fed.us

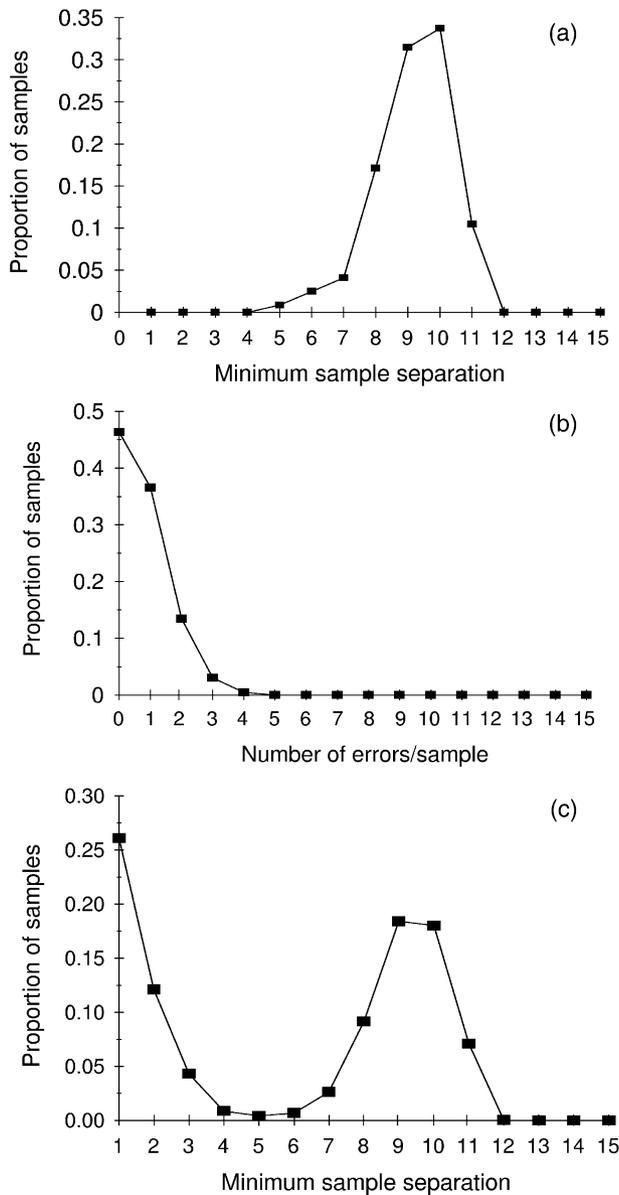


Fig. 1 Figure (a) shows the histogram of minimum sample separation for a simulated population with a genetic tag of 15 loci, average heterozygosity of 0.78, and no genotyping errors. Figure (b) shows the binomial expectations of number of errors per sample given a tag of 15 loci and a 5% per locus error. Figure (c) shows the histogram for the same population but, with a 5% per locus error rate. To simulate a capture-mark-recapture (CMR) experiment, the population was sampled five times and with an individual probability of detection of 20% per session. If uncorrected, these data, based on a six-locus tag, produce CMR population estimates of $1.75\times$ the real population size (McKelvey & Schwartz 2004).

contain errors, and the second demonstrates that, for CMR estimation, a group of samples is virtually error-free. Both of these tests only work if a sample contains significant numbers of recaptures, a situation anticipated in CMR. The

first test takes advantage of the fact that Mendelian inheritance patterns produce a bell-shaped relatedness distribution in an ideal population (Rousset 2002). Thus, in a population sampled in a manner that generates recaptures, some of the samples (i.e. recaptures) will be genetically identical, whereas the rest will be minimally separated from all other samples by several loci (Fig. 1a; McKelvey & Schwartz 2004). By increasing the molecular tag size, the distribution mode can be moved away from 0, such that there is a very low likelihood that individuals differ at only one or two loci.

Genotyping errors generally occur at low rates and most samples, therefore, contain errors at only a few loci. If these errors occur in a sample that should be classified as a recapture, it most likely differs from other samples from the same individual at one or two loci (Fig. 1b; Paetkau 2003; McKelvey & Schwartz 2004). A sample containing both recaptures and genotyping errors will, with a sufficiently large tag, show a bimodal minimal separation distribution (Fig. 1c). This approach is a formalization of the *ad hoc* methodologies used in several laboratories (see Paetkau 2003). DROPOUT therefore compares each sample with all other samples to find the sample that is most similar, builds the minimal separation distribution and reports the number of recaptures. Additionally, DROPOUT displays all sample pairs that differ at one to three loci, the loci they differ at, and the alleles associated with those loci. For each pair member, DROPOUT also reports the number of samples that are identical to the pair member, and the total number of missing loci associated with the pair comparison. While these data are often sufficient to determine which of the two samples likely contains errors, final determination of error requires additional laboratory checks.

DROPOUT's 'bimodal' test informs one as to whether a sample likely contains genotyping errors and provides information to determine the source of the errors. DROPOUT's second test, the 'difference in capture history' test, provides a check to determine whether error has been successfully removed. The probability of two individuals having the same alleles decreases multiplicatively with increasing loci. At some intermediate number of loci (with heterozygosity ≈ 0.75 this occurs with approximately six to seven loci), with no genotyping error, the probability of identifying new individuals through expanding or changing the composition of the genetic tag becomes infinitesimal (Mills *et al.* 2000). However, genotyping errors potentially occur at each locus added to the tag, creating genetic differences that are interpreted as new individuals (Waits & Leberg 2000; McKelvey & Schwartz 2004).

The second test therefore consists of the following steps: (1) amplify enough loci that all individuals are likely to be unique given a tag at least one locus shorter, (2) remove all samples with the same multilocus genotype (recaptures), (3) compute the probability of identity (*PI*; Paetkau &

Strobeck 1994) or the probability that siblings are identical (PI_{sib} ; Evett & Weir 1998; Waits *et al.* 2001), (4) compute PI (or PI_{sib}) for two through n loci and choose a base tag size such that PI and PI_{sib} are very small and all individuals should therefore be uniquely identified (Mills *et al.* 2000; L_{base}), (5) generate a list of unique individuals when the tag size is L_{base} and (6) compare this to the number of unique individuals generated through adding additional loci to the genetic tag and changing the composition of the tag. If the sample is free of errors and L_{base} is sufficiently large, no new individuals will be generated through this process. To evaluate errors generated by a particular locus, DROPOUT rotates the order of loci so that each locus is added individually to an L_{base} -sized tag. If new individuals are produced when a particular locus is added, then errors exist at that locus. If many new individuals are produced when a specific locus is rotated through the $L_{\text{base}} + 1$ position, that locus is likely problematic. This test, both in theory and in simulations, is extremely sensitive (McKelvey & Schwartz 2004). With an eight-locus tag, an average heterozygosity of 0.8, and 100 recaptures, a per-locus error rate of 0.5% will be detected 96% of the time (McKelvey & Schwartz 2004).

Given a sample of genetic tags, the 'difference in capture history test' in DROPOUT performs all of these steps. DROPOUT records the number of new individuals produced at each rotation through the $L_{\text{base}} + 1$ position, performs a chi-squared homogeneity test and, if significant, constructs simultaneous Bonferroni confidence intervals to determine which loci are contributing more new individuals than expectation (see: Sokal & Rohlf 1981 : 728).

DROPOUT is designed specifically for microsatellite markers on diploid organisms, but can be modified for other marker categories and genomes. DROPOUT uses an input format similar to GENEPOP (Raymond & Rousset 1995; detailed in the program's help files), with minor modifications to accommodate the CMR framework (sample input files are included with the program). Alleles must be designated by integers 0–999, and missing data are indicated by a 0/0 score at a locus.

Our experience, based on black bear (*Ursus americanus*) hair samples ($H \approx 0.75$), indicates that an efficient approach is to amplify a relatively large tag consisting of nine or more microsatellites once for all samples in a new data set. Next, focus on the difference in capture history test to identify the problem loci, and remove any loci that produce significantly more individuals than expected. Having removed the problem loci, and assuming that the tag size is still at least nine loci, we then use the bimodal test to determine which samples likely contain errors and apply the multiple-tube method to establish consensus genotypes (Taberlet *et al.* 1996). Lastly, we rerun DROPOUT examining

the results of both tests. This process is repeated until the sample is error free based on the difference in the capture history test.

DROPOUT is free and can be downloaded from <http://www.fs.fed.us/rm/wildlife/genetics/>.

References

- Creel S, Spong G, Sands JL *et al.* (2003) Population size estimation in Yellowstone wolves with error-prone noninvasive microsatellite genotypes. *Molecular Ecology*, **12**, 2003–2009.
- Evett IW, Weir BS (1998) *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists*. Sinauer, Sunderland, Massachusetts, USA.
- Ewen KR, Bahlo M, Treloar SA *et al.* (2000) Identification and analysis of error types in high-throughput genotyping. *American Journal of Human Genetics*, **67**, 727–736.
- McKelvey KS, Schwartz MK (2004) Genetic errors associated with population estimation using non-invasive molecular tagging: problems and new solutions. *Journal of Wildlife Management*, **68**, 439–448.
- Miller CR, Joyce P, Waits LP (2002) Assessing allelic dropout and genotyping reliability using maximum likelihood. *Genetics*, **160**, 357–249.
- Mills LS, Citta JJ, Lair KP, Schwartz MK, Talmon DA (2000) Estimating animal abundance using noninvasive DNA sampling: promise and pitfalls. *Ecological Applications*, **10**, 283–294.
- 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*). *Molecular Ecology*, **10**, 1835–1844.
- Paetkau D (2003) An empirical exploration of data quality in DNA-based population inventories. *Molecular Ecology*, **12**, 1375–1387.
- Paetkau D, Strobeck C (1994) Microsatellite analysis of genetic variation in black bear populations. *Molecular Ecology*, **3**, 489–495.
- Raymond M, Rousset F (1995) GENEPOP (version 1.2): population genetic software for exact tests and ecumenicism. *Journal of Heredity*, **86**, 248–249.
- Rousset F (2002) Inbreeding and relatedness coefficients: what do they measure? *Heredity*, **88**, 371–380.
- Sokal RR, Rohlf JF (1981) *Biometry*, 2nd edn. W.H. Freeman and Company, New York, USA.
- Taberlet S, Griffin B, Goossens S *et al.* (1996) Reliable genotyping of samples with very low DNA quantities using PCR. *Nucleic Acids Research*, **26**, 3189–3194.
- Valiere N (2002) Gimlet: a computer program for analysing genetic individual identification data. *Molecular Ecology Notes*, **2**, 377–379.
- Van Oosterhout C, Hutchinson WF, Willis DPM, Shipley P (2004) MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. *Molecular Ecology Notes*, **4**, 535–538.
- Waits JL, Leberg PL (2000) Biases associated with population estimation using molecular tagging. *Animal Conservation*, **3**, 191–199.
- Waits LP, Luikart G, Taberlet P (2001) Estimating the probability of identity among genotypes in natural populations: cautions and guidelines. *Molecular Ecology*, **10**, 249–256.