doi: 10.1111/j.1471-8286.2004.00797.x

PROGRAM NOTE

GENECAP: a program for analysis of multilocus genotype data for non-invasive sampling and capture-recapture population estimation

MICHAEL J. WILBERG and BRIAN P. DREHER*

Department of Fisheries and Wildlife, Michigan State University 13 Natural Resources Building East Lansing, Michigan 48824 U.S.A

Abstract

We created GENECAP to facilitate analysis of multilocus genotype data for use in non-invasive DNA sampling and genetic capture-recapture studies. GENECAP is a Microsoft EXCEL macro that uses multilocus genetic data to match samples with identical genotypes, calculate frequency of alleles, identify sample genotypes that differ by one and two alleles, calculate probabilities of identity, and match probabilities for matching samples. GENECAP allows the user to include background data and samples with missing genotypes for multiple loci. Capture histories for each user-defined sampling period are output in formats consistent with commonly employed population estimation programs.

Keywords: capture histories, mark-recapture, match probabilities, matching genotypes, non-invasive sampling, probability of identity

Received 12 February 2004; revision received 28 June 2004; accepted 9 August 2004

The use of microsatellite molecular markers to survey wildlife populations has been increasing for both conservation and management purposes, and for a wide range of species (e.g. Palsbøll *et al*. 1997; Woods *et al*. 1999). Although the species of survey may differ, the basic data structure is similar in studies using microsatellite or other Mendelian markers and non-invasive sampling methods to estimate population size. The challenge for researchers is to generate summary statistics that relate to statistical power of match assignment and convert multilocus genotypic data to a functional form required by widely used population estimation programs.

GENECAP is an executable macro within Microsoft EXCEL (versions 97, 2000, and XP) written in Microsoft Visual Basic. GENECAP compares each multilocus genotype with all other genotypes within the dataset to identify matching samples, while accounting for missing data. The program performs calculations of probability of identity (PI), allele frequencies, and match probabilities. GENECAP also creates capture histories from matching genotypes, identifies ambiguous samples, and outputs samples and accompanying genotypes that differ by one and two alleles. In addition to capture-recapture studies, GENECAP could be used for quality control of any dataset (even from standard tissue samples) to help, for example, identify potentially repeated (double-sampled) individuals or mislabelled tubes. GENECAP runs on Microsoft WINDOWS™-based (95, 98, ME, NT, 2000, XP) PCs. The program and user's manual can be downloaded free of charge at http://www.fw.msu.edu/labs/molecularecology/programs.htm.

GENECAP provides some of the same functions as API-CALC (Ayres & Overall 2004) and GIMLET (Valière 2002). GENECAP and API-CALC both calculate estimates of PI, but use different methods. GENECAP and GIMLET both find samples with matching genotypes within a dataset, provide estimates of allele frequencies and PI, and create capture histories. However, GENECAP allows the user to evaluate samples with similar genotypes for genotyping errors (Paetkau 2003), uses probability to determine matches (Woods *et al*. 1999), and easily handles missing data, while GIMLET does not. GENECAP also calculates allele frequency estimates differently than GIMLET because GENECAP excludes redundant samples when estimating the allele frequencies of the population, which subsequently affects the estimates of PI. Validation testing of GENECAP included comparisons with GIMLET and both produced the same estimates when redundant samples were removed.

Correspondence: M. J. Wilberg. *Current Address: Colorado Division of Wildlife, Southeast Region Service Center, 4255 Sinton Road, Colorado Springs, Colorado 80907, U.S.A. Fax: 517 4321699; E-mail: wilbergm@msu.edu

Program Inputs

The basic data requirements for each sample are a unique sample identification number, the diallelic genotype for up to 50 loci, sample type (e.g. hair, muscle tissue), the capture occasion that the sample was collected and the location (if available) where the sample was collected. One feature of GENECAP is the ability to account for missing genotype data, which can often result when genotyping with small quantities of DNA obtained through non-invasive techniques (Taberlet *et al*. 1999). Missing data are coded as '-99' and can be input for one or both alleles at a locus (i.e. non-amplification or a locus with high incidence of null alleles).

Calculations and Outputs

GENECAP output is provided to the user and includes: the number of records in the data; the number of matching genotypes; number of sample pairs that only differ by one allele; number of sample pairs that differ by two alleles; the number of unique genotypes (individuals) detected in the data, the sibling PI (Sib P(ID)); and the Hardy–Weinberg PI (HW P(ID)).

Matching Genotypes

Identifying matching genotypes (in pair wise comparisons) can be a difficult task when a dataset consists of many loci and samples. Manually matching sample genotypes can lead to errors. GENECAP identifies matching genotypes by comparing each allele of a sample to all other alleles in all other samples at a locus. The procedure is repeated for each locus. By comparing genotypes in this manner, it allows the user to include missing data within the dataset. Once a match has been determined the identification numbers of the matching genotypes are output to a worksheet with the type of sample, the capture occasion, and the match probabilities of a matching genotype based on allele frequencies.

Allele Frequencies

Because genotypic data are used to differentiate individuals, it is important to determine how informative the loci are in determining a match (Paetkau 2003). Before making statistical calculations about the discriminatory power of loci, allele frequencies for each locus must first be calculated. GENECAP calculates the allele frequencies by using the genotypes for each unique individual to prevent the bias of using duplicate genotypes of redundant samples.

Probability of Identity Calculations

GENECAP calculates the probability of two individuals within the population sharing the same genotype (PI) using two different formulations. The first formulation assumes Hardy–Weinberg equilibrium, HW P(ID), and is calculated for each locus with the following formula (Paetkau & Strobeck 1994):

$$HW\ P(ID) = \sum_i p_i^4 + \sum_i \sum_{j>i} (2p_i p_j)^2$$

where p_i and p_j are the frequencies of the ith and jth alleles. GENECAP calculates this value for each locus. To obtain the overall PI value across all loci, GENECAP assumes loci are independent and multiplies all locus-specific PIs to obtain the overall PI value.

Other researchers have demonstrated that the HW P(ID) can be biased and suggest using the sibling PI (Sib P(ID)) as an alternative because it is a more conservative measure of PI than HW P(ID) (Waits *et al*. 2001). GENECAP computes Sib P(ID) with the following formula (Evett & Weir 1998):

$$Sib\ P(ID) = 0.25 + \left(0.5 \sum p_i^2\right) + \left[0.5\left(\sum p_i^2\right)^2\right] - \left(0.25 \sum p_i^4\right)$$

where p_i is the frequency of the ith allele. GENECAP calculates the Sib P(ID) values for each locus. The overall value of the Sib P(ID) is calculated as the product of all locus values.

Match probabilities

Because there is a probability of two individuals sharing the same genotype by chance, statistical rigor must be used for match declarations (Woods *et al*. 1999). GENECAP calculates the probability that two individuals will have the same genotype under two different assumptions: (1) assuming Hardy–Weinberg equilibrium (P_{HW}), and (2) assuming that both individuals are siblings (P_{sib}). P_{HW} is calculated using the following equations:

$$P_{HW(Homozygote)} = p_i^2$$

$$P_{HW(Heterozygote)} = 2p_ip_j$$

where p_i and p_j are the frequencies of the ith and jth alleles. Calculations are made for each locus of a matching genotype, multiplied across all loci (which assumes all loci are independent), and output for each matching genotype.

P_{sib} is calculated with the following equations:

$$P_{sib(Homozygote)} = \frac{(1 + 2p_i + p_i^2)}{4}$$

$$P_{sib(Heterozygote)} = \frac{(1 + p_i + p_j + 2p_ip_j)}{4}$$

where p_i and p_j are the frequencies of the ith and jth alleles (Woods *et al*. 1999). These calculations are made for each locus and multiplied across all loci to calculate the final probability for each matching genotype. GENECAP allows the user to specify which match probability is used to identify a match, that is used in forming capture histories.

Capture Histories

Population estimation programs such as program MARK and program CAPTURE utilize a data structure referred to as capture histories to estimate population size (White & Burnham 1999). The basic format of a capture history is a series of 0's and 1's indicating the absence (0) or presence (1) of an individual for each capture occasion in the dataset. For example, the following capture history, 001011, can be interpreted as having 6 capture occasions and the individual was observed (indicated with the presence of a 1) on occasions 3, 5 and 6. The individual was not observed (indicated by the presence of a 0) on occasions 1, 2, and 4. All individuals in the dataset are assigned a capture history based on which samples matched that individual and the periods those samples were collected.

To create capture histories, GENECAP uses both the matching genotypes and match probabilities. Woods *et al*. (1999) suggested that an acceptable match probability was $P_{sib} < 0.05$. GENECAP allows the user to choose the type of match probability, P_{HW} or P_{Sib}, and the match probability level used to declare a match. If two genotypes match, but do not meet the match probability criterion, the sample listed second in the data is excluded from the capture history. This option can be ignored by setting the match probability criterion to 1. After evaluating each matching genotype for this criterion, GENECAP then creates capture histories for all unique genotypes within the dataset and outputs these capture histories to a worksheet.

Problem genotypes

One feature of GENECAP is the ability to include genotypes with missing information (i.e. no allele scores for some loci). Although this feature allows the inclusion of more samples to the analysis, it also adds an additional complication because it is possible for a genotype with missing data to match two or more other genotypes that do not match each other. GENECAP recognizes these ambiguous samples, excludes them from matching genotypes, and outputs their identification numbers to a worksheet for the user to examine.

One and Two Allele Miss-matches

One way to identify genotyping errors is to carefully scrutinize the genetic data for genotypes that are very similar and only differ at one or two loci (Paetkau 2003). To assist the researcher in screening data, GENECAP identifies the sample identification numbers and the genotypes of samples that differ by one and two alleles. GENECAP outputs the sample identification number, genotype, location, and sample period of both samples. This allows the user to examine these mismatches and correct for any potential errors. Also, the entire distribution of intergenotype differences is output to allow the user to compare the shape of the distribution to that expected from high-quality samples.

Acknowledgements

We thank K. Scribner for input in creating GENECAP and reviewing the program, user's manual, and this manuscript. K. Bennett-Filek and S. Libants reviewed the program and user's manual. G. Luikart provided helpful comments that improved the program and manuscript. This project was supported in part by the Federal Aid in Wildlife Restoration Act under Pittman-Robertson project W-147-R-109, the Michigan Department of Natural Resources, Michigan Sea Grant, and the Great Lakes Fish and Wildlife Restoration Program administered by U.S. Fish and Wildlife Service.

References

Ayres KL, Overall ADJ (2004) API-CALC 1.0: a computer program for calculating the average probability of identity allowing for substructure, inbreeding and the presence of close relatives. *Molecular Ecology Notes*, **4**, 315–318.

Evett IW, Weir BS (1998) *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists*. Sinnauer Associates Inc., Maine, USA.

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.

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

Taberlet P, Waits LP, Luikart G (1999) Noninvasive genetic sampling: look before you leap. *Trends in Ecology and Evolution*, **14**, 323–327.

Valière N (2002) GIMLET: a computer program for analysing genetic individual identification data. *Molecular Ecology Notes*, **2**, 377–379.

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.

White GC, Burnham KP (1999) Program MARK. *Survival Estimation from Populations of Marked Animals*. *Bird Study Suppl*., **46**, 120–138.

Woods JG, Paetkau D, Lewis D, McLellan BN, Proctor M, Strobeck C (1999) Genetic tagging of free-ranging black and brown bears. *Wildlife Society Bulletin*, **27**, 616–627.