COMPUTER PROGRAMS

ONeSAMP: a program to estimate effective population size using approximate Bayesian computation

DAVID A. TALLMON,* ALLY KOYUK,* GORDON LUIKART† and MARK A. BEAUMONT‡

**Biology Program, University of Alaska Southeast, 11120 Glacier Highway, Juneau, AK 99801, USA, †Division of Biological Sciences, DBS/HS 104, University of Montana, 32 Campus Drive, Missoula, MT 59812, USA, ‡School of Animal and Microbial Sciences, University of Reading, Whiteknights, Reading RG6 6AJ, UK*

Abstract

The estimation of effective population size from one sample of genotypes has been problematic because most estimators have been proven imprecise or biased. We developed a web-based program, ONeSAMP that uses approximate Bayesian computation to estimate effective population size from a sample of microsatellite genotypes. ONeSAMP requires an input file of sampled individuals' microsatellite genotypes along with information about several sampling and biological parameters. ONeSAMP provides an estimate of effective population size, along with 95% credible limits. We illustrate the use of ONeSAMP with an example data set from a re-introduced population of ibex *Capra ibex*.

Keywords: approximate Bayesian computation, bottleneck, conservation genetics, effective population size, microsatellite, *Ne*

Received 5 June 2007; revision accepted 29 August 2007

The effective size of a population (N_e) strongly influences the relative impact of different microevolutionary forces (e.g. drift vs. selection). Populations with recently reduced N_e may suffer increased risk of fitness loss or extinction (Newman & Pilson 1997). Consequently, a great deal of effort has been focused on estimating N_e. N_e estimators that require two temporally spaced genetic samples have been successful for a wide variety of organisms (Waples 1991). However, an obvious disadvantage of two-sample N_e estimators is that often, one must wait a generation or more between sampling events to estimate N_e. For long-lived species, this can be prohibitive. Although one-sample N_e estimators hold the advantage of requiring only a single sample, they have not been used much because they are often imprecise and biased (Waples 1991; England *et al*. 2006). Nonetheless, recent improvements in these estimators and the increasing availability of large, highly polymorphic data sets are likely to make one-sample N_e estimators more useful (Luikart & Cornuet 1999; Waples 2006).

We developed a novel one-sample estimator, ONeSAMP, which uses summary statistics and approximate Bayesian computation to estimate N_e from a single sample of microsatellite data. We have previously demonstrated how this statistical approach can be used to estimate N_e from two samples (Tallmon *et al*. 2004), and this statistical approach is well established for use in other population genetics problems (Beaumont *et al*. 2002). ONeSAMP should be of wide general interest because it requires summary statistics calculated only from a single sample, and can be accessed and used online (http://genomics.jun.alaska.edu/). The first of several inputs the user must provide are the numbers of individuals and loci genotyped from the target population with unknown N_e. There must be at least two polymorphic loci in the sample and all loci are assumed to be unlinked and neutral. The user must provide the repeat motif of each locus. In addition, the lower and upper bounds for N_e of the target population are required. In a Bayesian context, this information is known as the prior on N_e. This prior should include the user's best guess (or estimate using independent data) of the true N_e of the target population. For example, if the population size is approximately 50, a range of N_e from four to 100 might be appropriately conservative because N_e is generally much lower than the population size. The microsatellite genotypes for each individual must be provided in GENEPOP format (Raymond & Rousset 1995), along with a return e-mail address for results.

Correspondence: David A. Tallmon, Fax: 1(907)796 6447; E-mail: david.tallmon@uas.alaska.edu

ONESAMP calculates eight summary statistics from the target data set input by the user. We selected eight summary statistics for which population genetics theory or our own simulations established a relationship with N_e. The statistics include: the number of alleles divided by allele length range (Garza & Williamson 2001), the difference of the natural logarithms of variance in allele length and heterozygosity (King *et al*. 2000), expected heterozygosity (Nei 1987), number of alleles per locus, Wright's F_{IS} (Nei 1987), the mean and variance of multilocus homozygosity, and the square of the correlation of alleles at different loci (Hill 1981).

Using information provided by the user, ONESAMP creates 50 000 simulated populations. Each simulated population has an effective size drawn from a uniform random number between the lower and upper N_e specified by the user in the prior. Each population is assumed to come from a population with an initial level of genetic variation determined by theta: the product of its historic effective size and the mutation rate ($4N_e^*\mu$). This theta value is randomly drawn for each population from a uniform random number between two and 12. Each simulated population reproduces following a Wright–Fisher model for two to eight generations before being sampled. Again, the exact number of generations is drawn from a uniform random number between these values.

For each simulated population, ONESAMP draws samples with identical numbers of individuals and loci to those contained in the target data set. The N_e values from simulated populations with summary statistic values close to the summary statistic values from the target population are accepted. Then, the N_e values from the accepted simulated populations are used in a weighted local regression to infer the effective size of the target population.

This approximate Bayesian computation approach is especially useful when inferences about some parameter of interest, Φ, are difficult to make using full likelihoods. This approach can be described more formally as follows. In this method, J values of Φ_i are simulated from a prior distribution, $\Phi_i \sim P(\Phi)$. For each Φ_i, a data set, D_i, is simulated using a Wright–Fisher model. Summary statistics, S_i, are then calculated from the data and scaled to have unit variance. Thus, the S_i, and Φ_i are drawn from the joint distribution $P(S,\Phi)$. The posterior distribution $P(\Phi \mid S = S^*)$ is the conditional distribution of Φ given the target summary statistics S^*, calculated from the sample data. To approximate this, the simulated candidate value Φ_i and associated S_i are accepted when the Euclidean distance $\| S_i - S^* \| < g$, where g defines a distance such that a proportion d_g of points closest to S^* are accepted (Tavaré *et al*. 1997).

To improve the accuracy of the rejection sampling method, we follow the approach of (Beaumont *et al*. 2002). Each accepted Φ_i is given a weight that declines quadratically as a function of $\| S_i - S^* \|$ from 1 at distance 0 to 0 at distance g, and then weighted linear regression is used to adjust the values of Φ_i. The method fits a regression line such that each $\Phi_i = a + b\,S_i + e_i$, and then, assuming constant variance within the interval given by $\| S_i - S^* \| < g$, makes the adjustment $\Phi'_i = a + b(S^* - S_i)$. These Φ'_i are then assumed to be random samples from the posterior distribution $P(S,\Phi)$, which, depending on how close to sufficient are the summary statistics, is itself assumed to be close to $P(D \mid \Phi)$. We use a Box-Cox transformation ($\lambda = -0.2$) of N_e in all regressions in order to ensure that the values of Φ_i are robust to changes in g. Values of N_e accepted within $d_g = 0.02$, as described above, are then regarded as samples from the posterior distribution of N_e.

To illustrate the use of ONESAMP on a published data set of modest size, we estimated the effective size of the ibex *Capra ibex* population inhabiting the Belledone Mountains of the French Alps. This population was re-introduced in 1983 with 20 individuals, of which seven were males (Maudet *et al*. 2002). This population grew rapidly in the three to four generations following re-introduction to ~800 individuals before it was sampled. The sample of 26 individuals was genotyped at 19 loci (Maudet *et al*. 2002). We used this published data set, excluding two loci, SR-CSRP-24 and SR-CSRP-6, which were far out of Hardy–Weinberg proportions and were monomorphic, respectively. We used upper and lower bounds on the prior for N_e of two and 100, respectively.

The estimated mean $\hat{N}_e = 19.06$ (95% CL = 16.59–24.10) from the posterior distribution of N_e for the Belledone ibex population is consistent with the known history of this population (Maudet *et al*. 2002). In addition, the results for this data set are robust to changes in the prior. With priors on N_e of 2–400, and 4–1000, respective $\hat{N}_e$ values of 19.96 (95% CL = 17.02–26.92) and 18.35 (95% CL = 15.96–22.44) were obtained. These results and additional testing of ONESAMP using more extensive simulations provide further evidence that this approach is a useful one for estimating N_e.

Acknowledgements

D.A.T. was supported by a National Science Foundation postdoctoral grant (INT 0202707) and Alaska EPSCoR. G.L. was supported by the European Commission (Econogene QLK5-CT-2001–02461). M.A.B. is supported by a UK Natural Research Counsel Advanced Fellowship.

References

Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035.

England PR, Cornuet J-M, Berthier P, Tallmon DA, Luikart G (2006) Estimating effective population size from linkage disequilibrium: severe bias in small samples. *Conservation Genetics*, **7**, 303–308.

Garza JC, Williamson EG (2001) Detection of reduction in population size using data from microsatellite loci. *Molecular Ecology*, **10**, 305–318.

Hill WG (1981) Estimation of effective population size from data on linkage disequilibrium. *Genetical Research*, **38**, 209–216.

King JP, Kimmel M, Chakraborty R (2000) A power analysis of microsatellite-based statistics for inferring past population growth. *Molecular Biology and Evolution*, **17**, 1859–1868.

Luikart G, Cornuet J-M (1999) Estimating the effective number of breeders from heterozygote excess in progeny. *Genetics*, **151**, 1211–1216.

Maudet C, Miller C, Bassano B *et al*. (2002) Microsatellite DNA and recent statistical methods in wildlife conservation management: applications in Alpine ibex [*Capra ibex* (*ibex*)]. *Molecular Ecology*, **11**, 421–436.

Nei M (1987) *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Newman D, Pilson D (1997) Increased probability of extinction due to decreased genetic effective size: experimental populations of *Clarkia pulchella*. *Evolution*, **51**, 354–362.

Raymond M, Rousset F (1995) GENEPOP version 1.2: population genetics software for exact tests and ecumenicism. *Journal of Heredity*, **86**, 248–249.

Tallmon DA, Luikart G, Beaumont MA (2004) A comparative evaluation of a new effective population size estimator based on approximate Bayesian computation. *Genetics*, **167**, 977–988.

Tavaré SJ, Balding DJ, Griffiths RC, Donnelly P (1997) Inferring coalescence times from DNA sequence data. *Genetics*, **145**, 505–518.

Waples RS (1991) Genetic methods for estimating the effective size of cetacean populations. In: *Report of the International Whaling Commission* (ed. Hoezel AR), pp. 279–300. International Whaling Commission, UK.

Waples RS (2006) A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci. *Conservation Genetics*, **7**, 167–184.