

Effective population sizes: missing measures and missing concepts

Keith A. Crandall, David Posada and Dan Vasco

Department of Zoology, 574 Widtsoe Building, Brigham Young University, Provo, UT 84602-5255, USA

(Received 2 February 1999; accepted 26 April 1999)

Abstract

Genetic diversity within a population is a critical parameter for conservation biologists to assess. One approach for examining genetic diversity is through the determination of effective population sizes. *Animal Conservation* recently published a review of genetic approaches for estimating effective population sizes. We felt this review was incomplete both methodologically and conceptually. Here we extend that review to include coalescent-based approaches for estimating effective population sizes. We then discuss different kinds of effective population sizes, including inbreeding, variance and eigenvalue effective sizes.

The concept of effective population size has played an important role in the conservation genetics of endangered species. Therefore, it is indeed important for a conservation journal such as this to publish appropriate reviews of such a concept to keep its readers abreast of recent methodological and theoretical developments. Such a review was recently offered (Schwartz, Tallmon & Luikart, 1998). Unfortunately, this review falls short on three counts. First, those authors missed the entire field of coalescent theory, which has provided a number of estimators for inbreeding effective population sizes based on nucleotide sequence and microsatellite data. Second, those authors failed to make the distinction between methods that estimate effective population sizes for present-day populations *versus* those used to estimate historical effective sizes. Finally, those authors failed to distinguish among the various types of effective population sizes (inbreeding, variance, eigenvalue, etc). This led the authors to erroneous conclusions about the relationships between effective population sizes and census sizes. Here we summarize recent developments in population genetic theory as they pertain to the estimation of effective population size and then reiterate the distinctions among the different types of effective population sizes as previously outlined by Ewens (1982).

THEORETICAL APPROACHES TO ESTIMATING EFFECTIVE POPULATION SIZE

Quantifying the amount of genetic diversity within a population or species is of the utmost importance in

conservation genetics. With such information, better decisions can be made about breeding designs and artificial migration events. The population genetic parameter $\theta = 4N_e\mu$, where N_e is the effective population size and μ is the mutation rate per nucleotide (or per locus) per generation, leads to an expectation of the amount of genetic diversity in a population. With a known mutation rate and an estimate of θ , we can then estimate the effective population size of the population

$$N_e = \frac{\theta}{4\mu}.$$

Historically, genetic diversity has been estimated in two ways assuming a sample of n DNA sequences. First, Watterson (1975) estimated genetic diversity as a function of the number of segregating sites: $\hat{K} = K/a_n$, where a_n is given by:

$$a_n = 1 + \frac{1}{2} + \dots + \frac{1}{n-1} \quad (1)$$

Nei & Tajima (1981) offered an alternative estimator of genetic diversity based on the proportion of nucleotide differences per pair of sequences, $\hat{\pi}$. The variances of \hat{K} and $\hat{\pi}$ where given by Watterson (1975) and Tajima (1983), respectively, as:

$$\text{Var}(\hat{K}) = \frac{\theta}{a_n} + \frac{\theta^2}{a_n^2} \sum_{k=1}^{n-1} \frac{1}{k^2} \quad (2)$$

$$\text{Var}(\hat{\pi}) = \frac{n+1}{3(n-1)}\theta + \frac{2(n^2+n+3)}{9n(n-1)}\theta^2 \quad (3)$$

where a^n is given by eqn. (1). Under the assumptions of the infinite-sites Wright–Fisher model, including no recombination, no selection, and constant effective population size, these estimators are unbiased. The populations of inference for these two estimators are different. The population of inference for $\hat{\pi}$ is the current population whereas the population of inference for \hat{K} is the historical population. Thus, contrasting these two estimators can lead to insights in the change of population dynamics over time (Templeton, 1993). However, the variance associated with these estimates can be large. Therefore, researchers have turned to the information incorporated in the genealogical relationships among sequences for better estimates of effective population size.

Felsenstein (1992a) demonstrated both the inefficiency of the pairwise and segregating sites approaches (eqns (1–3)), using both theoretical arguments and computer simulations. Likewise, Fu & Li (1993) also demonstrated the inefficiencies of these two estimators. Both research teams incorporated genealogical information to obtain high efficiency estimators of genetic diversity relative to the pairwise and segregating sites methods (Felsenstein, 1992a,b; Fu, 1994a,b; Kuhner, Yamato & Felsenstein, 1995). The approach used by Fu (1994b) is an unbiased linear estimator of genetic diversity over a single genealogy. This estimator assumes the genealogy to be correct (topology and branch lengths), no selection, constant population size, no recombination and no migration. The method described by Kuhner *et al.* (1995), however, is a maximum likelihood approach that samples over a range of weighted genealogies (weighted by their likelihoods). This method also assumes no selection, constant population size, no recombination and no migration. Both approaches have recently been improved to account for growing (or declining) population sizes (Kuhner, Yamato & Felsenstein 1998; Vasco & Fu, 1999). These approaches have the historical population as the population of inference, that is, they estimate the amount of genetic diversity over evolutionary time by inferring genealogical relationships among genes and fitting these genealogical relationships to a model of genetic drift.

Griffiths & Tavaré (1994a) have developed an alternative maximum likelihood approach for estimating genetic diversity based on a Markov Chain Monte Carlo method. This approach is more general than those described above in that it can be applied to either genealogical models or to classical models based on allele frequencies. This method calculates the probabilities of an allele configuration given a population size and mutation rate. Therefore the population of inference for this method can be either the historical population or the current population. Recently, Nielsen (1997) developed a likelihood estimator of θ based on the model described by Griffiths & Tavaré (1994b) for microsatellite data. O’Ryan *et al.* (1998) also modified the Griffiths & Tavaré approach to estimate recent population sizes in South African buffalo populations using microsatellite data. These estimates of effective population size

using nucleotide sequence data and microsatellite data are clearly of great interest to conservation geneticists who are collecting such data at an ever-increasing pace. The ability to use different methods to obtain both historical and current estimates of effective population size allows conservation biologists to determine the effects of recent changes in demographics (e.g. due to habitat destruction, fragmentation, etc) on genetic diversity in populations.

DIFFERENT EFFECTIVE POPULATION SIZES

Assuming a standard Wright–Fisher model, we can describe the change in gene frequencies from generation t to generation $t + 1$ by a binomial random variable with index $2N$ and parameter $X(t)/2N$. Given that $X(t) = i$, the probability p^{ij} that $X(t+1) = j$ is given by (Ewens, 1979):

$$P_{ij} = \binom{2N}{j} \left(\frac{i}{2N} \right)^j \left\{ 1 - \left(\frac{i}{2N} \right) \right\}^{2N-j}, \quad i, j = 0, 1, 2, \dots, 2N. \quad (4)$$

From this basic model, Ewens (1979, 1982) reviewed three different measures of effective population size: inbreeding effective population size, variance effective population size and eigenvalue effective population size. Ewens (1982) also demonstrated that these measures can theoretically differ from one another. Caballero (1994) reviewed extensions of these concepts to specific genetic systems (e.g. X-linked genes). Subsequent work has extended these concepts to sex-linked genes (Caballero, 1995), genetic systems under selection (Santiago & Caballero, 1995) and linked neutral loci in populations under directional selection (Santiago & Caballero, 1998). Likewise, the behaviour of different effective population sizes has recently been explored in subdivided populations (Whitlock & Barton, 1997) and in hierarchically structured populations (Nunney, 1999). Whitlock & Barton (1997) demonstrated that these different estimates of effective population size were identical under equilibrium conditions with a constant and large population size. However, in conservation biology, we tend to study small populations that are not in equilibrium, thereby requiring the distinction among these different effective population sizes. The conservation community has largely ignored this distinction and the different calculations under different population histories. Below we demonstrate the practical importance of such distinctions.

A common misconception resulting from the failure to distinguish between different types of effective population sizes is that stated by Schwartz *et al.* (1998): ‘a decrease in N_c (the census size) is accompanied by a decrease in effective population size’. Templeton & Read (1994) have shown in a captive population of Speke’s gazelle (*Gazella spekei*) that the inbreeding effective population size can actually be larger than the census population size and through the avoidance of inbreeding can increase even while the census size decreases. Likewise, a study of the glade grasshopper

(*Trimerotropis saxatilis*) demonstrated again how the inbreeding effective size can be larger than the census size (Gerber & Templeton, 1996). Larger values of inbreeding effective population size relative to census size are also supported by theoretical arguments suggesting that different population sizes can vary by orders of magnitude (Templeton, 1980).

In both of these cases, the larger inbreeding effective population size was due to a historically large population and the low census size to recent population reductions. The variance and eigenvalue effective population sizes for these populations were low, on the order of the census size. Thus, recognizing the differences in effective population size measurements allows one to partition historical and current phenomena, similar to the distinction between the estimates of genetic diversity discussed above. The inbreeding effective population size is a backward looking statistic whereas the variance and eigenvalue effective population sizes reflect recent demographic/population genetic processes influencing genetic systems. Large inbreeding effective population sizes and small variance effective population sizes are indicative of recently reduced genetic variation due to decreases in population size or habitat fragmentation (Gerber & Templeton, 1996). In contrast, with a rapid increase in population size, theory predicts a small inbreeding effective population size and large variance and eigenvalue effective sizes (Templeton, 1980). Recognizing this distinction allows conservation geneticists insights into the past and a target for the future. Thus measuring different effective sizes becomes important not only for monitoring populations (variance effective size) but also as a guide for restoring populations to historical levels of genetic diversity (inbreeding effective population size).

Acknowledgements

We thank Alan Templeton and an anonymous reviewer for excellent suggestions for improving this manuscript. We also thank Michael Schwartz, David Tallmon and Gordon Luikart for initiating a helpful discussion on effective population size. This work was supported by an Alfred P. Sloan Foundation Young Investigator Award and NIH award number R01-HD34350-01A1.

REFERENCES

- Caballero, A. (1994). Developments in the prediction of effective population size. *Heredity* **73**: 657–679.
- Caballero, A. (1995). On the effective size of populations with separate sexes, with particular reference to sex-linked genes. *Genetics* **139**: 1007–1011.
- Ewens, W. J. (1979). *Mathematical population genetics*. Berlin: Springer-Verlag.
- Ewens, W. J. (1982). On the concept of the effective population size. *Theoret. Pop. Biol.* **21**: 373–378.
- Felsenstein, J. (1992a). Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. *Genet. Res. Camb.* **59**: 139–147.
- Felsenstein, J. (1992b). Estimating effective population size from samples of sequences: a bootstrap Monte Carlo integration method. *Genet. Res. Camb.* **60**: 209–220.
- Fu, Y.-X. (1994a). A phylogenetic estimator of effective population size or mutation rate. *Genetics* **136**: 685–692.
- Fu, Y.-X. (1994b). Estimating effective population size or mutation rate using the frequencies of mutations of various classes in a sample of DNA sequences. *Genetics* **138**: 1375–1386.
- Fu, Y.-X. & Li, W.-H. (1993). Maximum likelihood estimation of population parameters. *Genetics* **193**: 1261–1270.
- Gerber, A. S. & Templeton, A. R. (1996). Population sizes and within-deme movement of *Trimerotropis saxatilis* (Acrididae), a grasshopper with a fragmented distribution. *Oecologia* **105**: 343–350.
- Griffiths, R. C. & Tavaré, S. (1994a). Ancestral inference in population genetics. *Stat. Sci.* **9**: 307–319.
- Griffiths, R. C. & Tavaré, S. (1994b). Simulating probability distributions in the coalescent. *Theoret. Pop. Biol.* **46**: 131–159.
- Kuhner, M. K., Yamato, J. & Felsenstein, J. (1995). Estimating effective population size and mutation rate from sequence data using Metropolis–Hastings sampling. *Genetics* **140**: 1421–1430.
- Kuhner, M. K., Yamato, J. & Felsenstein, J. (1998). Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics* **149**: 429–434.
- Nei, M. & Tajima, F. (1981). DNA polymorphism detectable by restriction endonucleases. *Genetics* **97**: 145–163.
- Nielsen, R. (1997). A likelihood approach to populations samples of microsatellite alleles. *Genetics* **146**: 711–716.
- Nunney, L. (1999). The effective size of a hierarchically structured population. *Evolution* **53**: 1–10.
- O’Ryan, C., Harley, E. H., Bruford, M. W., Beaumont, M., Wayne, R. K. & Cherry, M. I. (1998). Microsatellite analysis of genetic diversity in fragmented South African buffalo populations. *Anim. Conserv.* **1**: 85–94.
- Santiago, E. & Caballero, A. (1995). Effective size of populations under selection. *Genetics* **139**: 1013–1030.
- Santiago, E. & Caballero, A. (1998). Effective size and polymorphism of linked neutral loci in populations under directional selection. *Genetics* **149**: 2105–2117.
- Schwartz, M. K., Tallmon, D. A. & Luikart, G. (1998). Review of DNA-based census and effective population size estimators. *Anim. Conserv.* **1**: 293–299.
- Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**: 437–460.
- Templeton, A. R. (1980). The theory of speciation via the founder principle. *Genetics* **94**: 1011–1038.
- Templeton, A. R. (1993). The ‘Eve’ hypothesis: a genetic critique and reanalysis. *Am. Anthropol.* **95**: 51–72.
- Templeton, A. R. & Read, B. (1994). Inbreeding: one word, several meanings, much confusion. In *Conservation Genetics*: 91–105. Loeschucke, V., Tomiuk, J. & Jain, S. K. (Eds). Basel, Switzerland: Birkhauser Verlag.
- Vasco, D. A. & Fu, Y.-X. (1999). A phylogenetic estimator of population growth rate. *Genetics* (In press).
- Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoret. Pop. Biol.* **7**: 256–276.
- Whitlock, M. C. & Barton, N. H. (1997). The effective size of a subdivided population. *Genetics* **146**: 427–441.