doi:10.1111/j.1558-5646.2010.00953.x

ESTIMATION OF PARAMETERS OF INBREEDING AND GENETIC DRIFT IN POPULATIONS WITH OVERLAPPING GENERATIONS

Jinliang Wang,[1,2] Patricia Brekke,[1] Elise Huchard,[3] Leslie A. Knapp,[4] and Guy Cowlishaw[1]

[1]*Institute of Zoology, Zoological Society of London, Regent's Park, London NW1 4RY, United Kingdom*

[2]*E-mail: Jinliang.wang@ioz.ac.uk*

[3]*CNRS – Institut des Sciences de l'Evolution, Université Montpellier II, Place Eugène Bataillon, CC 065, 34 095 Montpellier Cedex 5, France*

[4]*Department of Biological Anthropology, University of Cambridge, Downing Street, Cambridge CB2 3DZ, United Kingdom*

Received September 11, 2009
Accepted December 21, 2009

Many long-lived plant and animal species have nondiscrete overlapping generations. Although numerous models have been developed to predict the effective sizes (N_e) of populations with overlapping generations, they are extremely difficult to apply to natural populations because of the large array of unknown and elusive life-table parameters involved. Unfortunately, little work has been done to estimate the N_e of populations with overlapping generations from marker data, in sharp contrast to the situation of populations with discrete generations for which quite a few estimators are available. In this study, we propose an estimator (EPA, estimator by parentage assignments) of the current N_e of populations with overlapping generations, using the sex, age, and multilocus genotype information of a single sample of individuals taken at random from the population. Simulations show that EPA provides unbiased and accurate estimates of N_e under realistic sampling and genotyping effort. Additionally, it yields estimates of other interesting parameters such as generation interval, the variances and covariances of lifetime family size, effective number of breeders of each age class, and life-table variables. Data from wild populations of baboons and hihi (stitchbird) were analyzed by EPA to demonstrate the use of the estimator in practical sampling and genotyping situations.

KEY WORDS: **Effective population size, generation interval, genetic markers, overlapping generations, parentage assignments.**

Inbreeding and genetic drift are two closely related yet distinctive stochastic processes characterizing populations of finite sizes. They act and interact to cause the erosion of neutral genetic variation (e.g., Frankham et al. 2003), the decline in fitness, and the increase in extinction risk due to the excessive accumulation and expression of deleterious mutations (e.g., Lynch et al. 1995), and the loss of adaptive evolution due to reduced efficacy of positive selection (Crow and Kimura 1970). In real populations, many factors (such as population size, sex ratio, variance in fecundity and viability, and mating system) affect the strength of inbreeding and genetic drift processes, but can be conveniently summarized into a single parameter, the effective population size (Wright 1931, 1938). Using this parameter, one cannot only explain the current levels of neutral genetic variation and fitness of a population, but also predict their evolution in the future.

Because of the importance of effective size (N_e) in both population genetics theory and applied disciplines such as evolutionary biology, ecology and conservation, tremendous efforts have been made to predict the parameter from demographic models (reviewed by Caballero 1994; Wang and Caballero 1999) and to estimate the parameter from genetic marker data (reviewed by Schwartz et al. 1999; Beaumont 2003; Wang 2005). In the latter case, a number of methods have been proposed and applied based on the amount and pattern of genetic variation observed at

a number of marker loci in one or more samples of individuals. In particular, the heterozygote excess method (Pudovkin et al. 1996; Luikart and Cornuet 1999), linkage disequilibrium method (e.g., Hill 1981; Waples 2006), temporal method (e.g., Nei and Tajima 1981), and sibship assignment method (Wang 2009) have been developed in recent years to estimate the current or short-term effective size.

Most methods available for estimating N_e assume a population with discrete nonoverlapping generations. The assumption, however, is violated in many long-lived plant and animal species in which individuals can reproduce multiple times in their lifetime and individuals of different life stages coexist to breed. Realizing the importance of overlapping generations, theoretical population geneticists developed various demographic models to predict the N_e of populations with overlapping generations (Moran 1962; Kimura and Crow 1963; Nei and Imaizumi 1966; Felsenstein 1971; Crow and Kimura 1972; Hill 1972; Johnson 1977; Choy and Weir 1978; Emigh and Pollak 1979; Hill 1979; Pollak 1990; Charlesworth 2001). These models have greatly advanced our understanding of inbreeding and genetic drift in the case of overlapping generations, but are difficult to apply to natural populations because of the lack of information about the large array of life-history parameters (such as age- and sex-specific reproduction and survival rates) involved in the models. There is still, therefore, an urgent need to develop practical methods that can use genetic marker data to estimate the N_e of populations with overlapping generations. Unfortunately, however, little work has been done (but see Jorde and Ryman 1995; Waples and Yokota 2007) in this regard.

For populations with discrete generations, the most widely applied method for estimating N_e has been the temporal method (Krimbas and Tsakas 1971; Nei and Tajima 1981; Waples 1989), so called because it measures and uses the changes in allele frequencies between temporally spaced samples taken from the same population. The method should apply approximately to populations with overlapping generations when the time interval between samples is much longer than the generation interval (Nei and Tajima 1981), as verified by simulations (Waples and Yokota 2007). This is understandable because with a long sampling interval the observed changes in sample allele frequencies would come mainly from genetic drift rather than the internal age-sex structures and small sizes of the samples. However, a direct application of the temporal method to populations with overlapping generations has met several difficulties. First, species with overlapping generations generally have a long individual life expectancy and the minimum period of a single generation required for the temporal method may well exceed the life span of a typical research project. Second, a much longer sampling interval is required for populations with overlapping generations than those with discrete generations to reach the same accuracy of estimation under similar conditions (e.g., comparable values of N_e and sample size), because the heterogeneity of the age-structured samples incurs noises additional to those caused by small sample sizes, as shown by simulations (Jorde and Ryman 1995; Waples and Yokota 2007). Indeed the allele frequency difference among age classes, such as genetic drift, is a property of the population determined by its life-table, so its effect on N_e estimation cannot be eliminated or reduced by increasing sample size or the number of markers. Third, how to estimate allele frequencies from the heterogeneous samples (i.e., whether to weight individuals by their reproductive values or not) is a difficult problem but proves to be critical in determining the accuracy of N_e estimates (Waples and Yokota 2007).

Jorde and Ryman (1995) modified the temporal method to apply it to populations with overlapping generations. They used the observed allele frequency shifts between consecutive cohorts to measure genetic drift, so that the weighting problem mentioned above is avoided because the cohort-stratified samples are homogenous. They took the effect of age structure into account by a correction constant (C, in their notation), which is a function of the age-specific reproduction and survival rates. This method is shown to be almost unbiased (Jorde and Ryman 1995; Waples and Yokota 2007) when 5000 markers are used in the estimation in simulations. However, consecutive cohorts are expected to have very small differences in allele frequency even for a small population, because they share a substantial proportion of their genes coming from the same set of parents. Therefore, many loci (in the thousands) are necessary to measure genetic drift accurately. Furthermore, the calculation of C requires detailed life-table information such as the age-specific reproduction and survival rates, which are usually numerous and difficult to obtain from natural populations. In fact, the modified methods need such extensive information that it actually allows only one parameter in the model of overlapping generations, the absolute size of a single age class, to be unknown.

In this study, we propose a new method to estimate the effective size and the related causal parameters (such as generation interval and variance of lifetime family sizes) for populations with overlapping generations. The method uses the multilocus genotypes of a single sample of individuals taken at random from the population to assign parentage among sampled individuals. It then estimates N_e and its causal parameters from the inferred parentage by a likelihood approach. The only information additional to genotypes required by the method is the age and sex of each sampled individual, which is relatively easy to obtain for many species. We will first briefly describe the genetic model of populations with overlapping generations on which our method is based. We then propose a likelihood estimator of the parameters of the model based on parentage assignments between age classes in a sample of individuals. Simulations are conducted to

investigate the performance of the estimator in estimating N_e and its causal parameters. The simulation results are helpful in understanding the behavior of the estimator and the factors affecting its accuracy, providing useful information for practical applications. Finally, we analyze two empirical datasets to demonstrate the use of the proposed estimator in practical situations.

A Genetic Model of Populations with Overlapping Generations

Among the various models of inbreeding and genetic drift for populations with overlapping generations (e.g., Felsenstein 1971; Hill 1972; Johnson 1977), we focus on the inbreeding model of a dioecious diploid species proposed by Johnson (1977), on which our estimator of N_e is based.

The model assumes a population consisting of N_1 diploid males and N_2 diploid females distributed in n_1 and n_2 age classes, respectively. Age classes can be defined by years or any other time units. For convenience, we use year as the time unit so that individuals in age class i are i-year-olds for both males and females. The number of individuals in age class i and sex s ($s = 1, 2$; $1 \le i \le n_s$) is assumed to be fixed at $N_{i,s}$ in any year, so that $\sum_{i=1}^{n_s} N_{i,s} = N_s$. Each year $N_{1,s}$ 1-year-olds enter the population whereas death claims all $N_{n_s,s}$ n_s-year-olds and a random sample of $N_{i,s} - N_{i+1,s}$ ($1 \le i \le n_s - 1$) i-year-olds, for each sex $s = 1$, 2. The probability of survival to age i is then $l_{i,s} = N_{i,s}/N_{1,s}$ for $i = 1 \sim n_s$, and the age-specific survival rate is then $N_{i+1,s}/N_{i,s} = l_{i+1,s}/l_{i,s}$ and 0 for $i < n_s$ and $i = n_s$, respectively, of sex s.

Reproduction is assumed to occur at random within and between age classes. Between age classes, sampling of gametes is accommodated by a set of parameters $p_{rs,i}$, where $p_{rs,i}$ is the probability that a random gamete contributing to the conception of a newborn individual of sex r ($r = 1, 2$) in any year came from the ith ($1 \le i \le n_s$) age class of parents of sex s ($s = 1, 2$) in the previous year. Because half of the genes of an individual came from its mother and half from its father, we have

$$\sum_{i=1}^{n_1} p_{r1,i} = \sum_{i=1}^{n_2} p_{r2,i} = \frac{1}{2}, \quad r = 1, 2.$$

Within an age class, each individual is equally likely to contribute gametes to the next generation. Parameters $p_{rs,i}$ determine the parental age distribution of newborns, and can be estimated from age-specific reproduction rates and age class sizes in practice. If an i-year-old parent of sex s contributes on average $m_{rs,i}$ newborns of sex r, then $p_{rs,i}$ is calculated as $p_{rs,i} = \frac{1}{2} m_{rs,i} N_{i,s} / \sum_{i=1}^{n_s} m_{rs,i} N_{i,s}$.

For dioecious species, there are four pathways of genes: father to son, father to daughter, mother to son, and mother to daughter. The average lengths of the four pathways are

$$\begin{aligned} L_{11} &= 2\sum_{i=1}^{n_1} q_{11,i}, \quad L_{21} = 2\sum_{i=1}^{n_1} q_{21,i}, \\ L_{12} &= 2\sum_{i=1}^{n_2} q_{12,i}, \quad L_{22} = 2\sum_{i=1}^{n_2} q_{22,i}, \end{aligned} \tag{1}$$

where L_{rs} is the average age of parents of sex s when their offspring of sex r are born, and $q_{rs,i} = \sum_{j=i}^{n_s} p_{rs,j}$ is the reproductive value of age class i ($1 \le i \le n_s$) and sex s for producing offspring of sex r ($r, s = 1, 2$). Generation interval is the average of the lengths of the four pathways

$$L = \frac{1}{4}(L_{11} + L_{21} + L_{12} + L_{22}) = \frac{1}{2}\sum_{i=1}^{n_1} q_{1,i} + \frac{1}{2}\sum_{i=1}^{n_2} q_{2,i}, \tag{2}$$

where $q_{s,i} = q_{1s,i} + q_{2s,i}$ for $s = 1$, 2 and $1 \le i \le n_s$. Generation interval can also be written as $L = \frac{1}{2}(L_1 + L_2)$, where $L_s = \sum_{i=1}^{n_s} q_{s,i}$ is the paternal ($s = 1$) or maternal ($s = 2$) generation interval.

Under the above inbreeding model, Johnson (1977) and Emigh and Pollak (1979) derived the equation for effective size

$$\frac{1}{N_e} = \frac{1}{4L}\sum_{s=1}^{2}\left[\frac{1}{N_{1,s}} + \sum_{i=2}^{n_s} q_{s,i}^2\left(\frac{1}{N_{i,s}} - \frac{1}{N_{i-1,s}}\right)\right], \tag{3}$$

and the variances and covariances in lifetime family sizes

$$\begin{aligned} \sigma_{sr}^2 &= \frac{N_{1,r}}{N_{1,s}} + 4\frac{N_{1,r}^2}{N_{1,s}}\sum_{i=2}^{n_s} q_{rs,i}^2\left(\frac{1}{N_{i,s}} - \frac{1}{N_{i-1,s}}\right), \quad r, s = 1, 2 \\ \sigma_{s1,s2} &= 4N_{1,3-s}\sum_{i=2}^{n_s} q_{1s,i} q_{2s,i}\left(\frac{1}{N_{i,s}} - \frac{1}{N_{i-1,s}}\right), \quad s = 1, 2. \end{aligned} \tag{4}$$

In (4), σ_{sr}^2 is the variance of the lifetime number of offspring of sex r per parent of sex s, and $\sigma_{s1,s2}$ is the covariance between the lifetime numbers of sons and daughters per parent of sex s (r,$s = 1$, 2).

An Estimator of N_e for Populations with Overlapping Generations

The parameters (e.g., $N_{i,s}$, $q_{s,i}$) in (2–4) are difficult to acquire from a natural population without detailed long-term pedigree records. We propose to use marker-based parentage assignments to estimate these parameters and thus L and N_e.

We assume that a sample of individuals is taken at random (with respect to kinship) from the focal population that follows the assumptions of the genetic model described above. It is not necessary for sampling to be at random between sexes or between age classes, for example, one particular age class (say, 1-year-old males) may be sampled at a much higher or lower proportion than others. Each sampled individual is sexed, aged, and genotyped at a number of marker loci. Using methods already available (e.g., Marshall et al. 1998; Wang and Santure 2009), the genotype,

sex, and age information of sampled individuals can be used for parentage assignments, which are then employed by the estimator below for estimating the N_e of the population.

Let us consider the parameters for males. For clarity in this section, all subscripts denoting males are omitted so that, for example, n, $p_{r,i}$ and N_i refer to n_1, $p_{r1,i}$ and $N_{i,1}$, respectively. Let us denote the observed number of i-year-old males who are included in the sample as x_i, and the unknown parameter of the sampling proportion of this age class as S_i. Given the parameter set $\mathbf{S} = \{S_1, S_2, \ldots, S_n\}$ and $\mathbf{N} = \{N_1, N_2, \ldots, N_n\}$, the count data, $\mathbf{x} = \{x_1, \ldots, x_n\}$, follow the multivariate hypergeometric distribution,

$$\Pr(\mathbf{x} \mid \mathbf{S}, \mathbf{N}) = \frac{\left(\sum_{i=1}^{n} N_i S_i\right)!}{\left(\sum_{i=1}^{n} x_i\right)! \left(\sum_{i=1}^{n} N_i S_i - \sum_{i=1}^{n} x_i\right)!} \times \prod_{i=1}^{n} \frac{(N_i S_i)!}{x_i!(N_i S_i - x_i)!}. \quad (5)$$

Assuming that a newborn individual is equally probable to be either sex (i.e., unbiased primary sex ratio), we have $p_{1,i} \equiv p_{2,i}$ for age class $1 \le i \le n$. This assumption is plausible and should be valid for many natural populations. Let us denote the observed numbers of male and female j-year-olds whose paternity is unassigned and assigned to males of age i as $y_{j,n+1}$ and $y_{j,i}$, respectively, where $0 \le j \le i-1$. The total number of j-year-olds in the sample is then $Y_j = \sum_{i=j+1}^{n+1} y_{j,i}$. The vector $\mathbf{y}_j = \{y_{j,j+1}, y_{j,j+2}, \ldots, y_{j,n+1}\}$ follows the multinomial distribution with parameters Y_j and $\{b_{j,j+1}, b_{j,j+2}, \ldots, b_{j,n+1}\}$, where $b_{j,i}$ is the probability that the father of a j-year-old drawn at random from the population is found in age class i ($j+1 \le i \le n$) or not found ($i = n+1$) in the sample. $b_{j,i}$ can be calculated from the sampling proportion (S_i), size (N_i), and reproductive contribution ($p_{1,i}$ or $p_{2,i}$) of male age class i, $b_{j,i} = 2p_{1,i-j} S_i N_i / N_{i-j}$ for $j+1 \le i \le n$ and $b_{j,n+1} = 1 - \sum_{i=j+1}^{n} b_{j,i}$. The overall probability of the observed paternity assignments $\mathbf{y} = \{\mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_{n-1}\}$, given parameter sets $\mathbf{S}$, $\mathbf{N}$, $\mathbf{p} = \{p_{1,1}, p_{1,2}, \ldots, p_{1,n}\}$, is

$$\Pr(\mathbf{y} \mid \mathbf{S}, \mathbf{N}, \mathbf{p}) = \prod_{j=0}^{n-1} Y_j! \prod_{i=j+1}^{n+1} \frac{b_{j,i}^{y_{j,i}}}{y_{j,i}!}. \quad (6)$$

Let us denote the number of individuals of any sex and age whose paternity is assigned to male k ($1 \le k \le x_i$) of age i ($1 \le i \le n$) as $z_{k,i}$, and the total number of individuals whose paternity is assigned to males of age i as $Z_i = \sum_{k=1}^{x_i} z_{k,i}$. Approximately $z_{k,i}$ follows a Poisson distribution with parameter $\mu_i = Z_i/(N_i S_i)$. The probability of the observed counts $\mathbf{z} = \{\mathbf{z_1}, \mathbf{z}_2, \ldots, \mathbf{z_n}\}$ with $\mathbf{z}_i = \{z_{1,i}, z_{2,i}, \ldots, z_{x_i,i}\}$ is

$$\Pr(\mathbf{z} \mid \mathbf{S}, \mathbf{N}) = \prod_{i=1}^{n} \prod_{k=1}^{x_i} \prod_{l=0}^{\infty} \frac{\mu_i^l}{l! e^{\mu_i}}. \quad (7)$$

Maximum likelihood estimates of parameters $\mathbf{N}$ and $\mathbf{p}$ as well as $\mathbf{S}$ can be obtained by maximizing the product of (5–7), with further constraints of $N_{i+1} \le N_i$ for $1 \le i \le n-1$. When $\mathbf{S}$ is known in the data, only $\mathbf{N}$ and $\mathbf{p}$ are estimated. We use Powell's quadratic convergence method (Press et al. 1996) to solve this constrained maximization problem. Each dataset is analyzed with 1000 replicates, each with a different set of starting values for the parameters and searching directions. Among these replicates, the solution with the maximal likelihood is returned as the best estimate. The starting values for parameters $\mathbf{p}$ and $\mathbf{S}$ and for searching directions are determined at random for each replicate whereas the starting value for N_i ($1 \le i \le n$) in replicate k is set as $10^{k/200} u^{1-k/1000} - i$, where u is the maximal number of sampled individuals in an age class. The value of N_i is constrained to the range of $[1, 10^5]$ whereas that of $2p_i$ or S_i is constrained to the range of [0, 1]. Numerical examples show that replicates with different starting points and searching directions are not guaranteed to converge to the same solution with the same likelihood, especially when the population is not intensively sampled for all age classes. However, 1000 replicates seem to be sufficient, as more replicates do not change the results of many simulated datasets appreciably.

With the same procedure, we can obtain maximum likelihood estimates of parameters for females, which are $N_{i,2}$ and $p_{r2,i}$ for $r = 1, 2$ and $1 \le i \le n_2$. Once we have obtained estimates for $N_{i,s}$ and $p_{rs,i}$ ($r, s = 1, 2$; $1 \le i \le n_s$) for both sexes, these values can be inserted into (2–4) to yield estimates of L and N_e, as well as the variances and covariances of lifetime family sizes for both males and females.

Parametric bootstrapping can be used to estimate the confidence intervals of N_e and L, and other parameters. We simulate a population characterized by the estimated values of $N_{i,s}$, and $p_{rs,i}$ ($s = 1, 2$; $1 \le i \le n_s$). After a sufficient number of generations for the population to reach the steady state of inbreeding, a sample similar to the real one is taken and used for parentage assignments and parameter estimation. Repeating this process a large number of times (say, 1000), we obtain a 95% confidence interval for each of the parameters of interest.

Some Extensions to the Basic Model

The estimation procedure described above assumes the inbreeding model of random births and deaths. When the assumption is violated, such as when individuals within an age class have fertilities or viabilities different in expectation, the $N_{i,s}$ value estimated by the above procedure is just the census size of age class i of sex s. If the effective size of age class i and sex s, $N_{ei,s}$, can be

estimated, then N_e can still be obtained by (3), replacing $N_{i,s}$ by $N_{ei,s}$. Similar to N_e, $N_{ei,s}$ is determined by census size $N_{i,s}$ and the variance in reproductive contribution among the $N_{i,s}$ individuals. As an example, let us consider the estimation of the effective size of male age class i, $N_{ei,1}$. The frequency with which two randomly selected individuals of any age and sex whose paternity is assigned to males of age i share the same father is

$$\hat{f}_{i,1} = \frac{\sum_{k=1}^{x_i} z_{k,i}(z_{k,i}-1)}{Z_i(Z_i-1)}, \tag{8}$$

where $z_{k,i}$ and Z_i are defined in deriving (7). With random births and deaths as assumed in the inbreeding model, we have $E(\hat{f}_{i,1}) \equiv 1/x_i$. The value of $\hat{f}_{i,1}$ is expected to be larger (smaller) than $1/x_i$ when the variance of family size is larger (smaller) than that of a Poisson distribution. Therefore, the effective size of male age class i is estimated as

$$\hat{N}_{ei,1} = \frac{N_{i,1}}{x_i \hat{f}_{i,1}} = \frac{Z_i(Z_i-1)N_{i,1}}{x_i \sum_{k=1}^{x_i} z_{k,i}(z_{k,i}-1)}. \tag{9}$$

When $Z_i \leq 1$ or no offspring are observed to share the same father in age class i, (9) becomes undefined because the denominator is zero. In such a case, we assume a Poisson distribution of family size so that $\hat{f}_{i,1} = 1/x_i$ and $N_{ei,1} = N_{i,1}$.

The parameters ($N_{i,s}$, $p_{rs,i}$) obtained by the estimator using autosomal markers can also be used to calculate the generation intervals and effective sizes for X-linked, Y-linked, and mtDNA loci of the same population. For X-linked loci, generation interval is estimated by $L_X = \frac{1}{3}L_1 + \frac{2}{3}L_2$ as 1/3 and 2/3 of the genes come from males and females, respectively. The effective size is calculated by Pollak (1990) as

$$\frac{1}{N_{e(X)}} = \frac{1}{9N_{1,1}L_X}\left[1 + 2\sigma_{12}^2/\mu_{12}^2\right] + \frac{1}{9N_{1,2}L_X}\left[1 + \sigma_{22}^2 + 2\sigma_{21,22}/\mu_{21} + \sigma_{21}^2/\mu_{21}^2\right],$$

where σ_{sr}^2 and $\sigma_{s1,s2}$ are calculated by (4), and $\mu_{sr} = N_{1,r}/N_{1,s}(r,s = 1,2)$. For Y-linked and mtDNA loci, the haploid model applies and the generation interval and effective size are estimated by L_s and $N_e = N_{1,s}L_s/\sigma_{ss}^2$ (Hill 1972), respectively, where $s = 1$ and 2 for Y-linked and mtDNA loci, respectively.

Simulations

We conducted simulations to check the performance of the estimator, and to investigate the factors that affect the accuracy of the estimator. For a given set of demographic parameters, $N_{i,s}$ and $p_{rs,i}$ ($s = 1, 2$; $1 \leq i \leq n_s$), the initial population is generated in which all individuals are outbred and unrelated. The population then evolves for N_e generations, following the genetic model described above, to reach the steady state of inbreeding and genetic drift. In each year during the N_e generations, reproduction is followed by death and survival events. For reproduction, male and female gametes that unite to form a newborn of sex r are taken from age class i ($1 \leq i \leq n_1$) in males and j ($1 \leq j \leq n_2$) in females with probabilities $2p_{r1,i}$ and $2p_{r2,j}$, respectively. Within an age class, individuals are assumed to have either equal or differential fertility in expectation. Under equal fertility, a gamete from an age class is equally probable to come from each individual in the class. Under differential fertility, half of the newborns of each sex are marked as high fertile (HF) and the other half as low fertile (LF). An HF individual at any age is expected to contribute m times the number of gametes of an LF individual of the same age. This differential fertility model leads to a variance in family size larger than that of the binomial distribution within any age class, and to an even greater variance in lifetime family size.

At generation $N_e + 1$, a sample of individuals is taken at random from the population, and each sampled individual is sexed, aged, and genotyped at a number of loci. The parent—offspring (PO) relationships among sampled individuals were either assumed known, or inferred (at 95% confidence level) from the marker, sex, and age data using the method of Marshall et al. (1998) implemented in the computer program Colony (Wang 2004; Wang and Santure 2009). The known or inferred parentage is then used by estimator by parentage assignments (EPA) to estimate the demographic parameters and thus L and N_e.

For a given set of parameters, $R = 1000$ replicate datasets are simulated and analyzed. The estimation accuracy of a parameter x is measured by the bias ($B = \mu_x - \bar{x}$, where mean estimate $\bar{x} = \frac{1}{R}\sum_{i=1}^{R}\hat{x}_i$ and μ_x is the parameter value), standard deviation ($SD = \sqrt{\frac{1}{R}\sum_{i=1}^{R}\hat{x}_i^2 - \bar{x}^2}$), and root mean squared errors (RMSE $= \sqrt{SD^2 + B^2}$). For N_e, $N_{ei,s}$, and $N_{i,s}$, accuracy is measured on the reciprocal of the parameter. The parameter value of generation interval is calculated by (2). With equal fertility, the parameter values for the variances of lifetime family sizes and N_e are calculated by (4) and (3), respectively. With differential fertility, the parameter values for the variances of lifetime family sizes are estimated by the average simulated values, which are then inserted into Hill's (1972) equation to calculate the parameter value of N_e.

Results

KNOWN PARENTAGE

In the best scenario, where the parentage of sampled individuals is known (e.g., from behavioral observations or from many informative markers) without error, the accuracy of EPA for the estimation of N_e is compared among different sampling intensities and schemes. The simulation results in Table 1 verify that EPA

Table 1. Accuracy of $1/N_e$ estimates from EPA when parentage is known.

Fertility	Parameter value ($10^3/N_e$)	Sampling percentage	Mean	SD	RMSE
Equal	1.486	2,2,2,2,2	1.807	2.454	2.475
		4,4,4,4,4	1.771	1.216	1.249
		8,8,8,8,8	1.611	0.5119	0.5269
		16,16,16,16,16	1.521	0.2210	0.2237
		32,32,32,32,32	1.447	0.0853	0.0939
		64,64,64,64,64	1.451	0.0572	0.0671
		100,100,100,100,100	1.486	0.0032	0. 0032
		5,8,11,17,25	1.614	0.3874	0.4079
		25,17,11,8,5	1.475	0.2934	0.2935
		0,16,16,16,16	1.836	0.3444	0.4908
		16,0,16,16,16	1.441	0.3190	0.3222
		16,16,0,16,16	1.502	0.3177	0.3181
Different	2.148	2,2,2,2,2	1.791	2.654	2.678
		4,4,4,4,4	2.276	2.221	2.225
		8,8,8,8,8	2.729	1.982	2.065
		16,16,16,16,16	2.608	1.212	1.295
		32,32,32,32,32	2.186	0.3862	0.3874
		64,64,64,64,64	2.169	0.1317	0.1333
		100,100,100,100,100	2.161	0.0817	0.0825
		5,8,11,17,25	2.734	2.028	2.111
		25,17,11,8,5	2.547	1.108	1.177

Data were simulated with parameters n_1=3, n_2=4, $\{p_{r1,1}, p_{r1,2}, p_{r1,3}\}$=$\{0.10, 0.25, 0.15\}$, $\{p_{r2,1}, p_{r2,2}, p_{r2,3}, p_{r2,4}\}$=$\{0.10, 0.15, 0.15, 0.10\}$, $\{N_{0,1}, N_{1,1}, N_{2,1}, N_{3,1}\}$=$\{300, 200, 150, 100\}$, $\{N_{0,2}, N_{1,2}, N_{2,2}, N_{3,2}, N_{4,2}\}$=$\{300, 200, 150, 100, 50\}$, where r=1, 2. The sampling proportions of 0∼3 year old males and of 0∼4 year old females are listed in column 3: the five figures thus indicate the percentage of new borns, 1-year olds, 2-year olds, 3-year olds, and 4-year olds (females only) sampled for each sex. The sex and age of the sampled individuals, as well as their parentage relationships, are assumed known. The parameter value of L is 2.3 years. Here, we compare the estimated values with the actual parameter values of $10^3/N_e$, where N_e and $10^3/N_e$ are 673 and 1.486 for equal expected fertility among individuals, and are 466 and 2.148 for unequal expected fertility among individuals, respectively. With equal fertility, each individual within an age class has an equal probability of reproductive contribution. With unequal fertility, half the newborns are marked as high fertility and the other half as low fertility, and a high fertility individual is five times more likely to contribute reproductively than a low fertility individual of the same age class, at any age.

yields unbiased and accurate estimates of N_e when a medium-to-large proportion of the population is sampled. As expected, the accuracy of the estimator increases with an increasing sampling proportion under both fertility models.

Under both the equal and differential fertility models, the estimator is robust to disproportional sampling among age classes (Table 1). Old-biased sampling (sampling percentages 5, 8, 11, 17, 25 for ages 0∼4) and young-biased sampling (sampling percentages 25, 17, 11, 8, 5 for ages 0∼4) lead to similar accuracies, which are also comparable to that when 8∼16% of individuals are sampled from each age class. Even when no individuals are sampled from an age class for both males and females, and sampling effort is relatively low (e.g., 16%), the estimator still yields good estimates of N_e (Table 1 shows these results for equal fertility only, but comparable results were also obtained for unequal fertility conditions). Newborn individuals are relatively more important than other age classes, and the complete absence of newborns in a sample leads to an overestimation of generation interval and underestimation of N_e. However, the estimator can also become highly biased and imprecise when data from several age classes are missing (data not shown).

The results in Table 1 indicate that the inaccuracy of EPA is caused predominantly by the sampling variances rather than biases. In a very small sample due to a low sampling proportion, there are either very few or no PO dyads. In such a small sample, the presence or absence of PO dyads can have a disproportionate effect on the estimator: in their presence the proportion of PO dyads in the sample is too large, whereas in their absence the proportion is too small. In other words, N_e estimates are in a bimodal distribution when sampling proportion is small. This pattern is illustrated in Figure 1, which shows the EPA estimates of N_e from 500 replicate datasets under each sampling proportion (2%, 4%, 16%, 32%) when individuals within an age class have the same expected fertility. Therefore, for small samples due to

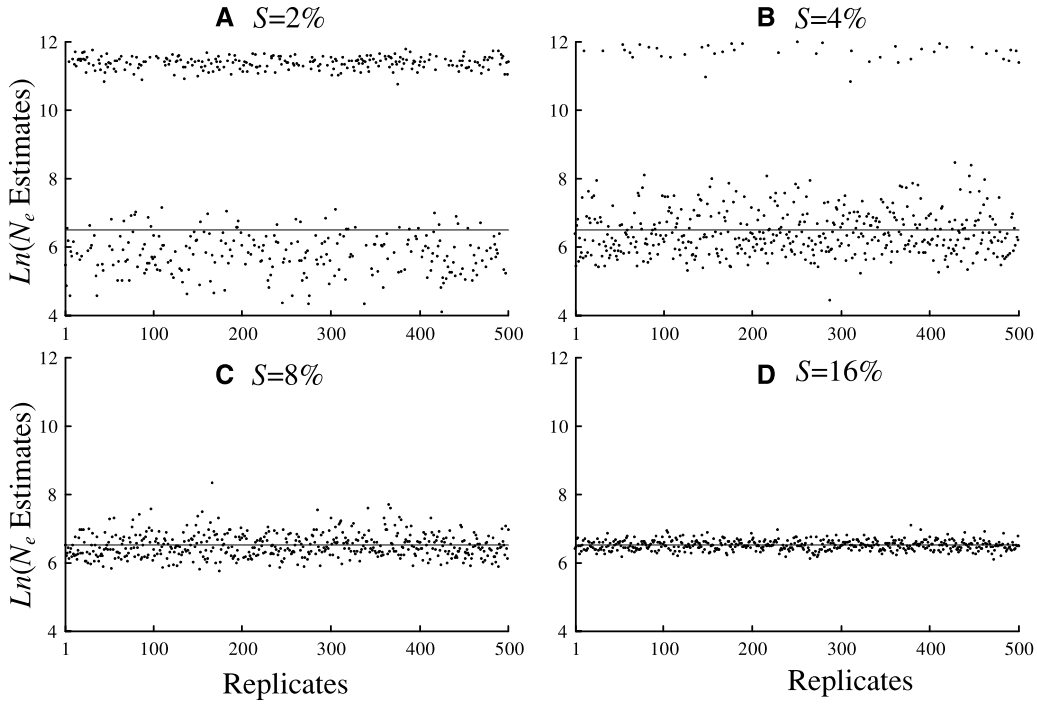

Figure 1. Distributions of N_e estimates (in natural logarithm) from EPA. Data were simulated with parameters $n_1 = 3$, $n_2 = 4$, $\{p_{r1,1}, p_{r1,2}, p_{r1,3}\} = \{0.10, 0.25, 0.15\}$, $\{p_{r2,1}, p_{r2,2}, p_{r2,3}, p_{r2,4}\} = \{0.10, 0.15, 0.15, 0.10\}$, $\{N_{0,1}, N_{1,1}, N_{2,1}, N_{3,1}\} = \{300, 200, 150, 100\}$, $\{N_{0,2}, N_{1,2}, N_{2,2}, N_{3,2}, N_{4,2}\} = \{300, 200, 150, 100, 50\}$, where $r = 1, 2$. The sampling proportion is 2% (A), 4% (B), 8% (C), and 16% (D) for each age class (0∼3 year old males, 0∼4 year old females). Within each age class, individuals are sampled at random, and the sex and age of and parentage among the sampled individuals are assumed known. The horizontal line in the graphs indicates the actual parameter value of N_e, which is 673 (6.51 in natural logarithm) for an equal expected fertility among individuals within an age class.

a very low sampling proportion, the estimates of N_e should be treated with caution.

THE TYPE, NUMBER AND POLYMORPHISM OF MARKERS

Given the sampling effort, the accuracy of EPA depends critically on the accuracy of parentage assignments, which is determined mainly by the information content of markers. Table 2 compares the estimates of L and N_e when different types, numbers, and polymorphisms of markers are used in the analysis. As can be seen, both L and N_e are accurately estimated by EPA using $\geq$8 microsatellites or $\geq$50 single nucleotide polymorphisms (SNPs). However, both L and N_e are overestimated when using only five microsatellites. This is because five loci do not usually provide sufficient information to assign parentage at the 95% confidence level. As a result, too few PO dyads are identified by colony, leading to an overestimation of N_e. It is also difficult to use dominant markers (random amplification of polymorphic DNA [RAPDs]) to assign parentage and thus estimate N_e accurately. A substantial bias of N_e is present even when hundreds of dominant markers are used in the analysis. The estimates of N_e obtained with dominant markers are fairly precise (SD small) but consistently smaller than the true value of 673 by about 50%. This is because dominant markers are particularly uninformative in distinguishing full-siblings and PO relationships, and roughly 1000 dominant markers (each having two equi-frequency alleles) are required to correctly distinguish the two competing relationships at a frequency of 80% (Wang 2006). Therefore, some full-sibling dyads were falsely assigned PO relationship when dominant markers were used in the parentage analysis, leading to an underestimation of N_e.

ESTIMATES OF OTHER PARAMETERS

In addition to L and N_e, EPA estimates other parameters that specify the inbreeding and genetic drift processes in a population with overlapping generations. These parameters are the reproductive contribution ($p_{rs,i}$), number of individuals ($N_{s,i}$), and proportion of individuals sampled ($S_{s,i}$) for each age-sex class, together with the generation interval (L_s) and variance of lifetime family size (σ^2_{sr}) for both sexes. Table 3 lists the simulation results for the population considered in Tables 1 and 2. As can be seen, all parameters are estimated with little bias. Compared with N_e and L, however, these component parameters are less accurately estimated under the same conditions. The RMSEs of the component parameters are often not much smaller than the mean estimates, in contrast to N_e and L in which the RMSE is about one-seventh and one-fifteenth of the mean estimates, respectively (Table 2). Similar to the estimates of N_e as listed in Table 1, the estimates of

Table 2. Effect of marker information on the L and $1/N_e$ estimates from EPA.

Marker	Number of loci	L			$10^3/N_e$		
		Mean	SD	RMSE	Mean	SD	RMSE
Microsatellites	5	3.14	0.52	0.99	0.020	0.039	1.471
	8	2.27	0.17	0.17	1.739	0.285	0.380
	10	2.32	0.14	0.15	1.595	0.211	0.237
	15	2.34	0.15	0.15	1.544	0.234	0.241
SNPs	50	2.26	0.15	0.16	1.645	0.221	0.271
	100	2.34	0.15	0.15	1.547	0.232	0.239
	150	2.35	0.15	0.15	1.516	0.224	0.227
RAPDs	150	2.14	0.15	0.22	1.867	0.244	0.451
	300	2.19	0.12	0.16	2.273	0.263	0.828
	450	2.18	0.12	0.17	2.407	0.284	0.962

Data were simulated with parameters n_1=3, n_2=4, $\{p_{r1,1}, p_{r1,2}, p_{r1,3}\}$=$\{0.10, 0.25, 0.15\}$, $\{p_{r2,1}, p_{r2,2}, p_{r2,3}, p_{r2,4}\}$=$\{0.10, 0.15, 0.15, 0.10\}$, $\{N_{0,1}, N_{1,1}, N_{2,1}, N_{3,1}\}$=$\{300, 200, 150, 100\}$, $\{N_{0,2}, N_{1,2}, N_{2,2}, N_{3,2}, N_{4,2}\}$=$\{300, 200, 150, 100, 50\}$, where r=1, 2 (following Table 1). The sampling proportion for each age class (0∼3 year old males and 0∼4 year old females) is 16%. Individuals within any age class are assumed to have the same expected fertility. The actual parameter values of L and $10^3/N_e$ are 2.30 years and 1.486, respectively. Each sampled individual is sexed, aged, and genotyped at a number of marker loci. Microsatellites, SNPs and RAPDs, respectively represent highly polymorphic codominant markers (10 alleles per locus), low polymorphic codominant markers (two alleles per locus), and dominant markers (two alleles per locus). All markers are assumed to have an initially equal allele frequency.

other parameters become less accurate when expected fertilities are very different among individuals within an age class (results not shown).

NUMBER OF AGE CLASSES

The EPA method applies to any age structure of a population, including the special case of discrete generations. In the latter case, both males and females reproduce at most just once in their lifetime, and the sampled individuals can be partitioned into two age classes, class 0 and 1 corresponding to the offspring and parental generation, respectively. Table 4 lists the simulation results for three age structures, which have respectively 1, 8, and 16 age classes for both nonnewborn males and females. Analyses were conducted assuming that the sampling proportion

Table 3. Estimates of all parameters by EPA.

Parameter set	Parameter	Males				Females			
		Actual value	Mean	SD	RMSE	Actual value	Mean	SD	RMSE
Reproduction	$p_{rs,1}$	0.10	0.11	0.04	0.04	0.10	0.09	0.04	0.04
	$p_{rs,2}$	0.25	0.25	0.06	0.06	0.15	0.13	0.05	0.05
	$p_{rs,3}$	0.15	0.14	0.05	0.05	0.15	0.17	0.06	0.06
	$p_{r2,4}$					0.10	0.11	0.06	0.06
Age class	$10^3/N_{s,1}$	5.00	4.90	1.10	1.11	5.00	5.81	1.21	1.46
Size	$10^3/N_{s,2}$	6.66	6.86	1.35	1.36	6.66	7.58	1.39	1.67
	$10^3/N_{s,3}$	10.00	11.14	2.23	2.50	10.00	9.22	2.54	2.66
	$10^3/N_{2,4}$					20.00	16.61	6.32	7.17
Sampling percentage	$S_{s,1}$	16	17.4	4.6	4.9	16	20.6	5.7	7.3
	$S_{s,2}$	16	16.7	4.2	4.3	16	18.9	5.3	6.1
	$S_{s,3}$	16	20.6	5.4	7.1	16	15.3	4.6	4.7
	$S_{2,4}$					16	16.8	6.4	6.5
Variance of lifetime	σ^2_{s1}	1.273	1.335	0.213	0.222	1.460	2.188	1.376	1.557
family size	σ^2_{s2}	1.273	1.096	0.512	0.542	1.460	1.398	0.227	0.236
	$\sigma_{s1,s2}$	0.273	0.265	0.160	0.161	0.460	0.526	0.349	0.355
Generation interval	L_s	2.10	2.05	0.16	0.17	2.50	2.61	0.25	0.27

Data were simulated with parameters n_1=3, n_2=4, $N_{0,1}$=$N_{0,2}$=300 and other parameter values shown in columns 3 and 7 (following Tables 1 and 2). Individuals within any age class are assumed to have the same expected fertility. Each sampled individual is sexed, aged, and genotyped at 10 microsatellite loci, each having 10 alleles at an initially equal frequency.

Table 4. Effect of the number of age classes on the L and $1/N_e$ estimates from EPA.

Sampling percentage	Number of age classes	L				$10^4/N_e$			
		Actual values	Mean	SD	RMSE	Actual values	Mean	SD	RMSE
Unknown	1	1.00	1.00	0.00	0.00	8.45	8.92	3.64	3.69
	8	5.18	5.28	0.19	0.22	1.94	2.29	1.05	1.10
	16	10.16	11.53	0.39	1.43	1.37	1.95	0.59	0.83
Known	1	1.00	1.00	0.00	0.00	8.45	9.27	3.68	3.80
	8	5.18	5.16	0.16	0.16	1.94	1.92	0.75	0.75
	16	10.16	10.29	0.31	0.33	1.37	1.41	0.65	0.65

Data were simulated assuming an equal number of male (n_1) and female (n_2) age classes. When n_s=1, the population consists of 1000 individuals of each sex and an individual reproduces at most only once in its lifetime (i.e., discrete generation model). When n_s=8, the parameters are {0.01, 0.02, 0.04, 0.08, 0.16, 0.08, 0.06, 0.05} for $\{p_{rs,i}\}$ where r,s=1,2 and i=1∼8, and are $N_{1,s}$=1000 and $N_{i+1,s}=N_{i,s}-50$ for s=1,2 and i=1∼7. When n_s=16, the parameters are {0, 0.01, 0.01, 0.02, 0.02, 0.03, 0.03, 0.04, 0.04, 0.05, 0.05, 0.05, 0.05, 0.04, 0.04, 0.02} for $\{p_{rs,i}\}$ where r,s=1,2 and i=1∼16, and are $N_{1,s}$=1000 and $N_{i+1,s}=N_{i,s}-50$ for s=1,2 and i=1∼15. In all three cases, the sampling proportion is 10% for each age class, and each sampled individual is genotyped at 15 microsatellite loci (each having 10 alleles at an initially equal frequency in simulations). Data were simulated under a differential fertility model, with half the newborns marked as high fertility and the other half as low fertility. A high-fertility individual is 10 times more likely to contribute reproductively than a low-fertility individual of the same age class at any time. Analyses of the simulated data were conducted assuming the sampling proportion (sampling%, first column) is either known or unknown (and estimated by EPA in the latter case).

is either known or unknown and estimated jointly with other parameters.

With discrete generations, N_e estimates are almost unbiased and reasonably precise no matter whether the sampling proportion is known or not. With overlapping generations, N_e is slightly underestimated when the number of age classes, n_s, is large. However, the overall accuracy as measured by RMSE is still reasonably good. An examination of the distribution of N_e estimates for the case of $n_s = 16$ shows that only 7% and 1% of the estimates are smaller than $1/2N_e$ (= 3645) and larger than $3/2N_e$ (= 10,937) respectively. With known sampling proportions, the estimates of N_e and L become unbiased and accurate.

The simulation results in Table 4 also show that EPA applies to large populations. For the case of $n_s = 16$, there are in total 20,000 individuals distributed in age classes 1–16 and the effective size is 7291. The effective size would be 12,202 if there were no difference in fertility among individuals in expectation. As long as a substantial proportion of the population is sampled and genotyped at a sufficient number of loci, accurate parentage assignments provide sufficient information for the estimation of L, N_e, and other parameters.

THE LIFE-HISTORY OF SPECIES

To investigate whether EPA applies to different species that may have dramatically different life histories as defined by age-specific survival and reproduction rates, we simulated and analyzed data for three species typical of the three types of survivorship. Following Waples and Yokota (2007): we chose humans to represent type I survivorship species, characterized by a high survival well into adulthood followed by a period of rapidly increasing mortality; white-crowned sparrows (*Zonotrichia leucophrys nuttalli*) to represent type II survivorship species, characterized by a constant survival rate after an episode of high early mortality; and the barnacle (*Balanus glandula*) to represent type III survivorship species, characterized by an extremely high early mortality followed by a constant survival rate. We used the age-specific survival and reproduction rates listed in Table 2 from Waples and Yokota (2007) to obtain the parameters of our model and analyze a population matching the life-history of each of these three species, as listed in Table 5. Each population is simulated and sampled at two intensities, and each sampled individual is sexed, aged, and genotyped at 10 microsatellites (each having 10 equi-frequency alleles initially in simulations). The quality of estimates of L and N_e from EPA is listed in Table 5.

As can be seen, EPA yields good estimates of both L and N_e for all three species. Unsurprisingly, barnacles represent the most difficult species for EPA, because the extremely high early mortality makes it difficult to capture the PO relationships in a sample with a low sampling intensity. At a low sampling intensity, therefore, N_e tends to be overestimated. This remains true when the PO relationships in a sample are known rather than inferred by genetic markers (data not shown).

ANALYSIS OF A BABOON DATASET

These data are taken from a long-term study of a wild population of chacma baboons (*Papio ursinus*) living at Tsaobis Leopard Park, on the edge of the Namib Desert in Namibia (Cowlishaw 1999). Chacma baboons typically live in social groups of 20–80 individuals, containing multiple males, multiple females, and offspring. Females are philopatric whereas males leave their natal group and join other groups on reaching adulthood. The study population consists of six groups. Four groups, F, G, H, and I,

Table 5. Effects of life histories on the L and $1/N_e$ estimates from EPA.

Age Class i	Human				Sparrow				Barnacle			
	$N_{i,s}$	$p_{rs,i}$	$S_{s,i}$ (1)	$S_{s,i}$ (2)	$N_{i,s}$	$p_{rs,i}$	$S_{s,i}$ (1)	$S_{s,i}$ (2)	$N_{i,s}$	$p_{rs,i}$	$S_{s,i}$ (1)	$S_{s,i}$ (2)
1	100	0	0.1	0.2	1000	0	0.05	0.1	10^5	0	0.001	0.002
2	98	0	0.1	0.2	180	0.229	0.1	0.2	62	0.112	0.25	0.5
3	98	0.008	0.1	0.2	95	0.131	0.1	0.2	34	0.117	0.25	0.5
4	97	0.141	0.1	0.2	51	0.074	0.1	0.2	20	0.091	0.25	0.5
5	97	0.167	0.1	0.2	27	0.042	0.1	0.2	16	0.080	0.25	0.5
6	97	0.104	0.1	0.2	14	0.023	0.1	0.2	11	0.055	0.25	0.5
7	96	0.055	0.1	0.2	–	–	–	–	7	0.035	0.25	0.5
8	96	0.022	0.1	0.2	–	–	–	–	2	0.010	0.25	0.5
9	95	0.003	0.1	0.2	–	–	–	–	–	–	–	–
L(True)			5.26	5.26			3.00	3.00			4.00	4.00
$\hat{L}$(Mean)			5.32	5.29			3.49	3.44			3.67	3.75
$\hat{L}$(SD)			0.46	0.23			0.38	0.23			0.28	0.21
$(\hat{L}<L/2)\%$			0.00	0.00			0.00	0.00			0.00	0.00
$(\hat{L}>3L/2)\%$			0.00	0.00			0.60	0.10			0.00	0.00
N_e(True)			1024	1024			740	740			240	240
$\hat{N}_e$(Mean)			951	979			622	682			345	277
$\hat{N}_e$(SD)			146	132			184	95			56	27
$(\hat{N}_e<N_e/2)\%$			0.20	0.10			0.40	0.00			0.00	0.00
$(\hat{N}_e>3N_e/2)\%$			0.60	0.90			0.70	0.30			32.6	0.62

Parameters $N_{i,s}$ and $p_{rs,i}$ for the three species are based on Tables 2 and 3 from Waples and Yokota (2007). For each age class i, it is assumed that $N_{i,1}=N_{i,2}$, $p_{11,i}=p_{12,i}=p_{21,i}=p_{22,i}$, and $S_{1,i}=S_{2,i}$. For humans, age classes are in units of 5 years and age classes $i>9$ are ignored because they do not contribute reproductively and thus have no effect on either L or N_e. Age classes are in units of 1 year for barnacles and sparrows. For each species, two sampling intensities (columns headed by $S_{s,i}$ (1) and $S_{s,i}$ (2)) are considered in the simulations. ($\hat{L}<L/2$)% and ($\hat{L}>3L/2$)% give the percentages that the estimates are smaller than $L/2$ and larger than $3L/2$, respectively. The quality of estimates for $\hat{N}_e$ is similarly measured and denoted.

were captured during the years 2000∼2001, and two extra groups, J and L, were captured in 2006. The numbers of captured animals are 17, 26, 59, 18, 55, and 32, and the group sizes at capture are 17, 27, 78, 19, 57, and 32, for groups F, G, H, I, J, L, respectively. For each captured animal, sex is identified and age is estimated through a dental examination (Huchard et al. 2009), and a tissue sample is taken for DNA analysis at 16 microsatellite loci and a major histocompatibility complex (MHC) locus (which contains multiple segments of the DRB region of the MHC) (Huchard et al. 2008). The number of alleles varies between 3 and 11 for the microsatellite loci, and there are 15 distinct MHC haplotypes.

For the current analysis, the sample is split into two subsamples, the first being the 120 animals captured from groups F, G, H, and I during the years 2000–2001 and the second being the 87 animals captured from groups J and L in the year 2006. The two subsamples are analyzed separately for estimating the generation intervals and effective sizes of the two clusters of groups. Because the sampling proportion is known (83% and 98% for group cluster F-G-H-I and J-L, respectively), it is fixed in the analysis. Because of the small sample size and the long life span (maximum age observed to be 19 years), we use a time unit of two years to partition the sampled individuals into 11 age classes, with class 0 being newborns younger than two years at capture. The minimum age for reproduction is approximately five years for both males and females (e.g., Altmann and Alberts 2003), so individuals in age class $i + 2$ or younger are excluded as candidate parents for individuals of age class i (= 0∼10).

As expected, consistent results are obtained from the separate analyses of the two subsamples (Table 6). First, males tend to have larger variances and covariances of lifetime family size than females. This reflects the high reproductive skew observed among male baboons: the alpha male monopolizes most mating opportunities in his group, thus siring most of the juveniles born during or shortly after his tenure (Altmann et al. 1996; Alberts et al. 2006), whereas most subordinate males cannot gain access to fertile females (Bulger 1993; Weingrill et al. 2000). In contrast, all females in the group reproduce, although with dominant females reproducing at a higher rate (e.g., Altmann and Alberts 2003). Second, the paternal generation interval is longer than the maternal generation interval. This is most likely because the male's reproductive output is concentrated during his alpha tenure that is typically reached in his prime (ca. 8–12 years old) (van Noordwijk and van Schaik 2004), whereas females start reproducing soon after sexual maturity with their first birth at about six years of age (e.g.,

Table 6. Parameter estimates for the baboon dataset.

Group cluster	L_1	L_2	L	σ^2_{11}	σ^2_{12}	$\sigma_{11,12}$	σ^2_{21}	σ^2_{22}	$\sigma_{21,22}$	N_e
F-G-H-I										
Estimate	13.9	9.4	11.7	5.10	9.48	5.75	1.05	1.67	0.48	84
95%CI_L	11.7	8.7	10.8	2.54	1.73	1.56	0.57	1.24	0.25	54
95%CI_U	15.4	13.3	13.9	11.53	39.80	14.86	5.65	4.59	3.14	155
J-L										
Estimate	15.3	12.7	14.0	9.85	18.03	12.14	2.00	3.39	1.74	60
95%CI_L	12.6	12.0	12.8	5.16	2.20	3.83	0.54	2.13	0.88	43
95%CI_U	17.8	14.6	15.8	25.99	116.3	36.27	13.65	5.94	5.44	128

The estimate, the lower (95%CI_L) and upper (95%CI_U) limits of the 95% confidence interval are listed for each parameter. The parameters are the paternal (L_1), maternal (L_2), and mean (L) generation intervals, the variances of the lifetime number of sons and daughters for males (σ^2_{11},σ^2_{12}) and females (σ^2_{21},σ^2_{22}), the covariances between the lifetime numbers of sons and daughters for males ($\sigma_{11,12}$) and females ($\sigma_{21,22}$), and the effective population size (N_e). Generation intervals (L_1, L_2, L) are in years.

Altmann and Alberts 2003). Finally, the estimate of the N_e/N ratio for group cluster F-G-H-I (0.54) is comparable to that for group cluster J-L (0.42). Both values are also similar to the value of 0.51 obtained for another baboon population (yellow baboons *Papio cynocephalus*) using extensive accumulated data on life-history variables (Storz et al. 2002). In calculating N_e/N, N is the number of adults present during the period equivalent to the generation interval, and is calculated as the product of the estimated number of 6- to 8-year-old individuals (age class 3) and L. The estimated values of N are 157 and 143 for group clusters F-G-H-I and J-L, respectively.

ANALYSIS OF A HIHI DATASET

These data are taken from a long-term study of a wild, reintroduced population of hihi, or stitchbird (*Notiomystis cincta*), a New Zealand endemic and endangered, forest-dwelling passerine. The population is found on the offshore island of Tiritiri Matangi (36°36′S, 174°53′E) which is a 220-ha island in the Hauraki Gulf. Hihi have a promiscuous mating system, and extra-pair copulations are frequent, which leads to substantial extra-pair paternity (Ewen et al. 1999). Hihi can breed within their first year of life, the minimal age at first breeding being roughly 10 and 11 months for males and females, respectively (Brekke 2009). For the current analysis, 261 hihi of the 305 recorded in the population were either caught as nestlings or adults in feeding cage traps during the Austral 2006/2007 breeding season (September–February). Blood samples were collected and genotyped at 19 microsatellite loci, with the number of alleles ranging from 2 to 10 per locus (Brekke et al. 2009). The sex of individuals was inferred from two sex-linked markers and/or from adult plumage morphology (Brekke 2009), whereas age was obtained from breeding records and annual census. The sampled males and females have maximal ages of 6 and 8 years, respectively, and are thus divided into 7 (0∼6) and 9 (0∼8) age classes, with class 0 being fledglings younger than six months at capture. Although the sampling proportion for each age class can be reliably estimated from the detailed census records for this dataset, it is assumed unknown in the analysis to demonstrate that our method does not rely on known sampling proportions.

The results are summarized in Table 7. In contrast to the baboon results, there is no obvious difference in the variances and covariances of lifetime family sizes between male and female hihi individuals. This is probably because the strong social structure seen in baboons, where a single dominant male monopolizes most breeding opportunities within a group, is absent from the hihi population, in which both territorial and floater males share the paternity of fledglings (Ewen et al. 1999). Similar to the baboon results, the paternal generation interval is longer than the maternal generation interval. However, this could be due to a sampling effect, because there are no females in age classes 4∼7 and there is just one female in age class 8. Because of the lack of old females in the sample, the maternal generation interval might be

Table 7. Parameter estimates for the hihi dataset.

	L_1	L_2	L	σ^2_{11}	σ^2_{12}	$\sigma_{11,12}$	σ^2_{21}	σ^2_{22}	$\sigma_{21,22}$	N_e
Estimate	3.59	2.93	3.26	1.29	2.35	1.01	1.63	2.05	0.90	111
95%CI_L	2.73	1.99	2.56	1.11	0.74	0.14	0.39	1.12	0.09	93
95%CI_U	4.13	4.08	3.88	1.85	6.18	1.07	5.89	6.52	6.52	258

Notation follows Table 6.

underestimated. The effective size of the population is estimated to be 111, with a 95% confidence interval of 93∼258. The total number of 1-year-old adults in the population at census is 50, so the number of adults during a generation interval is $N = 3.26 \times 50 = 163$. The N_e/N ratio is thus about 0.68, higher than that in the baboons, as expected given that hihi do not exhibit such strong male reproductive skew.

Discussion

In this study, we showed that the effective size of a population with overlapping generations can be inferred by an estimator that uses the sex, age, and multilocus genotype information of a single sample of individuals taken at random from the population. The estimator fits the observed numbers of parentage assignments among age classes to a genetic model to obtain estimates of N_e and L, as well as other important parameters such as the variances and covariances of lifetime family sizes. Simulations indicate that under realistic sampling and genotyping efforts, the estimator yields accurate estimates, especially for the composite parameters of N_e and L.

Similar to the case of discrete generations (Wang 2009), the N_e of a population with overlapping generations can also be estimated from sibship assignments, using essentially the same information as the EPA estimator. Following the approach adopted in Wang (2009), we obtained (derivation available upon request)

$$\frac{1}{N_e} = \frac{1}{16L}\sum_{s=1}^{2}\sum_{i=0}^{n_s-1}(2-\delta_i)(Q_{11,i,s} + Q_{22,i,s} + 2Q_{12,i,s}),$$

where $Q_{rt,i,s}$ is the probability that two individuals of sexes r and t ($r, t = 1, 2$) with an age difference of i years ($0 \le i \le n_s - 1$) taken at random from the population share the same parent of sex s ($s = 1, 2$), and δ_i is the Kronecker delta ($\delta_i = 1, 0$ if $i = 0$ and $i \ne 0$, respectively). All of the Q terms can be estimated from a sibship analysis of a single sample of individuals taken at random from a population, using the sex, age, and multilocus genotype of each sampled individual as information. When $n_1 = n_2 = 1$, the above equation for N_e reduces to that of a population with discrete generations (eq. 9 in Wang 2009) in the case of random mating, as expected. Our simulations show that this sibship-based estimator is equally or slightly more accurate than EPA when the actual sibship and parentage among sampled individuals are known without error. However, it usually underestimates N_e substantially when markers are used in sibship assignments in the case of a population with overlapping generations, where several relationships equivalent to half-sibling relationship (in the pattern and amount of relatedness) typically coexist. For instance, the avuncular and grandparent–grandoffspring relationships are indistinguishable from a half-sibling relationship for a pair of individuals, no matter how many unlinked autosomal markers are used (Epstein et al. 2000; Wang 2007). Such relationships inevitably cause an overestimation of half-sibling frequencies and an underestimation of N_e (simulation results not shown) in populations with overlapping generations. It thus seems rather difficult to estimate N_e from sibship frequencies for populations with overlapping generations, except when sibship is known (from behavioral observation) or can be accurately inferred (e.g., when most sampled individuals are in large sibships containing more than two individuals). In contrast, such across-generation relationships are absent from a sample of a single cohort of individuals, and therefore the sibship assignment method is accurate for a population with discrete generations (Wang 2009).

Fortunately, parentage assignments can be accurately inferred even in the presence of closely competing relationships such as full sibship. This is because the PO relationship is unique in the pattern of identical by descent (IBD). In a large population under random mating, a PO dyad always shares one pair of alleles IBD at each locus, meaning that the inherent variance of IBD among loci is zero (Wang 2002). The PO relationship is thus easily inferred and distinguished from other relationships. Indeed, as verified by simulations in the current study, the PO relationship can be reliably identified and distinguished from other relationships among a sample of individuals taken from a population with overlapping generations, using a realistic number of polymorphic markers (say, 10∼20 microsatellites). An advantage of the parentage-based estimator, EPA, is that it provides information on the generation interval, variance in family size, and effective number of breeders of each age class as well as N_e. In other words, EPA gives not only a simple estimate of N_e and L, but also the estimates of their causal parameters that facilitate the interpretations of the N_e and L estimates and, more importantly, provide useful information for population management to maintain genetic variation. For example, if a small N_e is estimated, and attributed to a large variance in family size among male breeders, then measures may be taken to minimize this variance, such as removal of the dominant males.

Throughout this study, we focused on the effective size at autosomal diploid loci in dioecious species. A great advantage of EPA is that it can use autosomal marker data to estimate N_e and L for X-linked, Y-linked, and mtDNA loci as well. Let us consider group cluster F-G-H-I of baboons as an example. Using the estimates listed in Table 6, we obtain an average generation interval of 10.9 years, 13.9 years, and 9.4 years, and an estimate of N_e of 68, 28, and 82, for X-linked, Y-linked, and mtDNA loci, respectively. The ratios $N_{e(X)}/N_e$, $N_{e(Y)}/N_e$, $N_{e(mt)}/N_e$ are 0.81, 0.33, and 0.98 respectively, deviating from the expected values of 0.75, 0.5, and 0.5 when both males and females have Poisson distributed family sizes (Caballero 1994). The deviations can be explained by the much larger variance of paternal family size than that of maternal family size as observed in the baboon dataset.

In spirit, sibship and parentage assignment methods for estimating N_e are analogous to the mark–recapture method in ecology for estimating the population size. This is especially obvious with the sibship assignment method. Two siblings in a sample mean a recapture of the sibship they represent, and the frequency of sibling dyads in a sample of individuals taken at random (with respect to kinship) from the population gives information about the effective number of breeders. A high sibling frequency would mean either a small number of actual breeders, or a high reproductive skew (a large variance of family size) among breeders, or both. The method using parentage assignments is more complicated, but is still similar to the mark–recapture method in principle. For example, the co-assignment of paternity of two individuals to the same or different individuals in age class *i* affects the recapture of reproduction events in this age class, and would indicate a high reproductive contribution and a high sampling proportion of age *i*.

Noticing the close analogy between the parentage assignment methods for estimating N_e and the mark–recapture method for estimating population size, it is understandable that the sampling proportion is critical in determining the accuracy of the estimates. When a very small proportion of individuals are sampled, the N_e estimates may become bimodally distributed (Fig. 1). It should be emphasized that other factors, such as the N_e/N ratio, also has a large impact on the accuracy of the methods. For a given small sampling proportion (say, 5%), a small N_e/N ratio means that there could still be a substantial number of parentage assignments among sampled individuals, which ensures a high accuracy. It is really the expected number of parentage assignments that determines the accuracy. The larger this expected number of assignments, the higher will be the accuracy.

Compared with the temporal method proposed to estimate N_e for populations with overlapping generations (Jorde and Ryman 1995), the parentage assignment method requires much less information. It only requires the information of sex, age, and multilocus genotypes of a sample of individuals taken at random (with respect to kinship) from the population. The age-specific survival and reproduction rates required by the temporal method are not necessary in EPA. This is a great advantage because these life-table variables are not only numerous but also notoriously difficult to estimate. In fact, when one knows the life-table of a population, one already knows its generation interval as well as, to a good extent, its effective size. In contrast, EPA estimates these life-table variables or their equivalents from parentage assignments, and thus the summary parameters of L and N_e. Although these estimates of individual life-table variables tend to be unbiased, their precision is usually low, and usually much lower than that of the summary parameters of L and N_e (Table 3). Therefore, except when the sampling proportion is high and the marker information sufficient, one should be cautious about the estimates of individual life-table variables from EPA.

EPA is based on genetic models of populations with overlapping generations (Felsenstein 1971; Hill 1972; Johnson 1977), and the assumptions of those models equally apply to EPA. The important assumptions are a constant size and age structure of a population under random mating and with random births and deaths. Of course no real populations satisfy these restrictive assumptions, but it seems quite reasonable that small departures from them should not affect the results much (Hill 1979). The additional assumption made by EPA is random sampling, which dictates that individuals must be taken at random with respect to kinship from the population. A sampling scheme favoring related (unrelated) individuals will cause an underestimation (overestimation) of N_e. Localized sampling, where individuals are sampled predominantly from a local area, may lead to excessive parentage assignments and thus an underestimate of N_e. EPA also assumes that all individuals in a sample should be taken at the same time so that the ages and relationships of the sampled individuals are based on the same time point. However, as long as the sampling length (during which all individuals are sampled) is short relative to the time unit adopted in the analysis, the validity of EPA should be little affected.

Another assumption made by EPA is that the age of each sampled individual is known without error. In practice, however, the age of an individual is usually estimated from morphological traits, such as dentition in mammals and scales and otoliths in fish. Therefore age estimates can be imprecise, and how robust EPA is to the errors of age estimation is of a concern. To investigate this, we conducted simulations in which the ages of a proportion of the sampled individuals were set to values drawn at random between the minimum (0) and maximum ages. In simulations based on 10 microsatellites and with other parameter values fixed as listed in Tables 1 and 2, the mean estimates of $1000/N_e$ (true value = 1.49) are 1.59, 1.54, 1.52, 1.46, 1.35, 1.20, and the RMSEs of $1000/N_e$ are 0.24, 0.25, 0.22, 0.23, 0.25, 0.35, when a percentage of 0, 4, 8, 16, 32, 64 of the sampled individuals are given a random age, respectively. It thus seems that EPA is fairly robust to estimation errors of age. EPA leads to a substantial overestimation of N_e and a much reduced accuracy only when a large proportion of the sampled individuals are assigned incorrect ages.

A computer program, AgeStructure, which implements the EPA method described in this study is posted on our website (http://www.zsl.org/science/research/software/) for free download and use.

ACKNOWLEDGMENTS

We are grateful to W. G. Hill for his insightful and stimulating discussion and helpful comments on the manuscript, and to the associate editor (C. Goodnight) and two anonymous referees for constructive comments on an earlier version of the manuscript. The modeling work was supported by a BBSRC research grant awarded to JLW. The baboon work was supported by an NERC project grant and advanced fellowship awarded to GC and a

Ministère de l'Education et de la Recherche studentship awarded to EH. The hihi work was supported by a NERC Ph.D. studentship to PB with additional research funding from a NERC Sheffield Molecular Genetics Facility access grant.

LITERATURE CITED

Alberts, S., J. C. Buchan, and J. Altmann. 2006. Sexual selection in wild baboons: from mating opportunities to paternity success. Anim. Behav. 72:1177–1196.

Altmann, J., and S. C. Alberts. 2003. Variability in reproductive success viewed from a life-history perspective in baboons. Am. J. Hum. Biol. 15:401–409.

Altmann, J., S. C. Alberts, S. A. Haines, J. Dubach, P. Muruthi, T. Coote, E. Geffen, D. J. Cheesman, R. S. Mututua, S. N. Saiyalel, et al. 1996. Behaviour predicts genetic structure in a wild primate group. Proc. Natl Acad. Sci. USA 93:5793–5801.

Beaumont, M. A. 2003. Conservation genetics. Pp. 751–766 *in* D. J. Balding, M. Bishop, and C. Cannings, eds. Handbook of statistical genetics, 2nd ed. John Wiley and Sons, West Sussex, U.K.

Brekke, P. 2009. Conservation genetics of an island endemic bird, the hihi (*Notiomystis cincta*). Ph.D. Dissertation, Imperial College of London, U.K.

Brekke, P., D. A. Dawson, G. J. Horsburgh, and J. G. Ewen. 2009. Characterization of microsatellite loci in the hihi *Notiomystis cincta* (*Notiomystidae*, AVES). Mol. Ecol. Resources 9:1255–1258.

Bulger, J. B. 1993. Dominance rank and access to estrous females in male Savannah baboons. Behaviour 127:67–103.

Caballero, A. 1994. Developments in the prediction of effective population size. Heredity 73:657–679.

Charlesworth, B. 2001. The effect of life-history and mode of inheritance on neutral genetic variability. Genet. Res. 77:153–166.

Choy, S. C., and B. S. Weir. 1978. Exact inbreeding coefficients in populations with overlapping generations. Genetics 89:591–614.

Cowlishaw, G. 1999. Ecological and social determinants of spacing behaviour in desert baboon groups. Behav. Ecol. Sociobiol. 45:67–77.

Crow, J. F., and M. Kimura. 1970. An introduction to population genetics theory. Harper and Row, New York.

———. 1972. The effective number of a population with overlapping generations: a correction and further discussion. Am. J. Hum. Genet. 24:1–10.

Emigh, T. H., and E. Pollak. 1979. Fixation probabilities and effective population numbers in diploid populations with overlapping generations. Theor. Pop. Biol. 15:86–107.

Epstein, M. P., W. L. Duren, and M. Boehnke. 2000. Improved inference of relationship for pairs of individuals. Am. J. Hum. Genet. 67:1219–1231.

Ewen, J. G., D. P. Armstrong, and D. M. Lambert. 1999. Floater males gain reproductive success through extrapair fertilizations in the stitchbird. Anim. Behav. 58:321–328.

Felsenstein, J. 1971. Inbreeding and variance effective numbers in populations with overlapping generations. Genetics 68:581–597.

Frankham, R., J. D. Ballou, and D. A. Briscoe. 2003. Introduction to conservation genetics. Cambridge Univ. Press, Cambridge, UK.

Hill, W. G. 1972. Effective size of populations with overlapping generations. Theor. Pop. Biol. 3:278–289.

———. 1979. A note on effective population size with overlapping generations. Genetics 92:317–322.

———. 1981. Estimation of effective population size from data on linkage disequilibrium. Genet. Res. 38:209–216.

Huchard, E., M. Weill, G. Cowlishaw, M. Raymond, and L. A. Knapp. 2008. Polymorphism, haplotype composition, and selection in the MHC-DRB of wild baboons. Immunogenetics 60:585–598.

Huchard, E., A. Courtiol, J. A. Benavides, L. A. Knapp, M. Raymond, and G. Cowlishaw. 2009. Can fertility signals lead to quality signals? Insights from the evolution of primate sexual swellings. Proc. R. Soc. Lond. B 276:1889–1897.

Johnson, D. L. 1977. Inbreeding in populations with overlapping generations. Genetics 87:581–591.

Jorde, P. E., and N. Ryman. 1995. Temporal allele frequency change and estimation of effective size in populations with overlapping generations. Genetics 139:1077–1090.

Kimura, M., and J. F. Crow. 1963. The measurement of effective population number. Evolution 17:279–288.

Krimbas, C. B., and S. Tsakas. 1971. The genetics of *Dacus oleae*. V. Changes of esterase polymorphism in a natural population following insecticide control: selection or drift? Evolution 25:454–460.

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

Lynch, M., J. Conery, and R. Burger. 1995. Mutation accumulation and the extinction of small populations. Am. Nat. 146:489–518.

Marshall, T. C., J. Slate, L. E. B. Kruuk, and J. M. Pemberton. 1998. Statistical confidence for likelihood-based paternity inference in natural populations. Mol. Ecol. 7:639–655.

Moran, P. A. P. 1962. The statistical processes of evolutionary theory. Clarendon Press, Oxford.

Nei, M., and Y. Imaizumi. 1966. Genetic structure of human populations. II: differentiation of blood group gene frequencies among isolated populations. Heredity 21:183–190.

Nei, M., and F. Tajima. 1981. Genetic drift and estimation of effective population-size. Genetics 98:625–640.

Pollak, E. 1990. The effective population size of an age structured population with a sex-linked locus. Math. Biosci. 101:121–130.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. 1996. Numerical recipes in Fortran 77. 2nd ed. Cambridge Univ. Press, Cambridge, UK.

Pudovkin, A. I., D. V. Zaykin, and D. Hedgecock. 1996. On the potential for estimating the effective number of breeders from heterozygote-excess in progeny. Genetics 144:383–387.

Schwartz, M. K., D. A. Tallman, and G. Luikart. 1999. Review of DNA-based census and effective population size estimators. Anim. Conserv. 1:293–299.

Storz, J. F., U. Ramakrishnan, and S. C. Alberts. 2002. Genetic effective size of a wild primate population: influence of current and historical demography. Evolution 56:817–829.

van Noordwijk, A. J., and C. P. van Schaik. 2004. Sexual selection and the careers of primate males: paternity concentration, dominance acquisition tactics and transfer decisions. Pp. 208–229 *in* P. M. Kappeler and C. P. van Schaik, eds. Sexual selection in primates: new and comparative perspectives. Cambridge Univ. Press, Cambridge, UK.

Wang, J. 2002. An estimator for pairwise relatedness using molecular markers. Genetics 160:1203–1215.

———. 2004. Sibship reconstruction from genetic data with typing errors. Genetics 166:1963–1979.

———. 2005. Estimation of effective population sizes from data on genetic markers. Phil. Trans. R. Soc. Lond. B 360:1395–1409.

———. 2006. Informativeness of genetic markers for pairwise relationship and relatedness inference. Theor. Pop. Biol. 70:300–321.

———. 2007. Parentage and sibship exclusions: higher statistical power with more family members. Heredity 99:205–217.

———. 2009. A new method for estimating effective population sizes from a single sample of multilocus genotypes. Mol. Ecol. 18:2148–2164.

Wang, J., and A. Caballero. 1999. Developments in predicting the effective size of subdivided populations. Heredity 82:212–226.

Wang, J., and A. W. Santure. 2009. Parentage and sibship inference from multi-locus genotype data under polygamy. Genetics 181:1579–1594.

Waples, R. S. 1989. A generalised approach for estimating effective population size from temporal changes in allele frequency. Genetics 121:379–391.

———. 2006. A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci. Conserv. Genet. 7:167–184.

Waples, R. S., and M. Yokota. 2007. Temporal estimates of effective population size in species with overlapping generations. Genetics 175:219–233.

Weingrill, T., J. E. Lycett, and S. P. Henzi. 2000. Consortship and mating success in chacma baboons (*Papio cynocephalus ursinus*). Ethology 106:1033–1044.

Wright, S. 1931. Evolution in Mendelian populations. Genetics 16:97–159.

———. 1938. Size of population and breeding structure in relation to evolution. Science 87:430–431.

Associate Editor: C. Goodnight