Estimation of Population Growth or Decline in Genetically Monitored Populations

Mark A. Beaumont[1]

School of Animal and Microbial Sciences, Reading RG6 6AJ, United Kingdom

Manuscript received July 30, 2002
Accepted for publication March 19, 2003

ABSTRACT

This article introduces a new general method for genealogical inference that samples independent genealogical histories using importance sampling (IS) and then samples other parameters with Markov chain Monte Carlo (MCMC). It is then possible to more easily utilize the advantages of importance sampling in a fully Bayesian framework. The method is applied to the problem of estimating recent changes in effective population size from temporally spaced gene frequency data. The method gives the posterior distribution of effective population size at the time of the oldest sample and at the time of the most recent sample, assuming a model of exponential growth or decline during the interval. The effect of changes in number of alleles, number of loci, and sample size on the accuracy of the method is described using test simulations, and it is concluded that these have an approximately equivalent effect. The method is used on three example data sets and problems in interpreting the posterior densities are highlighted and discussed.

THE effect of inbreeding on population fitness is currently the focus of many studies, both empirical (Saccheri *et al.* 1998) and theoretical (Lynch *et al.* 1995; Lande 1998). One motivation behind these studies is the need to investigate the genetic component of the threat to endangered species arising from low population size. The rate of inbreeding depends on N_e, which is generally much lower than the census size. If it can be assumed that the ratio of effective to census size is approximately constant, the detection of historical changes in N_e may indicate changes in census size. Similarly, if the ratio of effective to census size can be estimated for one population it can then be used to estimate census sizes in other populations for which only genetic information is available (Beaumont 2001).

Estimation of N_e is problematic. There are three general approaches. One way is to estimate it nongenetically from the mating system (Caballero 1994). However, this is generally unsatisfactory because detailed life-history information is required, as well as good estimates of census size, which is often unavailable with sufficient precision to make a good estimate of N_e (Frankham 1995). Furthermore, cross-generational effects that are difficult to measure, such as serial correlations in family size, may cause a substantial reduction in N_e from that expected purely from consideration of the variance in reproductive success (Austerlitz and Heyer 1998). An alternative approach is to use information from single genetic samples. For example, using a mutation model, N_e can be estimated from the variability in the sample (*e.g.*, Griffiths and Tavaré 1994a,b,c; Kuhner *et al.* 1995; Wilson and Balding 1998; Storz and Beaumont 2002). A problem with this approach is that the value that is estimated may have little relationship to current rates of inbreeding or any value of N_e that could be estimated from direct observation of the mating system of the population. This is because, over the timescale in which the observed variability is generated by mutation, the unknown details of population history, gene flow, and metapopulation structure will greatly influence estimates of N_e, which is then probably best regarded as simply a scaling coefficient in a coalescent model (Donnelly and Tavaré 1995; Nordborg 1997; Wakeley 1999; Wakeley and Aliacar 2001). Alternatively, genotypic disequilibria in single samples can be used to estimate N_e. This can be achieved by measuring departures from either Hardy-Weinberg equilibrium (Pudovkin *et al.* 1996; Luikart and Cornuet 1999) or linkage disequilibrium (Langley *et al.* 1978; Laurie-Ahlberg and Weir 1979; Hill 1981). These have the advantage that they measure N_e on a more recent timescale, but have generally low power and are susceptible to the influence of many other phenomena.

The most widely used method to estimate N_e from genetic samples is from the difference in gene frequency between serial samples taken from the same population. This is the "temporal method," first introduced by Krimbas and Tsakas (1971). Their method-of-moments estimator has been elaborated by Nei and Tajima (1981), Pollak (1983), and Waples (1989). More recently Williamson and Slatkin (1999), Anderson *et al.* (2000), Wang (2001), and Berthier *et al.* (2002) have developed likelihood-based estimators, which show modest to rather more substantial improvements in accuracy over the method-of-moments estimators. In addition, Williamson and Slatkin (1999) and Wang (2001) have been able to estimate change in population size, further

[1]*Address for correspondence:* School of Animal and Microbial Sciences, Whiteknights, PO Box 228, Reading RG6 6AJ, United Kingdom. E-mail: m.a.beaumont@reading.ac.uk

illustrating the flexibility of likelihood-based approaches. Williamson and Slatkin (1999) estimated likelihoods from a Wright-Fisher model in which any number of serial samples could be analyzed. Their method is practicable only for the biallelic case. More recently Anderson *et al.* (2000) used importance sampling to improve the speed of the approach, which makes it practicable to look at multiallelic data. Wang (2001) has suggested a further improvement in computational speed by approximating the probability of the data by the product of the marginal probabilities for each allele, thus reducing the problem to that studied by Williamson and Slatkin (1999), but solved substantially more efficiently. The method of Berthier *et al.* (2002) differs from the other three methods in that likelihoods are estimated from a coalescent model in which two samples are analyzed. Since only two samples are analyzed in their method it is not possible to make inferences about changes in population size.

The above methods all assume that the sampling period is sufficiently short that the effects of mutations can be safely ignored. Recently a strand of research that is independent of that initiated by Krimbas and Tsakas (1971) was identified and has been motivated by the need to study human immunodeficiency virus viral dynamics and evolution on the basis of sequence data from serial samples (Rodrigo *et al.* 1999; Fu 2001; Drummond *et al.* 2002). These methods use a coalescent model with mutations, and that of Drummond *et al.* (2002) allows for full Bayesian estimation of mutational, demographic, and genealogical parameters from sequence data.

This study assumes that the effects of mutations over the sampling period can be ignored and makes three contributions. First, it is shown how the Monte Carlo method of importance sampling can be used to update sets of genealogies in a Markov chain Monte Carlo simulation to estimate posterior distributions of parameters of interest. This method is very general and can be applied to all models of genealogical inference and may lead to increased efficiency in implementation and execution. This computational method is applied to the coalescent-based model of Berthier *et al.* (2002), described above. Second, the model of Berthier *et al.* (2002) is generalized to consider any number of samples in a temporal sequence rather than just the two previously considered. Third, the method is further extended to estimate parameters in a model of population growth and decline, similar to that studied in Beaumont (1999).

IMPLEMENTATION OF MARKOV CHAIN MONTE CARLO WITH INDEPENDENT SAMPLING OF GENEALOGICAL HISTORIES

Background and motivations: The potential for genetic data to shed light on the evolutionary history of populations has been well appreciated over the last decade, and in the development of the statistical methodology there has been a general interest in moving away from moment-based methods of estimation to the use of likelihood and Bayesian inference (Stephens 2001). Reflecting the youth of this field, the computational and technical details of the different approaches to inference tend to dominate much of the research. There are currently two interrelated computer-intensive methods to statistical inference—Markov chain Monte Carlo (MCMC) and importance sampling (IS). MCMC is a method for generating autocorrelated random samples from probability distributions. IS is generally used to approximate likelihoods based on independent samples. In population genetics these have been combined to give three main groups of methods, as distinguished by Stephens and Donnelly (2000): those in which the genealogical histories are independently sampled using IS, with likelihoods for specific parameters then computed from the sample of genealogical histories (Griffiths and Tavaré 1994a; Anderson *et al.* 2000), which are referred to here loosely as "pure IS" methods; those in which autocorrelated genealogical histories are sampled using MCMC, with likelihoods for specific parameters then computed from the sampled genealogical histories (Kuhner *et al.* 1995; Beerli and Felsenstein 2001); and those in which there is autocorrelated sampling of genealogical histories and demographic/mutational parameters. The latter approach yields samples from the posterior distribution of parameter values and leads naturally to Bayesian inference or the use of integrated likelihood (*e.g.*, Wilson and Balding 1998; Beaumont 1999; Nielsen and Wakeley 2001; Drummond *et al.* 2002). By contrast, the other two approaches, which use importance sampling to approximate likelihood surfaces, tend to lead to more classical likelihood-based inference.

In the MCMC methods that give autocorrelated samples of parameter values the necessary integration for Bayesian inference is intrinsic, and the only additional computation, if required, is to estimate the posterior densities from the sampled points. By contrast, with importance sampling methods that give approximations of likelihood surfaces directly, further complex procedures are necessary for Bayesian inference. Given that it is generally easier to estimate densities (thereby enabling the choice of classical likelihood-based estimation, integrated likelihoods, or fully Bayesian inference) than to manipulate the approximated-likelihood surfaces, it would seem that methods that give autocorrelated samples of parameter values offer the greatest flexibility. However, these methods have currently two main drawbacks. They involve making small modifications to the genealogical history, and because of this (a) they are generally more difficult to program than pure IS methods, and (b) they can move quite slowly through the space of possible genealogical histories, making them potentially inefficient. This article introduces a method for overcoming these two disadvantages and applies it to a specific problem, the estimation of effective population size, N_e, from temporally spaced genetic samples.

The general motivation behind these computer-intensive methods is that from coalescent theory it is straightforward to calculate the probability $p(D, G|\Phi) = p(D|G)P(G|\Phi)$ of any particular genealogical history, G, that gives rise to some data D, as a function of parameters specifying the demographic history and mutation model, Φ. There are a number of different representations of the genealogical history (see Stephens and Donnelly 2000), and in this article I consider it to be the timed sequence of coalescent and mutation events in the genealogical history of a sample, so that $p(D|G) = 1$ if the genealogical history can give rise to the data and 0 otherwise. Any particular data set can be obtained from very many different genealogical histories, and to calculate the likelihood we need to evaluate

$$p(D|\Phi) = \int p(D|G)p(G|\Phi)\,dG, \qquad (1)$$

where, following Stephens and Donnelly (2000), the integral denotes a summation over all discrete states (*e.g.*, pattern of coalescences and mutations) and integration over continuous states (*e.g.*, duration of intervals between events). Estimation of $p(D|\Phi)$ directly is most conveniently made using importance sampling (Griffiths and Tavaré 1994a; Stephens and Donnelly 2000). In importance sampling, the equation is rewritten as

$$p(D|\Phi) = \int p(D|G)\frac{p(G|\Phi)}{q(G|\Phi)}q(G|\Phi)\,dG,$$

and this is estimated by sampling G_j from $q(G|\Phi)$ and evaluating

$$\tilde{p}(D|\Phi) = 1/h\sum_{j=1}^{h} p(D|G_j)p(G_j|\Phi)/q(G_j|\Phi). \qquad (2)$$

Generally the sampling distribution is chosen such that $P(D|G_j) = 1$ for all G_j. In the ideal case that $q(G|\Phi) = p(G|D, \Phi)$, *i.e.*, the posterior distribution of genealogical histories given the data and parameters, the variance in the estimate of $p(D|\Phi)$ is zero because each term in (2) evaluates to the likelihood axiomatically. The ratio $p(G|\Phi)/q(G|\Phi)$ is called the importance ratio, or importance weight.

However, the evaluation or estimation of $p(D|\Phi)$ is not necessarily an ideal goal for population genetic inference. The problem is that Φ often has many components, and generally we wish to make inferences about one component (*e.g.*, growth rate) independent of the others. Furthermore, for most population genetic problems, the likelihood surfaces do not approximate that of a multivariate normal distribution, and therefore asymptotic theory and methods often do not apply. These problems can be side-stepped by taking a Bayesian approach to inference, which also has the advantage that background information can be incorporated into the model (Wilson and Balding 1998). In this case we estimate the posterior distribution

$$p(\Phi|D) = \frac{p(D|\Phi)p(\Phi)}{\int p(D|\Phi)p(\Phi)\,d\Phi}.$$

Inferences on particular parameters can be made from the marginal posterior distribution, where $p(\Phi|D)$ is integrated over all other parameters. If uniform improper priors are used $p(\Phi|D) \propto p(D|\Phi)$ and the methods used to obtain marginal posterior distributions will also give the integrated (relative) likelihood surface. Considerations of how best to make inferences on single parameters in multiparameter models has led, for example, Nielsen and Wakeley (2001) to advocate that there are many advantages to using integrated likelihoods even when a frequentist approach is preferred.

The only method currently used to perform fully Bayesian analyses for population genetic inference has been Metropolis-Hastings sampling of parameter values (*e.g.*, Wilson and Balding 1998; Beaumont 1999). Although the potential to use purely importance-sampling approaches for Bayesian analyses has been discussed (*e.g.*, Fearnhead and Donnelly 2001), no such analysis of genetic data based on importance sampling has yet been published, and there has been no proposal for how this could easily be done for a complex multiparameter model, such as a hierarchical Bayesian model (Storz and Beaumont 2002).

To perform Metropolis-Hastings sampling it is not necessary to evaluate $p(D|\Phi)$, and we can work with $p(D, G|\Phi)$, which is easily calculated from coalescent theory. Starting with any G_i such that $P(D|G_i) = 1$, modify $G_i \rightarrow G_{i+1}$ [where $P(D|G_{i+1}) = 1$] and $\Phi_i \rightarrow \Phi_{i+1}$ such that it is straightforward to calculate the probability, $p(G_{i+1}, \Phi_{i+1}|G_i, \Phi_i)$, of obtaining G_{i+1} and Φ_{i+1}, conditional on being at G_i, Φ_i, and the reverse. Then accept G_{i+1} and Φ_{i+1}, with probability

$$\min\left(1, \frac{p(D, G_{i+1}|\Phi_{i+1})}{p(D, G_i|\Phi_i)} \times \frac{p(G_i, \Phi_i|G_{i+1}, \Phi_{i+1})}{p(G_{i+1}, \Phi_{i+1}|G_i, \Phi_i)} \times \frac{p(\Phi_{i+1})}{p(\Phi_i)}\right); \qquad (3)$$

otherwise $G_{i+1} = G_i$, and $\Phi_{i+1} = \Phi_i$. The first term in the product is the likelihood ratio, the second is the Hastings term, and the third is the ratio of the priors. The Hastings term is the ratio of the probability of reaching the current state from the proposed state to that of the reverse and ensures a uniform coverage of the parameter space. This simulated Markov chain will then give a (serially autocorrelated) sample from $p(\Phi, G|D)$. Summaries of the marginal posterior density for a particular parameter or an estimate of the density itself can be obtained from the simulated sequence of values realized for that parameter, ignoring the others. The key point here is that it is possible to perform the simulation using $p(D, G|\Phi)$, which is easy to calculate, by updating the genealogical history G, and then the posterior distribution for the parameters of interest are obtained marginal to the genealogical histories. The price for this convenience is that the search space of the MCMC simu-

lation is greatly increased. If it were possible to evaluate Equation 1, then $p(\Phi|D)$ marginal to G could have been obtained by running the simulation with $p(D|\Phi)$ and updating $\Phi_i \rightarrow \Phi_{i+1}$ alone—*i.e.*, accepting Φ_{i+1} with probability

$$\min\left(1, \frac{p(D|\Phi_{i+1})}{p(D|\Phi_i)} \times \frac{p(\Phi_i|\Phi_{i+1})}{p(\Phi_{i+1}|\Phi_i)} \times \frac{p(\Phi_{i+1})}{p(\Phi_i)}\right), \quad (4)$$

and thus only Φ would have to be explored by the MCMC simulation.

Current methods that use independent sampling of genealogical histories within an MCMC framework: Hitherto it has been easier to run the MCMC using $p(G, D|\Phi)$ and hence autocorrelated sampling of genealogical histories, but, as discussed above, there are programming problems and problems of efficiency with this approach. Therefore it is tempting to consider the use of importance sampling to obtain an approximation, $\tilde{p}(D|\Phi)$, which can then be implemented in an MCMC simulation to incorporate prior information, and obtain marginal posterior distributions or integrated likelihoods, as discussed above. One advantage of importance sampling is that it is often very straightforward to implement in a computer program. Also, because the importance-sampling function uses heuristics from coalescent theory to attempt to generate genealogies from their posterior distribution, given the data, it is a potentially more efficient method for sampling genealogical histories in comparison with MCMC.

This approach has been used in a series of articles (O'Ryan *et al.* 1998; Ciofi *et al.* 1999; Chikhi *et al.* 2001; Berthier *et al.* 2002) to make inferences based on coalescent models of drift without mutations (reviewed in Beaumont 2001). A related method has been used by O'Neill *et al.* (2000) for an epidemiological model. The likelihood ratio

$$R = \frac{p(D|\Phi_{i+1})}{p(D|\Phi_i)}$$

in Equation 4 is replaced by

$$\hat{R} = \frac{\tilde{p}(D|\Phi_{i+1})}{\tilde{p}(D|\Phi_i)},$$

estimated (in the genealogical analyses) using Equation 2. Note that in normal MCMC, if the denominator $p(D|\Phi_i)$ in the likelihood ratio were known without error there would be no need to reevaluate it each time that R was evaluated. By contrast, with $\hat{R}$ there is a choice whether to make independent estimates of $\tilde{p}(D|\Phi_i)$ when it is evaluated at each update of the MCMC (evaluation of Equation 4) or to reuse the earlier estimate. Intuitively it seems reasonable, though more time consuming, to reevaluate it each time so that the estimates of $\hat{R}$ are independent of each other and so that unusually large ratios arising by chance do not lead to sticking of the simulated Markov chain. Furthermore, the results in O'Neill *et al.* (2000) concerning bias correction (discussed below) require independence of the estimates. The reevaluation approach has been taken in all the genealogical models that have used the method and also by O'Neill *et al.* (2000). Updates are required only for Φ and not for G as in the MCMC methods of Wilson and Balding (1998) and Beaumont (1999). This general method, where the MCMC uses an approximation to the likelihood, is abbreviated here as Monte Carlo within Metropolis (MCWM), following the terminology of O'Neill *et al.* (2000). A basic algorithm for MCWM is as follows:

1. Choose initial parameter values Φ_i with $i = 0$.
2. Sample h independent genealogical histories using importance-sampling function and calculate $\tilde{p}(D|\Phi_i)$ from Equation 2.
3. Draw $\Phi_{i+1} \sim p(\Phi_{i+1}|\Phi_i)$.
4. Sample h independent genealogical histories using importance-sampling function and calculate $\tilde{p}(D|\Phi_{i+1})$ from Equation 2.
5. Accept Φ_{i+1} with probability

$$\min\left(1, \frac{\tilde{p}(D|\Phi_{i+1})}{\tilde{p}(D|\Phi_i)} \times \frac{p(\Phi_i|\Phi_{i+1})}{p(\Phi_{i+1}|\Phi_i)} \times \frac{p(\Phi_{i+1})}{p(\Phi_i)}\right).$$

   Otherwise $\Phi_{i+1} = \Phi_i$.
6. Set $i = i + 1$ and go to 2.

Bias correction: Clearly, since $\hat{R}$ is based on an approximation of the likelihood ratio the posterior distribution will also be approximate. Simulation tests performed in O'Ryan *et al.* (1998) suggested that an IS size of 500 was sufficient to obtain accurate estimates of posterior distributions, and this number has been used for subsequent articles.

O'Neill *et al.* (2000) have carried out an analogous procedure where the likelihoods are estimated by a Monte Carlo (MC) method. They show that the method should be exact, independent of the sampling variance, providing that

$$\frac{E[\min(1, \hat{R})]}{E[\min(1, 1/\hat{R})]} = R,$$

where R is the true likelihood ratio and $\hat{R}$ is its estimate. They suggest using the estimator $R^* = \hat{R}^2/\tilde{E}[\hat{R}]$, where $\tilde{E}[\hat{R}]$ is an estimate of the expected value of the ratio, to correct for the bias in $\hat{R}$. Details of how R^* has been estimated for the genealogical model considered here are given in the appendix. In the results below, simulations carried out with this bias correction are referred to here as MCWM with bias correction and the earlier method as MCWM without bias correction.

Independence Metropolis-Hastings simulation: The methods described above all use importance sampling to approximate $p(D|\Phi)$ and, with or without bias correction, will lead the MCMC simulation to sample from an approximate posterior distribution. I now show how a small modification to the approach will guarantee that the MCMC will sample from the true posterior distribution.

Consider now an importance sample of size 1 (*i.e.*, $h = 1$ in Equation 2 above). The importance weight,

$$p(G, D|\Phi)/q(G, D|\Phi),$$

is an (admittedly very poor) estimate of $p(D|\Phi)$ as described above, but is also the ratio of the probability of sampling the genealogy under the coalescent to the probability of sampling the genealogy under the importance-sampling function. Supposing this were used in the Metropolis-Hastings simulation described by (3), the ratio of importance weights for the ith and $i + 1$th MCMC update is

$$\frac{p(D, G_{i+1}|\Phi_{i+1})}{p(D, G_i|\Phi_i)} \times \frac{q(D, G_i|\Phi_i)}{q(D, G_{i+1}|\Phi_{i+1})},$$

which, multiplied with the Hastings term for the parameter updates, $p(\Phi_i|\Phi_{i+1})/p(\Phi_{i+1}|\Phi_i)$, will give (3) above. Note that the Hastings term for updates to G is not conditional on any particular value of G. Thus, at least for G, this method is an example of the well-studied independence Metropolis-Hastings sampler (see, *e.g.*, Tierney 1996, pp. 69–70) and has been proposed as a possible approach for genealogical inference by Stephens and Donnelly (2000). The MCMC is sampling the posterior distribution of genealogical histories, as well as parameter values, and inference is performed in much the same way as in Wilson and Balding (1998) and Beaumont (1999). If the importance sampling is used in this way, then the MCMC will correctly sample from the posterior distribution of parameters provided that the current genealogical history and associated likelihood are *kept* like any other parameter rather than *resampled* at each evaluation of (3). The sampled genealogical history is treated as a parameter on an equal footing with Φ, and although it would be possible to update the genealogical history independently of the demographic parameters, they are all updated together in the following simulations. To reiterate, the difference between the independence Metropolis-Hastings sampler and MCWM is that in MCWM the importance weight is viewed as an estimate of $p(D|\Phi_i)$ in (4) and, although it would never be advisable to use it with a sample of size 1, is reevaluated with a new G_i at each evaluation of (4), whereas with the independence Metropolis-Hastings sampler the current G_i is retained at each evaluation of (3).

As is shown in the results, using a single genealogy in the independence sampler leads to very poor convergence of the MCMC, and, again this is a well-known property of the independence Metropolis-Hastings sampler when the sampling function is a poor approximation of the target density (Tierney 1996). Essentially the distribution of importance weights is very skewed so that the simulation will "stick" at the (very rarely obtained) high importance weights and then wait a long time for Equation 3 to be satisfied. By contrast, at the other extreme, if the importance sample size was very large so that independent estimates of the likelihood ratio R had negligible variance then convergence of the MCMC would depend only on Φ and would generally be very good. Intuitively, therefore, we should get better convergence if we take larger sample sizes, but then the question arises whether the MCMC will converge to the required target density exactly—*i.e.*, $p(\Phi, G|D)$.

As shown in the appendix the target density for the MCMC becomes rather more complicated when we consider importance sample sizes greater than one. However, it can be proved (see the appendix) that if Equation 2 is used with values of $h > 1$ then the independence sampling procedure will always give the correct posterior densities for the demographic and genealogical parameters for any importance sample size. Although the sample of h genealogical histories observed at any point in the simulated chain is not drawn from the posterior distribution, if we consider the sample of genealogies simulated by the importance sampling procedure to be ordered and keep a track of, say, the genealogies occupying the jth position throughout the MCMC simulation, then these genealogies will (in the long run) be sampled from the correct posterior distribution and, jointly with the parameters, will be sampled from $p(\Phi, G|D)$. As described below, simulation tests suggest that acceptance rates increase rapidly with larger importance sample sizes, and for adequate importance sample sizes this procedure is, in general, more efficient than the other methods. This method differs from the normal independence sampler because we are using a group of sampled genealogies rather than one and is called grouped independence Metropolis-Hastings (GIMH) to distinguish it from the normal independence Metropolis-Hastings and from MCWM, which involves reevaluation of the likelihood. Since the grouped independence Metropolis-Hastings sampler can be shown to converge to the target densities exactly, whereas this is only approximate in the case of MCWM, with or without bias correction, the bulk of the analyses performed in this article are carried out using this approach.

A basic algorithm for GIMH (or the standard independence sampler when $h = 1$) is as follows:

1. Choose initial parameter values Φ_i with $i = 0$.
2. Sample h independent genealogical histories using importance-sampling function and calculate $\tilde{p}_i(D|\Phi_i)$ from Equation 2.
3. Draw $\Phi_{i+1} \sim p(\Phi_{i+1}|\Phi_i)$.
4. Sample h independent genealogical histories using importance-sampling function and calculate $\tilde{p}_{i+1}(D|\Phi_{i+1})$ from Equation 2.
5. Accept Φ_{i+1} and $\tilde{p}_{i+1}(D|\Phi_{i+1})$ with probability

$$\min\left(1, \frac{\tilde{p}_{i+1}(D|\Phi_{i+1})}{\tilde{p}_i(D|\Phi_i)} \times \frac{p(\Phi_i|\Phi_{i+1})}{p(\Phi_{i+1}|\Phi_i)} \times \frac{p(\Phi_{i+1})}{p(\Phi_i)}\right).$$

   Otherwise $\Phi_{i+1} = \Phi_i$ and $\tilde{p}_{i+1}(D|\Phi_{i+1}) = \tilde{p}_i(D|\Phi_i)$.
6. Set $i = i + 1$ and go to 3.

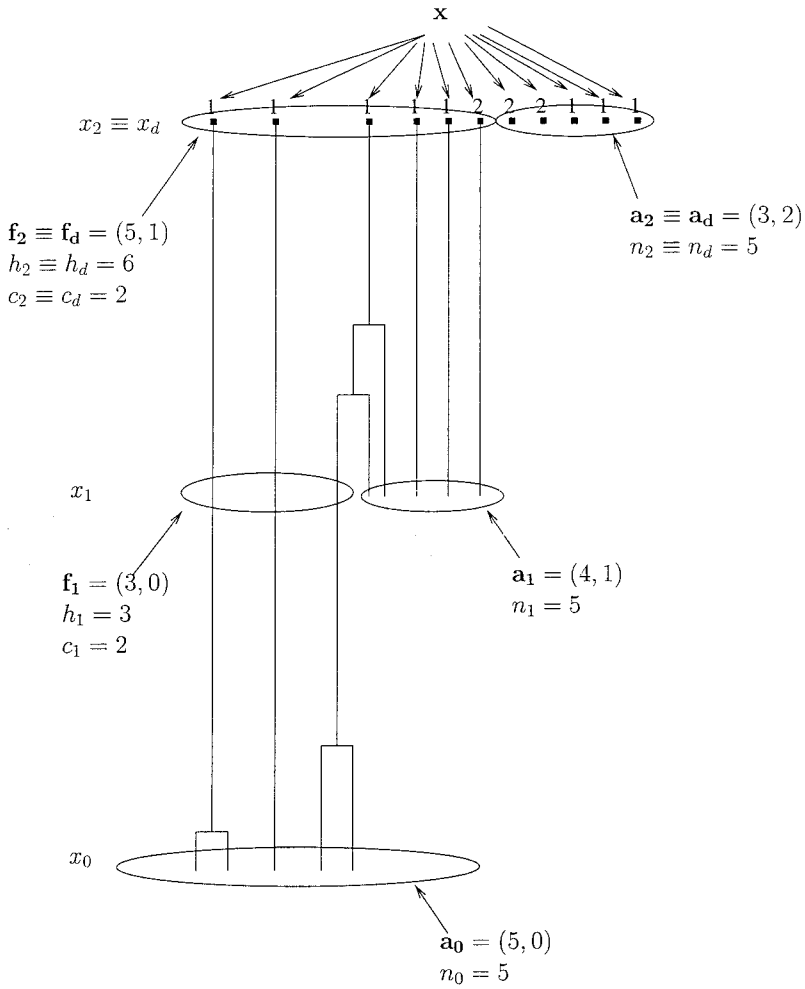

FIGURE 1.—Diagram illustrating the terminology used in the text.

In this algorithm the importance sampling calculations are explicitly indexed for clarity. The essential difference from MCWM is that the iterations start at step 3 and genealogical histories are simulated only for the trial values.

INFERENCE IN THE TEMPORAL METHOD BASED ON A COALESCENT MODEL WITH SAMPLES TAKEN AT MANY TIME POINTS

The data are assumed to be sampled at different times, given by the sequence $\mathcal{X} = (x_0, x_1, \ldots, x_d)$. Time is measured in units of generations, and the most recent sample is given subscript 0, and $x_0 = 0$. The population is changing exponentially in size from a previously constant ancestral size N_A at time X to size N_0 at time 0. Time is taken to be increasing into the past, and terms such as "earlier" and "later" refer, respectively, to times nearer or farther from the most recent sample (see Figure 1). Corresponding to each time point is a sequence of sample sizes (number of chromosomes) $\mathcal{N} = (n_0, n_1, \ldots, n_d)$, where lineages are added to the genealogy. The sequence of frequency counts of the different allelic types in each sample is given by $\mathcal{A} = (\mathbf{a_0}, \mathbf{a_1}, \ldots, \mathbf{a_d})$, where the vectors are of length k, the total number of different allelic types observed in the data. For times $> x_0$, at the time each set of lineages is added a number of lineages are present with descendants in earlier samples, lower down the genealogy (*i.e.*, lower down in Figure 1). The allele frequency counts among these base lineages are denoted here as the random variable $\mathcal{F} = (\mathbf{f_0}, \ldots, \mathbf{f_d})$, where, to ease the notation below, $\mathbf{f_0}$ is defined to be 0. The number of these lineages, also a random variable, $H = (h_1, \ldots, h_d)$, depends on the number of coalescences that occur in the intervals between sampling points. These are given by the sequence $\mathcal{C} = (c_1, \ldots, c_d)$. Thus, at the ith sample point, the number of lineages deriving from earlier samples is given by

$$h_i = \sum_{j=0}^{i-1} n_j - \sum_{j=1}^{i} c_j, \quad i \geq 1.$$

The notation used here is summarized in Figure 1.

The likelihood, assuming a model of drift without mutations, can be obtained as a straightforward extension of the two-sample case in BERTHIER *et al.* (2002) and is given by

$$p(\mathcal{A}/\mathcal{X}, N_0, N_A, X) = \sum_{\mathcal{C},\mathcal{F}} \Big[p(\mathbf{a_d} + \mathbf{f_d}) \prod_{i=0}^{d-1} p(\mathbf{f_i} + \mathbf{a_i} | \mathbf{f_{i+1}}, c_{i+1}) \times p(\mathbf{a_{i+1}}, \mathbf{f_{i+1}} | \mathbf{a_{i+1}} + \mathbf{f_{i+1}}) \, p\left(c_{i+1} \Big| \frac{x_{i+1} - x_i}{2\tilde{N}_{i+1}}\right) \Big], \quad (5)$$

where

$p(\mathbf{a}_d + \mathbf{f}_d)$ is the probability of sampling the gene frequencies in the lineages extant at the earliest sampling time (at the top of Figure 1);

$p(\mathbf{f}_i + \mathbf{a}_i|\mathbf{f}_{i+1}, c_{i+1})$ is the probability of obtaining the gene frequencies among the lineages extant at sample i given the base lineages at $i + 1$ and the number of coalescences within the interval;

$p(\mathbf{a}_{i+1}, \mathbf{f}_{i+1}|\mathbf{a}_{i+1} + \mathbf{f}_{i+1})$ is the hypergeometric sampling probability of obtaining the frequencies in the base lineages and the frequencies in the sample lineages, given the frequencies of the combined lineages; and

$p(c_{i+1}|(x_{i+1} - x_i)/(2\tilde{N}_{i+1}))$ is the probability of obtaining c coalescences in the sampling interval, over which the harmonic mean effective size is $\tilde{N}$

(see the APPENDIX for further details). The sum is over all possible numbers of coalescences between sampling intervals and all possible frequency counts among the base lineages at each interval. In the case of many unlinked loci, the likelihoods can be estimated for each locus separately and then multiplied together. Although in principle the possibilities can be straightforwardly enumerated, allowing Equation 5 to be solved, in practice there are far too many possibilities to make this useful. Instead, the importance sampling approach of GRIFFITHS and TAVARÉ (1994a) is applied to this problem, as in BERTHIER *et al.* (2002).

In this approach S independent sequences of coalescences of lineages are explicitly sampled by simulation (see the APPENDIX for details), and we obtain

$$\tilde{p}(\mathcal{A}/\mathcal{X}, N_0, N_A, X) = \frac{1}{S}\sum_{\kappa=0}^{S}\left[p(\mathbf{a}_d^\kappa + \mathbf{f}_d^\kappa)\prod_{i=0}^{d} p(\mathbf{a}_i^\kappa, \mathbf{f}_i^\kappa|\mathbf{a}_i^\kappa + \mathbf{f}_i^\kappa)\prod_{e=0}^{c_i^\kappa} w_{i(e+1)}^\kappa\right]. \quad (6)$$

Thus for the κth simulated sequence $p(\mathbf{f}_i + \mathbf{a}_i|\mathbf{f}_{i+1}, c_{i+1}) \times p(c_{i+1}|(x_{i+1} - x_i)/(2\tilde{N}_{i+1}))$ in (5) is replaced by $\prod_{e=0}^{c_i^\kappa} w_{i(e+1)}^\kappa$, which is the ratio of the probability of obtaining the sampled sequence of lineages under the coalescent model, independent of the data, to the probability of obtaining it from the importance-sampling function.

The c_i^κ coalescences are simulated using the coalescent model (see the APPENDIX for details of how this was done for a population of varying size). The distribution of the number of coalescences between data-sampling intervals is identical under the coalescent model and the importance-sampling function, and hence this term cancels out. This form of sampling is used for all the analyses described below. However, if importance weights are to be evaluated at parameter values other than those used to generate the samples, the terms in (6) need to be multiplied by a weight reflecting the different probability of obtaining the simulated number of coalescences under the coalescent compared to that under the importance-sampling function. This can be done in two ways. The weight $p(c_i,(x_{i+1} - x_i)/2\tilde{N}_{i+1})/p(c_i,(x_{i+1} - x_i)/2\tilde{N}^*_{i+1})$ can be used from TAVARÉ's (1984) Equation 6.1 for each interval between samples, where $\tilde{N}$ and $\tilde{N}^*$ are calculated for each interval from (A3), and $\tilde{N}^*$ is used to generate the importance samples. Alternatively, the simulated coalescence *times* can be recorded, and an equivalent ratio can be calculated from their joint density under the coalescent compared to their joint density under the importance-sampling function. The advantage of the former is that it is marginal to the coalescence times and should therefore be more efficient; however, it is computationally time consuming to calculate and numerically unstable, and the latter is probably more practicable.

In the results described in the next sections Equation 6 has been used on its own to estimate likelihoods and also incorporated into the MCWM procedure with and without bias correction and into the GIMH) sampler. In general when MCWM and GIMH are used in the analyses rectangular priors are assumed for each parameter, as in BEAUMONT (1999). In all of the analyses, X is assumed to be equal to x_d and not separately estimated. In the MCMC, the initial values of the parameters are taken uniformly randomly from the priors. They are updated from a lognormal distribution with the median centered on the current value of the parameter and standard deviation (on a log scale) of 0.5, unless otherwise stated. In all the MCMC analyses the parameters are updated simultaneously (with the genealogies, as discussed above). Comparisons among the various approaches are made to demonstrate the superiority of GIMH, and then this method is used for further investigations of the accuracy and coverage properties of the method using simulated data sets. Finally GIMH is applied to three published data sets to illustrate its utility.

SIMULATION TESTS

Comparison of MCWM and GIMH with pure IS estimation: To compare the accuracy of the three different MCMC approaches a data set was simulated from the model from a diploid population with effective size $N_e = 51.2$. The population did not change in size over the sampling period, and six samples each of size 20 chromosomes were taken at generations 0, 4, 8, 12, 16, and 20. The data set consisted of 10 loci each with five alleles in the population (although, due to sampling, some data sets had fewer than five alleles). The population frequencies were simulated from a uniform Dirichlet distribution, according to the assumption of the model.

The data set was then analyzed by four different approaches (in all models, $N_e = N_0 = N_A$):

i. The likelihoods for a grid of 81 values of N_e from 20 to 80 were evaluated using (6). The likelihoods were evaluated at each point independently, using an IS size of 40,000. The standard errors were estimated using (A2). The approximate likelihood surface was normalized to have unit volume, and the

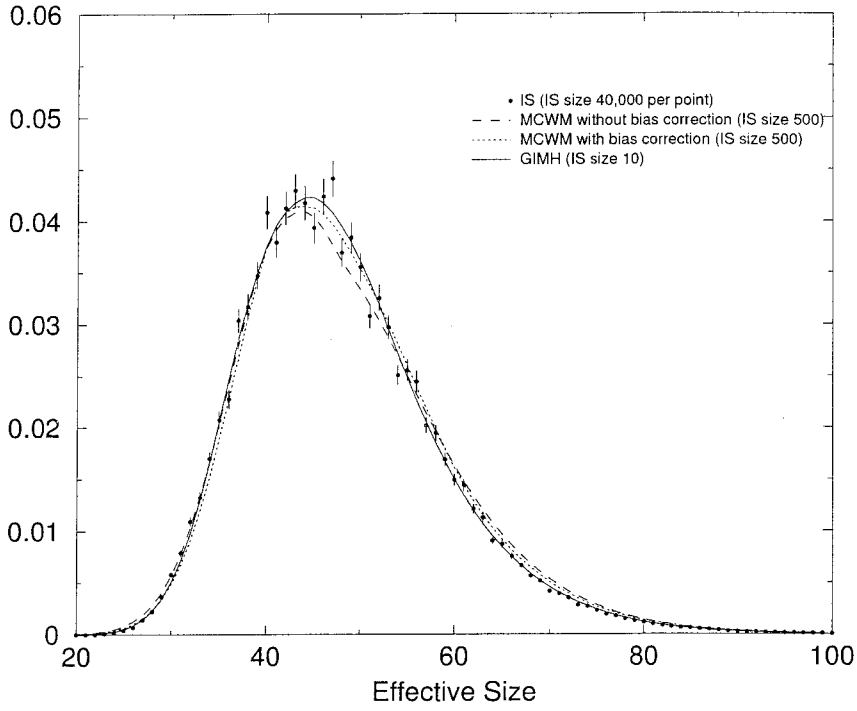

FIGURE 2.—Comparison of the posterior distributions obtained using MCWM without bias correction, MCWM with bias correction, GIMH, and pure importance sampling.

standard errors were scaled accordingly. The standard deviation was estimated from this distribution.

ii. Nine different simulations using MCWM without bias correction were carried out in which the IS sizes used in the evaluation of (4) were 5000, 1000, 500, 100, 50, 10, 5, 2, and 1. The simulations were run for 20,000 updates, which appeared to give good convergence (as judged by eye from the output traces), and densities and standard deviations of the posterior distribution were estimated from the values of N_e generated by the simulation.

iii. Eight simulations using bias-corrected MCWM were carried out as for MCWM. Simulations using an IS size of one were not performed because $\mathrm{SE}[\tilde{p}(D/\Phi)]$ cannot be estimated.

iv. Nine simulations were carried out using GIMH as for MCWM. However, with GIMH the rate of convergence is heavily dependent on the IS size used. In particular, with an IS size of one the MCMC procedure tends to mix very poorly because the simulated chain will stick at chance high values of $\tilde{p}(D/\Phi)$. The length of simulation, the thinning interval (the number of MCMC iterations between successive recordings of parameter values), and the standard deviation of the trial parameter updates were varied between simulations to achieve satisfactory convergence, judged by eye from output traces.

Estimated densities using the four approaches are shown in Figure 2. It can be seen that the standard errors for the IS method are still large, even with 40,000 points. However, the posterior distributions estimated by MCWM with and without bias correction are very similar to each other and to the distribution estimated from the pure IS method, despite the variability in the estimates of the likelihood (for an IS size of 500 the standard errors are expected to be around nine times larger than those shown in Figure 2). The distribution for GIMH with an IS size of just 10 per MCMC update (evaluation of Equation 4) is very close to that of the pure IS method.

Figure 3 shows how the width of the estimated posterior distribution varies with the IS size.

The standard deviation estimated from the pure IS method is 10.4. It can be seen that for MCWM with and without bias correction there is a strong relationship between the width of the distribution and the IS size. Bias correction does appear to ameliorate the problem to some extent, but still leads to inaccurate estimation of the posterior distribution when the IS size is small. For MCWM without bias correction, an IS size of $\sim$500 is the minimum required for accurate estimation, whereas $\sim$100 are needed with bias correction. The rate of convergence of the two MCWM methods appears to be independent of IS size.

GIMH is unaffected by the IS size, as expected from the result in the APPENDIX. This is not a free lunch, however. The tradeoff is that the amount of mixing is severely reduced when the IS size is low, and the length of time required to achieve convergence is correspondingly increased. For example, with an IS size of 1 the result shown in Figure 3 was obtained by pooling together results from seven independent simulations of 10^8 MCMC updates, thinned every 10,000 updates. Even in this case, there is still appreciable variability in the results between the independent simulations. The result for an IS size of 10 was obtained from a single simulation of 10^7 MCMC updates thinned every 200 updates. In

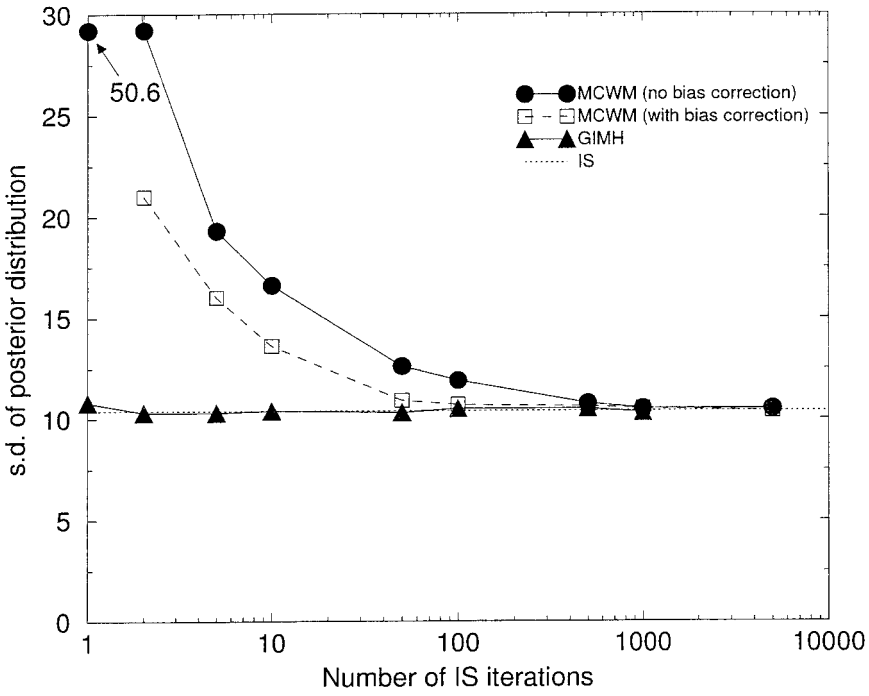

FIGURE 3.—The relationship between the IS size and the standard deviation of the posterior distribution for MCWM with and without bias correction and GIMH. The standard deviation estimated from pure IS is shown as a dotted line.

this case convergence appears very good, with a very uniform trace for N_e, and the resulting density is plotted in Figure 2.

These points are further illustrated in Figure 4, where the proportion of trial updates that are accepted in the MCMC, divided by IS size, is plotted against IS size. The acceptance rate varied rapidly from 0.0011 with an IS size of 1 to 0.15 with an IS size of 10 and then more slowly to 0.85 with an IS size of 5000. It can be seen in Figure 4 that the scaled acceptance rate has an optimum at an IS size of ~10.

In these simulations the standard deviation of the distribution of parameter updates was kept at 0.1 for all IS sizes, and therefore the scaled acceptance rate is a measure of efficiency—for a given number of accepted trial updates, the total required number of IS evaluations is at a minimum if an IS size of 10 is used. This optimum will vary for different data sets and sizes of the trial parameter updates, and for the remaining analyses described in this article, which mostly used a standard deviation of 0.5 for the trial parameters, unless otherwise stated, GIMH was used with an IS size of 100 and 10^5 MCMC updates, thinned every 10 updates to give 10,000 points.

Effect of sample size and numbers of alleles and loci on the estimation of N_A and N_0: To illustrate the effect of varying aspects of the sample, five independent simulations were performed for each combination of parameters in Table 1. As shown in the table the parameters were also summarized by a composite parameter SSAL = (sample size) × (k − 1) × (number of loci), where k is the number of alleles. Samples were simulated from populations that grew from N_A = 20 to N_0 = 200 and also populations that contracted from N_A = 200 to N_0 = 20. The samples were taken at six times, $\mathcal{H}$ = (0, 2, 4, 6, 8, 10). The population frequencies were simulated from a uniform Dirichlet distribution, as before. The aim of the analysis was to compare the effect of the parameters on the deviation of the joint posterior mode of N_A and N_0 from the value used in the simulations and also to illustrate typical posterior distributions obtained with different data sets. Ideally, of course, an analysis of the accuracy of estimators should use a larger number of replicates, but the time taken to run the MCMC precludes this. Five replicates are, however, sufficient to illustrate the general trend toward consistency in the estimator, as the amount of information in the data increases. This number was chosen because pairs of sets of five replicates could be run in parallel on a 10-node cluster of 700-Mhz Pentium 3 processors running under Linux. MCMC parameters are as described above for all simulations other than those with SSAL = 8000. In this case an IS size of 500 was used and the standard deviation (on a log scale) of the lognormal used for updating the demographic parameters was 0.1 rather than 0.5. The simulations took ~4 hr for SSAL = 800 and ~6 days for SSAL = 8000 (which has an IS size five times larger).

Examples of the joint posterior distributions for N_A and N_0 are shown in Figure 5. The posterior distributions are illustrated using highest posterior density (HPD) limits (as in BEAUMONT 1999). These are obtained from the simulations of growing and declining populations with SSAL = 800 and SSAL = 8000 (see Table 1). In each case, of the five replicate simulations, that where the mode for N_A and N_0 is the median distance away from the true value was chosen to be illustrated. It can be seen that there is a tendency for the larger population

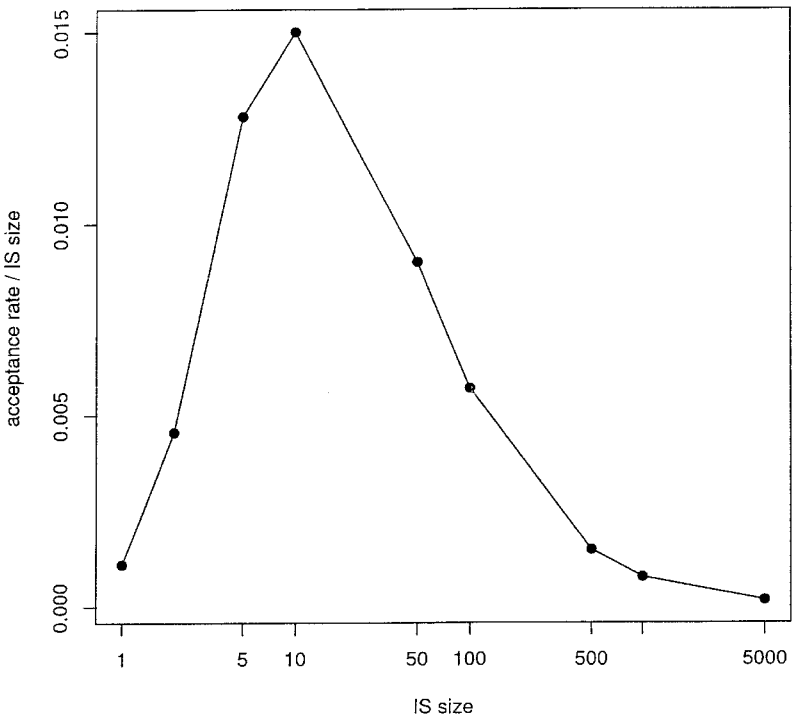

FIGURE 4.—The proportion of trial updates in the MCMC that are accepted, divided by IS size, is plotted against IS size.

size to be most poorly estimated, with substantial skew in the posterior density. In addition, there is a tendency (observed generally, as well as in the simulations that are illustrated) for the current population size to be well estimated by the joint mode (25 modes higher than true value and 35 lower out of 60 simulations) and the ancestral population size to be generally overestimated by the modes (44 modes higher and 16 lower).

Using the mode from the joint posterior distribution as an estimator for N_A and N_0, the square root of the relative square error [defined as $(\widehat{N_A} - N_A)^2/N_A^2 + (\widehat{N_0} - N_0)^2/N_0^2$], referred to here as the "relative error," was calculated for each simulation and is shown plotted against SSAL in Figure 6, a and b. Each point in the figure is the relative error for a data set plotted against the SSAL value given in Table 1. Although there is substantial variability a general trend toward a reduction of the relative error with increasing SSAL can be seen. Given the variability of the results, the logarithm of the relative errors was analyzed using a linear model. In the model growth/decline was specified as a factor and the covariates (log transformed) were number of loci, number of independent alleles at each locus ($k - 1$), and sample size. The coefficients and standard errors were 4.96 (1.30), −0.629 (0.217), −0.517 (0.325), −0.794 (0.314), and −0.686 (0.164) for the intercept, effect of growth, number of loci, $k - 1$, and sample size, respectively. The effect of growth/decline was significant at $P = 0.005$, and the effect of the three covariates was significant at $P = 0.0001$. The residuals from this model were roughly normal with no obvious heteroscedasticity, although the limit on the relative errors arising from the rectangular priors has some effect on the residuals. A model with the coefficients for the three covariates forced to be −1 did not fit significantly less well than a model where the three covariates were free to vary ($P = 0.21$). This simple model gives the equations Relative Error = 2128/SSAL for a declining population and Relative Error = 1135/SSAL for a growing population. The fitted values from this model are shown in Figure 6. Thus the main conclusions of this analysis are: (a) there is, at least at the level of precision in this simulation study, an equivalence between the number of independent alleles, number of loci, and sample size; and (b) for the same value of SSAL the relative error is almost twice as large in a declining population in comparison with a growing population.

The Bayes factor favoring a model of population growth *vs.* decline was calculated for each MCMC simu-

TABLE 1

Combinations of numbers of loci, numbers of alleles at each locus, and sample size used

SSAL	No. loci	No. alleles per locus	Sample size
800	10	5	20
4000	10	5	100
1600	10	9	20
8000	10	9	100
1600	20	5	20
4000	25	9	20

The sample size is the number of chromosomes taken at each of six time points. The composite parameter SSAL = (sample size) × (no. of alleles per locus − 1) × (no. of loci). This set of combinations was used for populations that grew from $N_A = 20$ to $N_0 = 200$ and contracted from $N_A = 200$ to $N_0 = 20$. Further details are in the text.

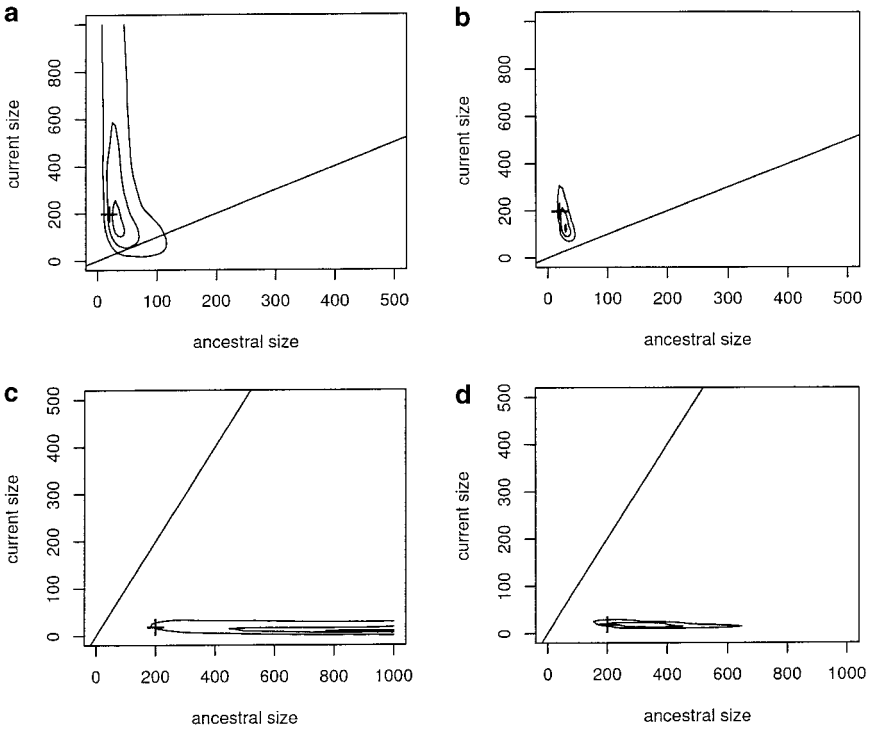

FIGURE 5.—Posterior distribution of N_A and N_0 for four different simulated data sets. The contour levels correspond to the 90, 50, and 10% HPD limits. The line where $N_A = N_0$ is shown. The values of N_A and N_0 used in the simulation are shown as a cross. (a) SSAL = 800, N_A = 20, N_0 = 200; (b) SSAL = 8000, N_A = 20, N_0 = 200; (c) SSAL = 800, N_A = 200, N_0 = 20; (d) SSAL = 8000, N_A = 200, N_0 = 20.

lation as the proportion of MCMC iterations where $N_0 > N_A$ divided by the proportion of iterations where $N_0 < N_A$ (each model has equal prior probability). The Bayes factor gives the relative likelihood of one model over the other (GELMAN *et al.* 1995). The logarithm of the Bayes factor is plotted against SSAL in Figure 7, a and b. Each point in the figure is the logarithm of the Bayes factor for a data set plotted against the SSAL value given in Table 1. Although some of the simulations with SSAL $\leq$ 1600 have $|\log(\text{Bayes factor})| < 2$ (*i.e.*, would be judged to be nonsignificant by conventional criteria), the great majority of results very strongly support the model under which they were generated. It should be noted that the Bayes factor is sensitive to the priors chosen, and this is discussed in more detail in the context of the example data sets analyzed below.

The bias in the joint estimation of N_A and N_0 apparent in Figure 5 does not appear to be caused by any systematic error in the estimation procedure, as judged by an examination of the coverage properties of the posterior distributions. The critical HPD *P* values corresponding to the true N_A and N_0 were estimated for each data set and plotted against the SSAL values given in Table 1 and Figure 8, a and b. If the posterior distribution was the same as the repeated sampling distribution the HPD *P* values should be uniformly distributed, irrespective of treatment. This will be true asymptotically when the posterior distribution approximates a multivariate normal. It can be seen that the estimated critical *P* values are broadly uniformly distributed, which is what would be expected under asymptotic theory. A Kolmogorov-Smirnoff one-sample test on the 60 *P* values shows no departure from a uniform ($P = 0.93$). There is no trend toward small *P* values with increasing SSAL, which would be expected if there was an error in the estimation procedure. The estimated critical *P* values for the examples in Figure 5, a–d, are, respectively, 0.57, 0.29, 0.11, and 0.44. Thus, overall, the method appears to estimate changes in effective population size satisfactorily, but it is preferable to present results for the full posterior distribution rather than rely on the mode as a point estimate.

ANALYSIS OF EXAMPLE DATA SETS

To illustrate the behavior of the method on real data sets, three examples have been chosen: data from a population of *Drosophila subobscura* surveyed by BEGON *et al.* (1980), data from a population of northern pike (*Esox lucius*) surveyed by MILLER and KAPUSCINSKI (1997), and data from the Mauritius kestrel surveyed by GROOMBRIDGE *et al.* (2000).

Drosophila: The data were sampled from a population on Mount Parnes, 40 km north of Athens. The flies were genotyped for nine allozyme loci. The study site occupied ~20,000 m^2 of fir woodland at an elevation of 900 m. BEGON *et al.* (1980) estimated the total suitable habitat to extend at least 10^7 m^2. Thus the population is clearly open, vitiating one of the assumptions of the temporal method. Samples were taken in September 1975 (190 individuals), September 1976 (250 individuals), and May 1977 (335 individuals). BEGON *et al.* (1980) estimated these corresponded to sampling intervals of nine and two generations, respectively. Using mark-release-recapture methods they estimated the census size in their study area to be ~150,000 individuals. These

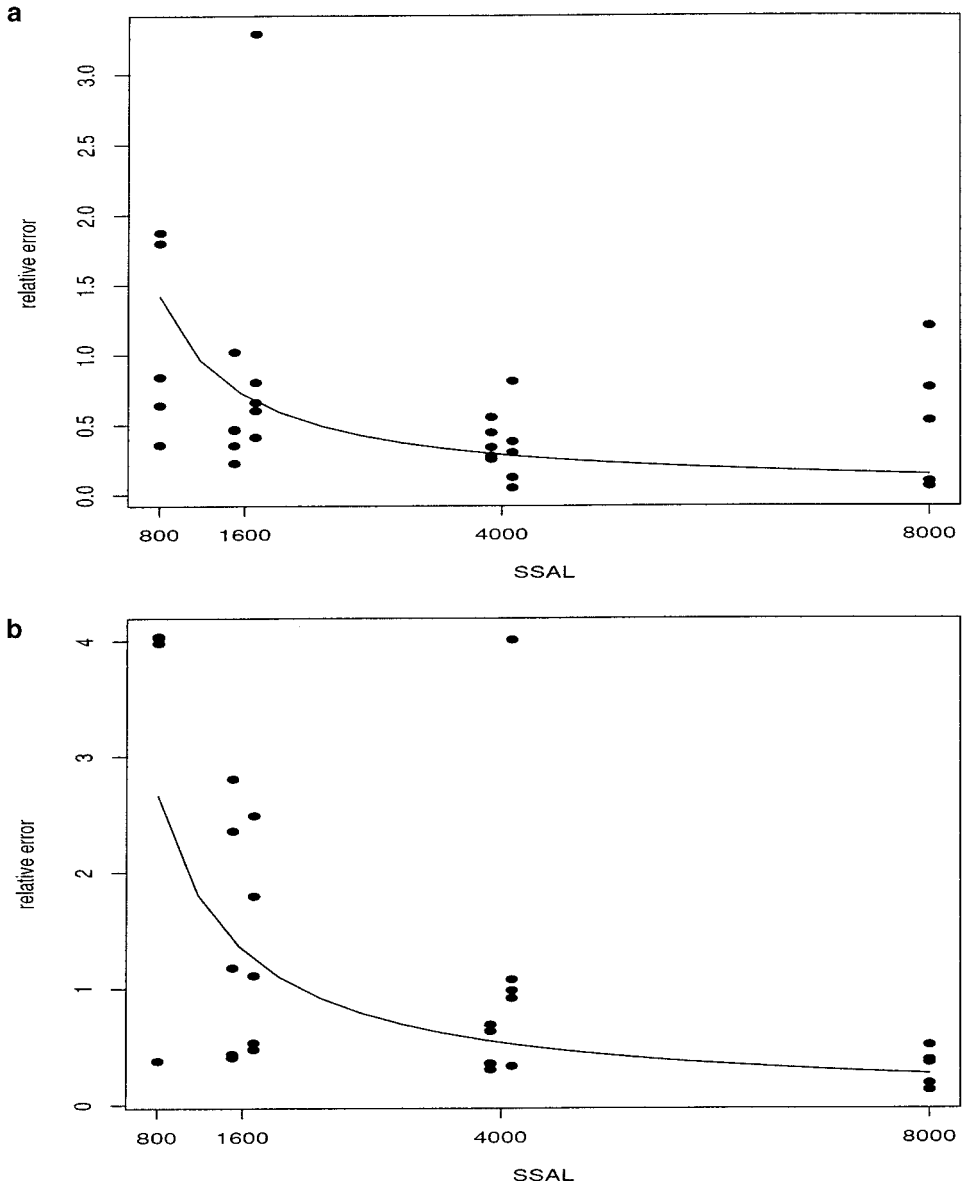

FIGURE 6.—A plot of the relative error in estimation of ancestral and current population size against SSAL, which is a summary of the number of loci, number of alleles, and sample size. Simulations with different configurations in Table 1 but the same SSAL have been shifted slightly so that those with a higher number of loci are on the right. The fitted line is obtained using the model described in the text. (a) Growing population; (b) declining population.

data have also been analyzed by ANDERSON *et al.* (2000), who noted that the frequencies at one locus (*Pgm*) appeared to be misreported in BEGON *et al.* (1980), and used only eight loci. For comparison, these same eight loci are analyzed here (input file kindly provided by Eric Anderson). The number of alleles at each locus varied from three to six. A rectangular prior of (0, 5000) was chosen for both N_A and N_0. The joint posterior distribution for N_A and N_0 is shown in Figure 9.

In addition, a separate analysis was carried out with $N_e = N_A = N_0$ to compare with the results obtained by ANDERSON *et al.* (2000), who obtained a maximum-likelihood estimate for N_e of 500 with support limits (log-likelihood 2 units less than the maximum) of 250–975. The posterior distribution obtained with the method described here should be directly comparable with the likelihood curve estimated by ANDERSON *et al.* (2000) because the limits of the rectangular priors, (0, 5000), are substantially wider than the posterior distribution obtained. The trace for N_e is illustrated in Figure 10 (with the initial 100 points discarded). The time taken to obtain 10,000 points on a 500-Mhz Pentium was 27 hr, although it can be seen from Figure 10 that good estimates of the posterior distribution can be obtained with substantially fewer points.

The mode of the posterior distribution is 449 with support limits of 253–925 (in this case the support limit corresponds to the 0.922 HPD limit), which is very similar to the result of ANDERSON *et al.* (2000). Interestingly, as noted by Anderson *et al.* this result is very different from that obtained by POLLAK (1983), who obtained estimates of 253 (±115) for the first interval and 244 (±123) for the second and an overall estimate of 251 (±115) for both intervals. The reason for the discrepancy between the results from the two likelihood-based approaches and that from the moment-based approach

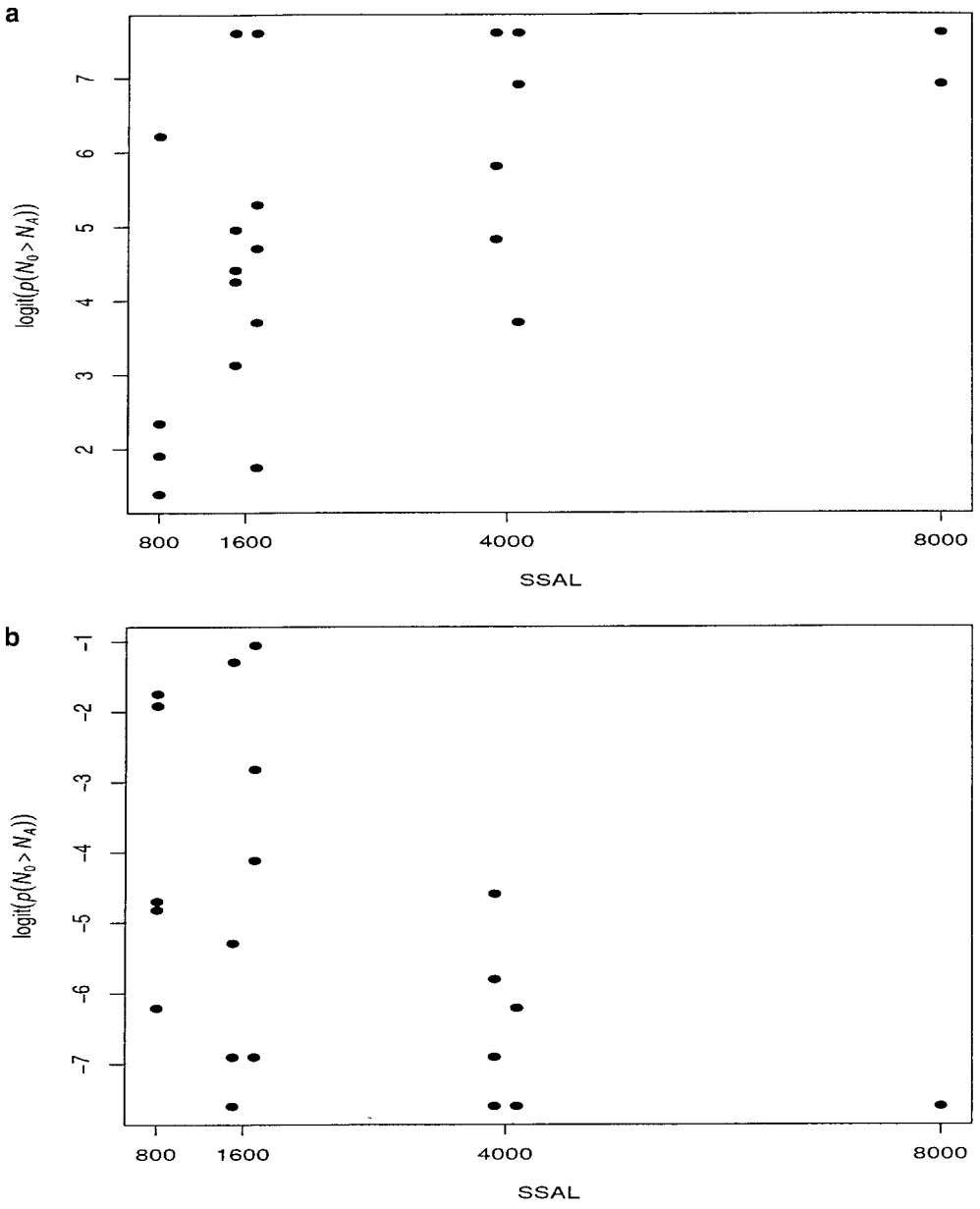

FIGURE 7.—A plot of the logarithm of the Bayes factor in supporting a model of population growth against SSAL. Other details are as in Figure 6.

of POLLAK (1983) is unclear (see ANDERSON *et al.* 2000 for discussion), although it is possible that omitting *Pgm* has some effect on the results.

The results for the varying population model are more in line with those of BEGON *et al.* (1980), who used the original method of KRIMBAS and TSAKAS (1971), and estimated N_e at 268 ($\pm$73) for the first interval and ∞ for the second. In Figure 9 it can be seen that there is very little evidence of a change in population size. The joint mode is at $N_A = 337$ and $N_0 = 890$. The line of equal population sizes is well within the 90% HPD limits. The Bayes factor in favor of growth is 5.4. The marginal modes and HPD limits are 196 (57–913) for N_A and 726 (112–4138) for N_0. As noted above, the Bayes factor is sensitive to the priors chosen. Thus, for example, widening the rectangular bounds equally for both N_A and N_0 will tend to increase the Bayes factor, and narrowing them will decrease it. In general the tendency for the posterior distributions to reach an asymptote for large N_A or N_0 will cause sample size to affect the inferences. Considering three samples, as here, if, for example, the most recent sample is smaller than the oldest sample there will be greater uncertainty in N_0 and the posterior distributions may be more likely to asymptote, and therefore there will be a tendency to suggest population growth, even if there is none. In fact, for the fly data, it is the oldest sample that is the smallest, and therefore this argument does not explain the broad posterior distribution for N_0.

Northern pike: Fish scale samples taken in 1961, 1977, and 1993 from Lake Escanaba, Wisconsin, were chosen from a collection of scales kept by the Wisconsin Depart-

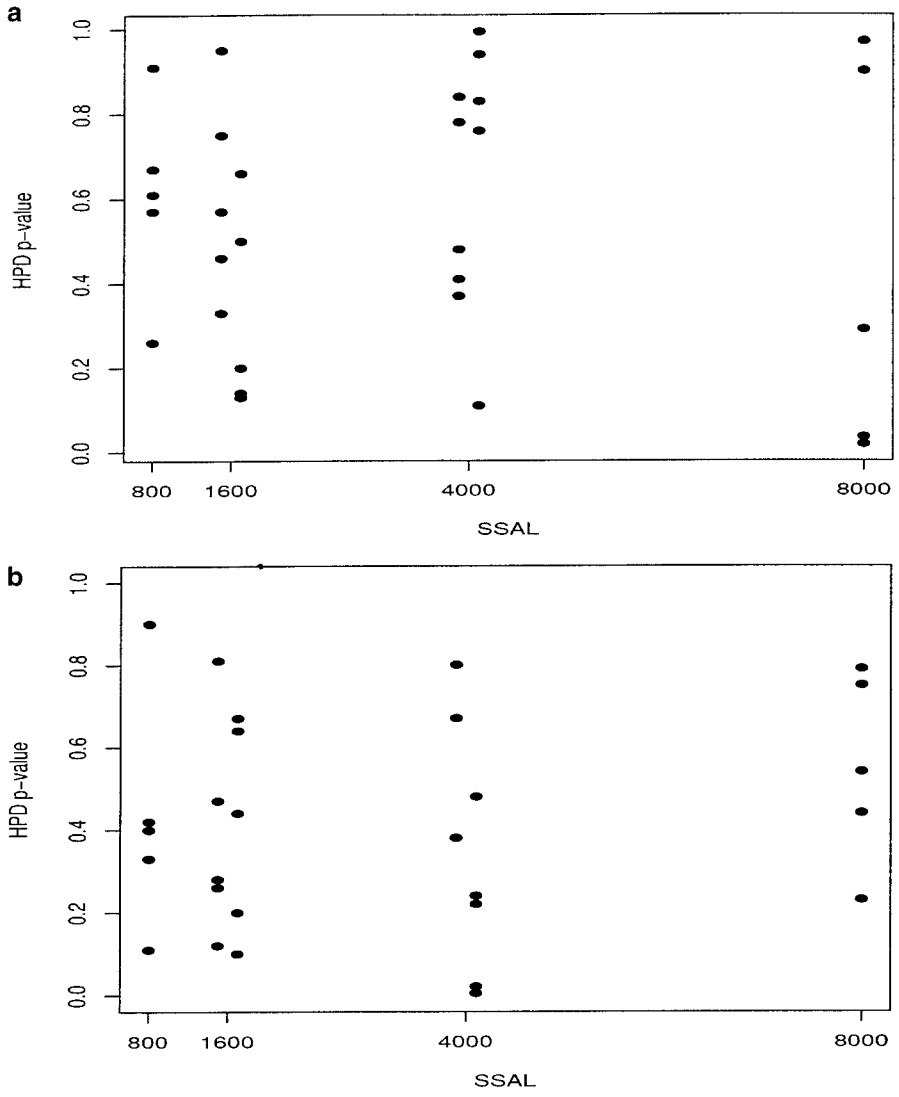

FIGURE 8.—A plot of the critical HPD P value of the true N_A and N_0 against SSAL. Other details are as in Figure 6.

ment of Natural Resources and were genotyped at seven microsatellite loci. Five of these loci were biallelic and the remaining two were triallelic. The allele frequency counts used in the following analysis were obtained from the relative frequencies in Table 3 of MILLER and KAPUSCINSKI (1997). There is good evidence that the population is closed and the last restocking of the lake was in 1941. The generation time was estimated by MILLER and KAPUSCINSKI (1997) to be 4 years. The same data were analyzed by WILLIAMSON and SLATKIN (1999). In their analysis, which was restricted to biallelic loci, the frequencies from two allelic classes at the two triallelic loci were combined.

The largest estimate of census size over the period 1961–1963 was 2300 individuals, and, assuming a ratio of effective to census size of <0.5, it seems reasonable to assume a rectangular prior of 0–1000 for both N_A and N_0. The results of the analysis of the gene frequency data are presented in Figure 11. It can be seen that there is good information on the ancestral effective population size and it is unlikely to be $> \sim 150$. There is less information on the current population size, which could be as high as 1000 or close to 0. The joint mode is at $N_A = 34.6$, $N_0 = 151$. The line of equal population size is well within the 90% HPD limits. The Bayes factor in favor of growth is 8.86. The modes and 90% HPD limits for the marginals are 20.0 (2.44–104) and 126 (8.88–766) for N_A and N_0, respectively. Thus, in conclusion, it is unlikely that the population is shrinking (although this depends on the priors chosen), but there is only very weak evidence of growth. The result here is similar to that obtained by WILLIAMSON and SLATKIN (1999) on the modified data, who estimated $N_A = 25$ and $N_0 = 107$. When interpreting the results it should be noted that the Bayes factor is comparing the posterior probabilities of growth *vs.* decline whereas the HPD analysis is asking whether a point on the line of equal population sizes is a reasonable draw from the posterior distribution. This latter question is more closely related to estimating a Bayes factor for growth *vs.* zero growth and involves comparing models of different dimensions. Although the implementation of reversible-jump MCMC (GREEN 1995) is relatively straightforward for this simple case, it is likely to increase convergence time and awaits further investigation.

Mauritius kestrel: The sample analyzed here consisted

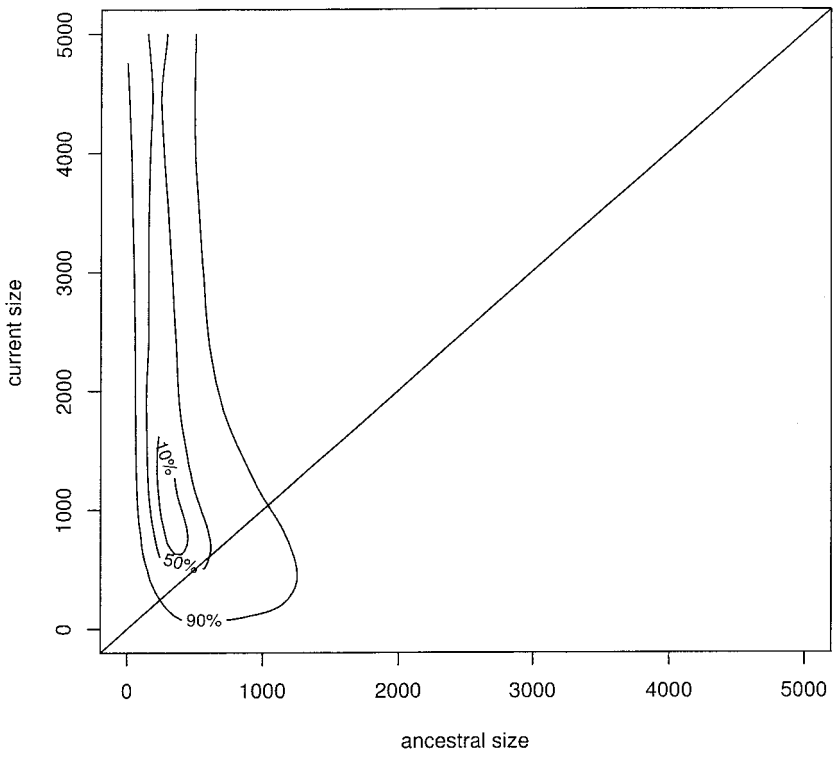

FIGURE 9.—Posterior distribution of N_0 and N_A for the fly data of BEGON *et al.* (1980). The contour levels are at the 0.1, 0.5, and 0.9 HPD limits, as in Figure 5.

of a number of individuals genotyped for 12 microsatellite loci, of which 7 were polymorphic. In this data set 75 individuals sampled in 1993 and (depending on the loci) up to 26 museum skins dating from 1829 to 1960 were genotyped. These data are described in GROOMBRIDGE *et al.* (2000) and the data were kindly given to me by Jim Groombridge. The population has undergone a dramatic decline over the 20th century and is believed to have been reduced to a single breeding pair in 1974. It now numbers some 200 pairs. Although this complex demography is not captured by the simple exponential model considered here, since there are no samples be-

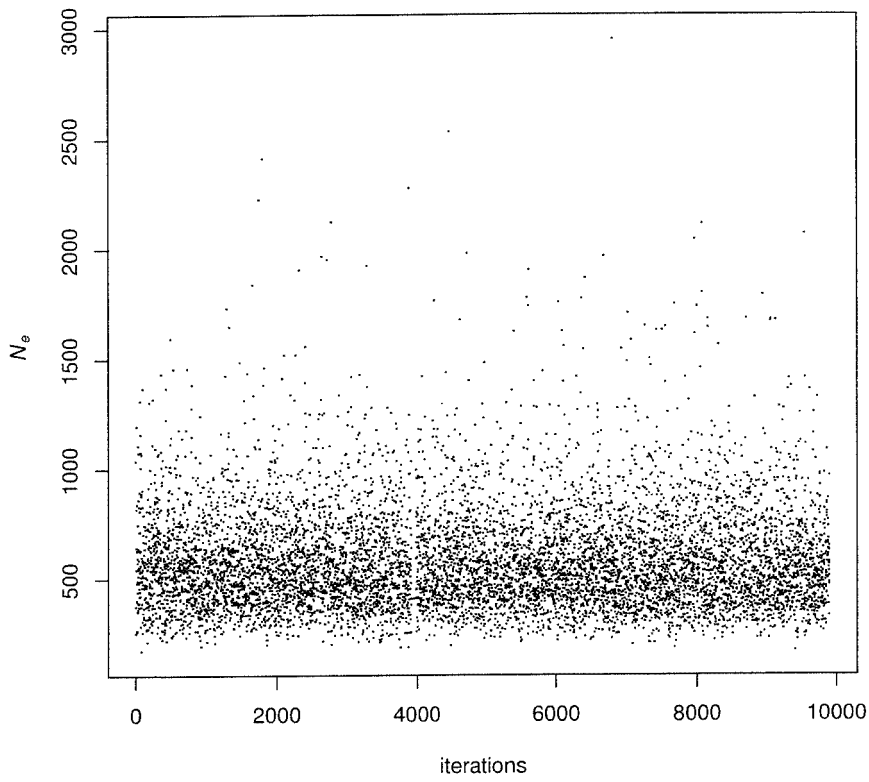

FIGURE 10.—Trace of the realized values of N_e during an MCMC run with the fly data of BEGON *et al.* (1980). The initial 100 points have been discarded.

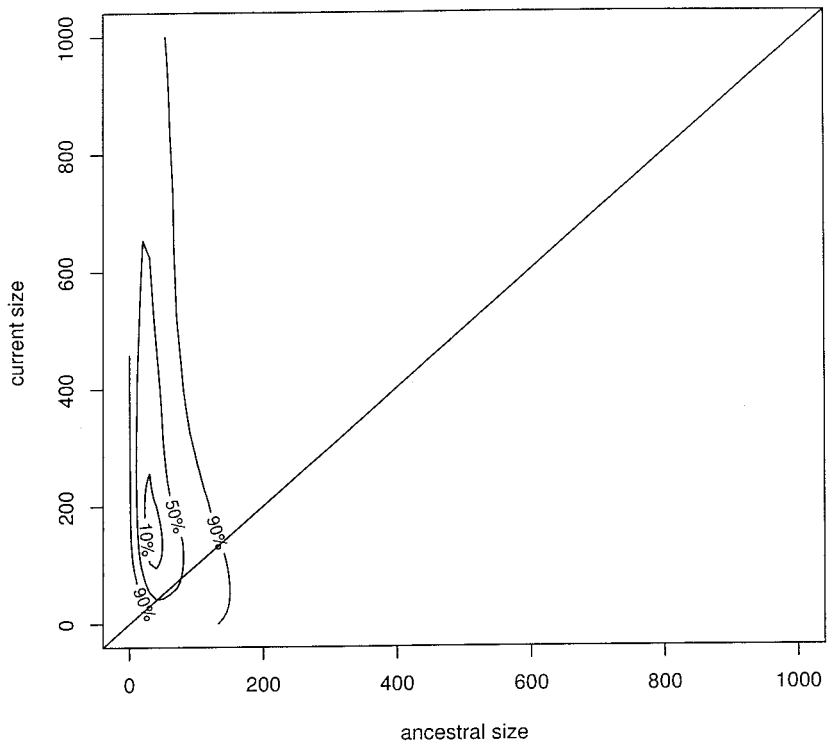

FIGURE 11.—Posterior distribution of N_0 and N_A for the northern pike data. The contour levels are at the 0.1, 0.5, and 0.9 HPD limits, as in Figure 5.

tween 1960 and 1993, the estimates of current effective population size will essentially reflect the effective size over this period, and this will be dominated by the 1974 bottleneck. For the analysis I assumed a generation time of 4 years and rectangular priors of 0–1000 for N_A and N_0. The posterior distribution is shown in Figure 12. There is very strong evidence of population decline, and N_A is unlikely to be $< \sim 300$ individuals and N_0 is unlikely to be $> \sim 10$ individuals. The joint mode from the density estimation is $N_A = 957$, $N_0 = 4.16$. The modes and 90% HPD limits for the marginals are 987 (390–1000) and 4.26 (2.17–9.78) for N_A and N_0, respectively. The Bayes factor in favor of decline is >9900. NICHOLS *et al.* (2001), analyzing the same data by different methods, suggested that they were incompatible with the known demographic history, with too much genetic variation still present. They proposed that this could be explained if the assumption of panmixia was invalid and that population structure would lead to the retention of more genetic variation than expected. It is not clear whether the results here contradict this conclusion. The value of N_0 should reflect the 1974 bottleneck, because the population subsequently grew after this period (*i.e.*, without mutation, the estimate of N_0 can be only the same as or lower than that if the sample had been taken immediately after the bottleneck). The 90% HPD limits exclude 2 individuals for N_0 and hence suggest that there was more than one breeding pair in 1974. However, (a) the exclusion is statistically borderline; (b) the demographic model is fitted over the whole data set, and thus a poor fit in one part may influence the estimate of N_0; (c) new mutations may lead to a tendency to overestimate population sizes; and (d) the model assumes an onset of population decline beginning in 1829 rather than in the 20th century, which will also lead to overestimation of N_0.

DISCUSSION

Estimation of change in population size: This article demonstrates that it is relatively straightforward to estimate change in population size using genetic samples taken over a time period, as also demonstrated by WILLIAMSON and SLATKIN (1999) and WANG (2001). Clearly a large sampling effort is needed to obtain accurate estimates. Although limitations on computer time preclude a thorough examination, this study suggests that there is an approximate equivalence of sample size, number of loci, and number of alleles toward the total sampling effort. The equivalence of number of loci and number of independent alleles on the variability of *F*-statistics was first noted by LEWONTIN and KRAKAUER (1973) and has been investigated using simulations by WAPLES (1989), who found that it is in general a very good approximation provided alleles are not close to fixation (this issue is also discussed in some detail in WANG 2001). The effect of sample size is not well established. For two samples, on the basis of an approximation obtained by POLLAK (1983), Waples suggests that when $x_1\tilde{n}/N_e \sim \sqrt{2}$, where $\tilde{n}$ is the harmonic mean of n_0 and n_1, there is a general equivalence among number of independent alleles at a locus, number of loci, sample size, and time between samples on the variance of estimates of N_e. For values $\ll \sqrt{2}$, change in sample size and time between samples has the greater effect, and for

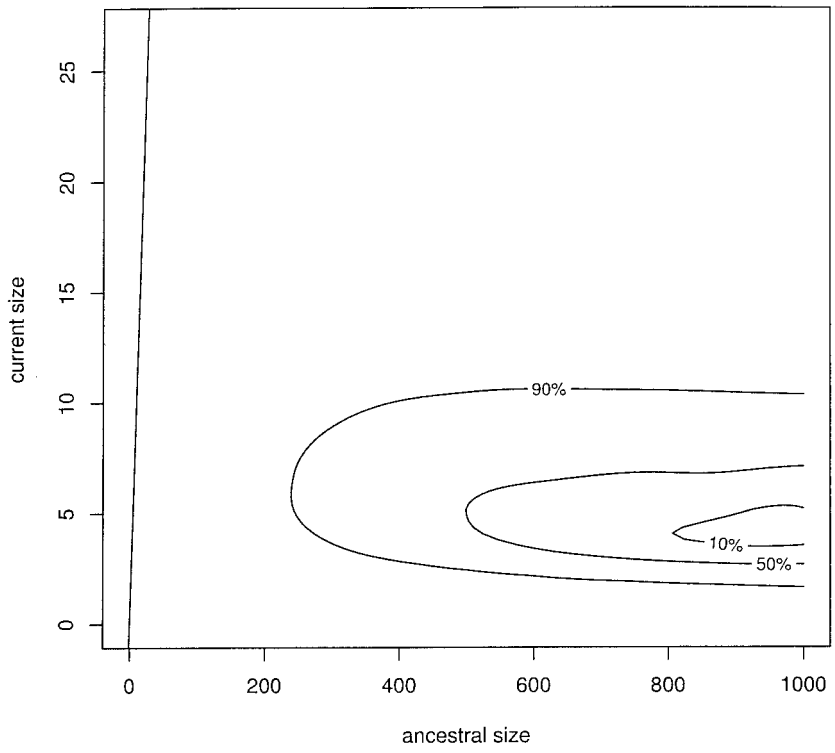

FIGURE 12.—Posterior distribution of N_0 and N_A for the Mauritius kestrel data.

values $\geqslant \sqrt{2}$ change in the total number of independent alleles, $(k - 1) \times$ (number of loci), has the greater effect. Obviously these results relate to the precision of moment-based estimates of N_e, and how well these results extend to the accuracy (as measured by the relative error) of likelihood-based estimates of changes in population size with time is unclear. An obvious variable that needs further investigation is the number and placement of sampling times.

The lack of precision in the estimation of ancestral and current population sizes can lead to problems when interpreting the results. It will often be the case that the likelihoods for either the current or the ancestral population sizes will asymptote for large values. In these cases, as demonstrated in the three examples, there is uncertainty in determining whether the population is actually changing in size because of the strong sensitivity on the prior assumptions. In the case of the Mauritius kestrel, even though the likelihood appears to reach an asymptote for N_A, it is reasonable to interpret the results as showing strong evidence of population decline for any reasonable prior. This is because there is little overlap between the marginal posterior distributions for N_A and N_0. For the other two cases, inferences are much less clear. The Bayes factor approach and the use of the HPD limits are both sensitive to the prior. For skewed posterior distributions the lower HPD limits are generally constrained by the mode and will therefore be less sensitive to the prior, but with rectangular priors there is the problem of the HPD limits becoming undefined when the likelihood surface becomes flat. Despite this problem (which can be avoided by using other prior distributions), it is probably preferable and more conservative to use the HPD limits to exclude the possibility of $N_A = N_0$ and reserve the use of the Bayes factor when it is important (*e.g.*, for management purposes) to distinguish between the possibility of growth or decline. Other approaches would be to use reversible-jump MCMC to compare different models or to directly estimate P(data) using the likelihood estimates from the MCMC run and use this to compare between models. This latter approach, while straightforward to perform, can be problematic because of the low accuracy in estimation of P(data) (PRITCHARD *et al.* 2000).

The compression of complex changes in population size into a simple model of exponential change in population size between the initial and final sampling periods may not give an accurate reflection of the complex demographic changes that might be involved. In this study, the assumption was made that $x_d = X$. It is straightforward, using the MCMC approach to also include X into the model at little extra computational cost. In general, the joint posterior distribution is complex, and the marginal posterior estimates of N_A and N_0 tend to be broader. This model awaits further investigation. An alternative to fitting a smooth demographic model is to look at the joint distribution of N_e's estimated for each sampling interval (as in WANG 2001). Again, there should be little computational cost to doing so, but this has not yet been studied. However, if there are many intervals, such an approach is unlikely to give a clear indication of the underlying broad changes in population size.

An assumption of the method is that no selection is operating. The use of temporal gene frequency data to detect selection by identifying discrepant loci was first

suggested by Lewontin and Krakauer (1973). This can be achieved by a relatively straightforward extension of the current model to use a hierarchical Bayesian approach as in Storz and Beaumont (2002). Here, each demographic parameter is allowed to vary between loci, and it is possible to test whether the posterior distribution of the variance includes zero with reasonable probability. This method of analysis is useful because it (a) effectively downweights discrepant loci and therefore gives more robust estimates and (b) allows discrepant loci to be identified.

Comparison with other temporal methods: The studies of Williamson and Slatkin (1999), Anderson *et al.* (2000), and Wang (2001) have estimated likelihoods from a Wright-Fisher model, and one question is whether the coalescent approach used here will give similar answers. This issue is also discussed in Berthier *et al.* (2002). Obviously, since the coalescent gives the limiting distribution of genealogies for the Wright-Fisher model, providing the population size is sufficiently large relative to the sample size there should be little difference between the two approaches. In the case of the data of Begon *et al.* very similar answers were obtained using the coalesent method to those obtained by Anderson *et al.* (2000). Using data simulated from a Wright-Fisher model Berthier *et al.* (2002) demonstrate that the median of point estimates obtained by the coalescent are generally very close to the true values for $N_e > \sim 20$. Of course, many species will not conform to a Wright-Fisher model anyway and therefore the question of which approach is more applicable may be difficult to judge.

The efficiency of the coalescent approach scales with the number of coalescences within the time interval, which will depend on sample size and X/N_e. Unlike the Wright-Fisher methods it does not scale with X and N_e independently (although this difference should disappear when X and N_e are large) and scales only weakly with the number of alleles or number of samples. Use of MCMC means that more complex demographic models can be handled with little extra computational burden. Potentially, the scaling of length of computation of Wright-Fisher methods with N_e is roughly quadratic for the biallelic case, and this increases very dramatically with increasing number of alleles. Generally, in terms of computational speed, it would appear that the coalescent method compares favorably with that of Anderson *et al.* (2000) or Williamson and Slatkin (1999). However, a weakness of the coalescent approach is its reliance (as also in Anderson *et al.* 2000) on Monte Carlo methods. By approximating the likelihood by the product of biallelic likelihoods and by using a number of computational approximations and improvements to the method of Williamson and Slatkin (1999), Wang's (2001) method appears to be substantially faster than either the coalescent method here or the other Wright-Fisher methods. For example, the method of Wang (2001) can be used to calculate a maximum-likelihood estimate for N_e and confidence limits, with the Begon *et al.* (1980) Drosophila data (either for the entire period or jointly for both periods) in a few seconds on a standard PC (J. Wang, personal communication). Although Wang (2001) demonstrated only relatively small discrepancies between the pseudo-likelihood method and the full-likelihood method in the three-allele case, it would clearly be useful to compare the different approaches when there are larger numbers of alleles.

The current method makes the twin assumptions of no mutation and no migration. The effect of migration has been recently analyzed by Wang and Whitlock (2003), who have extended the method of Wang (2001) to jointly infer immigration rate and N_e from temporal data. For populations at immigration-drift equilibrium, the effect of immigration, if not included in the model, is to produce underestimates in N_e for short intervals between samples and overestimates in N_e for longer intervals. The degree of underestimation for short sampling intervals is, however, slight when the populations are at equilibrium. The overestimation of N_e for larger sampling intervals also occurs with the method that includes immigration and appears unavoidable—essentially, in the limit of a long interval, one is estimating the metapopulation N_e.

This study follows a long line of articles from Krimbas and Tsakas (1971), which estimate N_e from changes in gene frequencies, ignoring mutation. The utility of these purely drift-based approaches lies in their relative simplicity of implementation and reasonable computational speed in comparison with models that include mutation. A further benefit is that they allow the same model to be applied to different classes of marker. When used with markers that have a low mutation rate, such as single-nucleotide polymorphisms, these drift-based models may be particularly useful in the analysis of human demographic history. The effect of ignoring mutation on population size estimates, and, in particular, to what extent it will lead to apparent changes in population size, awaits further investigation. This assumption is probably reasonable for studies conducted on an "ecological" timescale, even for microsatellite markers, which tend to have a high mutation rate. However, with the increasing ability to extract DNA from ancient samples, it is clearly desirable to take mutation into account. The method described by Drummond *et al.* (2002), which is suitable for sequence data, is an important step in this direction. All of the drift-based models could be incorporated in some general genealogical MCMC scheme, which would then naturally provide the prior for the baseline gene frequencies from the mutation model. For microsatellites, one route to incorporating the effects of mutations is to extend the MCMC model of Beaumont (1999) to allow for samples to be taken at different times. However, given the disadvantages of the MCMC approach it might be better to simply extend the GIMH method described in the current study to incorporate mutations.

Computational methods: The study described here uses a mixture of importance sampling and MCMC to obtain posterior distributions for demographic parameters, and it follows the basic methodology of O'Ryan *et al.* (1998), Ciofi *et al.* (1999), Chikhi *et al.* (2001), and Berthier *et al.* (2002), with one significant modification.

A minor additional modification is that, rather than integrating out the unknown population gene frequencies **x** using MCMC, as done in the earlier studies, the integration is performed analytically using the multinomial Dirichlet. Trial simulations suggest that this leads to a small improvement in efficiency.

The most important modification arises from the demonstration that GIMH can be used with IS sizes greater than one. This study suggests that GIMH should always be used in preference to MCWM, with or without bias correction. A particular problem with the latter two approaches is that there is no intrinsic way of determining (other than by trial simulations) whether the number of importance samples used for the determination of the likelihood is large enough. If the sample size is too small the posterior distribution may not be estimated correctly. By contrast, with GIMH, if the importance sample size is too small, the MCMC chain obviously does not mix well.

In the initial study that used MCWM (O'Ryan *et al.* 1998), which modeled the divergence of different populations through drift, trial simulations based on the data sample size involved in that study indicated that an IS size of 500 gave accurate estimates of the posterior distribution. Subsequent articles have tended to use this value for simulations. By comparison, for the model considered here, the result displayed in Figure 3 suggests that, for the data used in these simulations, 500 is the lowest possible for accurate estimation of the posterior distribution for N_e using MCWM. A feature of the model described in this article is that the importance sampling variance is higher than that in the other models. A large variance is associated with the simulation of the genealogical history when a number of different samples are taken at different times. This is because the importance-sampling function proceeds sequentially from the most recent to the oldest sample and does not take into account the frequencies of older samples. Thus Equation A5 often has low probability for any given realization of the importance sampling process. A similar phenomenon also occurs in the models of diverging populations (O'Ryan *et al.* 1998; Ciofi *et al.* 1999) when the number of populations is large and is a general problem of using current methods of importance sampling, even with the modifications of Stephens and Donnelly (2000), when diverging populations are modeled. Undoubtedly a different importance-sampling function, based on different heuristics, can circumvent this problem.

The tendency for GIMH to produce sticky simulated chains when the importance sample sizes are too low is probably exacerbated by the current updating procedure whereby new sets of genealogies are simulated each time the parameters are updated. A potentially large improvement would be to update the demographic parameters independently of the genealogical history. This would require modification of the importance weights as discussed after the presentation of Equation 6, either using Tavaré's (1984) Equation 6.1 for each interval between samples or using the densities for the time intervals. Overall, however, even without these potential improvements, the GIMH method offers a straightforward way to carry out a Bayesian analysis with a genealogical model, on the basis of independent sampling of genealogies.

I am grateful to David Balding, Claire Calmet, Lounès Chikhi, Jean-Marie Cornuet, Kevin Dawson, Richard Nichols, Geoff Nicholls, Jinliang Wang, and two anonymous reviewers for their helpful comments on previous versions of the manuscript. This work was supported in part by Natural Environment Research Council grant NER/B/S/2000/00669 awarded to Ken Norris, M.A.B., and Mike Bruford.

LITERATURE CITED

Anderson, E. C., E. G. Williamson and E. A. Thompson, 2000 Monte Carlo evaluation of the likelihood for N_e from temporally spaced samples. Genetics **156:** 2109–2118.

Austerlitz, F., and E. Heyer, 1998 Social transmission of reproductive behavior increases frequency of inherited disorders in a young-expanding population. Proc. Natl. Acad. Sci. USA **95:** 15140–15144.

Beaumont, M. A., 1999 Detecting population expansion and decline using microsatellites. Genetics **153:** 2013–2029.

Beaumont, M. A., 2001 Conservation genetics, pp. 779–812 in *The Handbook of Statistical Genetics*, edited by D. J. Balding, M. Bishop and C. Cannings. Wiley, Chichester, UK.

Beerli, P., and J. Felsenstein, 2001 Maximum likelihood estimation of a migration matrix and effective population sizes in *n* subpopulations by using a coalescent approach. Proc. Natl. Acad. Sci. USA **98:** 4563–4568.

Begon, M., C. B. Krimbas and M. Loukas, 1980 The genetics of *Drosophila subobscura* populations. XV. Effective size of a natural population estimated by three independent methods. Heredity **45:** 335–350.

Berthier, P., M. A. Beaumont, J.-M. Cornuet and G. Luikart, 2002 Likelihood-based estimation of the effective population size using temporal changes in allele frequencies: a genealogical approach. Genetics **160:** 741–751.

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

Chikhi, L., M. W. Bruford and M. A. Beaumont, 2001 Estimation of admixture proportions: a likelihood-based approach using Markov chain Monte Carlo. Genetics **158:** 1347–1362.

Ciofi, C., M. A. Beaumont, I. R. Swingland and M. W. Bruford, 1999 Genetic divergence and units for conservation in the Komodo dragon *Varanus komodoensis*. Proc. R. Soc. Lond. Ser. B **266:** 2269–2274.

Donnelly, P., and S. Tavaré, 1995 Coalescents and genealogical structure under neutrality. Annu. Rev. Genet. **29:** 410–421.

Drummond, A. J., G. K. Nicholls, A. G. Rodrigo and W. Solomon, 2002 Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics **161:** 1307–1320.

Fearnhead, P., and P. Donnelly, 2001 Estimating recombination rates from population genetic data. Genetics **159:** 1299–1318.

Frankham, R., 1995 Conservation genetics. Annu. Rev. Genet. **29:** 305–327.

Fu, Y. X., 2001 Estimating mutation rate and generation time from longitudinal samples of DNA sequences. Mol. Biol. Evol. **18:** 620–626.

Gelman, A., J. B. Carlin, H. S. Stern and D. B. Rubin, 1995 *Bayesian Data Analysis*. Chapman & Hall, London.

Green, P. J., 1995 Reversible jump Markov chain Monte Carlo com-

putation and Bayesian model determination. Biometrika **82:** 711–732.

Griffiths, R. C., and S. Tavaré, 1994a Simulating probability distributions in the coalescent. Theor. Popul. Biol. **46:** 131–159.

Griffiths, R. C., and S. Tavaré, 1994b Sampling theory for neutral alleles in a varying environment. Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci. **344:** 403–410.

Griffiths, R. C., and S. Tavaré, 1994c Ancestral inference in population genetics. Stat. Sci. **9:** 307–319.

Groombridge, J. J., C. G. Jones, M. W. Bruford and R. A. Nichols, 2000 'Ghost' alleles of the Mauritius kestrel. Nature **403:** 616.

Hill, W. G., 1981 Estimation of effective population size from data on linkage disequilibrium. Genet. Res. **38:** 209–216.

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.

Kuhner, M. K., J. Yamato and J. Felsenstein, 1995 Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics **140:** 1421–1430.

Lande, R., 1998 Anthropogenic, ecological and genetic factors in extinction and conservation. Res. Popul. Ecol. **40:** 259–269.

Langley, C. H., D. B. Smith and F. M. Johnson, 1978 Analysis of linkage disequilibrium between allozyme loci in natural populations of *Drosophila melanogaster*. Genet. Res. **32:** 215–229.

Laurie-Ahlberg, C. C., and B. S. Weir, 1979 Allozyme variation and linkage disequilibrium in some laboratory populations of *Drosophila melanogaster*. Genetics **92:** 1295–1314.

Lewontin, R. C., and J. Krakauer, 1973 Distribution of gene frequency as a test of the theory of selective neutrality of polymorphisms. Genetics **74:** 175–195.

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.

Marjoram, P., and P. Donnelly, 1997 Human demography and the time since mitochondrial Eve, pp. 107–131 in *Progress in Population Genetics and Human Evolution*, edited by P. Donnelly and S. Tavaré. Springer-Verlag, New York.

Miller, L. M., and A. R. Kapuscinski, 1997 Historical analysis of genetic variation reveals low effective population size in a northern pike (*Esox lucius*) population. Genetics **147:** 1249–1258.

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

Nichols, R. A., M. W. Bruford and J. J. Groombridge, 2001 Sustaining genetic variation in a small population: evidence from the Mauritius kestrel. Mol. Ecol. **10:** 593–602.

Nielsen, R., and J. Wakeley, 2001 Distinguishing migration from isolation: a Markov chain Monte Carlo approach. Genetics **158:** 885–896.

Nielsen, R., J. L. Mountain, J. P. Huelsenbeck and M. Slatkin, 1998 Maximum-likelihood estimation of population divergence times and population phylogeny in models without mutation. Evolution **52:** 669–677.

Nordborg, M., 1997 Structured coalescent processes on different time scales. Genetics **146:** 1501–1514.

O'Neill, P. D., D. J. Balding, N. G. Becker, M. Eerola and D. Mollison, 2000 Analyses of infectious disease data from household outbreaks by Markov chain Monte Carlo methods. Appl. Stat. **49:** 517–542.

O'Ryan, C., E. H. Harley, M. W. Bruford, M. Beaumont, R. K. Wayne *et al.*, 1998 Microsatellite analysis of genetic diversity in fragmented South African buffalo populations. Anim. Conserv. **1:** 85–94.

Pollak, E., 1983 A new method for estimating the effective population size from allele frequency changes. Genetics **104:** 531–548.

Pritchard, J. K., M. Stephens and P. Donnelly, 2000 Inference of population structure using multilocus genotype data. Genetics **155:** 945–959.

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.

Rice, J. A., 1995 *Mathematical Statistics and Data Analysis*. Duxbury Press, Belmont, CA.

Rodrigo, A. G., E. G. Shpaer, E. L. Delwart, A. K. N. Iversen, M. V. Gallo *et al.*, 1999 Coalescent estimates of HIV-1 generation time in vivo. Proc. Natl. Acad. Sci. USA **96:** 2187–2191.

Saccheri, I., M. Kuussaari, M. Kankare, P. Vikman, W. Fortelius *et al.*, 1998 Inbreeding and extinction in a butterfly metapopulation. Nature **392:** 491–494.

Saccheri, I. J., I. J. Wilson, R. A. Nichols, M. W. Bruford and P. M. Brakefield, 1999 Inbreeding of bottlenecked butterfly populations: estimation using the likelihood of changes in marker allele frequencies. Genetics **151:** 1053–1063.

Slatkin, M., 1996 Gene genealogies within mutant allelic classes. Genetics **145:** 579–587.

Stephens, M., 2001 Inference under the coalescent, pp. 213–238 in *The Handbook of Statistical Genetics*, edited by D. J. Balding, M. Bishop and C. Cannings. Wiley, Chichester, UK.

Stephens, M., and P. Donnelly, 2000 Inference in molecular population genetics (with discussion). J. R. Stat. Soc. B **62:** 605–655.

Storz, J. F., and M. A. Beaumont, 2002 Testing for genetic evidence of population expansion and contraction: an empirical analysis of microsatellite DNA variation using a hierarchical Bayesian model. Evolution **56:** 154–166.

Tavaré, S., 1984 Lines-of-descent and genealogical processes, and their application in population genetics models. Theor. Popul. Biol. **26:** 119–164.

Tierney, L., 1996 Introduction to general state-space Markov chain theory, pp. 59–74 in *Markov Chain Monte Carlo in Practice*, edited by W. R. Gilks, S. Richardson and D. J. Spiegelhalter. Chapman & Hall, London.

Wakeley, J., 1999 Nonequilibrium migration in human history. Genetics **153:** 1863–1871.

Wakeley, J., and N. Aliacar, 2001 Gene genealogies in a metapopulation. Genetics **159:** 893–905.

Wang, J., 2001 A pseudo-likelihood method for estimating effective population size from temporally spaced samples. Genet. Res. **78:** 243–257.

Wang, J., and M. C. Whitlock, 2003 Estimating effective population size and migration rates from genetic samples over space and time. Genetics **163:** 429–446.

Waples, R. S., 1989 A generalized approach for estimating effective population size from temporal changes in allele frequency. Genetics **121:** 379–392.

Williamson, E. G., and M. Slatkin, 1999 Using maximum likelihood to estimate population size from temporal changes in allele frequencies. Genetics **152:** 755–761.

Wilson, I. J., and D. J. Balding, 1998 Genealogical inference from microsatellite data. Genetics **150:** 499–510.

Communicating editor: G. Churchill

APPENDIX

Correcting likelihood ratio for bias: O'Neill *et al.* (2000) suggest using the estimator $R^* = \hat{R}^2/\tilde{E}[\hat{R}]$, where $\tilde{E}[\hat{R}]$ is an estimate of the expected value of the ratio, to correct for the bias in $\hat{R}$. For the genealogical model considered here, using the standard method for estimating the expected value of ratios,

$$\tilde{E}\left[\frac{\tilde{p}(D|\Phi_{i+1})}{\tilde{p}(D|\Phi_i)}\right] = \frac{\tilde{p}(D|\Phi_{i+1})}{\tilde{p}(D|\Phi_i)}\left(1 + \frac{\mathrm{SE}^2(\tilde{p}(D|\Phi_i))}{\tilde{p}^2(D|\Phi_i)}\right) \qquad \text{(A1)}$$

(see, *e.g.*, Rice 1995). When multilocus data are used, the likelihoods are multiplied over loci. In this case the standard error $\mathrm{SE}[\tilde{p}(D|\Phi)]$ is estimated recursively using standard methods for the variance of a product (see, *e.g.*, Rice 1995),

$$\underset{(1\ldots j)}{\mathrm{SE}^2}[\tilde{p}(D|\Phi)] = \underset{(1\ldots j-1)}{\mathrm{SE}^2}[\tilde{p}(D|\Phi)]\underset{j}{\mathrm{SE}^2}[\tilde{p}(D|\Phi)] + \underset{(1\ldots j-1)}{\tilde{p}^2}(D|\Phi)\ \underset{j}{\mathrm{SE}^2}[\tilde{p}(D|\Phi)]$$

$$+ \tilde{p}_j^2(D|\Phi) \underset{(1 \ldots j-1)}{\mathrm{SE}^2} [\tilde{p}(D|\Phi)]$$

$$\underset{(1 \ldots j)}{\tilde{p}}(D|\Phi) = \tilde{p}_j(D|\Phi) \underset{(1 \ldots j-1)}{\tilde{p}}(D|\Phi), \quad \text{(A2)}$$

where $j = 1 \ldots, k$ and $\mathrm{SE}[\tilde{p}(D|\Phi)] = \mathrm{SE}_{(1 \ldots k)}[\tilde{p}(D|\Phi)]$.

Proof that the "grouped"-independence Metropolis-Hastings sampler gives the correct marginal density for demographic parameters: The aim here is to show that implementation of GIMH samples demographic parameters, Φ, from the correct posterior density for any $n \geq 1$ sampled genealogical histories. To ease the notation I assume uniform improper priors on Φ and thus I wish to estimate the posterior distribution $p(\Phi|D)$, which is proportional to the likelihood $P(D|\Phi) = \int p(D, G|\Phi)\, dG$, where the integration is over all genealogical histories G that could have given rise to the data. Also, differing slightly from the notation in Equation 3, prime ($'$) is used to denote trial updates.

In the case of GIMH, the Metropolis-Hastings ratio is

$$\frac{\sum_{j=1}^{h} (p(D, G_j'|\Phi')/q(D, G_j'|\Phi'))}{\sum_{j=1}^{h} (p(D, G_j|\Phi)/q(D, G_j|\Phi))} \frac{p(\Phi|\Phi')}{p(\Phi'|\Phi)},$$

which can be simply rewritten as

$$\frac{\sum_j (p(D, G_j'|\Phi') \prod_{i \neq j} q(D, G_i'|\Phi'))}{\sum_j (p(D, G_j|\Phi) \prod_{i \neq j} q(D, G_i|\Phi))} \frac{\prod_i q(D, G_i|\Phi)}{\prod_i q(D, G_i'|\Phi')} \frac{p(\Phi|\Phi')}{p(\Phi'|\Phi)}.$$

If we regard the sampling procedure as ordered (*i.e.*, the sampled genealogies occupy "slots" $j = 1 \ldots, h$), then the two right-hand terms are the correct Hastings terms for the sampling process. Given that this is the case, it then follows that the target marginal density must be given by the numerator and denominator of the left-hand term. Looking at individual terms in the sum, for any jth position the density is proportional to

$$\int \ldots \int p(D, G_j|\Phi) \prod_{i \neq j} q(D, G_i|\Phi)\, dG_1 \ldots dG_h,$$

where the integration is over all genealogical histories in slots $1 \ldots, h$. This evaluates to

$$p(D|\Phi) \int \ldots \int \prod_{i \neq j} q(D, G_i|\Phi)\, dG_1 \ldots dG_h \quad \text{(excluding } dG_j),$$

which is

$$P(D|\Phi) \quad \text{since} \int q(D, G|\Phi)\, dG = 1, \quad \text{by construction.}$$

Since this proportionality is true for all terms, it will also be true for the sum. The key point is that if we concentrate on the jth slot, marginal to what is happening in the other slots, the MCMC is sampling from the joint distribution $p(\Phi, G|D)$, and this follows because the importance sampling function integrates to 1 over all genealogical histories, irrespective of Φ. This result is quite general and GIMH could be used to perform genealogical MCMC with an independence sampler on the full range of problems for which MCMC and importance sampling have previously been applied.

Demographic model: A model of exponential growth is assumed, where

$$N_x = N_0 e^{-bx},$$

x is the time measured in units of generations backward from the current time, b is the growth rate, and N_0 is the current population size. We assume throughout the article that the organisms are diploid. At time X in the past the population is assumed to have been at an ancestral size N_A. From this it is possible to reparameterize to give

$$N_x = N_0 r^{-x/X},$$

where $r = N_0/N_A$. In the case $N_0 = N_A$, the effective population is referred to as N_e. The harmonic mean population size over the interval $[x_{i-1}, x_i]$ is given by

$$\tilde{N}_i = \frac{x_i - x_{i-1}}{\int_{x_{i-1}}^{x_i} (1/N_0 r^{-y/X})\, dy} = \frac{(x_i - x_{i-1}) N_0 \log(r)\, r^{-(x_{i-1})/X}}{X(r^{(x_i - x_{i-1})/X} - 1)}, \quad \text{(A3)}$$

and when $x_{i-1} = 0$, and $x_i = X$,

$$\tilde{N} = \frac{N_0 \log(r)}{r - 1}.$$

Derivation of the likelihood: The derivation of the likelihood (5) follows that in Berthier *et al.* (2002) and is expanded to consider more than one interval between samples.

The probability of obtaining c_i coalescences in any interval $[x_{i-1}, x_i]$, $p(c_i|(x_i - x_{i-1})/2\tilde{N}_i)$, is given by Tavaré (1984, Equation 6.1). Although this was derived on the assumption of a stable population of size N, it is also applicable to populations whose size is changing because the distribution of waiting times for coalescence in this case is the same as that for a stable population once each infinitesimal of time is expressed as the reciprocal of the population size at that point (Griffiths and Tavaré 1994b; Marjoram and Donnelly 1997), and hence we need only replace N by $\tilde{N}_i$ from (A3) in the previous section.

Given $\mathbf{f}_{i+1}$, the probability of obtaining the allele frequency count among both the base lineages and the sample at the ith sample point (without regard to how they are partitioned) is given by

$$p(\mathbf{f}_i + \mathbf{a}_i|\mathbf{f}_{i+1}, c_{i+1}) = \frac{\prod_{j=1}^{k} \binom{f_{ij} + a_{ij} - 1}{f_{(i+1)j} - 1}}{\binom{h_i + n_i - 1}{h_{i+1} - 1}}, \quad 0 \leq i \leq d - 1 \quad \text{(A4)}$$

(Slatkin 1996; Nielsen *et al.* 1998; O'Ryan *et al.* 1998; Saccheri *et al.* 1999).

Given both sets of lineages at the ith sample point, the probability of partitioning the frequency count between the base lineages and the sample is given by the hypergeometric distribution

$$p(\mathbf{a}_i, \mathbf{f}_i|\mathbf{a}_i + \mathbf{f}_i) = \frac{n_i! h_i!}{(n_i + h_i)!} \prod_{j=1}^{k} \frac{(a_{ij} + f_{ij})!}{a_{ij}! f_{ij}!}, \quad 0 < i \leq d. \tag{A5}$$

At the final sample point, d, the sample and base lineages are taken to be a multinomial random draw from the population gene frequency distribution $\mathbf{x}$. In general, however, $\mathbf{x}$ is unknown, and it is preferable to assume that the sample has a marginal distribution over all possible values of $\mathbf{x}$, assuming a Dirichlet prior. This is given by the multinomial Dirichlet (obtained by integrating the product of the multinomial and the Dirichlet prior over $\mathbf{x}$),

$$p(\mathbf{a}_d + \mathbf{f}_d) = \frac{\Gamma(n_d + h_d)\Gamma(bk)}{\Gamma(n_d + h_d + bk)} \prod_{j=1}^{k} \frac{\Gamma(a_{dj} + f_{dj} + b)}{\Gamma(a_{dj} + f_{dj} + 1)\Gamma(b)}, \tag{A6}$$

where b is taken here to be 1 [equivalent to assuming a Dirichlet prior of $D(1, \ldots, 1)$]. In earlier articles using this methodology (O'Ryan *et al.* 1998; Ciofi *et al.* 1999; Chikhi *et al.* 2001; Berthier *et al.* 2002), the multinomial was used and then the integration was performed by Metropolis-Hastings simulation.

Importance sampling: Extending the approach of Berthier *et al.* (2002) to multiple samples, Equation A4, above, can be rewritten as

$$p(\mathbf{f}_i + \mathbf{a}_i|\mathbf{f}_{i+1}) = \sum_{\mathbf{g}_i} \left[p(\mathbf{f}_i + \mathbf{a}_i|\mathbf{g}_{i0}) \prod_{e=0}^{c_{i+1}-1} p(\mathbf{g}_{ie}|\mathbf{g}_{i(e+1)}) \right],$$

where $\mathbf{g}_{ie}$ gives the allele frequency count among lineages at the eth coalescent event after the ith sample point, and $\mathbf{g}_{i(c_{i+1})} = \mathbf{f}_{i+1}$. The term $p(\mathbf{f}_i + \mathbf{a}_i|\mathbf{g}_{i0}) = 1$ when $\mathbf{f}_i + \mathbf{a}_i = \mathbf{g}_{i0}$, and is 0 otherwise. Looking forward in time, whenever a coalescent event occurs a lineage is chosen at random and duplicated. Thus if the lineage is in the jth allelic class

$$p(\mathbf{g}_{ie}|\mathbf{g}_{i(e+1)}) = \frac{g_{i(e+1)j}}{s_{i(e+1)}} = \frac{g_{iej} - 1}{s_{ie} - 1},$$

where $s_{i(e+1)} = \sum_{l=1}^{k} g_{i(e+1)l}$.

To estimate the likelihood, the c_i are sampled using standard Monte Carlo coalescent simulations, given in the next section, and the genealogical history is sampled backward from the data using the method of Griffiths and Tavaré. The jth allelic class is chosen with probability

$$q(\mathbf{g}_{i(e+1)}|\mathbf{g}_{ie}) = \frac{g_{iej} - 1}{s_{ie} - m_i},$$

where m_i ($\leq k$) is the number of allelic classes in which at least one representative is in $\mathbf{f}_i + \mathbf{a}_i$. Individual terms of the importance ratio are then

$$w_{i(e+1)} = \frac{p(\mathbf{g}_{ie}|\mathbf{g}_{i(e+1)})}{q(\mathbf{g}_{i(e+1)}|\mathbf{g}_{ie})} = \frac{s_{ie} - m_i}{s_{ie} - 1}$$

(O'Ryan *et al.* 1998). Note that the importance ratio can be zero if genealogical histories are sampled with fewer lineages than alleles in the data.

The number of coalescent events occurring before the ith data sample, c_i, are sampled by simulating coalescence times using the model described in Beaumont (1999). The details are given in the next section. Thus, when the time of a coalescent event is generated that succeeds a sampling time, x_i, the time is set to x_i, the number of coalescent events between x_{i-1} and x_i is recorded as c_i, the data lineages $\mathbf{a}_i$ are added to the current lineages $\mathbf{f}_i$, and the partitioning probability (A5) is calculated. At the final data sample, the probability of the allele frequency count $\mathbf{a}_i + \mathbf{f}_i$ is given by (A6).

Simulation of coalescent times: The method for simulating coalescent times described here is similar to that of Marjoram and Donnelly (1997). Define $t_f = X/(2N_0)$, $t_i = x_i/(2N_0)$, and $r = N_0/N_A$. The uniform random variable U is simulated from (0, 1). Define $t' = -2\log(U)/(n_l(n_l - 1))$. To avoid the singularity at $r = 1$, if $|r - 1| < 10^{-5}$, $t_{i+1} \approx t' + t_i$. Otherwise, if $t_i \leq t_f$ and $t' \leq (r - r^{t_i/t_f})t_f/\log(r)$,

$$t_{i+1} = \log(t'\log(r)/t_f + r^{t_i/t_f})t_f/\log(r).$$

If $t_i \leq t_f$ and $t' > (r - r^{t_i/t_f})t_f/\log(r)$,

$$t_{i+1} = (t' - (r - r^{t_i/t_f})t_f/\log(r))/r + t_f.$$

Otherwise

$$t_{i+1} = t'/r + t_i.$$