

Estimation of population growth or decline in  
genetically monitored populations.

Mark A. Beaumont,

School of Animal and Microbial Sciences,

Whiteknights,

PO Box 228,

Reading RG6 6AJ,

UK

July 26, 2002

**Running title:** Measuring change in effective population size.

**Key words:**  $N_e$ , coalescent, likelihood, temporal method, MCMC, importance sampling.

**Corresponding author:**

Mark Beaumont,  
School of Animal and Microbial Sciences,  
Whiteknights,  
PO Box 228,  
Reading RG6 6AJ,  
UK  
Tel: +44 118 987 5123 extension 7707  
Fax: +44 118 931 0180  
E-mail: [m.a.beaumont@reading.ac.uk](mailto:m.a.beaumont@reading.ac.uk)

# Abstract

This paper introduces a new general method for genealogical inference that combines importance sampling (IS) with Markov chain Monte Carlo (MCMC). It is then possible to more easily utilise 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 using samples taken at a number of different times. 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.

# Introduction

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 main computer-intensive approaches to statistical inference — Markov chain Monte Carlo (MCMC) and importance sampling (IS). Both methods involve sampling possible genealogical histories of the sample, but MCMC does this through small modifications of an initial ‘guess’, whereas importance sampling, as usually implemented, yields completely independent genealogical histories. In general, MCMC methods tend naturally to lead to Bayesian inference or the use of integrated likelihood (*e.g.* WILSON and BALDING, 1998; BEAUMONT, 1999; NIELSEN and WAKELEY, 2001), whereas importance sampling tends to lead to more classical likelihood-based inference (BEERLI and FELSENSTEIN, 2001; ANDERSON *et al.*, 2000). MCMC yields samples from the posterior distribution and the chief problem is in obtaining likelihood surfaces or posterior densities via density estimation and related methods. Importance sampling, on the other hand, yields estimates of likelihood surfaces, and further complex procedures are necessary to carry out the necessary integration for obtaining posterior densities. 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 results from standard importance sampling, it would seem that MCMC offers the greatest flexibility. However MCMC methods have two main drawbacks: (a) they are generally more difficult to program than IS methods; (b) because they involve small modifications they can move quite slowly through the space of possible genealogical histories, making them potentially inefficient. This paper introduces a method for overcoming these two advantages and applies it to a specific problem, the estimation of effective population size,  $N_e$  from temporally spaced genetic samples.

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. Estimation of  $N_e$  is problematic. There are three general approaches. One way is to estimate it non-genetically 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 time scale 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, 2001). Alternatively, genotypic disequilibria in single samples can be used to estimate  $N_e$ . This can be achieved either by measuring departures from Hardy-Weinberg equilibrium (PUDOVKIN *et al.*, 1996; LUIKART and CORNUET 1999), or by measuring departures from linkage disequilibrium (LANGLEY *et al.*, 1978; LAURIE-AHLBERG *et al.*, 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 differ-

ence 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), 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 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 analysed. Their method is only practicable 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. Recently 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 that of the other three methods in that likelihoods are estimated from a coalescent model in which two samples are analysed. Since only two samples are analysed it is not possible to make inferences about changes in population size.

This study 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 generalised to consider any number of samples in a temporal sequence rather than just the two previously considered. Thirdly, 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 importance sampling

**Background and motivation:** The general motivation behind the computer intensive methods discussed in the Introduction 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,  $\Phi$ . There are a number of different representations of the genealogical history (see STEPHENS and DONNELLY, 2000), and in this paper I will 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 gives 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

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

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, 1994; STEPHENS and DONNELLY, 2000). In importance sampling, the equation is rewritten as

$$p(D|\Phi) = \int p(D|G)(p(G|\Phi)/q(G|\Phi))q(G|\Phi)dG$$

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

$$\tilde{p}(D|\Phi) = 1/h \sum_{j=0}^h p(D|G_j)p(G_j|\Phi)/q(G_j|\Phi). \quad (2)$$

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  $\Phi$  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 (*e.g.* the use of profile likelihoods) 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 wish to 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 multi-parameter 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 (*e.g.* WILSON and BALDING, 1998; BEAUMONT, 1999). Although the potential to use 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). (An extension of the bridge-sampling method used in FEARNHEAD and DONNELLY, 2001, may be a possibility.)

In order 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 probabil-

ity,  $p(G_{i+1}, \Phi_{i+1}|G_i, \Phi_i)$ , of obtaining  $G_{i+1}$  and  $\Phi_{i+1}$ , conditional on being at  $G_i, \Phi_i$ , and the reverse. Then accept  $G_{i+1}$  and  $\Phi_{i+1}$ , with probability

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

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. This 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 realised 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 simulation 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  $\Phi_{i+1}$  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  $\Phi$  would have to be explored by the MCMC simulation.

**Current methods that use MCMC with IS in population genetics:** Hitherto it has been easier to run the MCMC using  $p(G, D|\Phi)$ , but, as discussed in the Introduction, 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 generally very straightforward to implement importance sampling 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 papers (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’NEIL *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 the method of GRIFFITHS and TAVARÉ (1994). Note that in normal MCMC, because the denominator  $p(D|\Phi_i)$  in the likelihood ratio is known without error it is not re-evaluated each time that  $R$  is 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 (4)), or to re-use the earlier estimate. Intuitively it seems reasonable, though more time-consuming, to re-evaluate it each time so that the estimates of  $\hat{R}$  are independent of each other. Furthermore, the results in O’NEIL *et al.* (2000) concerning bias correction (discussed below) require independence of the estimates. The re-evaluation approach has been taken in all the genealogical models that have used the method and also by O’NEIL *et al.* (2000). Updates are only required for  $\Phi$ , and not for  $G$  as in the MCMC methods of BALDING and WILSON (1998) and BEAUMONT (1999). This general method, where the MCMC uses an approximation to the likelihood, will be abbreviated here as Monte Carlo within Metropolis, MCWM, following the terminology of O’NEIL *et al.* (2000).

**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 papers.

O’NEIL *et al.* (2000) have carried out an analogous procedure where the likelihoods are estimated by an 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 will 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 (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  $i$ th 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). 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 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 is *kept* like any other param-

eter rather than *resampled* at each evaluation of (3). The sampled genealogical history is treated as a parameter on an equal footing with  $\Phi$ , 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 re-evaluated 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 will be 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 only depend on  $\Phi$  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 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  $j$ th 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 Metropolis-Hastings sampler because we are using a group of sampled genealogies rather than one, and will be abbreviated to GIMH to distinguish it from the normal independence Metropolis-Hastings and from MCWM, which involves re-evaluation 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 paper are carried out using this approach. Throughout the following the grouped sampler will be abbreviated to

## Inference in the temporal method based on a coalescent model with samples taken at many time points

**The genealogical model** The data are assumed to be sampled at different times, given by the sequence  $\mathcal{X} = (x_0, x_1, \dots, x_d)$ . 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 further from the most recent sample. Corresponding to each time point there is a sequence of sample sizes (number of chromosomes)  $\mathcal{N} = (n_0, n_1, \dots, n_d)$ , where lineages are added to the genealogy. The sequence of frequency counts of the different allelic types in each sample is given by  $\mathcal{A} = (\mathbf{a}_0, \mathbf{a}_1, \dots, \mathbf{a}_d)$ , where the vectors are of length  $k$ , the total number of different allelic types observed in the data. For times greater than  $x_0$ , at the time each set of lineages is added there are a number of lineages present with descendants in earlier samples, lower down the genealogy. The allele frequency counts among these base lineages are denoted here as the random variable  $\mathcal{F} = (\mathbf{f}_0, \dots, \mathbf{f}_d)$ , where, to ease the notation below,  $\mathbf{f}_0$  is defined to be 0. The number

of these lineages, also a random variable,  $\mathcal{H} = (h_1, \dots, h_d)$ , depends on the number of coalescences that occur in the intervals between sampling points. These are given by the sequence  $\mathcal{C} = (c_1, \dots, c_d)$ . Thus, at the  $i$ th 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 summarised in Figure 1.

[Figure 1 about here.]

The likelihood can be obtained as a straightforward extension of the two sample case in Berthier *et al.* (2002) and is given by

$$p(\mathcal{A}|\mathcal{X}, N_0, N_A, X) = \sum_{\mathcal{C}, \mathcal{F}} \left[ p(\mathbf{a}_d + \mathbf{f}_d) \prod_{i=0}^{d-1} p(\mathbf{f}_i + \mathbf{a}_i | \mathbf{f}_{i+1}, c_{i+1}) p(\mathbf{a}_{i+1}, \mathbf{f}_{i+1} | \mathbf{a}_{i+1} + \mathbf{f}_{i+1}) \right. \\ \left. p\left(c_{i+1} \mid \frac{x_{i+1} - x_i}{2\tilde{N}_{i+1}}\right) \right] \quad (5)$$

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;

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

(see 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 Appendix for details), and we obtain

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

Thus for the  $\kappa$ th 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  $c_i^\kappa$  coalescences are simulated using the coalescent model (see 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. There are two ways this can be done. 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 (9), 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 also into the grouped independence Metropolis Hastings (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 median centred 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:** In order to compare the accuracy of the 3 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 6 samples each of size 20 chromosomes were taken at generations 0, 4, 8, 12, 16, 20. The data set consisted of 10 loci each with 5 alleles in the population (although, due to sampling, some data sets had fewer than 5 alleles). The population frequencies were simulated from a uniform Dirichlet distribution, according to the assumption of the model.

The data set was then analysed by 4 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 (8). The approximate likelihood

surface was normalized to have unit volume, and the 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 size used in the evaluation of (4) was 5000, 1000, 500, 100, 50, 10, 5, 2, 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  $\text{SE}[\tilde{p}(D|\Phi)]$  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, and the standard deviation of the trial parameter updates was varied between simulations to achieve satisfactory convergence, judged by eye from output traces.

Estimated densities using the 4 approaches are shown in Figure 2. It can be seen that the standard errors for the

[Figure 2 about here.]

IS method are still large, even with 40,000 points. However, the posterior distributions estimated by MCWM with and without bias correct 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 9 times larger than shown in the figure). The distribution for GIMH with an IS size of just 10 per MCMC update (evaluation of (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.

[Figure 3 about here.]

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 around 500 is the minimum required for accurate estimation, whereas around 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 7 independent simulations of  $10^8$  MCMC updates, thinned every 10000 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 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 around 10.

[Figure 4 about here.]

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 size of the trial parameter updates, and for the remaining

analyses described in this paper, 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  $N_0$ :** In order to illustrate the effect of varying aspects of the sample, 5 independent simulations were performed for each combination of parameters in Table 1.

[Table 1 about here.]

As shown in the table the parameters were also summarised by a composite parameter  $SSAL = (\text{sample size}) \times (k - 1) \times (\text{number of loci})$ . 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 6 times,  $\mathcal{X} = (0, 2, 4, 6, 8, 10)$ . 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 towards consistency in the estimator, as the amount of information in the data increases. This number was chosen because pairs of sets of 5 replicates could be run in parallel on a 10 node cluster of 700Mhz 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 approximately 4 hours for  $SSAL = 800$  and approximately 6 days for  $SSAL = 8000$  (which has an IS size 5 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).

[Figure 5 about here.]

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 5 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 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, 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 6a,b.

[Figure 6 about here.]

Although there is substantial variability it can be seen that there is a general trend towards a reduction of the relative error with increasing  $SSAL$ . Given the variability of the results, the logarithm of the relative errors was analysed 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), -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; 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 favouring a model of population growth versus decline was calculated for each simulation as the proportion of cases where  $p(N_0 > N_A)/p(N_0 < N_A)$  (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 7a,b.

[Figure 7 about here.]

Although some of the simulations with  $SSAL \leq 1600$  have  $|\log(\text{Bayes factor})| < 2$  (i.e. would be judged to be non-significant 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 analysed 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 from each simulation. These are plotted in Figure 8a,b.

[Figure 8 about here.]

Although the posterior distribution is only asymptotically the same as the repeated sampling distribution, 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 1-sample test on the 60 p-values shows no departure from a uniform ( $p=0.93$ ). There is no trend towards 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,b,c,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 to rely on the mode as a point estimate.

## Analysis of example data sets

In order to illustrate the behaviour 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 *et al.* (1997); and data from the Mauritius kestrel surveyed by GROOMBRIDGE *et al.* (2000).

The *Drosophila* data were sampled from a population on Mount Parnes, 40Km north of Athens. The flies were genotyped for 9 allozyme loci. The study site occupied around 20,000m<sup>2</sup> of fir woodland at an elevation of 900m. BEGON *et al.* (1980) estimated the total suitable habitat to extend at least 10<sup>7</sup>m<sup>2</sup>. 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.* estimated these corresponded to sampling intervals of 9 and 2 generations respectively. Using mark-release-recapture methods they estimated the census size in their study area to be around 150,000 individuals. These data have also been analysed 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 8 loci. For comparison, these same 8 loci are analysed here (input file kindly provided by Eric Anderson). The number of alleles at each locus varied from 3-6. 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.

[Figure 9 about here.]

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 10000 points on a 500Mhz Pentium was 27 hours, although it can be seen from Figure 10 that good estimates of the posterior distribution can be obtained with substantially fewer points.

[Figure 10 about here.]

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( $\pm 115$ ) for the first interval and 244 ( $\pm 123$ ) for the second, and an overall estimate of 251 ( $\pm 115$ ) for both intervals. The reason for the discrepancy between the results from the two likelihood-based approaches and that from the moment-based approach 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  $\infty$  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 favour 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 3 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  $N_0$ .

The northern pike sample came from Lake Escanaba in Wisconsin. Fish scale samples from 1961, 1977, and 1993 were chosen from a collection of scales kept by the Wisconsin Department of Natural Resources, and were genotyped at 7 microsatellite loci. Five of these loci were bi-allelic and the remaining two were tri-allelic. 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 was analysed by WILLIAMSON and SLATKIN (1999). In their analysis, which was restricted to bi-allelic loci, the frequencies from two allelic classes at the two tri-allelic 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.

[Figure 11 about here.]

It can be seen that there is good information on the ancestral effective population size and it is unlikely to be greater than around 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 favour 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 versus 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 versus 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.

The Mauritius kestrel sample analysed here consisted of a number of individuals genotyped for 12 microsatellite loci, of which 7 were polymorphic. In this data set 75 individuals sampled in 1993 were genotyped, and (depending on the loci) up to 26 museum skins dating from 1829 to 1960. These data are described in GROOMBRIDGE *et al.* (2000), and the data was 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 between 1960 and 1993, the estimates of current effective populations 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.

[Figure 12 about here.]

There is very strong evidence of population decline, and  $N_A$  is unlikely to be less than  $\sim 300$  individuals and  $N_0$  is unlikely to be greater than around 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 favour of decline is  $> 9900$ . NICHOLS *et al.* (2001), analysing 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 or not. 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 only be the same as, or lower than 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; d) the model assumes an onset of population decline since 1829 rather than in the 20th century, which will also lead to overestimation of  $N_0$ .

## Discussion

This paper is concerned with two issues: the use of a hybrid computational method based on importance sampling and MCMC; and an extension of a previously published method for estimating effective population to allow for multiple samples and to allow inference of change in effective size. The computational method has potential implications for likelihood-based approaches for inferring demographic history in general, and will be discussed first.

**Comparison of IS, MCWM, and GIMH:** 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  $\mathbf{x}$  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 modelled 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 papers 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 paper is that the importance sampling variance is higher than in the other models. There is a large variance associated with the simulation of the genealogical history when there are a number of different samples 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 11 often has low probability for any given realisation 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 (2001), when diverging populations are modelled. Undoubtedly a different importance sampling function, based on different heuristics, can circumvent this problem.

An alternative to the use of MCMC to obtain Bayesian summaries would be to use importance sampling directly on a grid of evaluation points to obtain the likelihood, and then use other numerical means to obtain summaries from the resulting approximated

function. This is the standard approach to the problem (*e.g.* BEERLI and FELSENSTEIN, 1999, 2001). In this approach a single parameter value (or vector for a multiparameter model) is used in the importance sampling function, and the importance weights are then evaluated at all the different points on the grid. Comparing the two approaches there are a number of considerations: a) As pointed out by STEPHENS and DONNELLY (2000), there is a general tendency to progressively underestimate likelihoods at grid points further away from that used in the IS function, leading to an underestimation of the posterior variance. By contrast, in the MCMC approach the parameters used in the IS function are the same as the evaluation points. b) Some experimentation with initial runs is needed to decide where to place the grid points, whereas the MCMC approach is inherently adaptive. c) The pure IS approach generates a single set of importance samples evaluated across the grid, whereas the MCMC evaluates a large number of importance samples at each point. Although this may make the IS/grid approach appear more efficient it should be noted that, although time is saved by only having to simulate one sample of genealogical histories for all the grid points, the importance weights still need to be evaluated at each point on the grid, and this evaluation will be computationally expensive for reasons discussed in the derivation of equation 6. d) Having obtained the likelihood surface from importance sampling further numerical processing is necessary to obtain Bayesian summaries, whereas all that is required from the MCMC approach is density estimation (and even this is not required if only marginal means/medians and equal-tail probability intervals are reported). e) Because of the need for complex processing, any addition of new parameters requires large changes to the original IS program, whereas only minor changes are needed in the MCMC method. However a clear drawback of the MCMC method is that complex posterior densities may not be well approximated by standard density estimation methods whereas this is automatically catered for in the IS/grid method, which produces smooth likelihood surfaces. One way to improve the estimation of posterior densities would be to use Rao-Blackwellisation (GELFAND and SMITH, 1990). For an example of its use in a genealogical model see BEAUMONT, p.642 in WILSON *et al.*, (2000).

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.

Thus, overall, it can be seen that the GIMH approach may well offer a number of advantages over other methods. Some of the potential disadvantages of earlier importance sampling approaches are addressed through the use of bridge-sampling in FEARNHEAD and DONNELLY (2001), which allows for importance samples generated under different parameter values to be mixed together. This circumvents some of the problems discussed above, and also makes a Bayesian approach easier. A substantial amount of processing of the importance samples is needed, however, and it will be interesting to see how GIMH, particularly if improved as discussed above, compares in general utility.

**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 coalescent 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  $N_e \gtrsim 20$ . 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 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 favourably 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 well under a minute on a standard PC (Jinliang Wang, *pers. comm.*). Although WANG (2001) demonstrated only relatively small discrepancies between the pseudo-likelihood method and the full likelihood method in the 3-allele case, it would clearly be useful to compare the different approaches when there are larger numbers of alleles.

**Estimation of change in population size:** This paper 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 in order 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 towards 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 between 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 much less than  $\sqrt{2}$ , change in sample size and time between samples has the greater effect, and for values much greater than  $\sqrt{2}$  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 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 RJMCMC to compare different models, or to directly estimate  $P(\text{data})$  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  $P(\text{data})$  (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 there is no selection 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.

The twin assumptions of no mutation and no migration will often be violated, particularly when long temporal sequences are analysed and when microsatellite data are used. The effect of migration will generally be to inflate estimates of the local population size. Intuitively, the effect of mutation should be similar, because, from a genealogical point of view, lineages that could otherwise have coalesced within the time interval will not be able to do so if there is a mutation to a new allelic type within the interval. The details

of these effects on population size estimates, and, in particular, whether they will lead to apparent changes in population size, await further investigation. For microsatellites, an obvious 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. Indeed, all of the drift-based models could be incorporated in such general genealogical MCMC schemes, which would then naturally provide the priors for the baseline gene frequencies. However, the utility of the drift-based approaches lies in their relative simplicity of implementation, and reasonable convergence times in comparison with a fully genealogical MCMC model. When used with markers that have a low mutation rate, such as SNPs, these drift-based models may be particularly useful in the analysis of human demographic history.

## Acknowledgements

I am grateful to David Balding, Claire Calmet, Lounes Chikhi, Jean-Marie Cornuet, Kevin Dawson, Richard Nichols, and Jinliang Wang for their helpful comments on previous versions of the manuscript. This work was supported in part by N.E.R.C. grant NER/B/S/2000/00669 awarded to Ken Norris, M.A.B., and Mike Bruford.

## References

- 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. U.S.A.* **95**: 15140–15144.
- BEAUMONT, M.A., 1999 Detecting population expansion and decline using microsatellites. *Genetics* **153**: 2013–2029.
- BEERLI P., 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. U.S.A.* **98**: 4563–4568.

- BEERLI P., J. FELSENSTEIN, 1999 Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics* **152**: 763–773.
- 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, 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. B* **266**: 2269–2274.
- DONNELLY, P., & S. TAVARÉ, 1995 Coalescents and genealogical structure under neutrality. *Annu. Rev. Genet.* **29**: 410–421.
- FEARNHEAD P., P. DONNELLY, 2001 Estimating recombination rates from population genetic data. *Genetics* **159**: 1299–1318.
- FELSENSTEIN J., M.K. KUHNER, J. YAMATO and P. BEERLI, 1999. Likelihoods on coalescents: a Monte Carlo sampling approach to inferring parameters from population samples of molecular data, pp. 163–185, in *Statistics in Molecular Biology*, IMS Lecture Notes — Monograph Series, Vol. 33, edited by F. SEILLIER-MOISEWITSCH. Institute of Mathematical Statistics and American Mathematical Society, Hayward, CA.

- FRANKHAM, R. (1995) Conservation Genetics. *Annu. Rev. Genet.* **29**: 305–327.
- GELFAND, A.E., and A.F.M. SMITH, 1990 Sampling based approaches to calculating marginal densities. *J. Amer. Statis. Ass.* **87**: 398–409.
- GELMAN, A., J.B. CARLIN, H.S. STERN and D.B. RUBIN, 1995 *Bayesian Data Analysis*. Chapman and Hall, London.
- GREEN, P.J., 1995 Reversible jump Markov chain Monte Carlo computation 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. B Biol. Sci.* **344**: 403–410.
- GRIFFITHS, R.C., and S. TAVARÉ, 1994c Ancestral inference in population genetics. *Statist. Sci.* **9**: 307–319.
- GROOMBRIDGE, J.J., C.G. JONES, M.W. BRUFORD, and R.A. NICHOLS, 1999 ‘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, 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, D. MOLLISON, 2000 Analyses of infectious disease data from household outbreaks by Markov chain Monte Carlo methods. *Appl. Statist.* **49**: 517–542.
- O’RYAN, C., E.H. HARLEY, M.W. BRUFORD, M. BEAUMONT, R.K. WAYNE, and M.I. CHERRY, 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, 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.
- Saccheri, I.J., Wilson, I.J., Nichols, R.A., Bruford, M.W. and P.M. Brakefield (1999) Inbreeding of bottlenecked butterfly populations: estimation using the likelihood of changes in marker allele frequencies. *Genetics* **151**: 1053–1063.
- SACCHERI I., M. KUUSSAARI, M. KANKARE, P. VIKMAN, W. FORTELIUS, I. HANSKI, 1998 Inbreeding and extinction in a butterfly metapopulation *Nature* **392**: 491–494
- SLATKIN, M., 1996 Gene genealogies within mutant allelic classes. *Genetics* **145**: 579–587.
- STEPHENS, M. and P. DONNELLY, 2000 Inference in molecular population genetics (with discussion). *J. R. Statist. 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**: (*in press*).
- 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 and Hall, London.
- WAKELEY J., 1999 Nonequilibrium migration in human history. *Genetics* **153**: 1863–1871.
- WAKELEY J., 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. Camb.* **78**: 243–257.
- 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.
- WILSON L., D.A. STEPHENS, R.M. HARDING, *et al.*, 2000 Inference in molecular population genetics — Discussion. *J. Roy. Stat. Soc. B* **62**: 636–655.

## 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 (see *e.g.* RICE, 1995),

$$\tilde{E} \left[ \frac{\tilde{p}(D|\Phi_{i+1})}{\tilde{p}(D|\Phi_i)} \right] = \frac{\tilde{p}(D|\Phi_{i+1})}{\tilde{p}(D|\Phi_i)} \left( 1 + \frac{\text{SE}^2(\tilde{p}(D|\Phi_i))}{\tilde{p}^2(D|\Phi_i)} \right) \quad (7)$$

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

$$\begin{aligned} \text{SE}_{(1\dots j)}^2[\tilde{p}(D|\Phi)] &= \text{SE}_{(1\dots j-1)}^2[\tilde{p}(D|\Phi)] \text{SE}_j^2[\tilde{p}(D|\Phi)] + \tilde{p}_{(1\dots j-1)}^2(D|\Phi) \text{SE}_j^2[\tilde{p}(D|\Phi)] + \tilde{p}_j^2(D|\Phi) \text{SE}_{(1\dots j-1)}^2[\tilde{p}(D|\Phi)] \\ \tilde{p}_{(1\dots j)}(D|\Phi) &= \tilde{p}_j(D|\Phi) \tilde{p}_{(1\dots j-1)}(D|\Phi) \end{aligned} \quad (8)$$

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

**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,  $\Phi$ , from the correct posterior density for any  $n \geq 1$  sampled genealogical histories. To ease the notation I will assume uniform improper priors on  $\Phi$  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

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

which can be simply rewritten as

$$\frac{\sum_j \left( p(D, G'_j|\Phi') \prod_{i \neq j} q(D, G'_i|\Phi') \right) \prod_i q(D, G_i|\Phi) p(\Phi|\Phi')}{\sum_j \left( p(D, G_j|\Phi) \prod_{i \neq j} q(D, G_i|\Phi) \right) \prod_i q(D, G'_i|\Phi') p(\Phi'|\Phi)}$$

If we regard the sampling procedure as ordered (*i.e.* the sampled genealogies occupy ‘slots’  $j = 1 \dots h$ ), 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  $j$ th position the density is proportional to

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

where the integration is over all genealogical histories in ‘slots’  $1 \dots h$ . This evaluates to

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

which is

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

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  $j$ th ‘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  $\Phi$ . 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 backwards from the current time,  $b$  is the growth rate, and  $N_0$  is the current population size. We assume throughout the paper 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

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

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

$$\tilde{N}_i = \frac{x_i - x_{i-1}}{\int_{x_{i-1}}^{x_i} \frac{1}{N_0 r^{\frac{-y}{X}}} dy} = \frac{(x_i - x_{i-1}) N_0 \log(r) r^{\frac{-x_{i-1}}{X}}}{X (r^{\frac{(x_i - x_{i-1})}{X}} - 1)}, \quad (9)$$

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), expanded to consider more than one interval between samples.

The probability of obtaining  $c_i$  coalescences in any interval  $[x_{i-1}, x_i]$ ,  $p(c_i | \frac{x_i - x_{i-1}}{2\tilde{N}_i})$  is given by TAVARÉ (1984, eq. 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 (9) 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  $i$ th sample point (without regard to how they are partitioned) is given by

$$p(\mathbf{f}_i + \mathbf{a}_i | \mathbf{f}_{i+1}, c_{i+1}) = \frac{\prod_{j=1}^k \binom{f_{ij} + a_{ij} - 1}{f_{(i+1)j} - 1}}{\binom{h_i + n_i - 1}{h_{i+1} - 1}} \quad 0 \leq i \leq d - 1 \quad (10)$$

(SLATKIN, 1996; NIELSEN *et al.*, 1998; O'RYAN *et al.*, 1998).

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

$$p(\mathbf{a}_i, \mathbf{f}_i | \mathbf{a}_i + \mathbf{f}_i) = \frac{n_i! h_i!}{(n_i + h_i)!} \prod_{j=1}^k \frac{(a_{ij} + f_{ij})!}{a_{ij}! f_{ij}!} \quad 0 < i \leq d. \quad (11)$$

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  $\mathbf{x}$ )

$$p(\mathbf{a}_d + \mathbf{f}_d) = \frac{\Gamma(n_d + h_d) \Gamma(bk)}{\Gamma(n_d + h_d + bk)} \prod_{j=1}^k \frac{\Gamma(a_{dj} + f_{dj} + b)}{\Gamma(a_{dj} + f_{dj} + 1) \Gamma(b)}, \quad (12)$$

where  $b$  is taken here to be 1 (equivalent to assuming a Dirichlet prior of  $D(1, \dots, 1)$ ). In earlier papers using this methodology (O'RYAN *et al.*, 1998; CIOFI *et al.*, 1999; CHIKHI *et al.* 2001; BERTHIER *et al.*, *in press*), the multinomial was used and then the integration performed by Metropolis-Hastings simulation.

**Importance sampling:** Extending the approach of Berthier *et al.* (2002) to multiple samples, equation (10), 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  $e$ th coalescent event after the  $i$ th 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  $j$ th 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 backwards from the data using the method of GRIFFITHS and TAVARÉ. The  $j$ th 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 there is at least one representative 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  $i$ th 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 (11) is calculated. At the final data sample, the probability of the allele frequency count  $\mathbf{a}_i + \mathbf{f}_i$  is given by (12).

**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

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

# List of Figures

1	Diagram illustrating the terminology used in the text. . . . .	45
2	Figure comparing the posterior distributions obtained using MCWM without bias correction, MCWM with bias correction, GIMH, and pure importance sampling. . . . .	46
3	The relationship between the IS size and the standard deviation of the posterior distribution for MCWM with and without bias correction and GIMH . . . . .	47
4	The proportion of trial updates in the MCMC that are accepted, divided by IS size, is plotted against IS size. . . . .	48
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 $N_0$ use 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, $N_A = 200$ , $N_0 = 20$ . . . . .	49
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 slightly shifted 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. . . .	50
7	A plot of the logarithm of the Bayes factor in supporting a model of population growth against SSAL. Other details as in Figure 6. . . . .	51
8	A plot of the critical HPD p-value of the true $N_A$ and $N_0$ against SSAL. Other details as in Figure 6. . . . .	52
9	Posterior distribution of $N_0$ and $N_A$ for the fly data of Begon <i>et al.</i> (1980). The contour levels are at the 0.1, 0.5 and 0.9 HPD limits, as in Figure 5 . .	53
10	Trace of the realised values of $N_e$ during an MCMC run with the fly data of Begon <i>et al.</i> (1980). The initial 100 points have been discarded. . . . .	54
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 . . . . .	55
12	Posterior distribution of $N_0$ and $N_A$ for the Mauritius kestrel data. . . . .	56

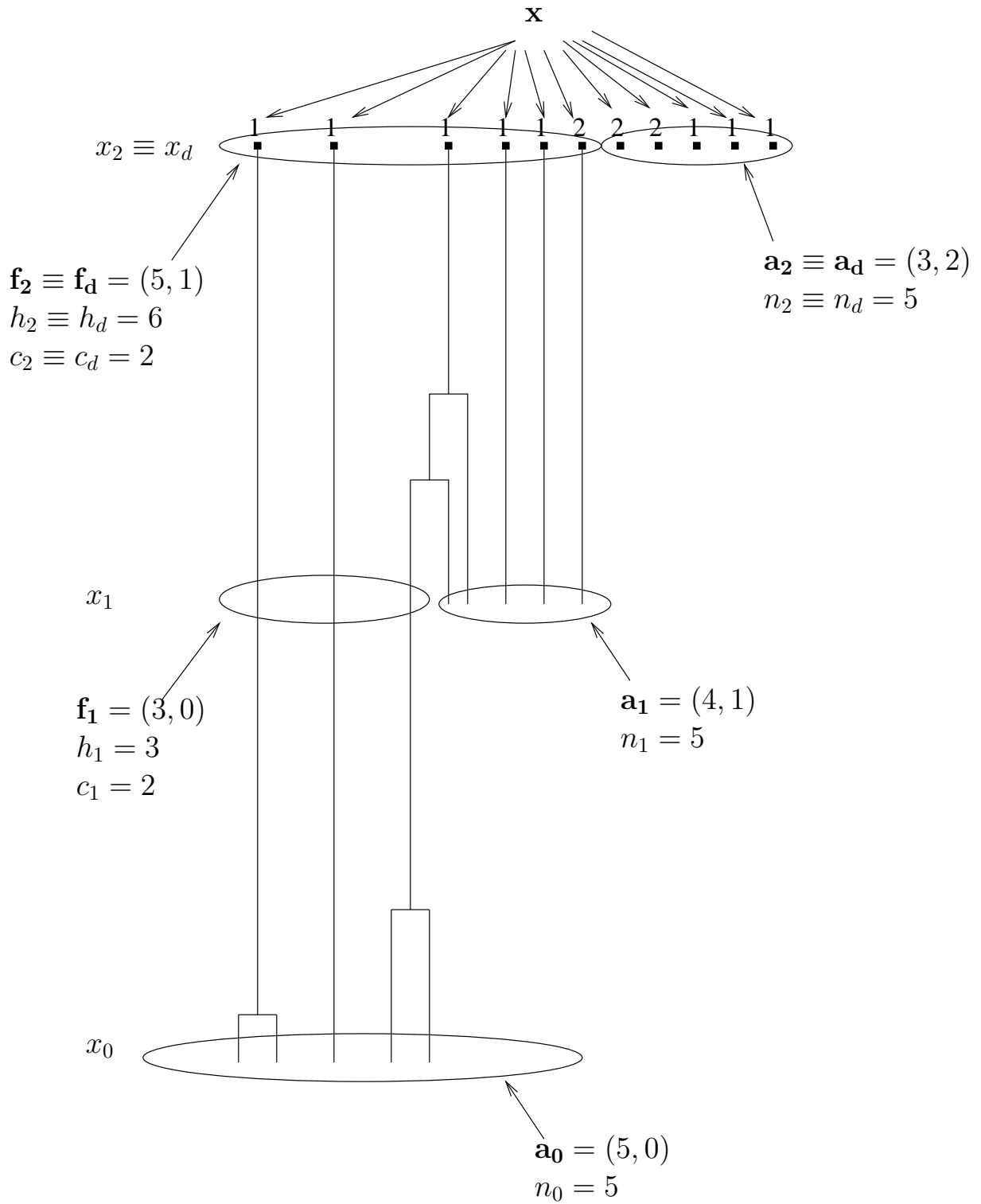


Figure 1:

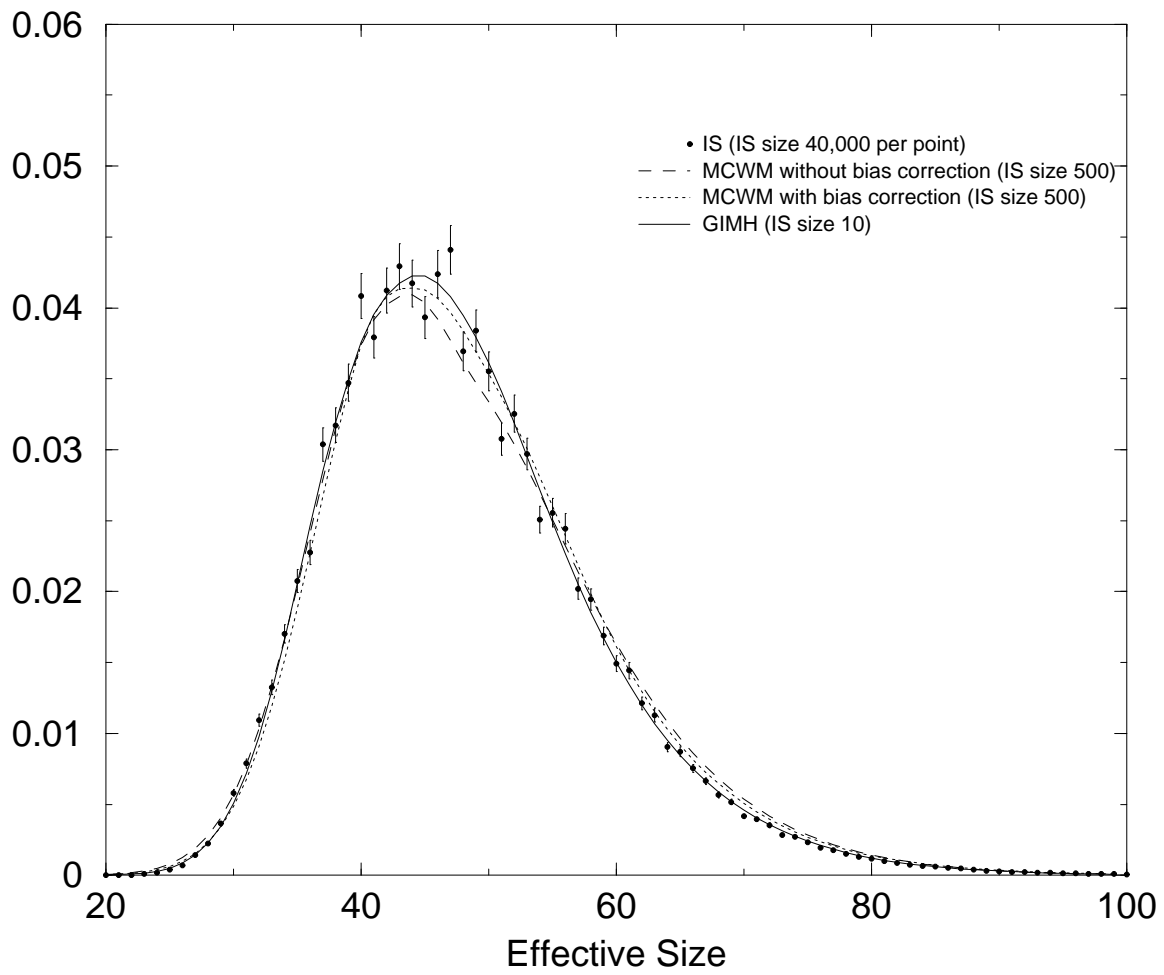


Figure 2:

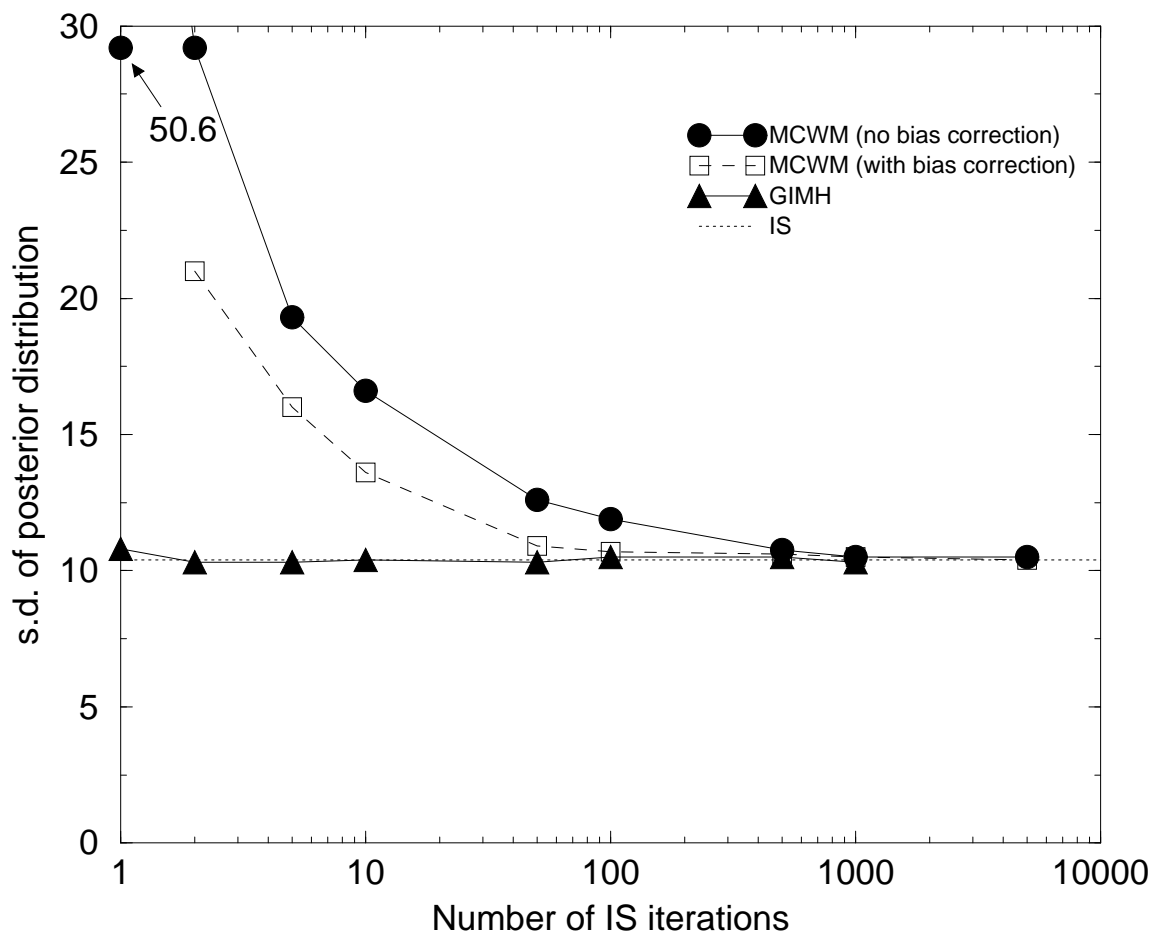


Figure 3:

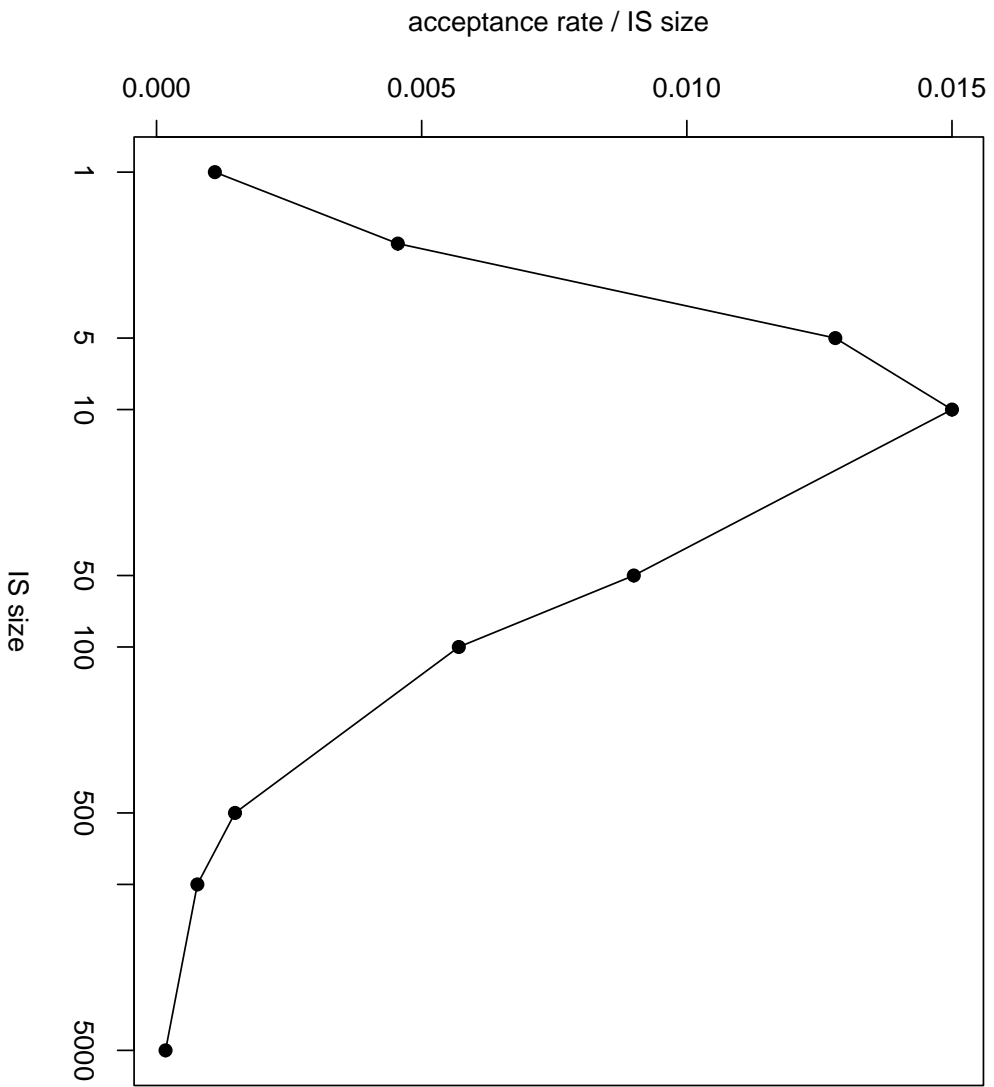


Figure 4:

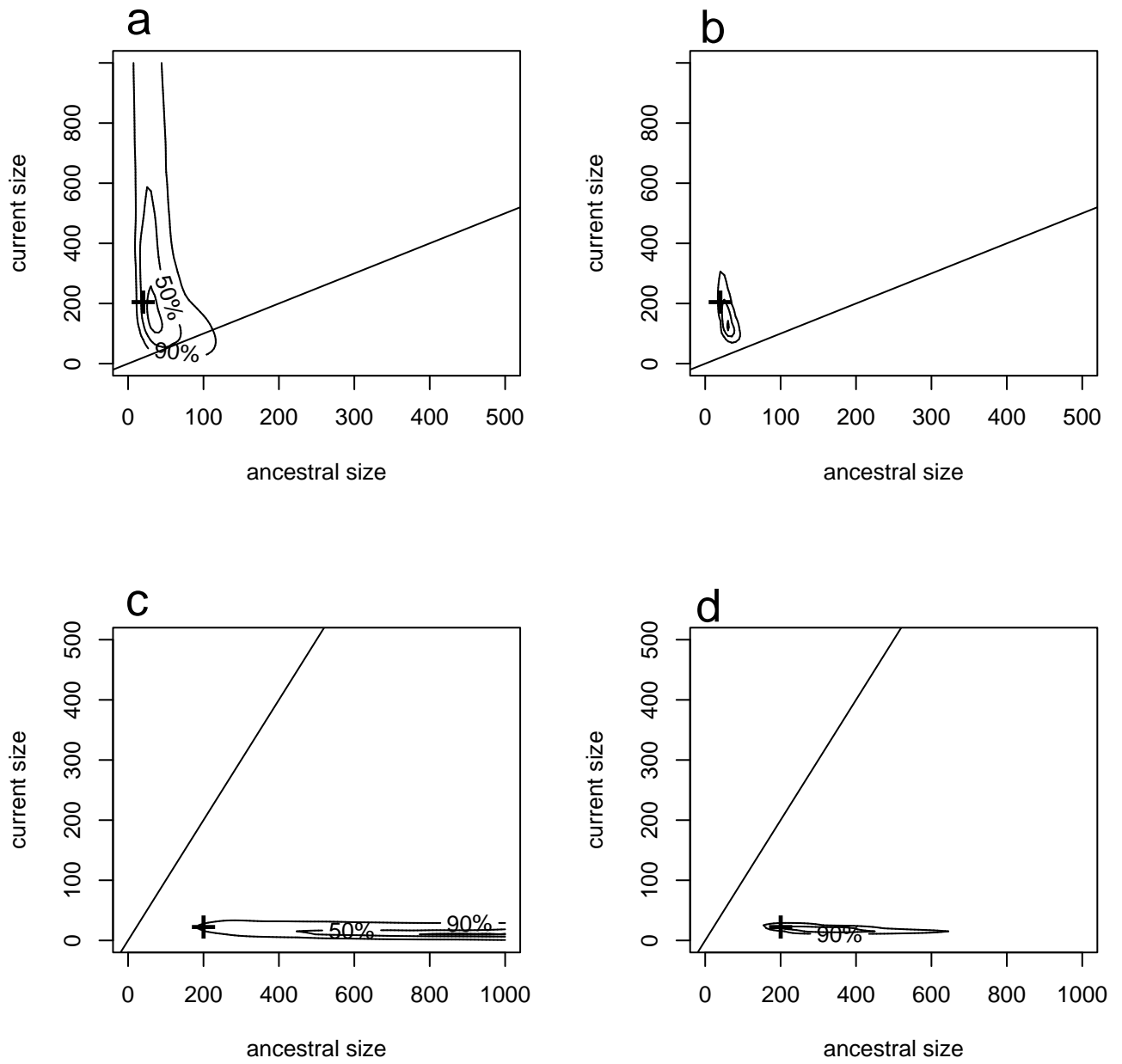


Figure 5:

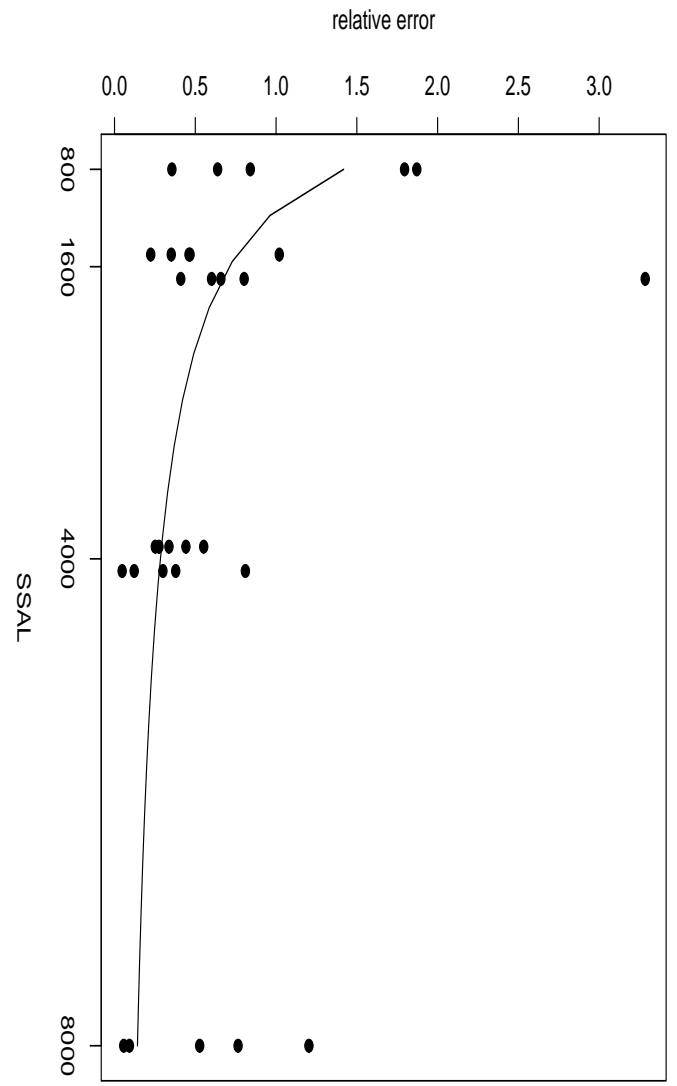
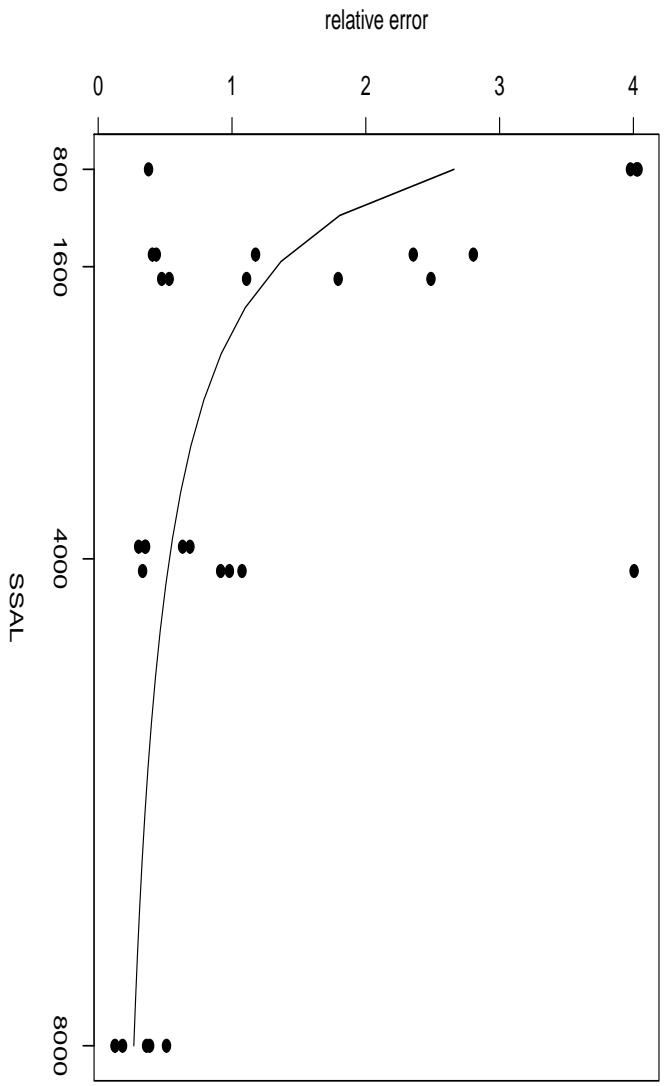
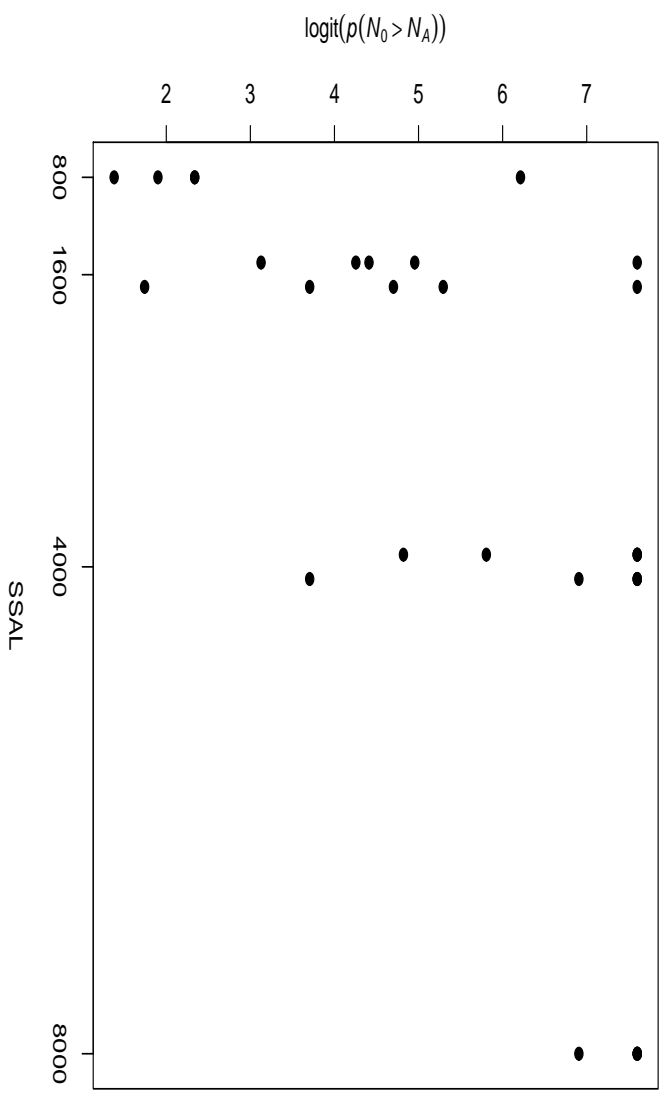


Figure 6:

a



b

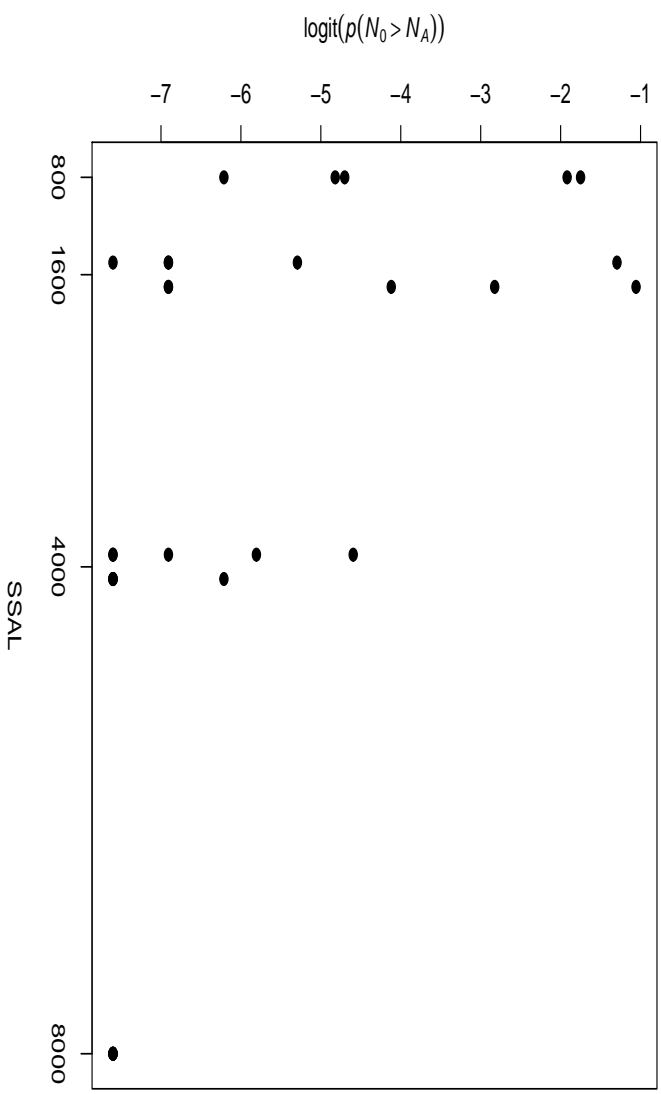


Figure 7:

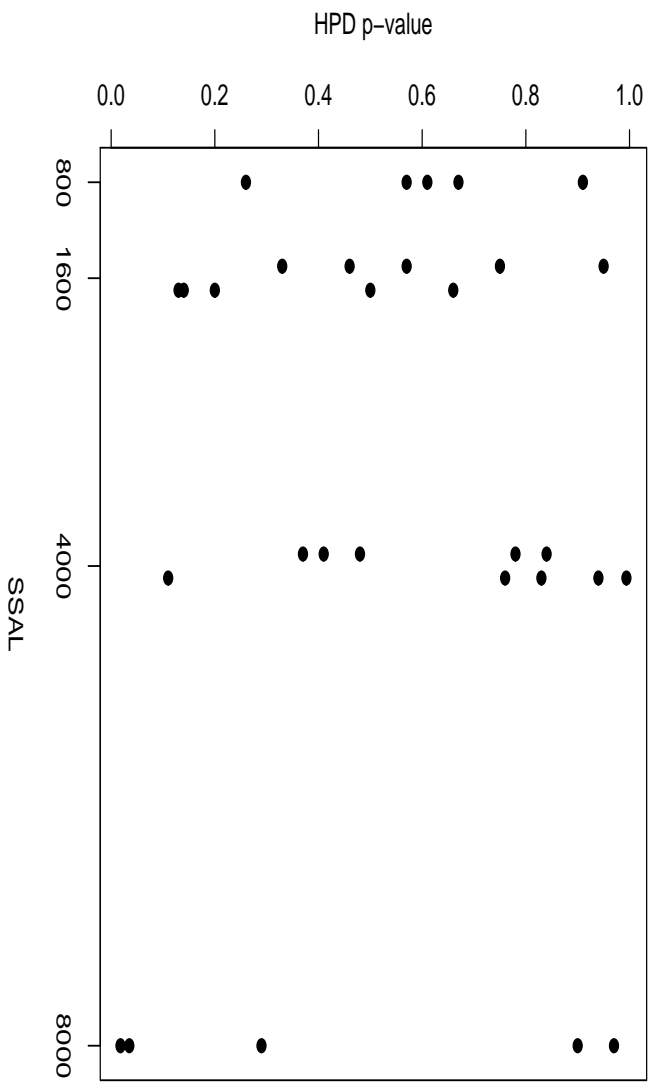
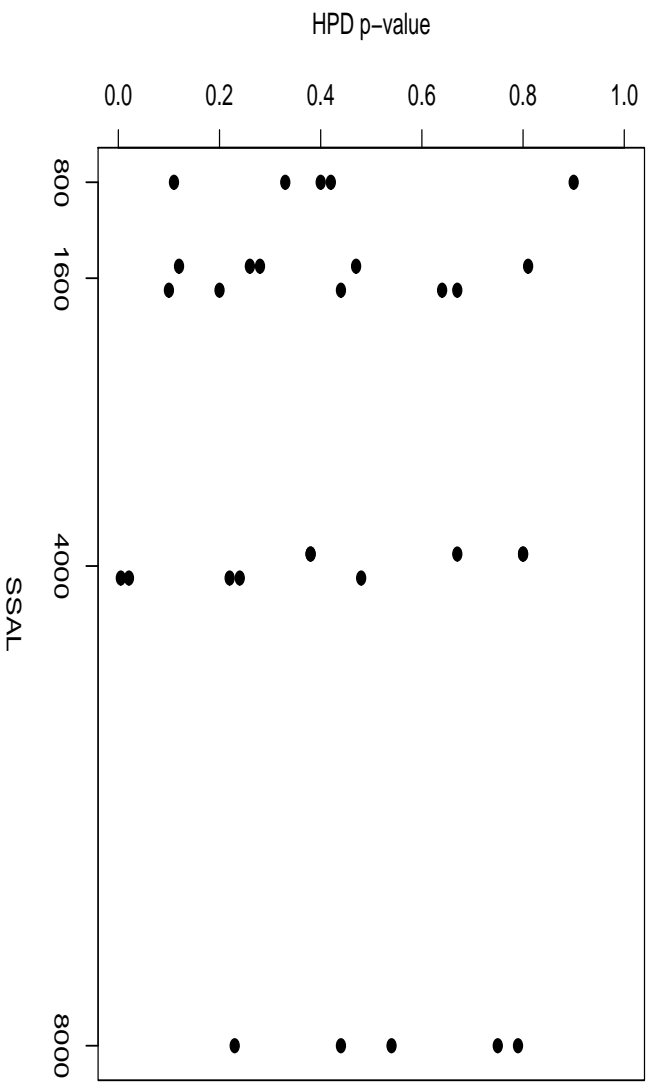


Figure 8:

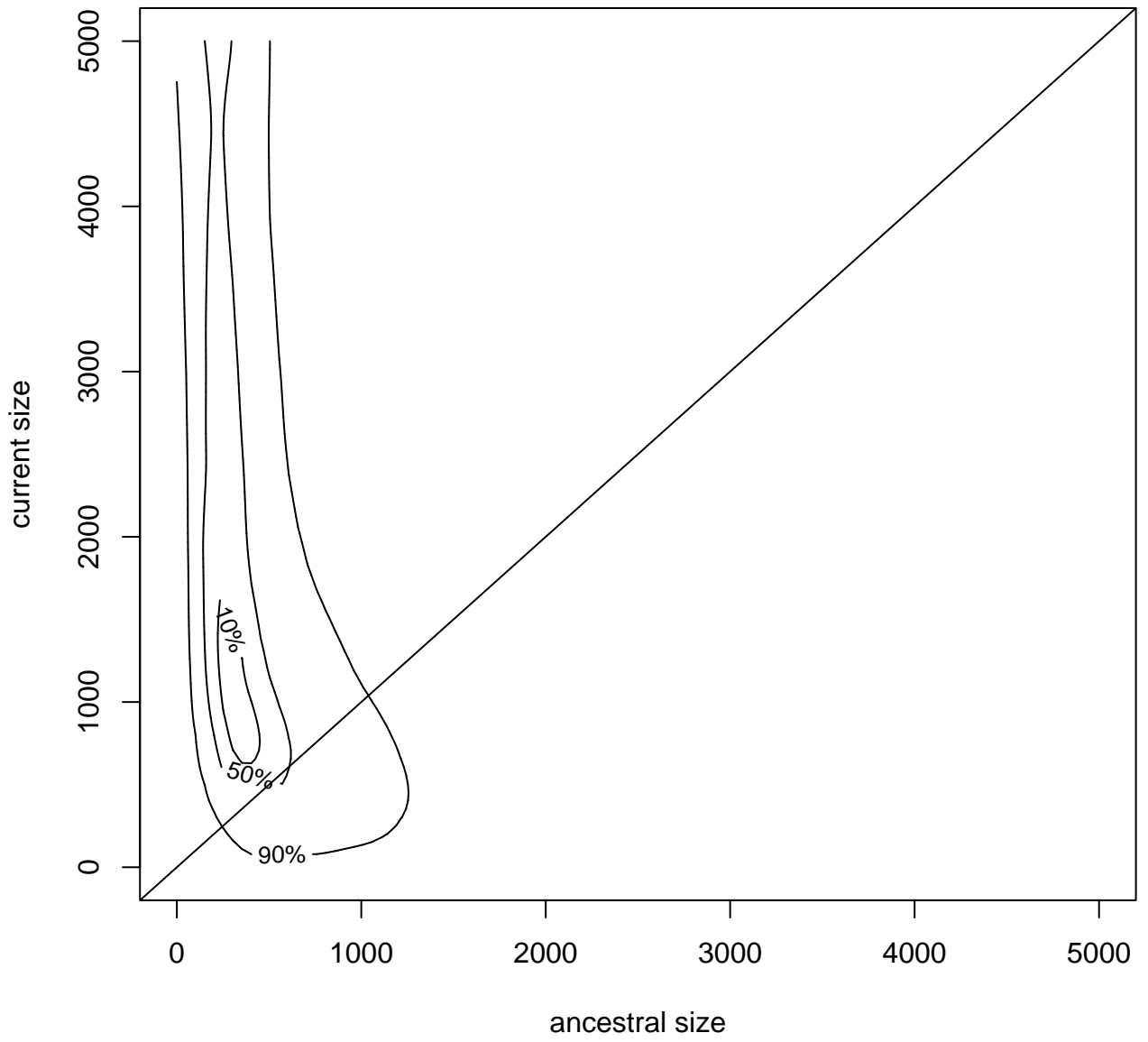


Figure 9:

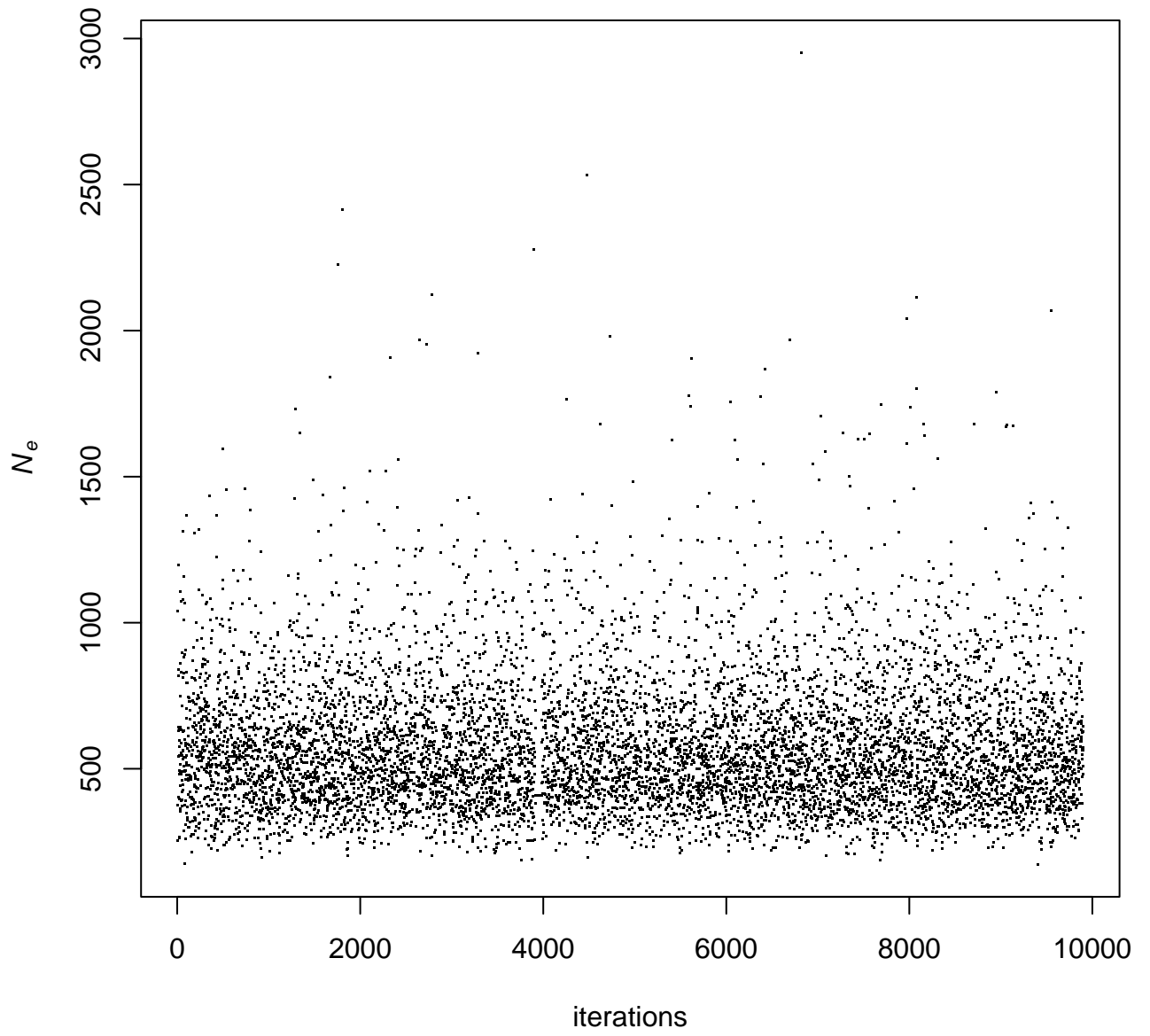


Figure 10:

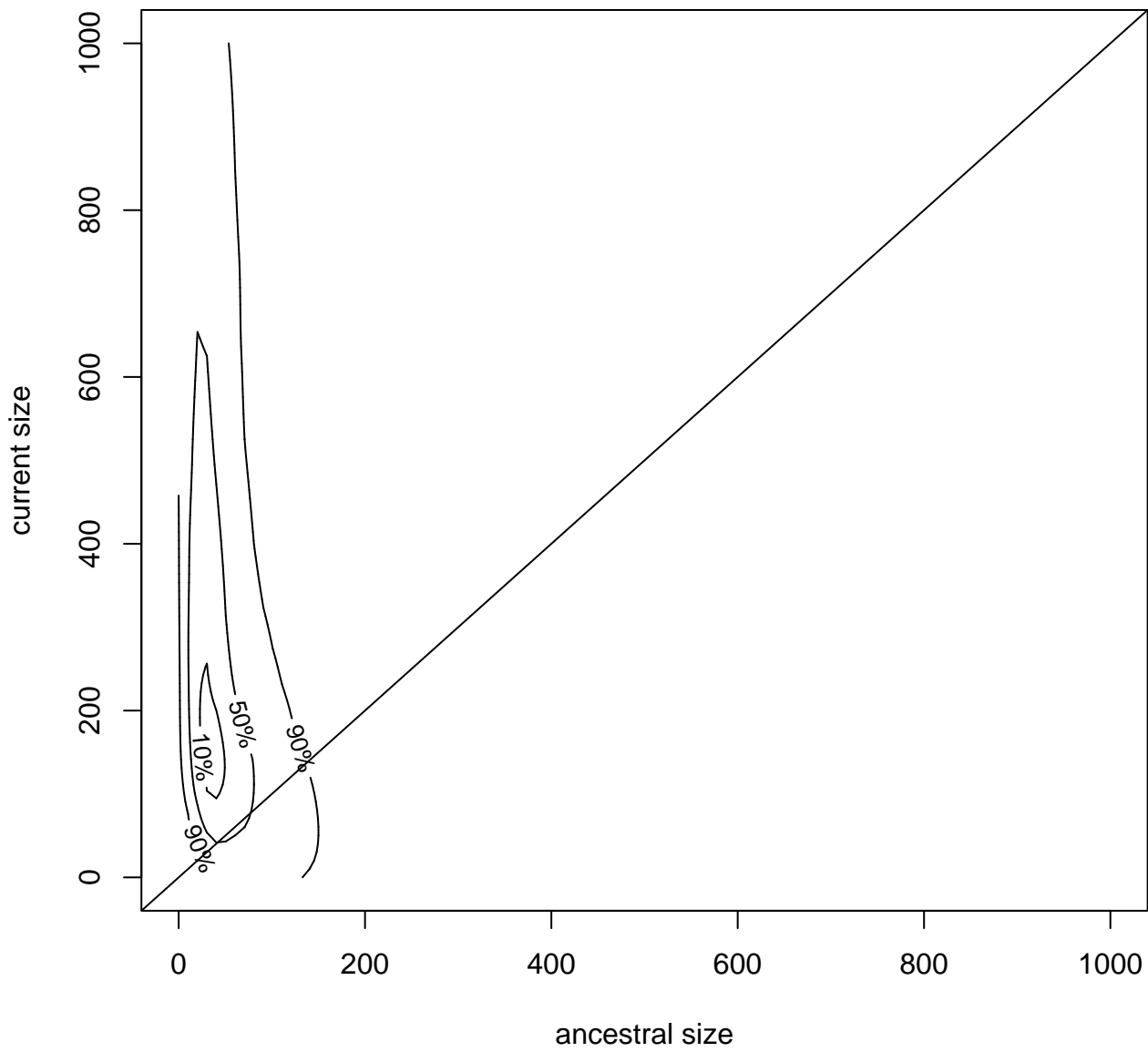


Figure 11:

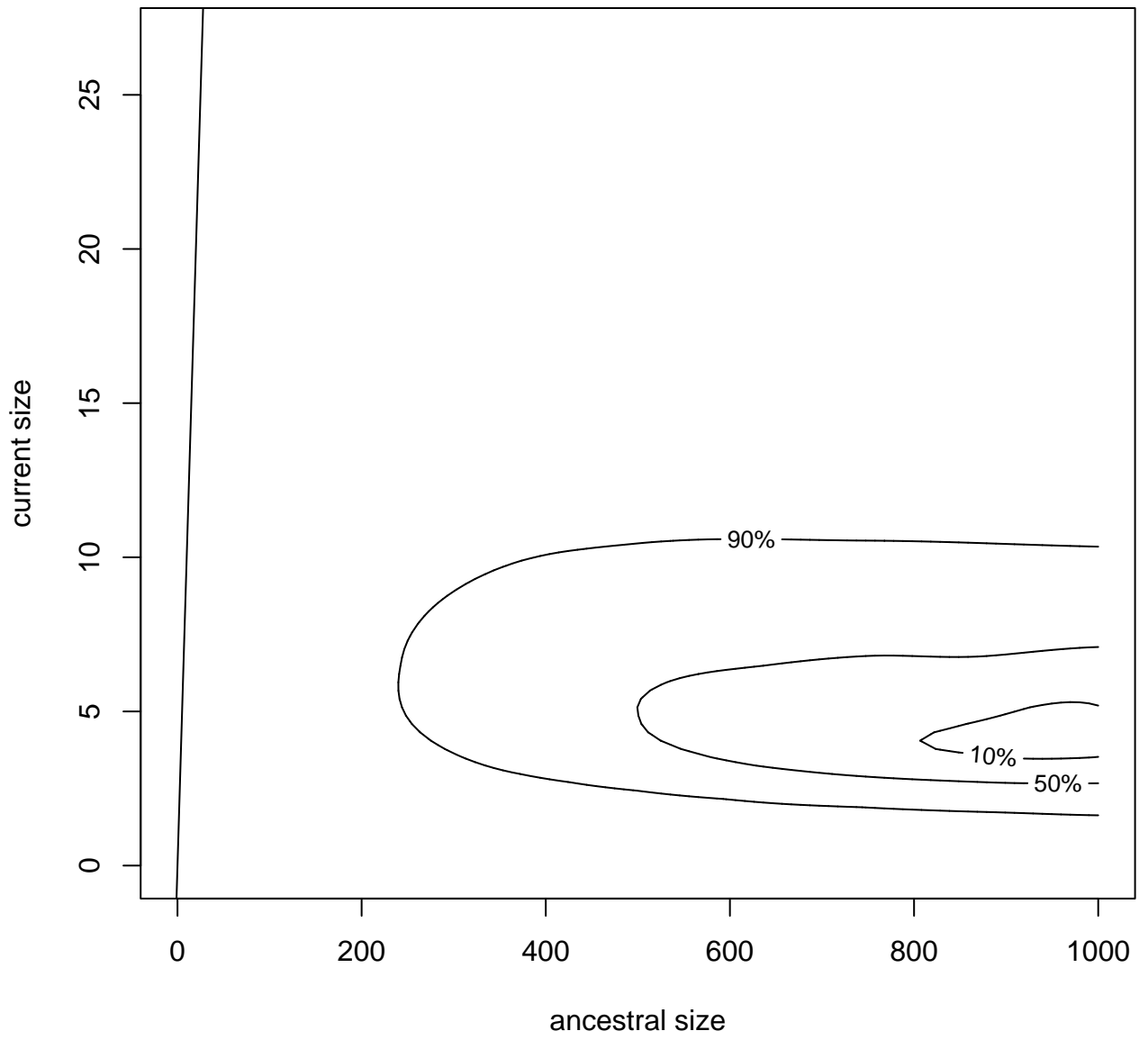


Figure 12:

# List of Tables

- 1 Table showing combinations of numbers of loci, numbers of alleles and sample size used in the simulations to test the accuracy of the method. The sample size is the number of chromosomes taken at each of 6 time points. 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. . . . . 58

SSAL	No. Loci	No. Alleles	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

Table 1: Table showing combinations of numbers of loci, numbers of alleles and sample size used in the simulations to test the accuracy of the method. The sample size is the number of chromosomes taken at each of 6 time points. 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.