Copyright © 2000 by the Genetics Society of America

Detecting Bottlenecks and Selective Sweeps From DNA Sequence Polymorphism

Nicolas Galtier, Frantz Depaulis and Nicholas H. Barton

Institute of Cell, Animal and Population Biology, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom

Manuscript received July 24, 1999
Accepted for publication February 29, 2000

ABSTRACT

A coalescence-based maximum-likelihood method is presented that aims to (i) detect diversity-reducing events in the recent history of a population and (ii) distinguish between demographic (*e.g.*, bottlenecks) and selective causes (selective sweep) of a recent reduction of genetic variability. The former goal is achieved by taking account of the distortion in the shape of gene genealogies generated by diversity-reducing events: gene trees tend to be more star-like than under the standard coalescent. The latter issue is addressed by comparing patterns between loci: demographic events apply to the whole genome whereas selective events affect distinct regions of the genome to a varying extent. The maximum-likelihood approach allows one to estimate the time and strength of diversity-reducing events and to choose among competing hypotheses. An application to sequence data from an African population of *Drosophila melanogaster* shows that the bottleneck hypothesis is unlikely and that one or several selective sweeps probably occurred in the recent history of this population.

LOW genetic variability in natural populations is not a rare feature: numerous examples have been reported in animals (*e.g.*, O'Brien and Evermann 1988), plants (Liu *et al.* 1998), and protists (Rich *et al.* 1998), among others. Low present-day levels of variation may reflect a persistent state maintained by, say, nonpanmictic mating systems (Charlesworth and Charlesworth 1995) or recurrent background selection (Charlesworth *et al.* 1995) at linked loci. In many cases, however, a recent event in the history of the population is invoked to explain reduced variability.

Such diversity-reducing events essentially fall into two categories: demographic factors and selective factors. Demographic factors include bottlenecks and population founder events; both involve a temporary reduction of population size resulting in an increased rate of genetic drift. Rapid fixation of a new, favorable allele through directional selection (a "selective sweep") also generates a sudden drop of genetic variability at linked loci by hitchhiking (Maynard-Smith and Haigh 1974). In this article, we address two questions: first, how recent diversity-reducing events can be detected, and second, how demographic and selective causes can be distinguished.

The issue of detecting diversity-reducing events is not trivial. Usually, a bottleneck (or a selective sweep) is invoked when the variability at some locus of some species is much lower than that usually observed in related species (or at distinct loci). However, mutation and drift processes have a high variance and may generate highly different patterns in distinct species or distinct loci just by chance. Additionally, some variance between observable patterns of polymorphism in distinct species is introduced by random sampling of individuals. Assessing the statistical significance of an "apparent" discrepancy between data sets is therefore essential.

Given that genetic variability has been reduced recently, the question of distinguishing between demographic and selective causes is a major one: the two kinds of events have different biological meanings. Detecting bottlenecks is relevant to conservation biology since the global reduction of genetic variability they induce may endanger populations. Detecting selective sweeps, on the other hand, is an important goal for the study of evolutionary mechanisms. In particular, a long-standing controversy persists about the relative importance of positive selection *vs.* neutral or nearly neutral evolution at the genomic level (*e.g.*, Gillespie 1991).

The theory of coalescence (Kingman 1982; Hudson 1991) provides a promising framework to address these questions. Variability-reducing events can be detected because they modify the shape of the genealogy of alleles. Basically, they tend to generate star-like (parts of) genealogies, as a consequence of a sudden increase of coalescence rate (Figure 1). Demographic events apply to the whole genome whereas selective events affect different regions of the genome to various extents thanks to recombination (*e.g.*, Hudson *et al.* 1987). This gives the possibility of distinguishing the two hypotheses by sampling several loci: a more or less common pattern is expected in the case of a bottleneck, while selective sweeps generate heterogeneity across loci.

Griffiths and Tavare (1994a) devised an efficient method to compute likelihoods under the coalescent model, which can be easily generalized to the case of

Corresponding author: N. Galtier, UPR 9060 "Génome, Populations, Interactions"-CC063, Université Montpellier 2, Place E. Bataillon, 34095 Montpellier, France. E-mail: galtier@crit1.univ-montp2.fr

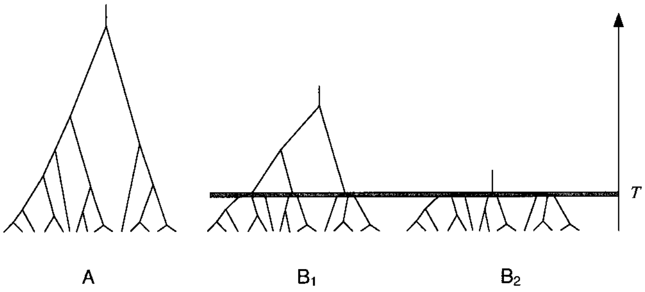

Figure 1.—Effect of bottlenecks on genealogies. (A) Standard genealogy under the neutral, constant-population size coalescent. (B_1) A short, moderate bottleneck occurs at time T (gray zone). Looking forward, all the individuals in the sample descend from three of the lineages that entered the bottleneck: today's sample includes three "gene families." Looking backward, three lineages survived the burst of coalescences generated by the bottleneck. (B_2) Strong bottleneck. Only one lineage survived it.

variable population size (Griffiths and Tavare 1994b). In this article, we implement a model of sudden reduction of genetic variability into Griffiths and Tavaré's scheme and devise likelihood-ratio tests to detect and discriminate between bottlenecks and selective sweeps. The new method is applied to polymorphism sequence data from an African population of *Drosophila melanogaster*.

METHODS

In this section, we first recall the main ideas of Griffiths and Tavare's (1994a,b, 1995) approach, then show how it can be used to model a bottleneck at one locus, and finally address the issue of hypothesis testing with multiple loci.

Griffiths and Tavaré's method: Consider a data set D_0 consisting of DNA sequences (genes) sampled at one locus in n individuals of a panmictic population of effective size $2N$. Assume that neutral mutations occur at rate μ. Assume that no recombination occurs within the locus. A fundamental recursion is

$$\Pr(D_0) = \sum_{D_1} \Pr(D_0|D_1)\cdot \Pr(D_1). \tag{1}$$

Here, D_1 is any state of the data one "step" before D_0, *i.e.*, any sample at a previous time that may be transformed into D_0 by either a mutation or a separation of lineages. The transition probabilities are (Griffiths and Tavare 1994a,b)

$$\Pr(D_0|D_1) = k_c \cdot \frac{n-1}{\theta + n - 1}$$

if the backward event is a coalescence

$$\Pr(D_0|D_1) = k_m \cdot \frac{\theta}{\theta + n - 1}$$

if the backward event is a mutation, (2)

where θ is the population mutation rate $4N\mu$. Coefficients k_c and k_m depend on how many distinct forward events can lead from D_1 to D_0 and on the mutation model (see below). The number of genes in D_1 is n if the event is a mutation and $n - 1$ if it is a coalescence. Recursion (1) can be expanded,

$$\begin{aligned}\Pr(D_0) &= \sum_{D_1} \Pr(D_0|D_1)\cdot \Pr(D_1)\\ &= \sum_{D_1}\sum_{D_2} \Pr(D_0|D_1)\cdot \Pr(D_1|D_2)\cdot \Pr(D_2)\\ &= \sum_{D_1}\sum_{D_2} \ldots \sum_{D_m} \Pr(D_0|D_1)\cdot \Pr(D_1|D_2) \ldots \Pr(D_{m-1}|D_m)\cdot \Pr(D_m),\end{aligned} \tag{3}$$

where m is the total number of events before reaching the common ancestor D_m of genes in the sample, and where $\Pr(D_m)$ is one. Transition probabilities $\Pr(D_i|D_{i+1})$ in (3) are given by adjusting n in (2) to the size of D_i. $\Pr(D_0)$ is the likelihood of parameter θ. It is expressed as a sum over all possible sets of ancestral states ($D_1, D_2, \ldots, D_m$), that is, the topology of the genealogy and the order of events.

For large data sets, one cannot compute (3) exactly: there are too many sets of ancestral states. Griffiths and Tavaré's idea was to randomly sample a reasonable number of sets of ancestral states and to estimate the likelihood from this sample. Let A be a set of ancestral states ($D_1, D_2, \ldots, D_m$), and let H be the hypothesis of the coalescent model (including parameter value θ). Equation 3 can be rewritten

$$\begin{aligned}\Pr(D_0|H) &= \sum_{A} \Pr(D_0 \,\&\, A|H)\\ &= \sum_{A} \frac{\Pr(D_0 \,\&\, A|H)\cdot \Pr(A|D_0, X)}{\Pr(A|D_0, X)}\\ &= E_X\left(\frac{\Pr(D_0 \,\&\, A|H)}{\Pr(A|D_0, X)}\right),\end{aligned} \tag{4}$$

where E_X means expectation with respect to X. Here, X is any sampling distribution of ancestral states of the data D_0. Equation 4 provides a method for estimating the likelihood in reasonable time: (i) sample one set A of ancestral states according to distribution X; (ii) calculate the expression in the expectation in (4) using the recursion (3) with that particular A (numerator) and the probability of A under X (denominator); and (iii) iterate (i) and (ii) several times and take the average.

Obviously, an efficient sampling process X is one that samples probable ancestral states (according to H) with high probability, *i.e.*, one as close as possible to the unknown "H given D_0" (the distribution of ancestral states under the coalescent conditional on D_0). For instance, a uniform sampling of sets of ancestral states is inefficient since most sets of ancestors have a very low probability, but a few of them have a high probability. Sampling uniformly, one would have to perform many replicates of X to get an accurate estimate of the likelihood. Griffiths and Tavare's importance sampling scheme X is a Markov chain: (i) start from D_0; (ii) for all possible D_1, compute $\Pr(D_0|D_1, H)$ according to (2); (iii) randomly sample D_1 with probability proportional to $\Pr(D_0|D_1, H)$,

$$\Pr(D_1|D_0, X) = \frac{\Pr(D_0|D_1, H)}{\sum_{D_1}\Pr(D_0|D_1, H)}; \tag{5}$$

and (iv) iterate until D_m.

This algorithm is presumably optimal if "Bayesian" probabilities $\Pr(D_{i+1}|D_i, X)$ equal the unknown $\Pr(D_{i+1}|D_i, H)$: in this case, X becomes identical to H given D_0. This requirement is met if, for any given state D_i, all the possible ancestral states D_{i+1} are equally probable under H.

The above equations hold whatever the assumed mutation model. In this article, we used the infinite-site model: it is assumed that no more than one mutation arises at any one site in the genealogy. A consequence is that no more than two distinct states should be observed at any site. Under this model, m equals $s + n - 1$ (where s is the number of segregating sites in the data set), and coefficients in (2) are $k_c = n_c/(n-1)$ and $k_m = n_m/n$, where n_c is the number of genes in D_1 leading to D_0 by splitting (in the case of a coalescent event), and n_m is the number of genes in D_1 leading to D_0 by mutating (in the case of a mutation event).

A bottleneck model: The above scheme can be used with models more complex than the standard coalescent (*e.g.*, Nielsen 1998). Griffiths and Tavare (1994b) showed how it can be generalized to account for variable population size. In this case, the relative probabilities of coalescence and mutation given in (2) depend on the time of the current state. This means that one has to keep track of the times of successive events when sampling ancestral states backward through the genealogy.

The bottleneck model we used has three parameters: population mutation rate θ, time *T* of occurrence of the bottleneck, and "strength" *S* of the bottleneck; all are scaled relative to a timescale set by 2*N*, which is the current number of genes. Looking backward in time, it is assumed that the population undergoes a drop of effective size at *T* (measured in units of 2*N* generations) during a short period of time and then recovers its initial size. If the duration of the bottleneck is short enough that one can neglect the occurrence of mutations during that period, the effect of the bottleneck depends only on the amount of coalescence it generates. Parameter *S* measures this coalescence pressure: it is the time that would be required for an equal expected amount of coalescence if the population size had not changed. Under these assumptions, the bottleneck model can be implemented under Griffiths and Tavare's scheme by keeping *N* constant, but changing the time scale during the bottleneck. The new Markov chain X' has three distinct phases: (i) starting from $t = 0$, recurse until *t* reaches *T* allowing coalescences and mutations, as in *X*; (ii) while $T < t < T + S$, recurse allowing coalescences only; and (iii) when $t > T + S$, switch on mutations again. In case of a severe bottleneck (high *S*), phase (iii) may not be reached for most realizations of X': only one lineage survives the bottleneck (backward), as in genealogy B_2 of Figure 1. This model reduces to the constant-size model by either setting $S = 0$ or $T = \infty$.

Maximizing the likelihood: The problem here is to find the values of θ, *T*, and *S* that maximize the likelihood for a given data set. Griffiths and Tavare (1994a) provide an efficient method for generating likelihood surfaces with respect to θ. The basic idea is to calculate the likelihoods of many θ's using a single sample of sets of ancestral states. This sample is obtained by performing *X* with transition probabilities computed from one particular value of θ called θ_0. Theoretically, this procedure may be used for all the parameters of any model. In practice, however, it does not work properly for *T* and *S* in the bottleneck model. The reason is that sets of ancestral states that have a high probability for some value T_0 (S_0) of the bottleneck time (strength) may have very low probability for other *T*'s (*S*'s). Using a common sample for all *T*'s (*S*'s) would therefore lead to variable accuracy in the estimation of the likelihood across parameter values. Therefore, a single sample of sets of ancestral states was used to generate a likelihood curve with respect to θ given *T* and *S*, but different samples were used for different (*T*, *S*) pairs.

We used a numerical technique to maximize the likelihood with respect to *T* and *S*. The problem with standard algorithms in this particular case is that the function to be maximized is "unstable": because of the stochastic process, several evaluations of the likelihood for a given (*T*, *S*) would return distinct numbers. The heuristic we used is a modification of the downhill simplex (Press *et al.* 1992). Details can be found in the help file of the program, both available on request.

The multilocus case and likelihood-ratio tests: We now turn to the problem of distinguishing selective sweeps from bottlenecks. We approximate the expected effect of a selective sweep at one neutral locus linked to the locus under selection by that of a population bottleneck: *T* is the time of fixation of the favorable mutation, and *S* depends on the ratio between the selection coefficient associated with this mutation and the recombination rate between the selected locus and the neutral locus. The discrepancy between the two hypotheses appears when several loci are considered: under the bottleneck model, all loci share a common *T* and *S*, while distinct loci have distinct *T*'s and *S*'s under the selective sweep hypothesis. In both cases, a specific mutation rate θ is assigned to each locus. Three nested models are therefore to be compared. Suppose that *p* loci are examined:

M_1 (no founder event), *p* parameters: $\theta_1, \theta_2, \ldots, \theta_p$
M_2 (bottleneck), $p + 2$ parameters: $\theta_1, \theta_2, \ldots, \theta_p, S, T$
M_3 (selective sweep), 3*p* parameters: $\theta_1, S_1, T_1, \theta_2, S_2,$
$T_2, \ldots, \theta_p, S_p, T_p$.

The likelihood for a data set of several independent loci is the product of the likelihoods for individual loci. Likelihood-ratio tests can be performed to detect a diversity-reducing event (M_2 *vs.* M_1 and M_3 *vs.* M_1) and to distinguish sweeps from bottlenecks (M_3 *vs.* M_2): twice the logarithm of the ratio of likelihoods of two competing models asymptotically follows a χ^2 distribution with *k* d.f., where *k* is the difference between the numbers of parameters of the two models.

SIMULATIONS

The reliability and efficiency of our method for detecting bottlenecks were assessed using simulated data sets, although an exhaustive power analysis was impossible because of the extensive running time. We first simulated 100 one-locus data sets under the null hypothesis of constant population size (eight genes, $\theta = 10$). The null hypothesis was rejected in 6 cases out of 100, suggesting that the test is reliable. Bottlenecked data sets were simulated under two conditions: old, strong *vs.* recent, weak bottlenecks (see Table 1). The number of rejections of the null hypothesis (power) and the mean and standard errors of estimates of parameters *T* and *S* are shown (Table 1). The power of the test was ~25%, and the parameter estimators were quite imprecise. The power, however, was significantly higher than that of Tajima's (1989) *D*-statistics, sometimes used to detect bottlenecks. When four-locus data sets (simulated under the same conditions) were analyzed, the power of the test and the accuracies of parameter estimates significantly increased (Table 2), suggesting that standard multilocus DNA sequence data sets are large enough to allow a reliable reconstruction of population history.

The above analyses address the M_2/M_1 test, *i.e.*, detecting bottlenecks. The power to detect selective sweeps is more difficult to assess since it depends much on how many loci depart from the null hypothesis. For example, when two-locus data sets including one "neutral" locus plus one bottlenecked locus were analyzed, the rejection rate (M_3/M_1) was 24 out of 100. This power would of course be increased by adding loci with their own *T* and *S*, but decreased by adding "neutral" loci. The M_3/

TABLE 1

Simulations: one-locus data sets

	Actual parameters				Estimations			Rejection rate[c]	
	n	θ	T[a]	S[a]	θ*	T*	S*	M_2/M_1[d]	D[e]
Old, strong Bottleneck	8	10	0.5	1	13.80[b] (6.64)	0.56 (0.40)	1.66 (0.99)	26/100	5/100
Recent, weak Bottleneck	8	10	0.1	0.5	11.35 (5.75)	0.27 (0.16)	0.72 (0.59)	23/100	14/100

[a] In units of 2N generations.
[b] Mean over 100 simulations, standard errors within parentheses.
[c] Number of data sets for which the null was rejected (5% level) out of 100.
[d] Likelihood-ratio test (this article).
[e] Tajima's D-test.

M_1 test is conservative: the maximum rejection rate under the null hypothesis is 5%.

DATA ANALYSIS

The above method has been applied to DNA sequence data obtained from an African population of *D. melanogaster* (Lamto, Ivory Coast). Three loci were used: *Fat Body Protein 2* (*Fbp2*, 2.15 kb, 10 individuals; Benassi *et al.* 1999), *Suppressor of Hairless* [*Su*(*H*), 1 kb, 20 individuals; Depaulis *et al.* 1999], and *Vacuolar H*$^+$ *ATPase* 68-1 (*Vha*, 1 kb, 20 individuals; Depaulis 1998). These genes are located near a region polymorphic for a chromosomal inversion on chromosome 2. This proximity increases the chances of detecting a selective sweep, if any (see Depaulis *et al.* 1999). Loci were sequenced in distinct but overlapping samples of a single population of *D. melanogaster*.

For each data set, sequences were truncated to fit the assumptions of infinite number of sites and no recombination: the largest segment of each locus showing no homoplasy was sought. Sites in such segments are phylogenetically compatible: one can find a genealogy for which the mutants at any site are a monophyletic group. Sites showing more than two distinct states were removed. This data-paring strategy reduced the number of variable sites from 64 to 19 (*Fbp2*), 44 to 40 [*Su*(*H*)], and 11 to 11 (*Vha*), respectively. Sites were oriented: the ancestral/derived status of character states was determined. Orienting sites allows one to sample rooted rather than unrooted genealogies during the likelihood estimation (Griffiths and Tavare 1995), greatly decreasing the running time. To orient sites, we first reconstructed a neighbor-joining phylogenetic tree (Saitou and Nei 1987; observed divergence) and located the root thanks to an outgroup sequence (*D. simulans*). Sites were oriented according to this tree: the monophyletic character state was said to be derived. When both character states defined a monophyletic group (*i.e.*, when the mutation occurred in the branch connected to the root), the state shared by the outgroup was supposed to be ancestral.

The maximum likelihood of the data under three competing models is given in Table 3, together with the parameter estimates. Likelihood-ratio tests favored the hypothesis of a selective sweep (M_3) *vs.* either the no-founder-event model (M_1) or the bottleneck model (M_2). A demographic event seems unlikely to explain the observed pattern, as indicated by the M_2 *vs.* M_1 comparison. The optimal times of occurrence and strengths of variability-reducing events were quite different among loci (model M_3): a very recent, weak sweep was detected for locus *Su*(*H*), a strong, recent one for locus *Vha*, while no significant sweep was found at locus *Fbp2*. As a consequence, the optimal T value under model M_2 is high: no recent bottleneck scenario was found that fits the data better than the simple no-founder-event model. By excluding demographic

TABLE 2

Simulations: four-locus data sets

	Actual parameters				Estimations			Rejection rate:
	n	θ	T	S	θ*	T*	S*	M_2/M_1
Old, strong Bottleneck	8	10	0.5	1	13.12 (4.83)	0.48 (0.18)	1.49 (0.54)	21/25
Recent, weak Bottleneck	8	10	0.1	0.5	10.23 (4.17)	0.17 (0.10)	0.56 (0.24)	20/25

TABLE 3

Analysis of three nuclear loci in an African population of *D. melanogaster*

	Fbp2	*Su*(*H*)	*Vha*	log(*L*)[f]	2 · log(likelihood ratio)[g]
M$_1$[a]	θ: 8.2	θ: 10.2	θ: 6.1	−78.5	
	(−24.7)[d]	(−38.2)	(−15.6)		
M$_2$[b]	θ: 9.8	θ: 12.6	θ: 6.1		
	T[e]: 1.9	*T*: 1.9	*T*: 1.9	−78.2	M$_2$ *vs.* M$_1$: 0.6 (2)
	S[e]: 1.0	*S*: 1.0	*S*: 1.0		
M$_3$[c]	θ: 8.6	θ: 22.0	θ: 8.6		
	T: 1.4	*T*: 0.0	*T*: 0.2	−69.4	M$_3$ *vs.* M$_1$: 18.2* (6)
	S: 1.1	*S*: 0.1	*S*: 1.2		M$_3$ *vs.* M$_2$: 17.6* (4)
	(−23.6)	(−34.8)	(−11.0)		

* Significant (5% level).

[a] No diversity-reducing event model.

[b] Bottleneck model.

[c] Selective-sweep model.

[d] Logarithm of the maximum likelihood for individual loci.

[e] Measured in units of 2*N* generations.

[f] Logarithm of the maximum likelihood for all three loci.

[g] To be compared to a χ^2 distribution (degrees of freedom within parentheses).

hypotheses, this result reinforces the hypothesis that one or several selective sweeps may have occurred recently in this region of chromosome 2 for this African population of *D. melanogaster* (Depaulis *et al.* 1999).

DISCUSSION

The new method presented in this article aims to reconstruct the recent history of a population. It allows detection of diversity-reducing events at one or several loci and bottlenecks to be distinguished from selective sweeps if more than one locus is available. The information dealt with to achieve the former goal is the distortion in gene genealogies generated by diversity-reducing events. The latter issue—distinguishing demographic from selective causes—is addressed by measuring the heterogeneity in time and strength of diversity-reducing events across loci. The maximum-likelihood approach allows one to test hypotheses and to estimate the times and strengths of diversity-reducing events. It is more efficient than methods based on pairwise differences (*e.g.*, Rogers and Harpending 1992) or test statistics (Tajima 1989), which do not make use of all the information contained in the data. Note that this method applies only to panmictic populations. Population structure is likely to introduce bias, especially if samples for distinct loci belong to distinct demes.

Theoretically, it should be possible to distinguish bottlenecks from selective sweeps using a single locus. This is because what is happening during the course of the event is different in both situations. Basically, a bottleneck applies identically to all the lineages that enter it. In the case of a partial selective sweep (where recombination occurred, so that more than one lineage escapes the sweep, *e.g.*, tree B$_1$, Figure 1), the lineage originally associated with the favorable mutation has a particular status: it is older than the other lineages that emerged thanks to recombination at various times during the sweep. Barton (1998) gives a detailed description of the properties of genealogies under a selective sweep. He shows that the discrepancy mentioned above results in distinct expected distributions of the size of "gene families" (see Figure 1 legend) under the two hypotheses. These can readily be distinguished statistically from a sample of 100 genes, with pairwise identity 0.1, provided that the genealogy is known with certainty (compare Figures 8 and 9 of Barton 1998). However, it is not known how far errors in estimating the genealogy from (say) infinite-sites mutation reduce the power of this method. We decided here to neglect this difference and to approximate the effect of a sweep at one locus by that of a bottleneck. We suspect that for many data sets the major part of the information lies in the heterogeneity between loci, rather than in the pattern at individual loci. This intuition would be worth verifying formally.

The method we present does not make use of data from an outgroup, in contrast with, say, the Hudson, Kreitman and Aguadé (HKA) test (Hudson *et al.* 1987). If it has some power in its current form—and our data analysis suggests it actually has some—then this property should be considered a strength. Using outgroup sequence data to estimate some neutral mutation rate involves making disputable assumptions. Any selective force having applied to some of the surveyed loci since the ingroup and the outgroup diverged may bias the estimation of mutation rates. Departure from the molecular clock has been observed in many genes and many taxonomic groups (*e.g.*, Li 1993). It may lead to significant HKA tests even if all the loci under consideration are currently neutrally evolving. If the user believes he has a reliable outgroup, information about it can be incorporated into our method. First, characters can be oriented by deciding that the state observed in the outgroup is the ancestral one. Second, the relative mutation rate of loci can be estimated from the ingroup/outgroup divergence. This would add valuable information and reduce the number of parameters of each model by $p - 1$, where p is the number of loci. Incidentally, this would significantly reduce the running time.

Two assumptions of the present method deserve discussion: the infinite-site mutation model and the no recombination assumption. Both are clearly violated by some data sets. One has to worry about them before using the method—Griffiths and Tavare's algorithm can be applied only if the data are consistent with these assumptions, *i.e.*, if distinct sites support phylogenetically compatible bipartitions of the individuals.

The assumption of an infinite number of sites can be avoided. Kuhner *et al.* (1995) compute the likelihood using a finite-site mutation model in the constant-population size case and use the Metropolis-Hastings algorithm (Hastings 1970) to find the value of θ that maximizes it. This, however, involves exploring a larger space of sets of ancestral states, *e.g.*, including genealogies where identical genes are not monophyletic. The Metropolis-Hastings Monte Carlo Markov chain algorithm is an interesting alternative to Griffiths and Tavare's method for computing likelihoods under the coalescent (*e.g.*, see Wilson and Balding 1998). In the case of the bottleneck model, it may provide an efficient way to maximize the likelihood with respect to T and S.

When its assumptions are more or less met, the infinite-site model (and DNA sequence data) is presumably preferable to the infinite-allele model (and, say, microsatellite data) for the purpose of detecting diversity-reducing events. The latter model is one where each mutation creates a new allele, but where successive mutations in the same lineage "hide" each other. The infinite-site model is better because, in addition to allele frequencies, the number of differences between alleles carries much information. Suppose that a moderate bottleneck occurred very recently in the history of a population, so that no mutations have arisen since T. The expected pattern of allele frequencies is identical to that expected under constant population size and low θ, since the shape of the observable genealogy is a standard one (Figure 2). Sequence data, however, would reveal a large number of segregating sites (*i.e.*, highly divergent alleles), incompatible with the hypothesis of low θ, and therefore would have some power to detect the bottleneck. Microsatellite data, however, are often more variable than sequence data and are more easily collected from a high number of loci. Cornuet and Luikart (1996) devised tests for detecting bottlenecks from allele-frequency data, using the sampling distribution of relevant statistics. Their power analysis indicates that very recent bottlenecks cannot be detected from allele frequencies, consistent with the above argument and with Maruyama and Fuerst (1985).

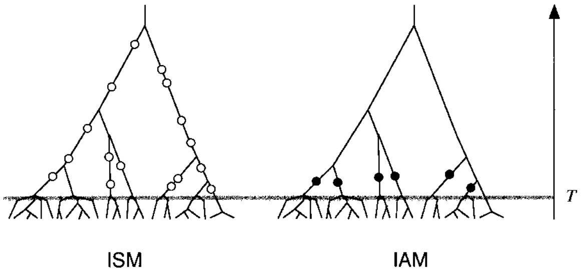

Figure 2.—Infinite-site (ISM) and infinite-allele (IAM) mutation models in the case of a recent bottleneck. (○) Actual mutations that occurred during the genealogy, *i.e.*, mutations dealt with by the ISM. (●) "Mutations" dealt with by the IAM. A method based on the IAM would "see" seven alleles, with frequencies compatible with a constant population size and a low mutation rate. In addition to allele frequencies, a method based on the ISM "sees" a high number of segregating sites, incompatible with a low mutation rate, and can detect the bottleneck.

The assumption of no recombination within loci during the genealogy is a major one. Meeting it involves reducing sequences to blocks whose sites share a unique genealogy and therefore losing information. Furthermore, one can hardly be sure that the length of sequence used is actually nonrecombined, even when no incompatibility between sites is found. Whether undetected recombination events do or do not bias the method is an open question that goes beyond the scope of this article. We doubt, however, that this issue has major practical consequences. This is because the bias, if any, must be higher when data strongly depart from the model assumptions, *i.e.*, when distinct fragments of the surveyed sequence have highly different actual genealogies. But important departures are likely to be detected by the four-gamete rule (see data analysis). Undetected recombination events are more likely to occur when distinct fragments have closely related genealogies, *i.e.*, when the bias is low.

For data sets showing a high number of recombination events, the present method is inapplicable. Actually, such data sets hardly include any genealogical information. Rather, the data can be recoded by pooling together sites of equal "size" (*i.e.*, the number of individuals carrying the mutation), irrespective of which individuals carry the mutation. Coalescence theory allows predictions about the frequency distribution of these classes of sites under various models of population history (Wakeley and Hey 1997). This approach may be applied to the three models we develop in this article, making it possible to detect diversity-reducing events from highly recombining sequence data.

Many thanks to Jody Hey and Kevin Dawson for helpful discussions. This work was supported by grant NERC R34365 and by a postdoctoral fellowship from the Federation of European Biochemical Societies.

LITERATURE CITED

Barton, N. H., 1998 The effect of hitch-hiking on neutral genealogies. Genet. Res. **72:** 123–133.

Benassi, V., F. Depaulis, G. K. Meghlaoui and M. Veuille, 1999 Partial sweeping of variation at the *Fbp2* locus in a West-African population of *Drosophila melanogaster*. Mol. Biol. Evol. **16:** 347–353.

Charlesworth, D., and B. Charlesworth, 1995 Quantitative genetics in plants: the effect of breeding systems on genetic variability. Evolution **49:** 911–920.

Charlesworth, B., D. Charlesworth and M. T. Morgan, 1995 The pattern of neutral molecular variation under the background selection model. Genetics **141:** 1619–1632.

Cornuet, J. M., and G. Luikart, 1996 Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. Genetics **144:** 2001–2014.

Depaulis, F., 1998 Auto-stop en liaison aux inversions chromosomiques chez *Drosophila melanogaster*. Ph.D. thesis, Université Paris 6, France.

Depaulis, F., L. Brazier and M. Veuille, 1999 Selective sweep at the *Drosophila melanogaster Suppressor of Hairless* locus and its association with the *In(2L)t* inversion polymorphism. Genetics **152:** 1017–1024.

Gillespie, J. H., 1991 *The Causes of Molecular Evolution*. Oxford University Press, Oxford.

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

Griffiths, R. C., and S. Tavare, 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. Tavare, 1995 Unrooted genealogical tree probabilities in the infinitely-many-sites model. Math. Biosci. **127:** 77–98.

Hastings, W. K., 1970 Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57:** 97–109.

Hudson, R. R., 1991 Gene genealogies and the coalescent process, pp. 1–44 in *Oxford Surveys in Evolutionary Biology*, Vol. 7, edited by D. Futuyma and J. Antonovics. Oxford University Press, London.

Hudson, R. R., M. Kreitman and M. Aguade, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

Kingman, J. F. C., 1982 The coalescent. Stoch. Proc. Appl. **13:** 235–248.

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.

Li, W. H., 1993 So, what about the molecular clock hypothesis? Curr. Opin. Genet. Dev. **3:** 896–901.

Liu, F., L. Zhang and D. Charlesworth, 1998 Genetic diversity in *Leavenworthia* populations with different inbreeding levels. Proc. R. Soc. Lond. Ser. B. **265:** 293–301.

Maruyama, T., and P. A. Fuerst, 1985 Population bottlenecks and non-equilibrium models in population genetics. II. Number of alleles in a small population that was formed by a recent bottleneck. Genetics **111:** 675–689.

Maynard-Smith, J., and J. Haigh, 1974 The hitch-hiking effect of a favourable gene. Genet. Res. **23:** 23–35.

Nielsen, R., 1998 Maximum likelihood estimation of population divergence times and population phylogenies under the infinite sites model. Theor. Popul. Biol. **53:** 143–151.

O'Brien, S. J., and J. F. Evermann, 1988 Interactive influence of infectious disease and genetic diversity in natural populations. Trends Ecol. Evol. **3:** 254–259.

Press, W. H., S. A. Teukolsky, W. T. Vetterling and B. P. Flannery, 1992 *Numerical Recipes in C*, Ed. 2. Cambridge University Press, Cambridge, United Kingdom.

Rich, S. M., M. C. Licht, R. R. Hudson and F. J. Ayala, 1998 Malaria's Eve: evidence of a recent population bottleneck throughout the world populations of *Plasmodium falciparum*. Proc. Natl. Acad. Sci. USA **95:** 4425–4430.

Rogers, A. R., and H. Harpending, 1992 Population growth makes waves in the distribution of pairwise genetic differences. Mol. Biol. Evol. **9:** 552–569.

Saitou, N., and M. Nei, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:** 406–425.

Tajima, F., 1989 Statistical methods for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.

Wakeley, J., and J. Hey, 1997 Estimating ancestral population parameters. Genetics **145:** 847–855.

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

Communicating editor: A. G. Clark