Ancestral inference on gene trees under selection

Graham Coop*, Robert C. Griffiths
Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK
Received 26 January 2004
Available online 2 September 2004

Abstract

The extent to which natural selection shapes diversity within populations is a key question for population genetics. Thus, there is considerable interest in quantifying the strength of selection. A full likelihood approach for inference about selection at a single site within an otherwise neutral fully linked sequence of sites is described here. A coalescent model of evolution is used to model the ancestry of a sample of DNA sequences which have the selected site segregating. The mutation model, for the selected and neutral sites, is the infinitely many-sites model where there is no back or parallel mutation at sites. A unique perfect phylogeny, a gene tree, can be constructed from the configuration of mutations on the sample sequences under this model of mutation. The approach is general and can be used for any bi-allelic selection scheme. Selection is incorporated through modelling the frequency of the selected and neutral allelic classes stochastically back in time, then using a subdivided population model considering the population frequencies through time as variable population sizes. An importance sampling algorithm is then used to explore over coalescent tree space consistent with the data. The method is applied to a simulated data set and the gene tree presented in Verrelli et al. (2002).

Keywords: Gene trees; Selection; Importance sampling; G6PD; Ancestral inference

1. Introduction

Patterns of genetic diversity within a population around a locus of known biological importance are often investigated to examine the possible role of natural selection in the evolution of the locus. A number of summary statistics have been developed to detect non-neutral evolution in a set of sequences sampled from a population (see Nielsen (2001) for a review). Understanding the strength of natural selection and estimating when the mutation arose can help reveal the historical forces that have been at work. There are only a small number of inference methods to quantify the strength of natural selection on an allele at a particular site that might underlie this non-neutral evolution (see for example Slatkin, 2001; Kim and Stephan, 2002; Przeworski, 2003).

A full likelihood coalescent inference method for a single selected site in an otherwise neutral non-recombining sequence is described in this paper. The joint likelihood curve of the selection parameter and the distributions of the times at which the mutations occurred and the time to the most recent common ancestor can be found by this method, allowing a better understanding of the history underlying the data. The method is flexible and can be extended to any type of single locus biallelic selection. In the process of describing the inference method a simulation scheme is also presented.

The coalescent (Kingman, 1982) presents an intuitive and elegant way to think about the genealogical history of a sample of DNA sequences. It is attractive as it allows the separation of the genealogical history and the neutral mutation process. A genealogy can be generated and then the mutations superimposed on this history (see Nordborg (2001) for a review). The coalescent has been extended to a large number of models including recombination (Hudson, 1983; Griffiths and Majoram, 2004).

*Corresponding author.
E-mail address: coop@stats.ox.ac.uk (G. Coop).
and Barton and Etheridge (2004). Selection significantly complicates the coalescent process as the genealogical and mutational processes are no longer independent. The rate of coalescence depends on the allelic state of the whole population with respect to the selected allele.

Current methods for including selection in to the coalescent can mainly be divided into two approaches that differ in how they deal with the problem of the coalescence rate depending on the frequency of the allele in the population. The first model, the Ancestral Selection Graph, developed by Krone and Neuhauser (1997) and Neuhauser and Krone (1997) effectively integrates out the population frequency by including branching events in the coalescent. However, the simulation of strong selection under this model is difficult. The Ancestral Selection Graph unfortunately can also lack the intuition that made the coalescent so appealing because of the branching events. Progress is being made on inference in this framework (Slade, 2000; Stephens and Donnelly, 2003) but full likelihood inference is not yet feasible.

The second approach explicitly uses the population frequency trajectory of the selected allele. This enables a more intuitive understanding of the process but leads in turn to its own problems. A deterministic approximation to model the sweep to fixation of a positively selected mutation is often used (see for example Przeworski, 2003). Models used by some authors have stochastic sections of trajectory at low frequencies (Kaplan et al., 1989). Kaplan et al. (1988) describes a method to evaluate the expectations of the coalescence times when selection acts at a single site. This approach has been extended to recombination and subdivision (Hudson and Kaplan, 1988; Kaplan et al., 1991), and made rigorous and more tractable by Barton et al. (2004) and Barton and Etheridge (2004).

In this paper the history of a sample of genes together with the population history of the frequency of the selected allele is considered. Our approach is similar to that of Slatkin (2001) in that it uses a fully stochastic trajectory generated backwards in time, and then generates a genealogy given this trajectory. In his approach the tree at the selected site is then used to assess the decline of some measure of linkage disequilibrium away from the selected site. By dealing with selection in a fully stochastic framework instead of a deterministic one, weaker selection can be dealt with. A weakly selected mutation might not allow neutrality to be rejected. However, if a priori it is known to have some biological effect, then its age and an indication of its selection coefficient are still of considerable interest.

A number of authors (see for example Rieder et al., 1999; Fullerton et al., 2000; Saunders et al., 2002; Verrelli et al., 2002) have used the full likelihood gene tree approach on sequence data believed to have experienced natural selection. The estimated age of the candidate selected mutation is examined for inconsistencies from the neutral model. However, the times are biased towards their values under neutrality as the coalescent prior does not include selection. Another method to assess the time of the most recent common ancestor in samples undergoing a selective sweep is the use of a star-like-tree. Under the assumption of a star-like-tree all the time in branches other than singletons is assumed to be negligible (see, for example, Hamblin and Di Rienzo, 2000). This approach can incur biases and underestimate the confidence interval (Rosenberg and Hirsh, 2003).

Our method belongs to a family of importance samplers developed by Griffiths and Tavaré (1994a) to deal with full information in both the finite and infinitely many-sites models of mutation. The method has been extended to variable population size models (Griffiths and Tavaré, 1994b), subdivided populations (Bahlo and Griffiths, 2000; De Iorio and Griffiths, 2004b; De Iorio et al., 2004), and recombination (Griffiths and Majoram, 1996; Fearnhead and Donnelly, 2001). The efficiency of the importance sampling algorithm was improved by Stephens and Donnelly (2000) and a general importance sampling framework based on an approximation of the diffusion generator is developed by De Iorio and Griffiths (2004a). Felsenstein and colleagues have used Markov Chain Monte Carlo (MCMC) techniques to examine similar problems, see for example Felsenstein et al. (1999). Full likelihood MCMC methods have also been developed by Drummond et al. (2002) and Wilson et al. (2003) among others.

The assumption of an infinitely many-sites model of mutation (Watterson, 1975) is made throughout this paper. Under this mutational model, a mutation can occur only once in the genealogical history of the sample. This scheme restricts the application of the method to DNA with a low mutation rate, but allows a unique history to be formed that provides a useful tool for inference. While in this paper the assumption of no recombination or recurrent mutation is made, we believe the approach represents a step forward in constructing a framework that includes selection and to which other processes such as recombination and recurrent mutation can be added at a later date.

The population frequency of the selected allele is modelled by a diffusion process. Simulating population frequency trajectories of a selected allele forward in time is inefficient. Many of the trajectories never reach a frequency where they have any appreciable probability of giving rise to the sample frequency of the selected mutation that is seen in the data. The diffusion process can be thought of back in time using a reversibility argument. The diffusion process is approximated by a
Moran Model, that then can be reversed, to simulate the population frequency trajectory back through time (Section 1.1). By considering the trajectory backward in time, the trajectory can be started from a population frequency simulated from the distribution of the population frequency conditional on its frequency in the sample. The trajectory is then used to simulate a subdivided coalescent using the frequencies through time of the allelic classes as variable population sizes (Section 1.2). This allows the probability of a particular coalescent history to be calculated. An importance sampling scheme is used to integrate over possible histories consistent with the data and find the likelihood with respect to a particular selection parameter (Section 1.5). Applications of the method to simulated data and a real data set are described in Section 3.

1.1. The trajectory of a selected mutation

Let \( \{X(t), t \geq 0\} \) be the trajectory of a diffusion process modelling the frequency of an allele forward in time from when it arose at \( t = 0 \) by a mutation. A general Wright-Fisher diffusion process model for the frequency has a generator

\[
\mathcal{L} = \frac{1}{2} \sigma^2(x) \frac{\partial^2}{\partial x^2} + \mu(x) \frac{\partial}{\partial x}, \tag{1}
\]

where \( \sigma^2(x) = x(1-x) \), and it is assumed that \( x = 0 \) and \( x = 1 \) are (the only) absorbing states. The reader is referred to Karlin and Taylor (1981) for an overview of diffusion processes. Expressing \( \mu(x) = \frac{1}{2} \beta(x)(1-x) \), the model is seen to be one of general frequency dependent selection, where the selection rate of the mutant allele is \( \beta(x) \), a function of \( x \). For example, in the usual genic selection model \( \beta(x) = \beta \), and in a model with heterozygote selection \( \beta(x) = \beta \cdot 2(x+h(1-2x)) \), with \( -\infty < \beta, h < \infty \).

A key idea in understanding the process is reversibility of the trajectory from a current time, before absorption or fixation, back to time 0. Reversibility of the process is studied by Griffiths (2003), and many earlier authors. The reversed process \( \{X^*(t), t \geq 0\} \) is a diffusion process with time measured back from the present time, \( t = 0 \), which has the same distribution as \( \{X(t), t \geq 0\} \) (with time measured forward) conditional on absorption at 0 (Nagasawa and Maruyama, 1979). The time when the mutation arose is the time to absorption in \( \{X^*(t), t \geq 0\} \) and will be denoted by \( \tau \). The generator of \( \{X^*(t), t \geq 0\} \) is similar to \( \mathcal{L} \) with \( \mu(x) \) replaced by

\[
\mu^*(x) = \mu(x) - \frac{\sigma(x)\sigma'(x)}{\int_{0}^{1} s(u) \, du}
\]

(see Ewens, 1973), where

\[
s(y) = \exp\left\{- \int_{0}^{y} \left[ 2\mu(\xi)/\sigma^2(\xi) \right] d\xi \right\}, \quad 0 < y < 1.
\]

For example, in the case of genic selection with \( \beta(x) = \beta \),

\[
\mu^*(x) = - \frac{1}{2} \beta x(1-x) \coth \left( \frac{1}{2} \beta (1-x) \right).
\]

The Moran model is a birth death process with a constant population size (Moran, 1958). The diffusion process can be approximated with a Moran model having a population size of \( N \) genes to simulate a trajectory, see for example Section 4.2 of Griffiths (2003). Then, forward in time, the number of copies of the mutant allele \( \{Z(t), t \geq 0\} \) is governed by a continuous time birth and death process. A fixed population size of \( N \) genes is kept so that at time \( t \) the number of mutant genes is \( Z(t) \) and the number of non-mutant genes is \( N - Z(t) \). The states \( z = 0 \) and \( z = N \) are absorbing states. The Moran model approximating the general diffusion process has birth and death rates that are respectively, with \( x = j/N \),

\[
\lambda_j = \frac{1}{2}N(N\sigma^2(x) + \mu(x)), \quad \mu_j = \frac{1}{2}N(N\sigma^2(x) - \mu(x)). \tag{2}
\]

In the case of frequency-dependent selection

\[
\lambda_j = \lambda^N(1 + \frac{1}{2}s_N(j))(N - j)/N, \quad \mu_j = \lambda^N(1 - \frac{1}{2}s_N(j))(N - j)/N,
\]

where \( \lambda^N = N/2 \) and \( s_N(j) = \beta(x)/2N \). Reproduction in this model is at rate \( \lambda^N(1 - s_N(j))/2 \) for non-mutant gene, where a non-mutant gene is chosen at random to reproduce and a gene (of either type) is chosen at random to die, and similarly with rate \( \lambda^N(1 + s_N(j))/2 \) for mutant genes. In the Moran model the probability of absorption into state 0 from an initial state \( z \) is

\[
u_z = \frac{\sum_{i=1}^{N-1} \prod_{j=1}^{i} \left( \frac{\mu_j}{\lambda_j} \right)}{1 + \sum_{i=1}^{N-1} \prod_{j=1}^{i} \left( \frac{\mu_j}{\lambda_j} \right)}; \tag{3}
\]

see, for example, Karlin and Taylor (1975). We choose to approximate \( \{X^*(t), t \geq 0\} \) by first approximating \( \{X(t), t \geq 0\} \) by a Moran Model as \( \{Z(t), t \geq 0\} \) and then reversing the process to \( \{Z^*(t), t \geq 0\} \). An alternative approach would be to approximate \( \{X^*(t), t \geq 0\} \) directly by simulating from a Moran Model with rates containing \( \mu^*(x) \) instead of \( \mu(x) \). The process \( \{Z^*(t), t \geq 0\} \) approximating \( \{X^*(t), t \geq 0\} \) backward in time has an identical distribution to the conditional process \( \{Z(t), t \geq 0\} \) Absorption at 0 forward in time. It is a birth and death process with rates

\[
\lambda_j = \lambda u_{j+1}/u_j, \quad \mu_j = \mu u_{j-1}/u_j, \tag{4}
\]

with \( u_1 = u_{N+1} = 0 \) by definition (Ewens, 1973). For example in the case of genic selection with \( s_N(j) = \beta/2N \),

\[
u_j = \frac{1 - z^{N-j}}{1 - z^N}, \quad \text{where} \quad z = \frac{1 - \frac{\beta}{2N}}{1 + \frac{\beta}{2N}}.
\]

Thus a method to simulate an approximate trajectory \( \{X^*(t), t \geq 0\} \) is to simulate a sample path of
\( \{Z^*(t), t \geq 0\} \) starting with \( Z^*(0) = [NX^*(0)] \). A continuous trajectory is formed by linear interpolation between points of change in the discrete sample path \( \{ Z^*(t)/N, t \geq 0, N \} \). From this point on we are only concerned with reversed processes, so abuse notation by omitting the \("*\) superscript. Now \( \{ Z(t), t \geq 0\} \) and \( \{ X(t), t \geq 0\} \) will denote reversed processes.

1.2. Simulating a coalescent history

This section describes a method to generate a coalescent genealogy for a sample of \( n + m \) fully linked sequences where \( n \) sequences contain a selected allele with selection coefficient \( \beta \), and \( m \) sequences are selectively neutral. The mutation rate in the complete population is \( \theta = 4N_{eu} \) where \( N_e \) is the effective population size and \( \mu \) is the mutation rate per sequence per generation. Time is measured in a coalescent scale of \( 2N_e \) generations. Note that \( N_e \) is different from the population size \( N \) in an approximating Moran model. \( N \) is chosen to be large enough to obtain a good diffusion process approximation, consistent with the amount of computation and storage required. A simple ‘rule of thumb’, automated in the software, is to use an \( N \) that gives a \( \beta/N \ll 1 \), with a minimum \( N \) of 300. In general, too smaller \( N \) will not approximate the diffusion well, especially near the boundaries, and may lead to an incorrect likelihood surface.

Let \( X \) be the population frequency of the selected allele at the present time. Griffiths (2003) shows that the density of \( X \), conditional on \( n \) and \( m \) is

\[
f_{n,m}(x) = \frac{\binom{n+m}{n} x^n (1-x)^m m(x) u_0(x)}{\int_0^1 \binom{n+m}{n} x^n (1-x)^m m(x) u_0(x) \, dx},
\]

where \( u_0(x) \) is the density of \( \{ X(t), t \geq 0\} \) absorbed at 0 when \( X(0) = x \), and \( m(x) \) is the speed of density of the process. Explicit expressions are \( u_0(x) = \int_x^1 s(y) \, dy / \int_0^1 s(y) \, dy \), and \( m(x) = [\sigma^2(x)s(x)]^{-1} \). The distribution of \( \{ X(t), t \geq 0\} \) is thus a diffusion process with a random initial frequency \( X(0) \) having density (5). The coalescent structure of the two subgroups of \( n \) and \( m \) genes behaves as a subdivided population model with respective variable population size proportions of \( X(t) \) and \( 1 - X(t) \) at time \( t \) back from the current time (Kaplan et al., 1988).

That is, if \( a \) and \( b \) are the number of ancestors of the two groups of genes at time \( t \), then coalescence rates in the two groups are respectively \( \frac{a}{2} X(t)^{-1} \) and \( \frac{b}{2} (1 - X(t))^{-1} \).

While the genealogy is independent of the neutral mutation process, to provide a better understanding of the recursion discussed later, we present here the process to simultaneously generate the genealogy and the pattern of mutation. Events in a coalescent history can be either mutations or coalescences within the respective selected and nonselected ancestral subsamples. The time to the next event \( t \) given that the last event occurred at time \( u \) has a non-homogeneous exponential distribution, because of the dependence on the trajectory, with a density function of

\[
g_{12}(t \mid n, m, u) = \gamma(t, n, m) \exp \left( - \int_u^t \gamma(s, n, m) \, ds \right),
\]

\( u < t < \infty \),

where at time \( s \), \( \gamma(s, n, m) = \gamma_1(s, n) + \gamma_2(s, m) \) is the total event rate and \( \gamma_1(s, n) = \left( \frac{n}{2} \right) X(s)^{-1} + n\theta/2 \), is the event rate in the selected subsample, \( \gamma_2(s, m) = \left( \frac{m}{2} \right) (1 - X(s))^{-1} + m\theta/2 \) is the event rate for the non-selected subsample. An event time is generated by decomposing the distribution into the time to the next event in each subsample, then taking the minimum of these times. The density of the time \( s \) back to the next event in a subsample is

\[
g_i(s \mid u, n) = \gamma_i(s, n) \exp \left( - \int_u^s \gamma_i(v, n) \, dv \right),
\]

\( u < s < \infty \), \( i = 1, 2 \).

To generate a time to the next event in either subsample (7) is further decomposed. There are two simulated times for each subsample. For example, in the selected subsample times to the next mutation or coalescence have respective rates at a time \( v \) of

\[
n\theta \frac{n}{2} \frac{1}{X(v)}.
\]

The time to the next mutation is simple to generate. For the time of the next coalescence, the appropriate cumulative distribution function is partially inverted and then solved numerically. The integral \( \int_0^s X(t)^{-1} \, dt \) is stored at jump positions \( v \) in the approximating Moran model as a lookup table. A uniform random number \( U \) in \([0,1]\) is generated and the value of \( s \) found by solving

\[
-\ln(1 - U) = \int_0^s \frac{1}{X(v)} \, dv - \int_0^u \frac{1}{X(v)} \, dv,
\]

by a bisection lookup routine, then interpolating between jumps of the Moran model. When an event time \( t \) has been generated, an event is chosen to happen with probability proportional to its relative rate at the time \( t \). The probabilities that the event is a mutation or a coalescence in the selected subsample are respectively

\[
\frac{\theta \frac{n}{2} \frac{1}{X(t)}}{\gamma(t, n, m)} \cdot \frac{\left( \frac{n}{2} \right) X(t)^{-1}}{\gamma(t, n, m)},
\]

while the probabilities of a mutation or a coalescence in the non-selected subsample are

\[
\frac{\theta \frac{m}{2} \frac{1}{1 - X(t)}}{\gamma(t, n, m)} \cdot \frac{\left( \frac{m}{2} \right) (1 - X(t))^{-1}}{\gamma(t, n, m)}.
\]

If a mutation occurs, one of the lineages in the subsample is chosen uniformly at random to have a
mutation placed on it. If a coalescence occurs, two lineages are chosen uniformly at random in the subsample to coalesce, decreasing the number of lineages by one.

The selected subsample is guaranteed to find a common ancestor before \( \tau \), the time of the selected mutation, as \( X(\tau) = 0 \) and \( \int_{\tau}^{\infty} X(t)^{-1} dt = \infty \), \( u \geq 0 \). When the trajectory of the selected mutation finishes at time \( \tau \), the selected mutation is placed on the common ancestor lineage of the selected subsample. Once the selected mutation has been added to the genealogy, the one remaining lineage of the selected subsample is added to the non-selected sample. The coalescent process returns to a neutral process for the remainder of the time back to the common ancestor.

The simulation scheme described is for an infinitely many-sites model but can be extended to finitely many-sites models. A method to simulate a recombining sequence based on the stochastic trajectory method described in this and the previous sections has been implemented (Spencer and Coop, 2004).

1.3. Gene trees

A sample of DNA sequences contains information about their ancestry. A unique gene tree can be constructed from the configuration of mutations arising at sites under the assumption of no parallel or back mutation. The gene tree is a perfect phylogeny representing the mutation history back in time of the ancestry of the sample. The tree is equivalent to the DNA sequence data and because of ancestry there is much interest in thinking of the DNA sequence data as a tree.

Gusfield (1991) has an efficient algorithm for constructing a perfect phylogeny from a \( p \times q \) 0-1 incidence matrix \( S \) describing which of \( p \) sequences contain \( q \) observed mutations. Thus \( S_{ij} = 1 \) if sequence \( i \) contains mutation \( j \), or 0 otherwise. Examples of gene trees constructed with Gusfield’s algorithm are shown in Griffiths (2001, 2002).

1.4. Models and inference problems

There are three cases to consider in a gene tree with a single selected mutation.

(a) The selected site is segregating in the sample. This is the main case considered in this paper.
(b) The selected mutation has fixed in the population at some time in the past. Perlitz and Stephan (1997) have dealt with inference of this type based on the number of segregating sites in an infinitely many-sites model with no recombination.
(c) The data have been collected in the subpopulation of sequences with a given mutation of known frequency, or the extension to case-control data. The mutation may be under selection. These data would be inappropriate for a coalescent analysis that assumes random sampling. The methods here can be used to correctly find the likelihood of a gene tree under this method of sampling. For example, consider the case where a sample is taken from a haplogroup defined by a particular neutral mutation (i.e. with \( n > 0 \) and \( m = 0 \)), found at a frequency of \( p \) in the population. To find the likelihood of the gene tree, say as a function of \( \theta \), trajectories of a neutral allele could be simulated back from the frequency \( p \) and then the importance sampling algorithm used to sample histories given these trajectories.

In this paper an algorithm for computing the likelihood surface \( L(\beta) \) from a gene tree is described, where \( \beta \) is a set of selection parameters. We focus on \( \beta \) being a single parameter, either in a genic selection model, or in a heterozygote advantage model, when the parameter \( h \) is known. This surface is specific to a particular mutant site, or sites if more than one mutation occurs on an edge in a gene tree, due to the unidentifiability of mutations along an edge. The maximum likelihood estimate of \( \beta \), denoted by \( \hat{\beta} \), can thus be found. If the effective population size is known then the selection coefficient \( \hat{s} \) can be calculated. Otherwise, \( N_e \) and \( s \) cannot be determined separately. However, \( \beta \) itself is of interest as it gives an indication of the relative effect of genetic drift compared to selection.

The hypothesis that \( \beta \neq 0 \) can be tested using the likelihood ratio statistic

\[
A(\beta) = -2 \ln \left( \frac{L(\hat{\beta})}{L(\beta = 0)} \right). \tag{12}
\]

This may however not be distributed as \( \chi^2 \), due to the highly correlated nature of the data through a common ancestry. The approximate rejection region for the likelihood ratio statistic can be found through simulation under the neutral hypothesis that \( \beta = 0 \).

The distributions of ages of mutations and the time to the most recent common ancestor of the sample (TMRCA) can be found. These times are of interest, particularly the age of the selected mutation.

1.5. The likelihood of a gene tree

The likelihood function \( L(\beta) = P(\mathcal{D} \mid \beta) \) is the probability of the data \( \mathcal{D} \), represented as a gene tree, treated as a function of \( \beta \). In this paper we assume for simplicity that \( \theta \) is known and that the emphasis is on estimating selection. In reality, likelihood surfaces for \( \beta \) for fixed values of \( \theta \) may need to be considered.
The likelihood can be expressed as
\[
L(\beta) = \int P(\mathcal{D} \mid \mathcal{H}) P_\beta(\mathcal{H}) \, d\mathcal{H},
\]
where \( \mathcal{H} \) is the full coalescent history including information about the time of mutations and coalescences. \( \mathcal{H} \) can be regarded as missing data in representation (13). \( P(\mathcal{D} \mid \mathcal{H}) \) is an indicator function of whether \( \mathcal{H} \) is compatible with the gene tree. The distribution of a coalescent history \( P_\beta(\mathcal{H}) \) for a given \( \beta \) value is described in Section 1.2.

A naive Monte Carlo approach to this would be to simulate \( M \) trees as described in Section 1.2, then count how many match the data to find a likelihood estimate
\[
L(\beta) \approx \frac{1}{M} \sum_{i=1}^{M} P(\mathcal{D} \mid \mathcal{H}^{(i)}),
\]
where \( \mathcal{H}^{(1)}, \ldots, \mathcal{H}^{(M)} \) are independent samples from \( P_\beta(\mathcal{H}) \). However, the space of possible genealogies is vast and to evaluate the likelihood of a particular parameter by simply sampling from tree space and see how many are compatible with the data is not feasible. Most if not all of the histories generated would be incompatible with the data, thus this method would be highly inefficient. A computationally efficient method is to focus on genealogies that are compatible with the data. This approach is followed in importance sampling, see Stephens (2001) for an introduction. The likelihood can be expressed as
\[
L(\beta) = \int \frac{P_\beta(\mathcal{H})}{Q_\beta(\mathcal{H})} Q_\beta(\mathcal{H}) \, d\mathcal{H},
\]
where \( Q_\beta(\mathcal{H}) \) is a proposal distribution that has weight only on genealogies where \( P(\mathcal{D} \mid \mathcal{H}) = 1 \). Thus samples from \( Q_\beta(\mathcal{H}) \) are always compatible with the data. The ratio \( P_\beta(\mathcal{H})/Q_\beta(\mathcal{H}) \) is the importance weight. An approximation to (15) is
\[
L(\beta) \approx \frac{1}{M} \sum_{i=1}^{M} \frac{P_\beta(\mathcal{H}^{(i)})}{Q_\beta(\mathcal{H}^{(i)})},
\]
where \( \mathcal{H}^{(1)}, \ldots, \mathcal{H}^{(M)} \) are independent samples from \( Q_\beta(\mathcal{H}) \). \( Q_\beta(\mathcal{H}) \) is chosen as the distribution of a reverse time Markov process on histories with an initial state the sample data. Sample paths under \( Q_\beta(\mathcal{H}) \) are thus always compatible with the observed data configuration. If the imbedded Markov chain of history changes is \( \{\mathcal{H}_j; j = 0, 1, \ldots, m\} \), with \( m \) the number of events in the history, then the importance weight can be expressed as
\[
\frac{P_\beta(\mathcal{H})}{Q_\beta(\mathcal{H})} = \left[ \prod_{j=1}^{m} \frac{P_\beta(\mathcal{H}_{j-1} \mid \mathcal{H}_j)}{P_\beta(\mathcal{H}_{j-1} \mid \mathcal{H}_{j-1})} \right] \cdot P(H_m),
\]
where \( \mathcal{H}_0 = \mathcal{D} \), the initial data (Griffiths and Tavaré, 1994a). \( P_\beta(\mathcal{H}_{j-1} \mid \mathcal{H}_j) \) and \( Q_\beta(\mathcal{H}_{j-1} \mid \mathcal{H}_{j-1}) \) are transition probabilities of history state changes. \( P_\beta(\mathcal{H}_{j-1} \mid \mathcal{H}_j) \) are the transition probabilities of a Markov chain forward in time. While the coalescent is usually considered backwards in time, the transition probabilities of the history state chain are only known forward in time. \( Q_\beta(\mathcal{H}_{j-1} \mid \mathcal{H}_{j-1}) \) are transition probabilities in reverse time, taking the form of approximately time-reversed distributions \( P_\beta(\mathcal{H}_{j-1} \mid \mathcal{H}_j) \). See Stephens and Donnelly (2000) and Sections 2 and 3 of De Iorio and Griffiths (2004a) for discussion. The ratios in (17) are sequential importance sampling weights. We can regard the stochastic process generating \( \mathcal{D} \) as consisting of two levels of missing data:

(a) the trajectory of the selected mutation \( \{X(t), t \geq 0\} \); and

(b) the coalescent history \( \mathcal{H} \) giving rise to the data.

That is, from (a),
\[
P(\mathcal{D} \mid \beta) = \mathbb{E}[P(\mathcal{D} \mid \{X(t), t \geq 0\})]
\]
\[
\approx \frac{1}{R} \sum_{i=0}^{R} P(\mathcal{D} \mid \{X(t), t \geq 0\}^{(i)}),
\]
where \( \{X(t), t \geq 0\}^{(i)} \) are independent copies of the selected allele’s trajectory, generated by the technique in Section 1.1 with a given value of \( \beta \). The distribution of \( \mathcal{D} \) given \( \{X(t), t \geq 0\} \) does not depend on \( \beta \); all the information about \( \beta \) is contained in the trajectory \( \{X(t), t \geq 0\} \). In a further decomposition from (b) into the coalescent history
\[
P(\mathcal{D} \mid \beta) = \mathbb{E}[P(\mathcal{D} \mid \mathcal{H})P(\mathcal{H} \mid \{X(t), t \geq 0\})]
\]
\[
\approx \frac{1}{RM} \sum_{i=0}^{R} \sum_{j=1}^{M} P(\mathcal{D} \mid \mathcal{H}^{(i,j)})
\]
\[
\times P(\mathcal{H}^{(i,j)} \mid \{X(t), t \geq 0\}^{(i)}),
\]
for \( R \) independent trajectory copies, and \( M \) independent history copies for each trajectory. The probability \( P(\mathcal{D} \mid \{X(t), t \geq 0\}) \) is calculated by sequential importance sampling on the coalescent history \( \mathcal{H} \) given \( \{X(t), t \geq 0\} \).

### 1.6. History states of a gene tree

In the evolution of a gene tree a history state \( \mathcal{H}_j \) is described by \( (\mathcal{T}, n, m) \), where \( \mathcal{T} \) is the ancestor’s gene tree topology back in time from when \( \mathcal{H}_j \) is observed, and \( n = (n_1, \ldots, n_k) \) and \( m = (m_1, \ldots, m_l) \) are multiplicities of \( k \) types with the selected mutation and \( l \) types without the selected mutation. Fig. 1a illustrates a gene tree configuration. Mutation 12 is the selected mutation, \( n = (1, 9) \) and \( m = (1, 2, 3, 2, 2) \). Mutations on an edge are exchangeable, for example mutation 12 could be above or below mutation 10. While the full method accounts for the exchangeability by summing over the possible positions of the selected mutation on the edge, for the sake of clarity we assume throughout this paper that the lowest mutation on the edge is the selected mutation.
The topology $\mathcal{T}$ is described by mutation paths from the leaves of the tree at the current time, to the ancestral root of the tree, labelled 0. In Fig. 1a the path to the root for the first type on the left of the tree is (4,2,0). Let $(\mathcal{T}(t),n(t),m(t))$ denote the ancestral gene tree at time $t$ back from the current time and

$$p(\mathcal{T},n,m,t) = \mathbb{P}[(\mathcal{T}(t),n(t),m(t)) = (\mathcal{T},n,m,t)].$$

The generator $\mathcal{A}$ for the Markov process $\{(\mathcal{T}(t),n(t),m(t)), t \geq 0\}$ satisfies

$$\frac{d}{dt}p(\mathcal{T},n,m,t) = \mathcal{A}p(\mathcal{T},n,m,t).$$

The right-hand side of (20) can be expressed as

$$\mathcal{A}p(\mathcal{T},n,m,t) = \gamma(t,n,m)(\mathcal{A} - I)p(\mathcal{T},n,m,t).$$

noting that the total rate of change at time $t$ is $\gamma(t,n,m)$. $\mathcal{A}$ is an operator related to the transition probability matrix of the embedded jump chain of the history and $I$ denotes the identity operator.

Let $P((\mathcal{T},n,m) | (\mathcal{T},n,m),t]$ be the transition probability matrix of a history change forward in time in the tree from a state $(\mathcal{T},n,m)$, given an event at $t$. Then

$$\mathcal{A}p(\mathcal{T},n,m,t) = \sum_{(\mathcal{T},n,m')} p((\mathcal{T},n,m) | (\mathcal{T},n,m),t] \times p(\mathcal{T},n,m,t+0).$$

Considering $u$ as the time to the next event back in time from a current time $t$, then

$$p(\mathcal{T},n,m,u) = \int_0^\infty \mathcal{A}p(\mathcal{T},n,m,t)g_{12}(t | n,m,u) dt.$$  

An integral recursion follows from (23), similar to Eq. (14) in Griffiths and Tavaré (1994b).

Summing over possible one-step history changes at time $t$ which lead to a current configuration of $(\mathcal{T},n,m)$, as in (22), gives

$$\mathcal{A}p(\mathcal{T},n,m,t) = \gamma(t,n,m)\frac{1}{2} \left\{ \begin{array}{c} \frac{n}{2} \sum_{i \leq j} \frac{1}{X(i)} \sum_{\{m_j \geq 2\}} (m_i - 1) m_j - 1 p(\mathcal{T},n-e_i,m,t) \\ + \frac{m}{2} \sum_{\{m_j \geq 2\}} (m_i - 1) \sum_{\{m_j \geq 2\}} (m_i - 1) m_j - 1 p(\mathcal{T},n-m-e_i,m,t) \end{array} \right\},$$

where $e_i$, the $i$th unit vector, represents a multiplicity of 1 for an allele of type $i$. An abbreviated notation $n_i,m_i$ (and similar other notation) is used to denote either $n_i$ or $m_i$ depending on whether haplotype $i$ belongs to the selected or neutral class of sequences. The probability of a mutation or a coalescence given that an event occurred is found by a relative rate argument as before; see Section 1.2. Examples of possible state changes of a tree back in time from a history state Fig. 1a are shown in Fig. 1b–d. The first and second terms on the right of (24) relate to removal of a mutation. There are two cases to consider when a mutation is removed. In the first case removal of the mutation on the singleton lineage of the $i$th haplotype leaves the lineage distinct. The resulting gene tree topology is denoted by $\mathcal{T}_{i-}$. The probability that a mutation occurred on this particular lineage forward in time, given that a mutation did occur, is $1/[n,m]$. Removal of mutation 4 in the example data set leaves the lineage distinct and results in the gene tree shown in Fig. 1b. In the second case, which is more complex, removal of the mutation on the singleton lineage of the $i$th haplotype results in a lineage that is non-distinct from the lineages of haplotype $j$. The resulting gene tree topology is denoted by $\mathcal{T}_{i-j+}$. The probability of this configuration change, conditional on
a mutation occurring, is the probability that prior to
mutation the $i$th haplotype had multiplicity $\{n_i, m_i\} + 1$
and was chosen to mutate. In the example data set,
removal of mutation 14 leaves a lineage identical to the
lineages subtended by mutation 12, illustrated in Fig. 1c.
The third and fourth terms on the right of (24) relate to
coalescence in the selected and non-selected groups. The
probability of coalescence occurring within the group of
haplotype $i$, conditional on coalescence occurring, is the
probability that prior to coalescence the $i$th haplotype
had multiplicity $\{n_i, m_i\} - 1$, and that one of these
lineages from a possible $\{n, m\} - 1$ was a parent in the
coalescent event. This is illustrated in Fig. 1d. The
boundary condition is the tree $((0), \{n, m\} = e)$, a
singleton root, and with probability 1 this is of the type
ancestral to the sample.

Bahlo and Griffiths (2000) study the distribution of
gene trees in a subdivided population model with
recurrent migration. They obtain an expression similar
to (24); however, there is no concept of migration
between the two selective classes in our case. If
recombination or repeat mutation were to be included
they would act as migration between the two classes as
discussed in Barton and Etheridge (2004).

2. Importance sampling conditional on a trajectory

The procedure to generate importance samples from
the proposal distribution is detailed in this section. It
follows a scheme similar to that described in Section 1.2,
with sequential importance sampling weights depending
on the time of history changes. The first step is
generating a trajectory for the selected allele back in
time, conditional on observing $n$ sequences of the
selected type, and $m$ neutral sequences. The initial
frequency $X(0) = x$ is chosen from the posterior density
$f(x \mid n, m, \beta)$; however, an importance weight is necessa-
riely generated from this choice of $X(0) = x$. Calculation
of the weight is detailed in the Appendix A. A Moran
model sample path approximation to a diffusion
trajectory $\{X(t), t \geq 0\}$, conditional on absorption at 0,
is simulated. This supplies the trajectory as missing data
for the gene tree evolution back in time. The distribution
of the trajectory is a Markov process with generator $L$
described in Section 1.1.

A subdivided population coalescent history $\mathcal{H}$
describing a gene tree sample path $\{(\mathcal{F}(t), n(t), m(t)),
\ t \geq 0\}$ is simulated back in time conditional on the data $\mathcal{D}$,
and the trajectory $\{X(t), t \geq 0\}$. The joint density of the
time and the subpopulation is detailed in column one of
Table 1. Informally, from a given state with $n, m$
sequences potential times to the next events in each
subpopulation are generated, and then the next event is
chosen to occur in the subpopulation with the minimum
time generated. This generates a time $t$ from $g \mathcal{H}(t \mid
n, m, u)$ as in Section 1.2. By simulating a time from $g \mathcal{H}(t \mid
n, m, u)$ we take a sample from the distribution being
integrated over in Eq. (23) and thus the proposal on a
gene tree sample path is correct in respect to the times.

Table 1

<table>
<thead>
<tr>
<th>Joint density of $t$ and subsample</th>
<th>$\mathcal{H}_k$</th>
<th>$Q(\mathcal{H}<em>k \mid \mathcal{H}</em>{k-1})$</th>
<th>$P(\mathcal{H}_{k-1} \mid \mathcal{H}_k)$</th>
<th>Importance weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A) Events within the selected subsample $n$</td>
<td>$\mathcal{F}_{n, n, m}$</td>
<td>$1/r_1$</td>
<td>$\frac{s_n}{2Li(\alpha_0, \beta_0, \theta)} g \mathcal{H}(t \mid n, m, u)$</td>
<td>$\frac{r_0}{2Li(\alpha_0, \beta_0, \theta)}$</td>
</tr>
<tr>
<td>$\mathcal{T}_{n-1, n, m - e_i, m}$</td>
<td>$1/r_1$</td>
<td>$\frac{s_n}{2Li(\alpha_0, \beta_0, \theta)} g \mathcal{H}(t \mid n, m, u)$</td>
<td>$\frac{r_0}{2Li(\alpha_0, \beta_0, \theta)}$</td>
<td></td>
</tr>
<tr>
<td>$\mathcal{T}_{n, n - e_i, m}$</td>
<td>$n_i/r_1$</td>
<td>$\frac{s_n}{2Li(\alpha_0, \beta_0, \theta)} g \mathcal{H}(t \mid n, m, u)$</td>
<td>$\frac{r_0}{2Li(\alpha_0, \beta_0, \theta)}$</td>
<td></td>
</tr>
<tr>
<td>(B) Events within the unselected subsample $m$</td>
<td>$\mathcal{F}_{m, n, m}$</td>
<td>$1/r_2$</td>
<td>$\frac{s_m}{2Li(\alpha_0, \beta_0, \theta)} g \mathcal{H}(t \mid n, m, u)$</td>
<td>$\frac{r_0}{2Li(\alpha_0, \beta_0, \theta)}$</td>
</tr>
<tr>
<td>$\mathcal{T}_{n, m - e_i, m}$</td>
<td>$1/r_2$</td>
<td>$\frac{s_m}{2Li(\alpha_0, \beta_0, \theta)} g \mathcal{H}(t \mid n, m, u)$</td>
<td>$\frac{r_0}{2Li(\alpha_0, \beta_0, \theta)}$</td>
<td></td>
</tr>
<tr>
<td>$\mathcal{T}_{n - 1, n, m - e_i}$</td>
<td>$m_i/r_2$</td>
<td>$\frac{s_m}{2Li(\alpha_0, \beta_0, \theta)} g \mathcal{H}(t \mid n, m, u)$</td>
<td>$\frac{r_0}{2Li(\alpha_0, \beta_0, \theta)}$</td>
<td></td>
</tr>
</tbody>
</table>

Subtable (A) gives details of events within the selected subsample $n$. Subtable (B) gives details of events in the non-selected subsample $m$. The first column gives the joint proposal density of time $t$ and the subsample. The second gives the new history state. The proposal probability of the event
within the subsample is shown in the third. The forward joint density of the event and the event time $t$ (i.e. the term from (24)) are in the fourth. Finally, the fifth column is the importance weight associated with that particular event and time. The time is distributed as $g \mathcal{H}(t \mid n, m, u)$ for both the proposal and the true density and thus no term for the time appears in the weight.
An event is chosen to occur in the subpopulation with the minimum generated time to the next event. This is equivalent to choosing a subpopulation 1 or 2 with probability proportional to rate $\gamma_1(t,n)$ or $\gamma_2(t,n)$, respectively at time $t$.

As before the time of the removal of the selected mutation from the tree is fixed at $\tau$ as, under the infinitely many-site assumption, the mutation must have occurred at the time in the history of the sample that it occurred in the population history. A special case occurs when at a time $t<\tau$ there are no possible events in the proposal distribution for either one or both subsamples, until removal of the selected mutation at time $\tau$. This can occur when the most recent common ancestor of the selected subsample has been reached or no further coalescense or mutation events can occur in the non-selected subsample compatible with the gene tree topology. The distribution of the time of the event is then no longer correctly generated by $g_{12}(t \mid n,m)$, and a correction importance sampling weight, detailed in the Appendix B, is required for this case. The next event after removal of the selected mutation is generated starting from time $\tau$.

The infinitely many-sites proposal distribution of Stephens and Donnelly (2000) is used within the chosen subsample to choose an event. The proposal distribution of Stephens and Donnelly (2000) is uniform on the choice of available lineages. An available lineage is one where there is more than one of that type, or a singleton where the removal of the mutation would not violate the gene tree topology. Denote the number of available lineages in the selected and neutral subpopulations by $r_1$ and $r_2$. Under the infinitely many-sites assumption once a lineage has been chosen only one event is possible: if there are others of the same type a coalescense must take place; or if the lineage is a singleton then the mutation is removed. The importance weights associated with sampling are shown in Table 1.

After the removal of the selected mutation, $m$ is set to $m+1$, and $n=0$ as the previously selected lineage becomes neutral. The neutral gene tree ancestral process then behaves as a constant-sized population process, with the coalescence rate while $l$ lineages being $\binom{l}{2}$, and the infinitely many-sites importance sampling algorithm of Stephens and Donnelly (2000) is used to complete the sample path simulation.

2.1. Computational features

A number of features have been incorporated to improve the computational performance of the likelihood algorithm. The independence of the history under the selected site and the rest of the history given the trajectory subdividing the two populations, is used to significantly lower the variance of the estimated likelihood. A number of trees are generated conditional on the same trajectory to improve computational efficiency. The usual importance sampling procedure to find the likelihood (19) would be to average the importance weight $W(\cdot \mid \cdot)$ of each history

$$L(\beta) \approx \frac{1}{RM} \sum_{i=1}^{R} \sum_{j=1}^{M} W(\mathscr{H}^{(i)} \mid [X(t), t \geq 0]^{(i)}),$$

(25)

where $M$ trees are sampled per trajectory and there are $R$ trajectories in total. However, using the independence of the two subtrees, we instead take the sum

$$L(\beta) \approx \frac{1}{RM} \sum_{j=1}^{M} \left( \sum_{i=1}^{R} W(\mathscr{H}^{(ij)} \mid [X(t), t \geq 0]^{(ij)} \right) \times \left( \sum_{j=1}^{M} W(\mathscr{H}^{(u,j)} \mid [X(t), t \geq 0]^{(u,j)}) \right),$$

(26)

where $\mathscr{H}^{(ij)}$ is the subtree history under the selected mutation not including the removal of the selected mutation, and $\mathscr{H}^{(u,j)}$ is the complementary neutral subtree history which includes removal of the selected mutation. The approach in effect permutes all possible combinations of the histories under the selected mutation and the rest of the history. This results in a reduced sampling variance for $P(\mathscr{D} \mid [X(t), t \geq 0])$ allowing (26) to converge more quickly than (25).

3. Results

Using the simulation method described in Section (1.2) a data set was simulated with $\beta = 20$ and $\theta = 5$. The gene tree for this data set, with the selected mutation 13, is shown in Fig. 2. The profile likelihood curve for $\beta$ with $\theta = 5$ is shown in Fig. 3. Each point on the curve was independently evaluated with 5 million iterations of the importance sampling algorithm. Each point takes 5 min to evaluate on a 2.4-GHz machine. Independent runs to evaluate the curve
Malaria is believed to have become endemic in human populations in the past 10,000 years. Thus it has been suggested that the $A -$ $G6PD$ allele has arisen recently and is currently sweeping through the sub-Saharan Africa population. A number of population genetics studies have focussed on the $G6PD$ region to examine the effect of this putative selective sweep in progress (Tishkoff et al., 2001; Saunders et al., 2002; Verrelli et al., 2002; Sabeti et al., 2002). Here we focus on the gene tree for the African data presented in Verrelli et al. (2002), reproduced in Fig. 5. Verrelli et al. (2002) removed a haplotype that appeared twice in the data set to make the data compatible with the infinitely many-sites and no recombination assumption. The haplotype removed was not from the $A -$ clade nor does it appear to be a recombinant with a sequence from the $A -$ clade. Thus the removal of this haplotype will have little effect on the estimated age of the $A -$ allele. The $A -$ allele is the result of a single amino acid change, mutation $2$ in Fig. 5. For simplicity, we shall ignore any possible negative selection against $A -$ and take the selective advantage of $A -$ to be the same in hemizygous males and both heterozygous and homozygous females. This is the model suggested by Ruwende et al. (1995) of dominant selection in females and genic selection in males. This gives

$$
\mu_F(x) = \beta x (1 - x)^2, \quad \mu_M(x) = \frac{\beta}{2} x (1 - x),
$$

where $\mu_F(x)$ and $\mu_M(x)$ are infinitesimal drift parameters in females and males respectively. Ignoring any negative selection against $A -$ is reasonable, since given its probable recent increase in frequency there will be little information about its effects in the sequence diversity. However, it is worth noting that the general selective scheme possible using this method would allow this to be explored in a more complete analysis. The X chromosome spends on average $\frac{3}{4}$ of its time in the female and $\frac{1}{4}$ in the male; thus a heuristic assumption is made that the total infinitesimal drift coefficient is

$$
\mu(x) = \frac{2}{3} \mu_F(x) + \frac{1}{3} \mu_M(x).
$$

The likelihood of the selection parameter under this selective scheme was evaluated for the Verrelli et al. (2002) gene tree using the importance sampling method described in this paper. The full maximum likelihood estimate of $\theta$ is $5.35$ when $\beta = 0$. The log likelihood curve of $\beta$ was evaluated when $\theta = 5.35$ and is shown in Fig. 4. A more complete analysis would explore a likelihood surface over $\theta$ as well, but given the low frequency of the $G6PD$ allele its selection coefficient is likely to have a weak effect on the total time in the tree, and hence will have little effect on the estimate of $\theta$. Each point on the curve was independently evaluated by 5 million iterations of the importance sampling
selection coefficient of the allele. Under the assumption recently there is relatively little information about the allele that the allele segregates, as no mutations have occurred on the background and thus, given that the allele segregates, it seems to have a horizontal asymptote. This is unsurprising as no mutations have occurred on the allele background and thus, given that the allele segregates, \( \beta \) could be the maximum likelihood estimate. The hypothesis that \( \beta = 0 \) is unlikely to be rejected, but given that the allele \( A^- \) is at low frequency and thus arose recently there is relatively little information about the selection coefficient of the allele. Under the assumption of 20 years per generation and \( N_e = 19,800 \), the age in years of the various events in the genealogical history may be estimated. The mean and standard error of the age of the selected mutation, and the TMRCA, are shown in Table 2 in coalescent and year time units for \( \beta = 0, 50, 100, 200 \). The full gene tree with estimated times using a selection parameter \( \beta = 200 \) is shown in Fig. 5. The estimate of the age of the \( A^- \) allele is strongly affected by selection. The TMRCA is relatively unaffected by selection on the \( A^- \) allele supporting the assumption that selection on a low-frequency allele has little effect on the time scale of the tree, and hence the estimate of \( \theta \).

### 4. Discussion

A gene tree is completely equivalent to the data, the configuration of mutations in a sample of DNA sequences, under the assumptions of infinitely many-sites mutation model and no recombination. The method presented here allows full likelihood inference to be performed on a gene tree with a single selected mutation and all other mutations evolving neutrally.

In this paper, the main focus has been on obtaining the likelihood of the selection parameter \( \beta \). However, there is considerable interest in estimating features of the genealogical history such as times underlying the gene tree. The empirical distributions of ages of the mutations, of clades under mutations and the most recent common ancestor of the sample can be evaluated by constructing histograms of the weights (Griffiths, 2001).

As in many population genetics inference problems it is likely that the information increases more rapidly with
$\theta$, i.e. with the sequence length than with the sample size. Additional mutations on the tree help to better resolve the shape of the underlying tree while additional sample members will coalesce rapidly with the existing genealogy providing little information. When the selected allele is at low frequency in the population there will be little to distinguish it from a neutral allele as both will be young. In the case of a partial selective sweep, if $\theta$ is known, then the approach will be most powerful when the allele has just reached high frequency in the population. However, when $\theta$ is unknown there is a compromise between the information about $\theta$ provided by the non-selected subsample and the information provided by the allele being at a high frequency.

A general framework for inference on parameters in a subdivided random environment coalescent process has been developed here. This approach allows an intuitive understanding of the incorporation of natural selection into the coalescent. The generality of the diffusion process and the coalescent would allow a wide variety of random background models, where the generator of the diffusion is known, to be simulated from and inference to be performed. For example, the method offers an inference framework into which recombination and recurrent mutation could be incorporated at a later date.

The importance of fluctuations in frequency away from the deterministic approximation to a trajectory of a selected allele is not well understood and will often vary depending on the type of selection considered. The computationally efficient fully stochastic treatment of selection and the coalescent described in this paper allows these fluctuations to be incorporated. The method is valid for a large number of selective schemes and strengths of selection, thus the method offers a viable alternative to the ancestral selection graph and deterministic trajectories.

4.1. Software

An implementation of the importance sampling algorithm described here is available from http://www.stats.ox.ac.uk/mathgen/programs.html A web-based program for simulation using the selected allele trajectory scheme described here for recombining sequences can also be found at this site.

Acknowledgments

The authors would like to thank Simon Myers and Chris Spencer for helpful discussions throughout this research, and the referees for valuable comments. GC and RCG were supported by an EPSRC Grant GR/N02153.

Appendix A

A.1. The trajectory starting frequency

The initial trajectory frequency $X(0)$ is sampled from density (5), conditional on $n$ sample sequences from a sample of $n+m$ sequences containing the selected mutation. However, the joint probability distribution of $X(0)$ and a sample with $(n,m)$ such sequences is required. The initial unconditional prior distribution of the frequency of the selected mutation can be taken as the frequency spectrum of the mutation in the population,

$$g(x) \propto m(x)u_0(x).$$

This is an improper prior. The correct joint probability of the sample and the density of the population frequency is then proportional to

$$\binom{n+m}{n} x^n (1-x)^m m(x)u_0(x). \quad (29)$$

The importance weight when $X(0) = x$ is then $f_{n,m}(x)$, which is equal to

$$\binom{n+m}{n} \int_0^1 x^n (1-x)^m m(x)u_0(x) \, dx. \quad (30)$$

One interpretation is that the selected site is a random choice of site chosen from many selected segregating sites along the sequences. Another is that the selected site is a specific site, and sampling is chosen at a uniform random time between when the mutation arose and before it became lost or fixed in the population. Then, conditioning on the selected site as a specific site segregating in the sample the frequency spectrum of the selected site in the sample is

$$\frac{\int_0^1 \binom{n+m}{n} x^n (1-x)^m m(x)u_0(x) \, dx}{\int_0^1 (1-x)^{n+m} m(x)u_0(x) \, dx}, \quad (31)$$

and the importance weight is this sample frequency spectrum. The general sample frequency spectrum is studied in Griffiths (2003).

Appendix B

B.1. Removal of the selected mutation

There may be no possible events in the proposal distribution before the selected mutation is removed at time $t$ if either $r_1 = 0$ or $r_2 = 0$. Event times can no longer be generated from the distribution $g_5(t \mid n, m, u)$ as there are no proposed events in one or both of the subsamples. If $r_i = 0$ then we generate a time from $g_i(t \mid n, u)$ $i \neq j$. When an event time is chosen an event is chosen using the Stephens and Donnelly (2000) proposal scheme as before. This sampling of $t$ must be corrected.
The different ways either or both subtrees can run out of events at a time $u$ are given in the first column. The history one event back in time is shown in the second column ($\mathcal{F}, \mathbf{n}, \mathbf{m}$ and $\mathcal{F}, \mathbf{n}, \mathbf{m}$ denote the histories given in parts A and B of Table 1, respectively). The time chosen for the next event is also given in the second column ($t < \tau$). The proposal density of the time is shown in the third. The joint forward history probability of the event and the time density is given in the fourth ($P(\cdot | \cdot)$ and $W(\cdot | \cdot)$ refer to the $P(\mathcal{F}_{k-1} | \mathcal{F}_k)$ and the importance weights entries in Table 1 respectively). Finally the importance weights associated with the different situations are given in the last column.

The importance weights for the two cases when either $r_1$ or $r_2$ are zero, these are detailed in the first two rows of Table 3. If an event in the non-selected subsample is proposed at a time $> \tau$, then the selected mutation must be removed at this time. This case is detailed in the third row. The fourth row has details of the case when there are no proposed events to perform ($r_1 = r_2 = 0$), and the next event is removal of the selected mutation.

For example, if the selected subsample has found its common ancestor at $u$ and the non-selected subsample still has events compatible with the gene tree topology, to be performed, event times will be generated only from $g_2(t | u, m)$ and events chosen in that subsample. Every new $t$ generated incurs the importance weight given in the final column of row 1 of Table 3. If the time $t$ is $> \tau$ then the mutation is removed and the weight given in the fourth row is incurred.

The importance weights described in this section also have a clear probabilistic interpretation. Informally, the weights can viewed as the probability of no events occurring in the subsample history in a subsample that has performed all the events permitted to it.


