

# *Statistical Applications in Genetics and Molecular Biology*

---

*Volume 7, Issue 1*

2008

*Article 32*

---

## Importance Sampling for the Infinite Sites Model

Asger Hobolth\*

Marcy K. Uyenoyama<sup>†</sup>

Carsten Wiuf<sup>‡</sup>

\*Aarhus University, asger@daimi.au.dk

<sup>†</sup>Duke University, marcy@duke.edu

<sup>‡</sup>Aarhus University, wiuf@daimi.au.dk

# Importance Sampling for the Infinite Sites Model\*

Asger Hobolth, Marcy K. Uyenoyama, and Carsten Wiuf

## Abstract

Importance sampling or Markov Chain Monte Carlo sampling is required for state-of-the-art statistical analysis of population genetics data. The applicability of these sampling-based inference techniques depends crucially on the proposal distribution. In this paper, we discuss importance sampling for the infinite sites model. The infinite sites assumption is attractive because it constrains the number of possible genealogies, thereby allowing for the analysis of larger data sets. We recall the Griffiths-Tavaré and Stephens-Donnelly proposals and emphasize the relation between the latter proposal and exact sampling from the infinite alleles model. We also introduce a new proposal that takes knowledge of the ancestral state into account. The new proposal is derived from a new result on exact sampling from a single site. The methods are illustrated on simulated data sets and the data considered in Griffiths and Tavaré (1994).

**KEYWORDS:** ancestral inference, coalescent, importance sampling, infinite sites

---

\*We are grateful to Jeff Thorne and Ben Redelings for valuable discussions and helpful comments on earlier versions of the manuscript. Asbjørn Brask is thanked for programming assistance. AH is financially supported by Danish Research Council grant 5111-95094832 and National Institute of Health grant R01 GM070806. CW is supported by the Danish Cancer Society and The Danish Research Council. Public Health Service grant GM 37841 (MKU) also provided support for this research.

# 1 Introduction

A crucial aspect of importance sampling and Markov Chain Monte Carlo (MCMC) sampling is the choice of proposal distribution. In principle, both methods can approximate a desired integral (e.g., the likelihood function) to any level of accuracy, but in practice it is important to choose a proposal distribution that promotes efficient search of the state space.

In this paper, we discuss proposal distributions for the infinite sites model (ISM), which is used for analysis of DNA sequence data sampled from a population of organisms. We introduce the ISM in terms of Ethier and Griffiths's (1987) algorithm for simulating samples of genes. The algorithm builds up a sample forward in time by duplicating an existing gene or by replacing an existing gene with a mutated gene. Based on Ethier and Griffiths's algorithm, a recursion can be constructed to calculate the likelihood of a data set. Unfortunately, it is not feasible to solve the recursion for data sets of useful size, and sampling-based techniques are required to calculate the likelihood.

Griffiths and Tavaré (1994) described a method for approximating the likelihood under the ISM, and Felsenstein et al. (1999) recognized that the Griffiths-Tavaré (GT) procedure is in essence importance sampling. Proceeding backward in time, a proposal distribution suggests histories of the sample by stepwise reduction of the data set, either by coalescence of two identical genes (the time-reversal of duplication) or by removal of a mutation unique to a single gene. Stephens and Donnelly (2000, Theorem 1) characterized the optimal proposal distribution for a large class of models, including the ISM, and constructed reasonable approximations to the optimal proposal. Their approximation for the ISM is based on the optimal proposal for parent-independent mutation models, first derived by Hoppe (1987) for the Infinite Alleles Model (IAM). We recall Hoppe's work and its relation to the Stephens and Donnelly (SD) proposal distribution for the ISM.

Although neither the GT nor the SD proposal takes into account the number of mutations carried by genetic lineages, one expects that those lineages that have experienced more evolutionary events within a given time period (the time since the most recent common ancestor of the sample) have a higher likelihood of having experienced the most recent evolutionary event. Here, we derive explicit expressions for the probability that the most recent event occurred in a given allele in a sample containing a single segregating site or in a sample of size two. We use these results to develop a new proposal distribution. Our comparison of the importance weights of histories sampled under the GT, SD, and new proposal schemes applied to actual and simulated data sets indicates that the new proposal generally shows greater efficiency.

## 2 Models of genetic evolution

Our objective is to use the pattern of genetic variation observed in the sample as a basis for inferring characteristics of the evolutionary process that produced it. Our sample comprises  $n$  nucleotide (DNA) sequences, all derived from a particular genomic region (locus). Mutation causes the replacement of a nucleotide base (A, C, G, or T) by a base of a different kind, and allelic classes correspond to distinct sequences. We refer to a unit of transmission from parent to offspring as a gene, and an allele as a particular sequence of bases.

To illustrate the use of sampling-based methods to infer characteristics of the evolutionary process, we address the estimation of the scaled mutation rate,

$$\theta = 2Nu, \tag{1}$$

for  $N$  the effective number of genes in the entire population of organisms (rather than in the sample) and  $u$  the probability that any offspring gene bears a newly-arisen mutation. We assume the absence of crossing-over within the sampled genomic region and that all mutational events in the history of the sample are observed in the DNA sequences.

### 2.1 The standard neutral model

Random transmission to offspring of one gene from each parent causes fluctuations in allele frequencies between generations (genetic drift). Selective neutrality of the segregating mutations implies that for all genes, irrespective of allelic class, the numbers transmitted to the offspring generation have independent identical distributions.

Each nucleotide site in a sampled haplotype has a genealogical history, reflecting a direct line of descent from the most recent common ancestor (MRCA) of the sample. In the absence of genetic recombination, all sites share a single genealogy. We impose the large-population assumption that two offspring genes descend from a common parent gene at rate  $1/N$  per generation and that the rate of common descent of more than two in a given generation ( $O(1/N^2)$ ) is negligible. As a consequence, the sample genealogy corresponds to a binary tree.

We assume that the waiting time to the next evolutionary event (common descent or mutation) has an exponential distribution. On level  $l$  of the genealogy of the sample (the segment of the gene tree in which  $l$  lineages ancestral to the sample exist), common descent occurs at a per-generation rate of  $\binom{l}{2}/N$  and mutation  $lu$ . It is natural to scale time in units of  $N$  generations, so that common descent happens at rate  $\binom{l}{2}$  and mutation  $lNu = l\theta/2$ .

## 2.2 An algorithm for evolutionary change

Ethier and Griffiths (1987) have described an algorithm for generating a sample of size  $n$  under mutation and genetic drift in an unstructured population:

ALGORITHM 1

- (1) Start with one gene, which immediately duplicates.
- (2) When there are  $l$  genes, the time until the next event is exponentially distributed with parameter  $l - 1 + \theta$ . Upon an event, choose a gene at random from the  $l$  genes. With probability  $(l - 1)/(l - 1 + \theta)$ , duplicate the gene; otherwise, with probability  $\theta/(l - 1 + \theta)$ , add a mutation to the gene.
- (3) If the number of genes is less than  $n + 1$ , return to (2); otherwise, delete the last gene and stop.

Figure 1 (left) shows an example of a genealogy generated under this algorithm, together with mutations (labeled dots). In the infinite sites model, multiple mutational events never occur at any nucleotide position within the locus. Distinct labels for the tips (leaves) of the tree indicate distinct nucleotide sequences, each of which arises through the accumulation of mutations along its line of descent from the root (MRCA): for example, haplotype  $b$  carries mutations 2 and 3, but not mutations 1, 4, or 5.

A convenient description of a data set lists the distinct nucleotide sequences (alleles) together with their multiplicities. We assume the ancestral state (0) is known, perhaps from information from an outgroup. Element  $s_{km}$  of matrix  $\mathbf{S}$  is 1 if allele  $k$  ( $k = 1, \dots, K$ ) carries mutation  $m$  ( $m = 1, \dots, M$ ) and is 0 otherwise, and vector  $\mathbf{n}$  gives the multiplicities of each allele in the sample. Figure 2 presents  $\mathbf{S}$  and  $\mathbf{n}$  for the genealogical history in Fig. 1.

Algorithm 1 proceeds forward in time, sampling a descendant state ( $D$ ) given the parental state ( $A$ ) from  $P(D|A)$ . Given any particular history (*e.g.*, Fig. 1), the likelihood of any value of  $\theta$  can easily be determined at each successive level from

$$P(D) = \sum_A P(D|A)P(A),$$

with the transition  $P(D|A)$  proceeding forward in time. Because the number of histories grows rapidly as sample size increases (see Song, Lyngsø, and Hein 2006), consideration of only a subset of histories is practicable. Furthermore, because only a very small fraction of histories (an enormous number, nevertheless) are compatible with the observed pattern of nucleotide variation,

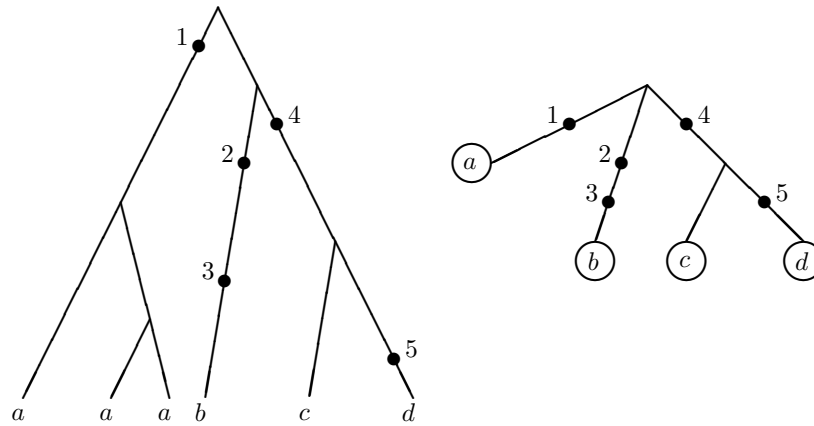


Figure 1: *Left: Genealogical history of six randomly-sampled haplotypes, with numbered dots representing mutations and letters alleles (distinct nucleotide sequences). Right: Perfect phylogeny constructed from the sequence information.*

many procedures sample histories by constructing them backward in time, beginning with the observed sample. Proceeding from descendant to ancestral states (a time-reversal of the evolutionary process) requires efficient proposals of  $P(A|D)$ . We address the backward-in-time construction of genealogical history under two classical forward-in-time descriptions of evolution: the infinite alleles model (IAM) and the infinite sites model (ISM).

### 2.3 A perfect sampler for the IAM

Under the infinite alleles model, any nucleotide substitution defines a new allele. Our description of the sample includes only the number of alleles and their multiplicities. From  $\mathbf{n}$ , we can represent the sample by  $\mathbf{a} = (a_1, \dots, a_n)$ , for  $a_j$  the number of alleles represented  $j$  times in the sample and  $n$  the total number of genes in the sample (see Karlin and McGregor 1972; Kingman 1978). For the data set in Fig. 1, for example, we have  $\mathbf{a} = (3, 0, 1, 0, 0, 0)$ . For  $K$  the number of alleles (distinct nucleotide sequences) in the sample,

$$\sum_{j=1}^n ja_j = n \quad \text{and} \quad \sum_{j=1}^n a_j = K.$$

		site							
$\mathbf{S} =$	allele	1	2	3	4	5	$\mathbf{n} =$	allele	multiplicity
	$a$	1	0	0	0	0		$a$	3
	$b$	0	1	1	0	0		$b$	1
	$c$	0	0	0	1	0		$c$	1
	$d$	0	0	0	1	1		$d$	1

Figure 2: Description of the data set with the genealogical history shown in Fig. 1. Matrix  $\mathbf{S}$  presents the DNA sequences of the alleles, with 0 denoting the original (ancestral) state and 1 the mutant state. Vector  $\mathbf{n}$  gives the multiplicity of each allele. Only segregating sites, those at which two bases exist in the sample, are shown.

The Ewens sampling formula (ESF; Ewens 1972) gives the probability of the sample:

$$P_n(\mathbf{a}) = \frac{n!}{\theta_{(n)}} \prod_{j=1}^n \binom{\theta}{j}^{a_j} \frac{1}{a_j!} \quad (2)$$

for

$$\theta_{(n)} = \theta(\theta + 1) \dots (\theta + n - 1).$$

This expression represents the marginal probability of all complete histories consistent with the observed number of alleles and their multiplicities.

Ewens’s derivation proceeded from the “remarkable intuitive insight” (Karlin and McGregor 1972) that the probability, for a sample of size  $l$ , that the next gene sampled represents an allelic type not yet observed is  $\theta/(\theta + l)$ . Coalescence arguments making explicit reference to the genealogy of the sample have yielded elegant combinatorial derivations of the ESF (Kingman 1978; Donnelly 1986; Fu 1995; Griffiths and Lessard 2005). Of particular relevance to the present investigation is the recursion of Karlin and McGregor (1972):

$$\begin{aligned} P_n(\mathbf{a}) &= \frac{\theta}{(n-1+\theta)} \frac{a_1}{n} P_n(\mathbf{a}) \\ &+ \sum_{j=1}^{n-1} \frac{\theta}{(n-1+\theta)} \frac{(j+1)(a_{j+1}+1)}{n} P_n(\mathbf{a} - \mathbf{e}_1 - \mathbf{e}_j + \mathbf{e}_{j+1}) \quad (3) \\ &+ \sum_{j=1}^{n-1} \frac{(n-1)}{(n-1+\theta)} \frac{j(a_j+1)}{(n-1)} P_{n-1}(\mathbf{a} - \mathbf{e}_{j+1} + \mathbf{e}_j), \end{aligned}$$

for  $\mathbf{e}_j$  the  $n$ -dimensional vector with 1 in position  $j$  and 0 elsewhere. The first term on the right of (3) represents a history in which the most recent event occurred in a singleton allele, with probability  $a_1/n$ , and was a mutation, with probability  $\theta/(n-1+\theta)$ . This event (E1) creates a new singleton allele from an existing singleton allele, preserving the allele spectrum  $\mathbf{a}$ . The second term includes cases in which an allele with multiplicity  $(j+1)$  in the parental generation receives a mutation. This event (E2) creates a new singleton allele, decrements the number of alleles with multiplicity  $(j+1)$  by one, and increments the number of alleles with multiplicity  $j$  by one. With probability  $(n-1)/(n-1+\theta)$ , the most recent event corresponds to a duplication (E3), causing the number of lineages to increase from  $(n-1)$  in the previous generation to  $n$  in the present generation. With probability  $j(a_j+1)/(n-1)$ , the duplication occurred in an allelic class comprising  $j$  replicates, resulting in a decrement of the number of alleles with multiplicity  $j$  and an increment in the number of alleles with multiplicity  $(j+1)$ . Substitution of (2) into (3) confirms the ESF.

Knowledge of the full solution (2) permits explicit specification of the distribution of the ancestral state ( $A$ ) given the descendant state ( $D = \mathbf{a}$ ). Only the ancestral configurations shown on the right side of (3) can have given rise to  $D$ . Bayes formula produces the exact transition probabilities:

$$P(A|D) = \begin{cases} P_n(\mathbf{a}|\mathbf{a}) = \frac{\theta}{n-1+\theta} \frac{a_1}{n} \text{ (E1)} \\ P_n(\mathbf{a} - 2\mathbf{e}_1 + \mathbf{e}_2|\mathbf{a}) = \frac{a_1-1}{n-1+\theta} \frac{a_1}{n} \text{ (E2, } j=1) \\ P_n(\mathbf{a} - \mathbf{e}_1 - \mathbf{e}_j + \mathbf{e}_{j+1}|\mathbf{a}) = \frac{j a_j}{n-1+\theta} \frac{a_1}{n}, \text{ for } j \geq 2 \text{ (E2)} \\ P_n(\mathbf{a} - \mathbf{e}_{j+1} + \mathbf{e}_j|\mathbf{a}) = \frac{(j+1)a_{j+1}}{n}, \text{ for } j \geq 1 \text{ (E3)}. \end{cases} \quad (4)$$

These expressions also follow from Theorem 1 of Stephens and Donnelly (2000). The probability that the next event backward in time corresponds to a mutation is  $a_1/n$ , the sum of the first three cases (all but E3). This probability and its complement, corresponding to a coalescence event (E3), agree with expressions given by Hoppe (1987).

One implementation of a perfect backward sampler using (4) begins with choosing a gene in  $D$  uniformly at random. If it belongs to an allelic class represented more than once (one of  $a_{j+1}$  genes, with  $j \geq 1$ ), assign the most recent event as a duplication (E3). Otherwise, with probability  $a_1/n$ , assign the ancestral state  $A$  as identical to the descendant state  $D = \mathbf{a}$  (E1) with probability  $\theta/(n-1+\theta)$ , as  $\mathbf{a} - 2\mathbf{e}_1 + \mathbf{e}_2$  with probability  $(a_1-1)/(n-1+\theta)$ , and as  $\mathbf{a} - \mathbf{e}_1 - \mathbf{e}_j + \mathbf{e}_{j+1}$  with probability  $j a_j/(n-1+\theta)$ .

## 2.4 The ISM

We now consider the infinite sites model, under which we base the estimation of  $\theta$  (1) on the sequences  $\mathbf{S}$  in addition to the allele frequency spectrum  $\mathbf{n}$  (recall Fig. 2), with  $D = (\mathbf{S}, \mathbf{n})$ .

Each mutation partitions the sample into genes that carry and those that do not carry the mutation. The four-gamete test of Hudson and Kaplan (1985) for the detection of crossing-over uses that sets  $A$  and  $B$ , associated with two distinct mutations, can only be either disjoint ( $A \cap B = \emptyset$ ) or nested ( $A \subseteq B$ ). Gusfield (1991) provided an efficient algorithm for constructing a perfect phylogeny, a summary of binary trees compatible with the observed pattern of mutations in the absence of recombination. A one-to-one correspondence exists between the segregating sites matrix  $\mathbf{S}$  and the perfect phylogeny. Figure 1 (right) shows the perfect phylogeny for the data set in Fig. 2. It differs from the complete genealogical history tree (Fig. 1 left) in not specifying the relative order of all mutation and coalescence events and collapsing some branches. Further, the relative order of mutations on a given branch (*e.g.*, 2 and 3) is not identifiable. A great many number of fully resolved trees may be compatible with the perfect phylogeny, especially for data sets in which the sample size exceeds the number of segregating mutations.

In reconstructing the history of the sample backward in time, we note that just as in the IAM, the descendant configuration  $D = (\mathbf{S}, \mathbf{n})$  can have derived from only three ancestral configurations. If the most recent event is a duplication, with probability  $(n-1)/(n-1+\theta)$ , the  $\mathbf{S}$  matrix of the ancestral configuration ( $A$ ) is identical to that of the descendant ( $D$ ), while the entry in  $\mathbf{n}$  corresponding to the duplicated allele decreases by one. Alternatively, if the most recent event is a mutation,  $\mathbf{S}$  of  $A$  reflects the removal of the corresponding column from  $\mathbf{S}$  of  $D$ . This mutation can have arisen either in an allele present in  $D$  or in an allele not in  $D$ . In the latter case,  $\mathbf{n}$  of  $A$  and  $D$  are identical. In the former case, the number of distinct haplotypes in  $A$  declines by one, entailing the removal of the row corresponding to the allele in  $D$  that bears the new mutation, and the multiplicity of the allele in which the mutation arose is increased by one. Note that in the example shown in Fig. 1, allele  $c$  cannot have been involved in the most recent event because it is a singleton allele, which precludes duplication, and does not carry a unique mutation.

These considerations imply a recursion in the probability of a data set (see

Griffiths and Tavaré, 1994):

$$\begin{aligned}
 P(\mathbf{S}, \mathbf{n}) = & \sum_{k:n_k \geq 2} \frac{(n-1)}{(n-1+\theta)} \frac{(n_k-1)}{(n-1)} P(\mathbf{S}, \mathbf{n} - \mathbf{e}_k) \\
 & + \sum_{k:n_k=1, k \in \mathcal{M}, \forall j: s_k^m \neq s_j} \frac{\theta}{(n-1+\theta)} \frac{1}{n} P(\mathcal{C}_m \mathbf{S}, \mathbf{n}) \\
 & + \sum_{k:n_k=1, k \in \mathcal{M}, \exists j: s_k^m = s_j} \frac{\theta}{(n-1+\theta)} \frac{(n_j+1)}{n} P(\mathcal{R}_k \mathcal{C}_m \mathbf{S}, \mathcal{R}_k(\mathbf{n} + \mathbf{e}_j)),
 \end{aligned} \tag{5}$$

(compare (3)), in which the first summation represents cases in which the most recent event is a duplication of allele  $k$ , the second a new mutation in an allele not present in  $D$ , and the third a new mutation in an allele present in  $D$ .

This recursion uses the notation of Song et al. (2006), in which  $\mathcal{M}$  denotes the set of row indices that correspond to alleles bearing at least one singleton mutation, a mutation borne by exactly one allele. For  $k \in \mathcal{M}$ ,  $s_k^m$  denotes the sequence associated with allele  $k$  with the singleton mutation  $m$  removed:  $s_k^m$  represents the ancestral allele from which allele  $k$  was generated through the acquisition of mutation  $m$ . If ancestral allele  $s_k^m$  does not occur in the sample ( $s_k^m \neq s_j$  for all  $j$ ), then the state of the sample immediately ancestral to the new mutation has the same number of alleles and the same vector of multiplicities  $\mathbf{n}$  and a sequence matrix  $\mathbf{S}$  without the column representing the new mutation. Accordingly,  $\mathcal{C}_m$  in (5) is an operator that removes the column with the mutation corresponding to the last event being a mutation in allele  $k$ . Alternatively, the occurrence of the ancestral sequence in the sample ( $s_k^m = s_j$  for some  $j$ ) implies changes in both  $\mathbf{S}$  and  $\mathbf{n}$ .

Proceeding backward in time, application of the recursion eventually reduces the configuration to a single ancestral lineage bearing no mutations. Song et al. (2006) provide a way of counting all possible ancestral configurations, of which an extremely large number exist even for moderately sized data sets (e.g., 50 alleles and 20 segregating sites). Monte Carlo methods offer a feasible alternative to a full recursive determination of the probabilities of all distinguishable ancestral configurations.

Markov Chain Monte Carlo (MCMC) sampling and importance sampling (IS) represent two major strategies for Monte Carlo methods for the analysis of genetic data. Kuhner, Yamato, and Felsenstein (1995) applied MCMC to estimate  $\theta$  (1). Felsenstein et al. (1999) recognized the method of Griffiths and Tavaré (1994) as a form of importance sampling, and Stephens and Donnelly (2000) addressed the construction of optimal and more efficient proposal distributions. Here, we introduce a new IS proposal distribution, suitable for the

infinite sites model.

### 3 Importance sampling

Importance sampling is based on the approximation

$$\begin{aligned} L(\theta) &= P_\theta(\text{Data}) = \sum_{\mathcal{H}} P_\theta(\text{Data}, \mathcal{H}) = \sum_{\mathcal{H}} P_\theta(\text{Data}, \mathcal{H}) Q(\mathcal{H}) / Q(\mathcal{H}) \\ &= E_Q[P_\theta(\text{Data}, \mathcal{H}) / Q(\mathcal{H})] \approx \frac{1}{R} \sum_{r=1}^R w_r, \end{aligned}$$

for  $\mathcal{H}$  a genealogical history,  $Q$  a proposal distribution for histories, and importance weight  $w_r = P_\theta(\text{Data}, \mathcal{H}_r) / Q(\mathcal{H}_r)$  for the  $r^{\text{th}}$  history ( $r = 1, 2, \dots, R$ ). Because all three schemes considered here propose genealogical histories backward in time, beginning with the observed sample, all histories are necessarily compatible with the sample ( $P_\theta(\text{Data}, \mathcal{H}) > 0$ ; see Felsenstein et al. 1999). Backward reconstruction entails choosing two identical genes to coalesce or one to mutate, with corresponding updates to  $\mathbf{S}$  and  $\mathbf{n}$ . Histories simulated under  $Q$  can be used to estimate the likelihood of any value of  $\theta$ .

The optimal proposal distribution  $Q(\mathcal{H})$  is the posterior distribution  $P(\mathcal{H}|\text{Data})$ :

$$\frac{P(\text{Data}, \mathcal{H})}{Q(\mathcal{H})} = P(\text{Data}|\mathcal{H}) \frac{P(\mathcal{H})}{Q(\mathcal{H})} = P(\text{Data}|\mathcal{H}) \frac{P(\mathcal{H})}{P(\mathcal{H}|\text{Data})} = P(\text{Data}).$$

Because all weights are the same in this case, we need sample only one history to obtain an exact estimate of  $P(\text{Data})$ . Unfortunately,  $P(\mathcal{H}|\text{Data})$  is in general unknown. In fact, these expressions illustrate that knowledge of  $P(\mathcal{H}|\text{Data})$  is equivalent to knowledge of  $P(\text{Data})$ . Section 2.3 provides the optimal proposal distribution (4) for the IAM, under which  $P(\text{Data})$  corresponds to the ESF (2).

#### 3.1 The GT and SD proposals

Griffiths and Tavaré (1994) and Stephens and Donnelly (2000) developed their proposal distributions by considering all possible events that could have occurred in the immediately preceding generation.

Using recursion (5), the GT proposal chooses allele  $k$  to be involved in the most recent evolutionary event (coalescence or mutation) with probability

$$Q_{\theta_0}^{\text{GT}}(k|\mathbf{S}, \mathbf{n}) \propto \begin{cases} (n_k - 1) & n_k \geq 2 \\ \theta_0/n & n_k = 1, k \in \mathcal{M}, \forall j : s_k^m \neq s_j \\ \theta_0(n_j + 1)/n & n_k = 1, k \in \mathcal{M}, \exists j : s_k^m = s_j \\ 0 & n_k = 1, k \notin \mathcal{M}, \end{cases} \quad (6)$$

for  $\theta_0$  the so-called driving value of  $\theta$ . Here, we have chosen to assign  $\theta_0$  as the Watterson estimator

$$\theta_W = \frac{M}{\sum_{j=1}^{n-1} 1/j}, \quad (7)$$

for  $M$  the number of segregating sites in the data set (number of columns in  $\mathbf{S}$ ). The GT proposal (6) gives more weight to states that comprise higher numbers of ancestral alleles (larger  $n_j$ ).

Under the SD proposal, the form of the exact sampler of ancestral states conditional on descendant states under the IAM (4) suggests a proposal mechanism: choose an allele uniformly at random and perform the unique update implied by the choice of allele. Thus, the SD proposal chooses an allele  $k$  with probability

$$\begin{aligned} Q^{\text{SD}}(k|\mathbf{S}, \mathbf{n}) &\propto \begin{cases} n_k & n_k \geq 2 \\ 1 & n_k = 1, k \in \mathcal{M}, \forall j : s_k^m \neq s_j \\ 1 & n_k = 1, k \in \mathcal{M}, \exists j : s_k^m = s_j \\ 0 & n_k = 1, k \notin \mathcal{M} \end{cases} \\ &= \begin{cases} n_k & \text{if } n_k \geq 2 \text{ or } k \in \mathcal{M} \\ 0 & \text{if } n_k = 1 \text{ and } k \notin \mathcal{M}. \end{cases} \end{aligned} \quad (8)$$

The SD proposal is computationally simpler than the GT proposal (6) and does not require a driving  $\theta_0$  value.

### 3.2 A proposal distribution for the ISM

Under both the infinite alleles model and the infinite sites model, each mutation generates a novel allele. While the IAM records only the number of alleles and their multiplicities ( $\mathbf{n}$  or  $\mathbf{a}$ ), the ISM specifies in addition the sequences ( $\mathbf{S}$ ), which contains information about evolutionary relationships among the alleles. Our new proposal distribution for the ISM draws upon both  $\mathbf{S}$  and  $\mathbf{n}$ .

### 3.2.1 New proposal

Allele  $k$  carries the mutation associated with column  $m$  in  $\mathbf{S}$  only if  $\mathbf{S}_{km} = 1$ . This mutation occurs in a total of  $d_m = \sum_k \mathbf{S}_{km} n_k$  alleles. Let  $p_\theta(d_m)$  denote the probability that an allele that bears this mutation is involved in the most recent evolutionary event (duplication or mutation). For a given mutation, we choose allele  $k$  to be involved in this event with probability

$$u_{km}(\theta) = \begin{cases} p_\theta(d_m) n_k / d_m & \text{if } \mathbf{S}_{km} = 1 \\ (1 - p_\theta(d_m)) n_k / (n - d_m) & \text{if } \mathbf{S}_{km} = 0. \end{cases} \quad (9)$$

Considering all mutations by summing over all columns of  $\mathbf{S}$ , our new proposal distribution chooses allele  $k$  with probability

$$Q_{\theta_0}^{\text{new}}(k|\mathbf{S}, \mathbf{n}) \propto \begin{cases} \sum_m u_{km}(\theta_0) & \text{if } n_k \geq 2 \text{ or } k \in \mathcal{M} \\ 0 & \text{if } n_k = 1 \text{ and } k \notin \mathcal{M}. \end{cases} \quad (10)$$

As for the GT proposal, we assign the driving value  $\theta_0$  as the Watterson estimator (7).

To motivate this distribution and obtain an expression for  $p_\theta(d_m)$ , we start by recapitulating and extending some known results for backward probabilities.

### 3.2.2 Derivation

**A single segregating site:** We consider a data set containing a single segregating mutation and describe how exact sampling from the infinite sites model can be performed under arbitrary  $\theta$ .

**THEOREM 1.** *Consider the data set  $\mathbf{S} = (1, 0)$  and  $\mathbf{n} = (d, n - d)$ , which we denote by  $\mathcal{M}_d^n$ . The conditional probability that the most recent event increases the number of mutant genes in the sample by one is*

$$p_\theta(d) = P(\mathcal{M}_{d-1}^{n-1} | \mathcal{M}_d^n) = \frac{\sum_{k=2}^{n-d+1} \frac{d-1}{n-k} \frac{1}{k-1+\theta} \binom{n-d-1}{k-2} \binom{n-1}{k-1}^{-1}}{\sum_{k_0=2}^{n-d+1} \frac{1}{k_0-1+\theta} \binom{n-d-1}{k_0-2} \binom{n-1}{k_0-1}^{-1}}. \quad (11)$$

A proof appears in Appendix 1.

**COROLLARY 1.** *In the limit  $\theta \rightarrow 0$ ,*

$$p_\theta(d) \rightarrow \frac{d}{n-1}, \quad (12)$$

and in the limit  $\theta \rightarrow \infty$ ,

$$p_\theta(d) \rightarrow \frac{d+1}{n}. \quad (13)$$

A proof appears in Appendix 2.

REMARK 1. Theorem 1 and Corollary 1 also hold for  $d = 1$ , with  $p_\theta(1)$  being the probability that the most recent event is the origin of the mutation (see Appendix 1). Equation (11) becomes

$$p_\theta(1) = P(\mathcal{M}_0^{n-1} | \mathcal{M}_1^n) = \frac{\frac{1}{n-1+\theta}}{\sum_{k_0=2}^n \frac{1}{k_0-1+\theta} \frac{k_0-1}{n-1}} \quad (14)$$

(see Appendix 1).

This expression converges to  $1/(n-1)$  for  $\theta \rightarrow 0$  and to  $2/n$  for  $\theta \rightarrow \infty$ . As in the general case (Corollary 1), the probability that the allele that carries the mutation is involved in the most recent event exceeds  $d/n$ , the value expected under a uniform random choice of alleles.

REMARK 2. The case  $\theta \rightarrow 0$  was also studied in Wiuf and Donnelly (1999), and our equation (12) agrees with Wiuf and Donnelly's Lemma 1. It is worth pointing out, that equation (13) for the case  $\theta \rightarrow \infty$  appears similar to Wiuf and Donnelly's Lemma 4. However, we condition on the presence of a mutation in the sample, whereas Lemma 4 is about the genealogy of a subsample of size  $d \leq n$  without a mutation.

**Two genes:** We now consider a data set containing two genes ( $n = 2$ ) and arbitrarily many segregating sites.

THEOREM 2. Consider a data set with  $\mathbf{n} = (1, 1)$ , for which  $m_1$  mutations occur only in one allele and  $m_2$  only in the other. Because both alleles are singletons, the existence of any mutation ( $m_1 + m_2 > 0$ ) entails that the most recent event must have been the origin of a mutation. The conditional probability that this event occurred in the gene that carries  $m_1$  mutations is

$$P(\mathbf{S} = (m_1 - 1, m_2) | \mathbf{S} = (m_1, m_2)) = \frac{m_1}{m_1 + m_2}. \quad (15)$$

This theorem can be found, for example, in Ethier and Griffiths (1987); for completeness, we recall the proof in Appendix 3.

Although the distribution of the number of mutations in the sample depends on  $\theta$ , the identity of the allelic class that is involved in the most recent

event is independent of  $\theta$ , given  $m_1$  and  $m_2$ . It is the allele that carries more mutations that is more likely to be involved in the most recent event.

REMARK 3. For the case specified in Theorem 2 ( $n = 2$  and  $n_k = d_m = 1$  for every mutation and allele), the proposal function  $Q_{\theta_0}^{\text{new}}(k|\mathbf{S}, \mathbf{n})$  (10) gives the exact ancestral probability. For each mutation,  $n_k/d_m = p_\theta(1) = 1$  in (9), implying a proposal probability of an allele proportional to the number of segregating mutations it carries, in agreement with (15).

In contrast, both the GT and SD methods propose the two alleles with equal probability, provided that neither allele is the ancestor of the other ( $m_1, m_2 > 0$ ; for both alleles,  $k \in \mathcal{M}, \forall j : s_k^m \neq s_j$ ). All three schemes propose the derived allele with probability 1 in the remaining case ( $m_1 + m_2 = 1$ ).

### 3.3 Example

For  $\theta \rightarrow 0$ , we have that an allele which bears the derived type at site  $m$  is involved in the most recent event with probability  $d_m/(n - 1)$ . For allele  $k$ , which either carries ( $\mathbf{S}_{km} = 1$ ) or does not carry ( $\mathbf{S}_{km} = 0$ ) this mutation, we obtain from (9) the weights

$$u_{km}(0) = \begin{cases} \frac{n_k}{n-1} & \text{if } \mathbf{S}_{km} = 1 \\ \binom{n-1-d_m}{n-1} \binom{n_k}{n-d_m} & \text{if } \mathbf{S}_{km} = 0. \end{cases}$$

For allele  $a$  in Figure 1, for example,

$$Q_0^{\text{new}}(k|\mathbf{S}, \mathbf{n}) \propto 3/5 + (4/5)(3/5) + (4/5)(3/5) + (3/5)(3/4) + (4/5)(3/5) = 2.49.$$

Table 1 presents the proposal probabilities under the various schemes for the data set shown in Fig. 1, with the driving value for  $\theta$  under the GT and new schemes assigned as the Watterson estimator ( $\theta_0 = 5/(1 + 1/2 + 1/3 + 1/4 + 1/5) = 2.19$ ; see (7)).

**Table 1: Proposal probabilities**

Allele ( $k$ )	Site ( $m$ )					$n_k$	$Q_{\theta_0}^{\text{GT}}$	$Q^{\text{SD}}$	$Q_0^{\text{new}}$	$Q_{\theta_0}^{\text{new}}$	$Q_\infty^{\text{new}}$
	1	2	3	4	5						
$a$	1	0	0	0	0	3	0.646	0.6	0.595	0.567	0.529
$b$	0	1	1	0	0	1	0.118	0.2	0.201	0.219	0.244
$c$	0	0	0	1	0	1	0	0	0	0	0
$d$	0	0	0	1	1	1	0.236	0.2	0.204	0.214	0.227
$d_m$	3	1	1	2	1	$n = 6$					

GT tends to favor alleles for which the ancestral state immediately preceding the most recent event has higher multiplicity (larger  $n_j$  in (6)). Proposal of allele  $d$  implies an ancestral state comprising two copies of allele  $c$  while proposal of allele  $b$  implies descent from a single ancestor that bears one mutation.

In this example, the new proposal closely resembles SD, but unlike both GT and SD, the proposal probability of an allelic class increases with the number of mutations it carries. Alleles  $b$  and  $d$ , which carry two mutations, have higher proposal probabilities under the new scheme than under SD and allele  $a$ , which carries only one mutation, has a correspondingly lower probability. This trend appears more pronounced under higher driving  $\theta_0$  values. Compared to allele  $d$ , proposal of allele  $b$  appears to increase faster with  $\theta$ , reflecting that allele  $b$  possesses more unique mutations (2 and 3), which promote the proposal of  $b$  alone, while the higher weighting of mutation (4) is shared by  $c$  and  $d$ .

## 4 Performance on actual and simulated data

### 4.1 The Griffiths and Tavaré (1994) data set

Griffiths and Tavaré (1994) examined a data set comprising 18 single nucleotide polymorphism (SNP) sites in a 360 base pair segment within the control region observed in a sample of 55 human mitochondrial genomes. For this data set, the alleles and their multiplicities ( $\mathbf{n}$ ) correspond to

allele	$a$	$b$	$c$	$d$	$e$	$f$	$g$	$h$	$i$	$j$	$k$	$l$	$m$	$n$	$\sum$
multiplicity	2	2	1	3	19	1	1	1	4	8	5	4	3	1	55

Figure 3 shows the perfect phylogeny associated with this data set, with mutations 1 through 5 representing A/G SNPs and the remaining mutations C/T SNPs.

#### 4.1.1 Likelihood curves

For this data set, Figure 4 presents the likelihood surfaces (full line) and standard errors (dashed line) estimated under the three sampling schemes (compare Figure 7 of Stephens and Donnelly 2000). Comparison of the left and middle plots indicates that SD has a smaller variance than GT, and the right plot indicates that our new proposal scheme in turn has a smaller variance than SD.

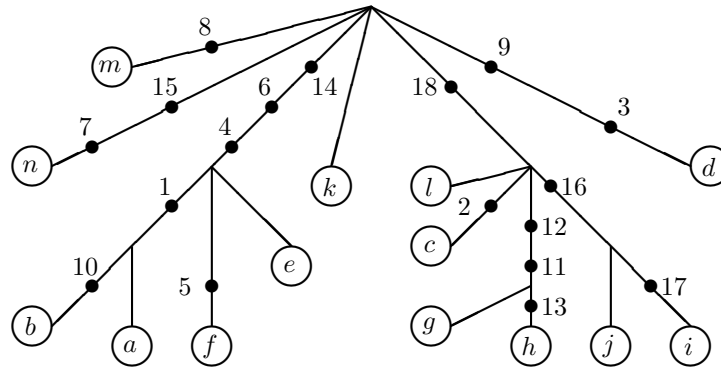


Figure 3: *Perfect phylogeny of the Griffiths and Tavaré (1994) data set.*

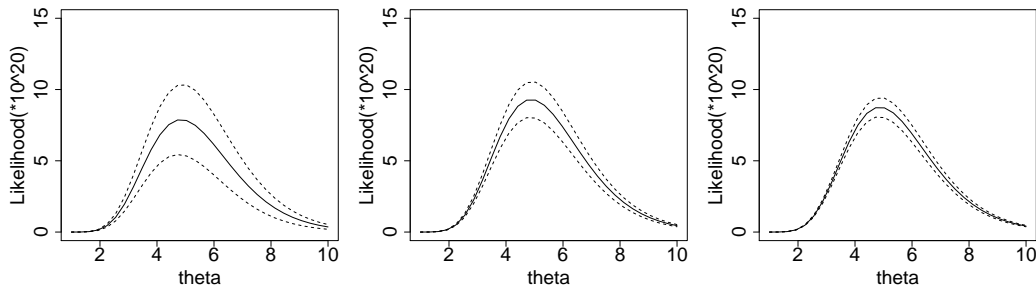


Figure 4: *Comparison of estimated likelihood surfaces (full line) and standard errors (dashed line) for the Griffiths and Tavaré (1994) data set, based on 100,000 samples from the GT (left), SD (middle), and new (right) proposal distributions.*

#### 4.1.2 Effective sample size

Effective sample size, defined by

$$\text{ESS} = \frac{R}{1 + \text{Var}(w)} \quad (16)$$

(Liu, 2001, Section 2.5), for  $R$  the number of samples and  $w = (w_1, \dots, w_R)$  their importance weights, provides a measure of proposal efficiency. To explore the relative efficiencies of the methods in proposing genealogical histories under a known  $\theta$ , we computed ESSs for  $\theta = \theta_0 = 5$ . From 100,000 samples, we obtained  $\text{ESS}_{\text{GT}} = 40$ ,  $\text{ESS}_{\text{SD}} = 215$ , and  $\text{ESS}_{\text{new}} = 548$ , suggesting a 5-fold

increase in efficiency of SD relative to GT and a greater than 2-fold increase of the new proposal relative to SD.

### 4.1.3 Posterior distribution of the level of origin of a mutation

As the perfect phylogeny (*e.g.*, Fig. 3) determines only the order of nested events, proposal of a full genealogical history entails specifying an order of coalescence and mutational events in disjunct clades. To explore the differences in efficiency among the methods, we compared the level of the genealogy on which particular mutations were proposed to have originated.

We borrow concepts from adaptive importance sampling techniques (*e.g.*, Givens and Raftery 1996) to assess bias in the level of the gene genealogy on which particular mutations arise. Comparison of the proposal distribution of genealogical history to the posterior distribution is the main idea behind adaptive importance sampling techniques, under which the proposal mechanism is adjusted according to the samples and their corresponding weights.

Let  $A$  be the event that the mutation at site  $m$  occurs on one of the  $l$  branches on level  $l$ . The probability of this event is given by

$$\begin{aligned} P(A|\text{Data}) &\propto \sum_{\mathcal{H}} P(\mathcal{H})1_A(\mathcal{H}) \\ &= \sum_{\mathcal{H}} \frac{P(\mathcal{H})1_A(\mathcal{H})}{Q(\mathcal{H})} Q(\mathcal{H}) \\ &= E_Q \left[ \frac{P(\mathcal{H})1_A(\mathcal{H})}{Q(\mathcal{H})} \right], \end{aligned}$$

for  $1_A(\mathcal{H})$  equal to unity for any history  $\mathcal{H}$  containing  $A$  and zero otherwise. We can easily obtain both a Monte Carlo estimate of the last term, the posterior distribution of the level of origin of any given mutation, and compare it to the proposal distribution  $Q(\mathcal{H})$ .

For the Griffiths and Tavaré (1994) data set (Fig. 3), Fig. 5 compares the posterior densities (solid red lines) of the levels of origin of mutations 5, 8, and 11 to the proposal probabilities under the GT (dotted), SD (dashed), and new (solid black) schemes. GT expresses a stronger preference than the posterior and the other two schemes for a more recent (higher level number) origin of the mutation at site 5 (the only A/G SNP shown). As observed for Table 1, GT (6) favors events for which the ancestral state occurs in higher multiplicity: the mutation at site 5 arose on allele  $e$ , of which 19 copies occur in the sample, while the other mutations arose on an ancestral allele with low multiplicity.

Figure 5 illustrates that while the SD and the new proposals resemble the posterior, the GT scheme tends to give a more peaked proposal distribution,

expressing more intense preferences, which can lead to underproposal of levels both too ancient (site 5) and too recent (site 8). The flatter proposal distributions of the SD and new schemes may promote more efficient exploration of the space of genealogical histories.

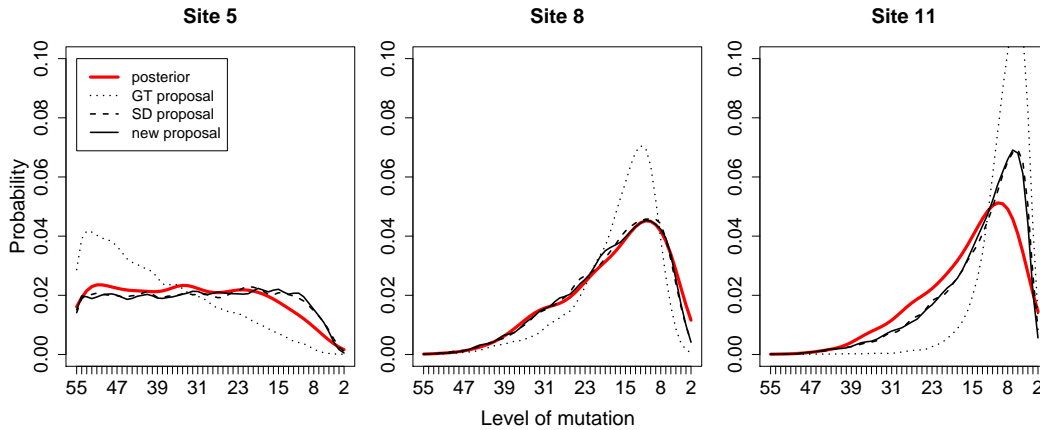


Figure 5: *Smoothed histograms for the level of the mutation at sites 5, 8, and 11 for 10,000 samples for the Griffiths and Tavaré (1994) data set under the GT (dotted), SD (dashed), and new (solid black) proposal schemes. Solid red lines represent the posterior distribution of the level of origin of the indicated mutation.*

## 4.2 Simulation study

We present ESS values (16) for the GT, SD, and new proposals applied to simulated data sets. Twenty-seven data sets were simulated using Hudson’s (2002) `ms` program, which generates samples under the ISM. We generated three data sets under each combination of three sample sizes ( $n = 50, 75, 100$ ) and three levels of mutation ( $\theta = 1, 3, 5$ ). Watterson’s estimator of  $\theta$  (7) uses that the expected number of segregating sites in a sample ( $M$ , or number of columns in  $\mathbf{S}$ ) is given by  $\theta \sum_{i=1}^{n-1} 1/i$ . The sum is 4.48 for  $n = 50$ , 4.89 for  $n = 75$  and 5.18 for  $n = 100$ , and in the simulated data sets, the number of segregating sites varied from around 2–10 for  $\theta = 1$  to around 15–35 for  $\theta = 5$ .

For the GT and the new proposal, we assigned the driving value  $\theta$  as the Watterson estimator (7). Figure 6 shows that the new proposal generally outperforms the other two proposals, but that sampling efficiency varies among data sets generated under the same evolutionary parameters. The new sampler

can be 5 times more efficient, but the typical increase in efficiency is around 1.2 for  $\theta = 1$  and 2.0 for  $\theta = 5$ .

We also applied the new proposal scheme under driving values  $\theta_0 = 0$  and  $\theta_0 \rightarrow \infty$ . Generally, we found that for data sets generated under  $\theta = 1$ , using a driving value of  $\theta_0 = 0$  yields results about as good as those obtained under the Watterson estimator (7). Similarly, we found that for data sets generated under  $\theta = 5$ , the method performs similarly under driving values  $\theta_0 \rightarrow \infty$  and (7).

## 5 Discussion

We have addressed the likelihood-based estimation of a population parameter (1) through marginalization over possible genealogical histories of an observed sample of genes. Within the importance sampling framework of Griffiths and Tavaré (1994), a genealogical history corresponds to an ordered list of two kinds of evolutionary events: mutation and splitting of lineages (coalescence under time-reversal). For a sample of  $n$  genes, a history comprises  $n - 1 + M$  evolutionary events, including  $n - 1$  coalescence events and  $M$  mutation events. Under the infinite sites model, all mutations are distinguishable, implying that  $M$  corresponds to the number of nucleotide sites at which more than one nucleotide segregates in the sample. We have here introduced a new method (10) for proposing genealogical histories and compared it to those of Griffiths and Tavaré (1994) and of Stephens and Donnelly (2000).

All three methods considered construct genealogical histories backward in time, beginning with the observed sample, and assume the absence of recombination. At any point in the history, only those genes that represent alleles present in more than one copy or that carry a singleton mutation (borne by exactly one gene) can have been involved in the most recent evolutionary event.

Recently, De Iorio and Griffiths (2004) showed that the SD proposal can be derived from arguments relating to the generating diffusion equation underlying the ISM. This diffusion equation is equivalent to recursion (5), in which the terms on the right correspond to a coalescence event or a removal of a singleton mutation. The SD proposal (8) is obtained by equating each term on the right with a term that is derived from a decomposition of the term on the left side.

In choosing a gene to be involved in the most recent evolutionary event, the new proposal (10) differs from both GT (6) and SD (8) by taking into account all mutations, not only singletons. Accordingly, it does not derive from the diffusion equation or (5).

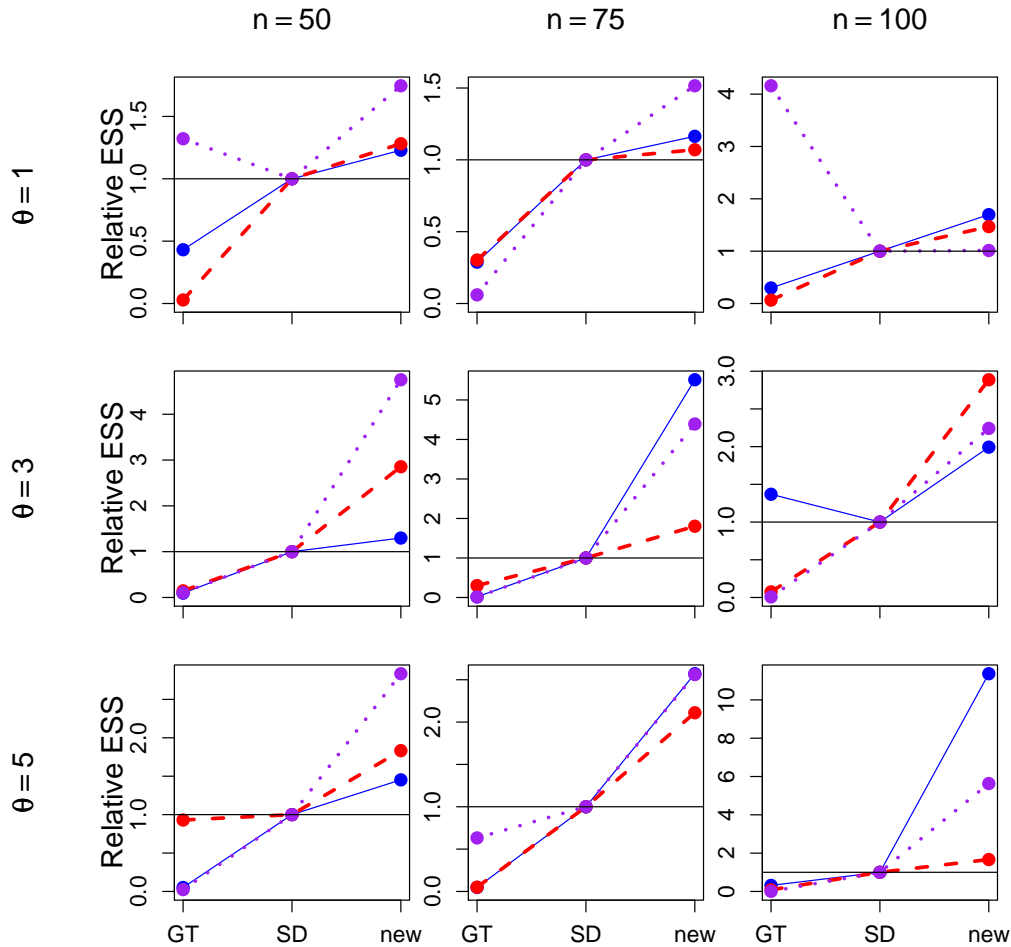


Figure 6: *Effective sample size (ESS) estimates for three data sets simulated under each combination of three mutation rates ( $\theta=1, 3, 5$ ) and three sample sizes ( $n=50, 75, 100$ ). All values are scaled to the ESS of the SD proposal, which corresponds uniformly to 1. The upward trend indicates that the new proposal generally performs better than the SD proposal, which in turn performs better than the GT proposal. Differences among lines within each plot illustrate a large variation in ESS among data sets, independently generated under the same sample size and mutation rate.*

Our proposal generally shows comparable or greater efficiency than do the GT and SD proposals. Applied to the data set of Griffiths and Tavaré (1994), for example, Fig. 4 shows that the new proposal (right plot) has a smaller variance than the GT proposal (left plot) and SD proposal (middle plot), based on 100,000 samples from each proposal. With respect to effective sample size (ESS), another measure of proposal efficiency, we obtained ESS values of 40, 215, and 548 for the GT, SD and new proposals, respectively.

In general, the performance of all proposals fluctuates substantially among data sets representing independent realizations of the same evolutionary process. We expect that further improvement can be achieved through consideration of other schemes for weighting mutations or otherwise taking into account information implicit in the total pattern of genetic variation.

## Appendices

### Appendix 1 Proof of Theorem 1

We first find the probability that the most recent event increased the number of genes with the mutation by one, given the level  $k$  on which the mutation arose. Second, we use Stephens (2000, Theorem 3.1) to give the full probability of this most recent event.

Let  $\mathcal{I}_k^n$  be the event that a single mutation occurred at level  $k$ , with no mutations on any other level of the genealogy. Let  $\mathcal{J}_k$  ( $\mathcal{I}_k^n \subseteq \mathcal{J}_k$ ) denote the event that a single mutation occurred at level  $k$  and no mutations occurred on any level more ancient. From Algorithm 1,

$$P(\mathcal{I}_k^n | \mathcal{J}_k) = \frac{k}{k + \theta} \cdots \frac{n - 1}{n - 1 + \theta}.$$

Let  $\mathcal{M}_d^n$  be the event that  $d$  of  $n$  individuals in the sample have the derived allele:  $\mathcal{M}_d^n$  corresponds to the data set  $\mathbf{S} = (0, 1)$  and  $\mathbf{n} = (d, n - d)$ . The probability that  $d$  of the  $n$  genes in the sample have the mutation, given that it occurred at level  $k$ , is

$$P(\mathcal{M}_d^n | \mathcal{I}_k^n) = \binom{n - d - 1}{k - 2} \binom{n - 1}{k - 1}^{-1}$$

(Fu 1995; Stephens 2000), and we get

$$\begin{aligned} P(\mathcal{M}_d^n | \mathcal{J}_k) &= P(\mathcal{I}_k^n | \mathcal{J}_k) P(\mathcal{M}_d^n | \mathcal{I}_k^n) \\ &= \frac{k}{k + \theta} \cdots \frac{n - 1}{n - 1 + \theta} \binom{n - d - 1}{k - 2} \binom{n - 1}{k - 1}^{-1}. \end{aligned}$$

Conditional on the sample ( $\mathcal{M}_d^n$ ) and the mutation having occurred on level  $k$  ( $\mathcal{J}_k$ ), we can now find the probability that the most recent event added one more gene with the mutation:

$$P(\mathcal{M}_{d-1}^{n-1} | \mathcal{M}_d^n, \mathcal{J}_k) = P(\mathcal{M}_d^n | \mathcal{M}_{d-1}^{n-1}, \mathcal{J}_k) \frac{P(\mathcal{M}_{d-1}^{n-1} | \mathcal{J}_k)}{P(\mathcal{M}_d^n | \mathcal{J}_k)} \quad (\text{A.1a})$$

$$\begin{aligned} &= \frac{(n-1)}{(n-1+\theta)} \frac{(d-1)}{(n-1)} \frac{\prod_{l=k}^{n-2} \frac{l}{l+\theta} \binom{(n-1)-(d-1)-1}{k-2} \binom{(n-1)-1}{k-1}^{-1}}{\prod_{l=k}^{n-1} \frac{l}{l+\theta} \binom{n-d-1}{k-2} \binom{n-1}{k-1}^{-1}} \\ &= \frac{d-1}{n-k}. \end{aligned} \quad (\text{A.1b})$$

This probability increases as the level of  $k$  increases (a more recent origin of the mutation). Also note that the conditional probability does not depend on  $\theta$ .

Stephens (2000, Theorem 3.1) shows that the probability of the mutation occurring at level  $k$  ( $k = 2, 3, \dots, n-d+1$ ) is given by

$$P(\mathcal{J}_k | \mathcal{M}_d^n) = P(\mathcal{I}_k^n | \mathcal{M}_d^n) = \frac{\frac{1}{k-1+\theta} \binom{n-d-1}{k-2} \binom{n-1}{k-1}^{-1}}{\sum_{k_0=2}^{n-d+1} \frac{1}{k_0-1+\theta} \binom{n-d-1}{k_0-2} \binom{n-1}{k_0-1}^{-1}}. \quad (\text{A.2})$$

We now get the desired backward probability (11) from

$$\begin{aligned} P(\mathcal{M}_{d-1}^{n-1} | \mathcal{M}_d^n) &= \sum_{k=2}^{n-d+1} P(\mathcal{M}_{d-1}^{n-1}, \mathcal{J}_k | \mathcal{M}_d^n) \\ &= \sum_{k=2}^{n-d+1} P(\mathcal{M}_{d-1}^{n-1} | \mathcal{M}_d^n, \mathcal{J}_k) P(\mathcal{J}_k | \mathcal{M}_d^n), \end{aligned} \quad (\text{A.3})$$

by using (A.1b) and (A.2).

For  $d = 1$ ,  $P(\mathcal{M}_{d-1}^{n-1} | \mathcal{M}_d^n) = P(\mathcal{M}_0^{n-1} | \mathcal{M}_1^n)$  represents the case in which the origin of the mutation corresponds to the most recent event in the full sample. That the mutation arose on level  $n$  implies

$$P(\mathcal{M}_1^n | \mathcal{M}_0^{n-1}, \mathcal{J}_k) = \begin{cases} 0 & \text{if } k \neq n \\ 1 & \text{if } k = n \end{cases}$$

and  $P(\mathcal{M}_1^n | \mathcal{J}_n) = P(\mathcal{M}_0^{n-1} | \mathcal{J}_n) = 1$  in (A.1a). Substitution of these expressions into (A.3) produces (14).

## Appendix 2 Proof of Corollary 1

We can rewrite (11) as

$$\frac{(d-1) \sum_{k=2}^{n-d+1} \frac{1}{k-1+\theta} (k-1) \frac{(n-k-1)!}{(n-d+1-k)!}}{\sum_{k_0=2}^{n-d+1} \frac{1}{k_0-1+\theta} (k_0-1) \frac{(n-k)!}{(n-d+1-k)!}}. \quad (\text{A.4})$$

In the limit  $\theta \rightarrow 0$ , the numerator becomes

$$(d-1) \sum_{k=2}^{n-d+1} \frac{(n-k-1)!}{(n-d+1-k)!} = (n-d)(n-d+1) \cdots (n-2),$$

and the denominator becomes

$$\sum_{k_0=2}^{n-d+1} \frac{(n-k)!}{(n-d+1-k)!} = \frac{1}{d} (n-d)(n-d+1) \cdots (n-1),$$

and we obtain (12).

When  $\theta \rightarrow \infty$ , (A.4) converges to

$$\frac{(d-1) \sum_{k=2}^{n-d+1} (k-1) \frac{(n-k-1)!}{(n-d+1-k)!}}{\sum_{k_0=2}^{n-d+1} (k_0-1) \frac{(n-k)!}{(n-d+1-k)!}}.$$

The numerator reduces to  $(n-1)(n-2) \cdots (n-d)/d$  and the denominator reduces to  $n(n-1) \cdots (n-d)/(d(d+1))$ , and we obtain (13).

## Appendix 3 Proof of Theorem 2

For a sample of two genes ( $n = 2$ ), the probability of having a total of  $(m_1 + m_2)$  mutations in the two lineages is geometrically distributed with parameter  $\theta/(1 + \theta)$ :  $(m_1 + m_2) \sim \text{Geo}(\theta/(1 + \theta))$ . Each mutation appears in the left lineage with probability  $1/2$ , so the probability of having  $m_1$  mutations in the left lineage conditional on  $(m_1 + m_2)$  lineages in total is binomially distributed with parameter  $1/2$ :  $m_1 | (m_1 + m_2) \sim \text{Bin}(1/2, m_1 + m_2)$ . The probability of observing  $m_1$  mutations in one gene and  $m_2$  in the other is therefore

$$P(m_1, m_2) = \frac{1}{\theta} \left( \frac{\theta}{1 + \theta} \right)^{(m_1 + m_2)} \binom{m_1 + m_2}{m_1} (1/2)^{m_1 + m_2}. \quad (\text{A.5})$$

We now find the backward probability from

$$\begin{aligned} P((m_1 - 1, m_2) | (m_1, m_2)) &= \frac{P((m_1, m_2) | (m_1 - 1, m_2)) P(m_1 - 1, m_2)}{P(m_1, m_2)} \\ &= \frac{m_1}{m_1 + m_2}, \end{aligned}$$

in which we have inserted (A.5) twice and used

$$P((m_1, m_2)|(m_1 - 1, m_2)) = \frac{\theta}{(1 + \theta)^2}.$$

## References

- DE IORIO, M. and GRIFFITHS, R.C. (2004) Importance sampling on coalescent histories. *Adv. Appl. Prop.* **36** 417–433.
- DONNELLY, P. (1986). Partition structures, Polya urns, the Ewens sampling formula, and the ages of alleles. *Theor. Popul. Biol.* **30** 271–288.
- ETHIER, S.N. and GRIFFITHS, R.C. (1987). The infinitely-many-sites model as a measure valued diffusion. *Ann. Probab.* **15** 414–545.
- EWENS, W.J. (1972). The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* **3** 87–112.
- FELSENSTEIN, J., KUHNER, M.K., YAMATO, J. and BEERLI, P. (1999). Likelihoods on coalescents: a Monte Carlo sampling approach to inferring parameters from population samples of molecular data. *IMS Lect. Notes Monogr. Ser.* **33** 163–185.
- FU, Y.-X. (1995). Statistical properties of segregating sites. *Theor. Popul. Biol.* **48** 172–197.
- GIVENS, G.H. and RAFTERY, A.E. (1996). Local adaptive importance sampling for multivariate densities with strong nonlinear relationships. *J. Amer. Statist. Assoc.* **91** 132–141.
- GRIFFITHS, R.C. and TAVARÉ, S. (1994). Ancestral inference in population genetics. *Statistical Science* **9** 307–319.
- GRIFFITHS, R.C. and LESSARD, S. (2005). Ewens’ sampling formula and related formulae: combinatorial proofs, extensions to variable population size and applications to ages of alleles. *Theor. Popul. Biol.* **68** 167–177.
- GUSFIELD, D. (1991). Efficient algorithms for inferring evolutionary trees. *Networks* **21** 19–28.
- HOPPE, F.M. (1987). The sampling theory of neutral alleles and an urn model in population genetics. *J. Math. Biol.* **25** 123–159.

- HUDSON, R. R. (2002). Generating samples under a Wright-Fisher neutral model. *Bioinformatics* **18** 337–338.
- HUDSON, R.R. and KAPLAN, N.L. (1985). Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, **111** 147–164.
- KARLIN, S. and MCGREGOR, J. (1972). Addendum to a paper of W. Ewens. *Theor. Popul. Biol.* **3** 113–116.
- KINGMAN, J.F.C. (1978). Random partitions in population genetics. *Proc. R. Soc. London Ser. A* **361** 1–20.
- KUHNER, M.K., YAMATO, J. and FELSENSTEIN, J. (1995). Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings. *Genetics* **140** 1421–1430.
- LIU, J.S. (2001). *Monte Carlo strategies in scientific computing*. Springer-Verlag New York.
- SONG, Y.S., LYNGSØ, R. and HEIN, J. (2006). Counting all possible ancestral configurations of sample sequences in population genetics. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **3** 239–251.
- STEPHENS, M. (2000). Times on trees, and the age of an allele. *Theor. Popul. Biol.* **57** 109–119.
- STEPHENS, M. and DONNELLY, P. (2000). Inference in molecular population genetics. *J. R. Statist. Soc. B* **62** 605–635.
- WIUF, C. and DONNELLY, P. (1999). Conditional genealogies and the age of a neutral mutant. *Theor. Popul. Biol.* **56** 183–201.