

# *Statistical Applications in Genetics and Molecular Biology*

---

*Volume 11, Issue 1*

2012

*Article 9*

---

## Stopping-Time Resampling and Population Genetic Inference under Coalescent Models

**Paul A. Jenkins**, *University of California, Berkeley*

**Recommended Citation:**

Jenkins, Paul A. (2012) "Stopping-Time Resampling and Population Genetic Inference under Coalescent Models," *Statistical Applications in Genetics and Molecular Biology*: Vol. 11: Iss. 1, Article 9.

**DOI:** 10.2202/1544-6115.1770

**Available at:** <http://www.bepress.com/sagmb/vol11/iss1/art9>

©2012 De Gruyter. All rights reserved.

# Stopping-Time Resampling and Population Genetic Inference under Coalescent Models

Paul A. Jenkins

## Abstract

To extract full information from samples of DNA sequence data, it is necessary to use sophisticated model-based techniques such as importance sampling under the coalescent. However, these are limited in the size of datasets they can handle efficiently. Chen and Liu (2000) introduced the idea of *stopping-time resampling* and showed that it can dramatically improve the efficiency of importance sampling methods under a finite-alleles coalescent model. In this paper, a new framework is developed for designing stopping-time resampling schemes under more general models. It is implemented on data both from infinite sites and stepwise models of mutation, and extended to incorporate crossover recombination. A simulation study shows that this new framework offers a substantial improvement in the accuracy of likelihood estimation over a range of parameters, while a direct application of the scheme of Chen and Liu (2000) can actually diminish the estimate. The method imposes no additional computational burden and is robust to the choice of parameters.

**KEYWORDS:** coalescence, importance sampling, resampling, infinite sites

**Author Notes:** This research benefitted from the helpful comments of Jotun Hein, Chris Holmes, Rune Lyngsø, Yun Song, and Carsten Wiuf. Above all, I thank my DPhil supervisor Bob Griffiths for his guidance. The work was carried out as a member of the University of Oxford Life Sciences Interface Doctoral Training Centre, funded by the EPSRC, and completed at the University of California, Berkeley, supported in part by an NIH grant R01-GM094402.

# 1 Introduction

Central to many applications in population genetic analyses is the problem of computing the likelihood of a sample of genetic data. To capture the full complexity of forces that shaped the patterns we see in contemporary samples, models of reproduction should incorporate the effects of genetic drift, mutation, recombination, selection, and so on. However, computing such likelihoods remains intractable for samples of moderate size, even under the simplifying coalescent limit. To deal with this, researchers have applied a number of computationally-intensive Monte Carlo techniques to a variety of coalescent-based models; these techniques include rejection algorithms (Tavaré, Balding, Griffiths, and Donnelly, 1997, Beaumont, Zhang, and Balding, 2002), importance sampling (Griffiths and Marjoram, 1996, Stephens and Donnelly, 2000, Fearnhead and Donnelly, 2001, De Iorio and Griffiths, 2004, De Iorio, Griffiths, Leblois, and Rousset, 2005, Griffiths, Jenkins, and Song, 2008, Hobolth, Uyenoyama, and Wiuf, 2008, Jenkins and Griffiths, 2011), and Markov chain Monte Carlo (Kuhner, Yamato, and Felsenstein, 2000, Wilson, Weale, and Balding, 2003, Hey and Nielsen, 2004, Wang and Rannala, 2008). These methods typically compute the full-likelihood of the data or an approximation thereof. As a wealth of new genetic data is being discovered, full-likelihood methods are also becoming limited. One solution is to seek even simpler models under which inference is tractable (e.g. Paul and Song, 2010), but there is still a great need to improve the efficiency of full-likelihood methods, both because they can be used as a gold standard for evaluating the performance of approximate methods (Paul and Song, 2010) and because approximate methods dispense with an interpretation of the genealogy associated with the data, which may be of direct interest.

The efficiency of sequential importance sampling (SIS) techniques can be improved by resampling; indeed, resampling is essential in the framework of sequential Monte Carlo (SMC) (Doucet and Johansen, 2011). Under the coalescent model, SIS proceeds by sampling genealogies consistent with the observed data according to a *proposal distribution*. Provided such genealogies can be reconstructed in parallel, resampling methods can be applied. Reconstruction is halted at some fixed time, and a resampling algorithm then duplicates those partially reconstructed genealogies which are ‘promising’ and discards those which are not. In principle, this ought to guide reconstruction towards genealogies with higher posterior probability given the observed data, as estimated by their current SIS weight, thus improving the Monte Carlo estimate of the likelihood. Although this scheme is standard in SMC, resampling can in fact be detrimental when applied to the coalescent model, since the current SIS weight is a poor predictor of final SIS weight in this setting (Fearnhead, 2008). Chen and Liu (2000) and Chen, Xie, and Liu (2005) showed that a clever remedy is to resample not at fixed times but at *stop-*

*ping*-times. A judicious choice of stopping-time makes the comparison between partially-reconstructed genealogies more meaningful. Chen et al. (2005) illustrate their method by applying it to genetic data from a fully linked locus under a finite-alleles model.

Unfortunately, extending the method of Chen et al. (2005) to more general coalescent models is not straightforward. As is shown in this paper, applying their stopping-time resampling method to data generated under other models of mutation can actually be detrimental to Monte Carlo estimation. In this paper a new framework is introduced for designing stopping-times applicable to more general coalescent models. The method is illustrated by incorporating both infinite sites and stepwise models of mutation, as well as crossover recombination. I perform a simulation study in order to quantify the improvement resulting from the new method and also to investigate the effect of choice of tuning parameter  $B$  (defined below). Finally, I briefly discuss how the problems we encounter under the coalescent model might also arise for other problems in which SIS is used.

## 2 Population genetic inference

### 2.1 The coalescent

Throughout we will assume the standard, neutral, coalescent model, which may be obtained as follows. Suppose haplotypes are sampled from a large, stationary, randomly mating population of diploid effective size  $N$ . Reproduction occurs in discrete generations. Haplotypes composing the present generation are formed by sampling a parent uniformly at random; with probability  $1 - u$  the inherited allele is identical to the parent, otherwise it differs by a mutation occurring under some specified model of allelic changes. For concreteness we will focus on the *infinite sites* model: A haplotype is idealized as a unit interval  $[0, 1]$ , and a mutation event occurs at a random site never previously mutant, which can be encapsulated by assigning a  $\text{Uniform}[0, 1]$  random variable to the location of the mutation. The haplotype is specified by the list of locations of mutant alleles that it carries. Crossover recombination can be incorporated by saying that with probability  $1 - r$  a haplotype inherits its alleles from a single parent in this manner, otherwise a recombination event occurs at a random breakpoint. Then the alleles to the left and to the right of this breakpoint are inherited independently by sampling two parents uniformly at random from the previous generation.

Take the usual diffusion limit in which the population mutation and recombination parameters  $\theta = 4Nu$  and  $\rho = 4Nr$  are held fixed as  $N \rightarrow \infty$ , and rescale time in units of  $2N$  generations. A random sample  $\mathcal{D}$  of  $n$  haplotypes can be drawn

from this model by picking  $n$  haplotypes from the present and tracing their ancestral lineages back in time until we reach a most recent common ancestor (MRCA) for every position along the haplotype. On the coalescent timescale, backwards in time each pair of lineages coalesces to find a common ancestor at rate 1, each lineage mutates at rate  $\theta/2$ , and each lineage recombines at rate  $\rho/2$ . When a recombination occurs, a lineage splits into two going backwards in time. The resulting genealogy is a random *ancestral recombination graph* (ARG) (Griffiths and Marjoram, 1996), embedded in which is a coalescent tree for each position along the haplotype. The ARG determines the allelic states at the leaves, which in turn specifies the configuration  $\mathcal{D}$  of our sample. By taking a slice of the ARG at a fixed time between each coalescence, mutation, and recombination event, we also specify a sequence  $\mathcal{H} = (H_0, H_{-1}, \dots, H_{-m})$  of successive configurations going back in time. The sequence proceeds from the present-day sample,  $H_0 = \mathcal{D}$ , back to a grand MRCA of the sample  $H_{-m}$  comprised of a single haplotype carrying no mutant alleles. When branch lengths between events are not recorded and lineages are labelled by the present-day samples to which they are ancestral, each ARG determines a sequence  $\mathcal{H}$  and vice versa. This notation follows Stephens and Donnelly (2000).

## 2.2 Sequential importance sampling (SIS)

The ARG associated with a given dataset is of course unobserved, and to compute the likelihood we must sum over all ARGs that could have given rise to the data. The state space of such ARGs is vast, and it is this that motivates the Monte Carlo approach. For parameters  $\Psi = (\theta, \rho)$ , a naïve Monte Carlo estimate of the likelihood is:

$$L(\Psi) = p(\mathcal{D}; \Psi) = \sum_{\mathcal{H}} p(H_0 = \mathcal{D} \mid \mathcal{H}) p(\mathcal{H}; \Psi) \approx \frac{1}{M} \sum_{j=1}^M p(H_0^{(j)} = \mathcal{D} \mid \mathcal{H}^{(j)}). \quad (1)$$

Here,  $p(H_0 = \mathcal{D} \mid \mathcal{H})$  is an indicator for whether the ARG  $\mathcal{H}$  gives rise to the configuration  $\mathcal{D}$ , and  $(\mathcal{H}^{(1)}, \dots, \mathcal{H}^{(M)})$  is a sample of ARGs drawn from the coalescent prior  $p(\mathcal{H}; \Psi)$ . This approach is doomed since, even for samples of moderate size, the overwhelming majority of draws  $\mathcal{H}^{(j)}$  will not give rise to the observed data and the Monte Carlo estimate is 0 (Stephens and Donnelly, 2000).

Importance sampling is an alternative that can guarantee each draw is compatible with  $\mathcal{D}$ , and can be used to try to sample ARGs of high posterior probability via a user-specified *proposal distribution*. Furthermore, it is useful to specify the IS proposal distribution *sequentially*; that is, we write:

$$\begin{aligned}
 p(\mathcal{D}; \Psi) &= \sum_{\{H_{-1}\}} p(H_0 | H_{-1})p(H_{-1}) = \sum_{\{H_{-1}\}} \frac{p(H_0 | H_{-1})}{q(H_{-1} | H_0)} q(H_{-1} | H_0) p(H_{-1}) \\
 &= \dots = \mathbb{E} \left[ \frac{p(H_0 | H_{-1}) \dots p(H_{-m+1} | H_{-m})}{q(H_{-1} | H_0) \dots q(H_{-m} | H_{-m+1})} p(H_{-m}) \right] \\
 &\approx \frac{1}{M} \sum_{j=1}^M \prod_{k=1}^m \frac{p(H_{-k+1}^{(j)} | H_{-k}^{(j)})}{q(H_{-k}^{(j)} | H_{-k+1}^{(j)})}. \tag{2}
 \end{aligned}$$

The first summation is over all possible previous configurations  $H_{-1}$  that could have given rise to  $H_0 = \mathcal{D}$ . Iterating for  $p(H_{-1}), p(H_{-2})$ , and so on, we sum over all paths  $(H_{-1}, H_{-2}, \dots, H_{-m})$  compatible with  $\mathcal{D}$ . (Dependence on  $\Psi$  is suppressed for convenience.) The summation is re-expressed as an expectation with respect to a *sequential proposal distribution*  $q(H_{-k} | H_{-k+1})$ : given a partially reconstructed genealogy back to a configuration  $H_{-k+1}$ ,  $q$  specifies a distribution over the next event back in time among all possible coalescences, mutations, and recombinations. A Monte Carlo estimate of this expectation is then taken by drawing ARGs not from  $p$  but from  $q$ . The likelihood is estimated from the mean of the importance weights of each genealogy, where draw  $\mathcal{H}^{(j)}$  from  $q$  contributes an importance weight  $w_m^{(j)}$  defined by

$$w_K^{(j)} = \prod_{k=1}^K \frac{p(H_{-k+1}^{(j)} | H_{-k}^{(j)})}{q(H_{-k}^{(j)} | H_{-k+1}^{(j)})}. \tag{3}$$

The final term,  $p(H_{-m})$ , does not appear in the last line of (2) because it is unity for the mutation models considered in this paper.

An efficient choice of  $q$  can make the variance of the estimate (2) considerably lower than that of (1). The optimal choice of  $q(\mathcal{H})$  is the posterior  $p(\mathcal{H} | \mathcal{D})$  (Stephens and Donnelly, 2000), but this is not known in general. There has been much research in the design of efficient proposal distributions (Stephens and Donnelly, 2000, Fearnhead and Donnelly, 2001, De Iorio and Griffiths, 2004, De Iorio et al., 2005, Griffiths et al., 2008, Hobolth et al., 2008, Jenkins and Griffiths, 2011).

### 2.3 Sequential importance resampling

The set  $\{(H_0^{(j)}, H_{-1}^{(j)}, \dots, H_{-m}^{(j)}) : j = 1, \dots, M\}$  can be constructed either serially or in parallel. Provided the resources are available to store each reconstruction, or *stream* (Liu and Chen, 1998), simultaneously, the latter approach can be advantageous. Intuitively, reconstructions that are doing badly can be discarded and replaced by

duplicating those whose performance is currently superior. Suppose we have the partial reconstructions  $\{\mathcal{H}_K^{(j)} : j = 1, \dots, M\}$  as far back as step  $K$ , where  $\mathcal{H}_K^{(j)} = (H_0^{(j)}, \dots, H_{-K}^{(j)})$  has current SIS weight  $w_K^{(j)}$  [equation (3)]. A vanilla resampling algorithm is:

1. Multinomially draw  $M$  new samples with replacement from  $\{\mathcal{H}_K^{(j)} : j = 1, \dots, M\}$  specified by the probabilities  $\{a^{(j)} : j = 1, \dots, M\}$ .
2. Set the weight of a stream resampled with probability  $a^{(j)}$  to be  $w_K^{(j)} / (Ma^{(j)})$ .

Typically one chooses  $a^{(j)} \propto w_K^{(j)}$ ; the performance of a stream is proportional to its current SIS weight. Step 2 of the algorithm ensures we properly account for the fact that only surviving streams are ultimately used in Monte Carlo estimation.

A number of improvements to this basic algorithm are available (for recent review see Doucet and Johansen, 2011). In particular:

1. Variation introduced as a consequence of multinomial resampling can be reduced without disrupting the expected number of resampled copies of each stream; alternatives include *systematic resampling* and *residual resampling*.
2. Resampling need not take place according to a deterministic schedule, say every  $K$  steps. An alternative, *dynamic* approach, is to resample only when the variation in the weights exceeds some threshold.
3. Resampling reduces the diversity of the current set of streams for an expected improvement at future steps. Thus, there is no gain in resampling at the final step, and we prohibit resampling here.

In this article we will use residual resampling (Liu and Chen, 1998) and a dynamic resampling schedule following Chen et al. (2005), by resampling only if the coefficient of variation of the SIS weights exceeds some threshold parameter  $B \geq 0$ . The idea here is that resampling is necessary when only a few streams become important to the likelihood estimate; we have some streams from a region of relatively high posterior probability with large SIS weights, and the remainder with weights effectively zero. Conversely, when the coefficient of variation of the SIS weights is low then the utility of each stream to the likelihood estimate is comparable, and resampling is unlikely to offer further improvement.

## 2.4 Stopping-time resampling

The resampling algorithm as described above relies on the idea that the current SIS weight is in positive correlation with—and therefore a good predictor of—the final

SIS weight. In most SMC applications this is indeed the case, but for the coalescent model the correlation between current and final SIS weight may be weak or even negative (Fearnhead, 2008). One reason for this reduction is that at stage  $K$  a stream could have a low SIS weight because it is already quite close to the MRCA. The total number of steps  $m$  required by each stream is really a random quantity, and those streams destined to require fewer steps overall may be punished by a resampling algorithm that fails to take this into account. Chen and Liu (2000) and Chen et al. (2005) provide a way around this problem by showing that resampling can still be carried out when streams are allowed to proceed a varying number of steps before being compared by the resampling algorithm. A summary of the SIS procedure with dynamic resampling and general stopping-times  $T_1, T_2, \dots$ , is as follows:

1. Set  $l = 0$ ,  $T_0 = 0$ , and  $w_0^{(j)} = 1$  for each  $j = 1, \dots, M$ .
2. If the coefficient of variation of  $\{w_{T_l}^{(j)} : j = 1, \dots, M\}$  exceeds  $B$  then apply the resampling algorithm described above.
3. For each  $j$ , continue to reconstruct the genealogy  $\mathcal{H}_{T_l}^{(j)}$  back in time according to the proposal distribution  $q(H_{-k} | H_{-k+1})$  until we reach  $k = T_{l+1}$ , and compute  $w_{T_{l+1}}^{(j)}$ . Increment  $l$ .
4. If the streams have not yet reached their MRCA then go to step 2.
5. Estimate the likelihood with the sample mean  $\sum_{j=1}^M w_m^{(j)} / M$ .

For genetic data drawn from a nonrecombining locus under a finite-alleles model, Chen and Liu (2000) and Chen et al. (2005) suggest a more natural timescale for reconstruction to be the stopping-times  $T_1^C, T_2^C, \dots, T_{n-2}^C$ , where

$$T_l^C := \inf\{k \in \mathbb{N} : C_k^{(j)} \geq l\}, \quad l = 1, \dots, n-2, \quad (4)$$

and  $C_k^{(j)}$  is the number of coalescence events encountered by stream  $j$  after  $k$  steps. (Henceforth this choice of stopping-time is referred to as *stopping scheme C*, denoted by the superscript C.) By comparing streams only after they have encountered the same number of coalescence events, the SIS procedure with stopping-times (4) operates on a more natural timescale. Notice that the definition of stopping-times is independent of the choice of proposal distribution  $q$ , and can be designed without a particular  $q$  in mind. One should note, however, that a given stopping scheme will perform better with some proposal distributions than with others. For example, if we had access to the optimal proposal distribution then resampling under any stopping scheme cannot improve our estimate and will typically make it worse.

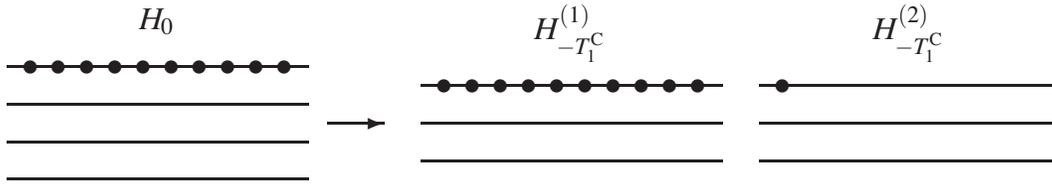


Figure 1: An example dataset,  $H_0$ , of four haplotypes with ten segregating sites (black dots). It is evolved back in time according to a given proposal distribution. Shown on the right are two possible configurations at the first stopping-time  $T_1^C$ .

### 3 A new stopping-time for sequence data

Stopping-time resampling has been applied without success to data assumed to conform to the infinite sites model [Larribe (2003), also mentioned in preprints of Fearnhead and Donnelly (2001) and Hobolth et al. (2008)], suggesting that the number of coalescence events may not be the most effective timescale under this model. In fact, as is demonstrated below, this resampling scheme can perform rather worse than if resampling is not used at all. Some intuition for this is as follows.

In many SMC applications the (normalized) sequence of SIS weights is a martingale (Kong, Liu, and Hong, 1994), so we should expect the effect of ‘bad’ decisions taken by a stream to be propagated to the future. Thus, its current and future weight are in positive correlation, and resampling can intercept these poor streams early. I refer to this component of the effect on SIS weight as *extrinsic*. For the coalescent model there is also an *extrinsic* component: an example would be for  $q$  to propose recombination events when  $\rho$  is small and recombination events need not be invoked to explain the data. However, for the coalescent model there is also an *intrinsic* component to the SIS weight, which depends upon decisions that are in some sense unavoidable. Regardless of how a stream reconstructs a genealogy, there are some events it must propose in order for the resulting ARG to be compatible with the data. An example is the effect on the SIS weight caused by proposing a mutation event when such a mutation is required at least somewhere in the genealogical history. Suppose for example that  $\theta$  is small, so the prior probability of seeing a mutation under  $p$  is small and hence proposing a mutation event generally causes the SIS weight of a stream to be reduced. Importantly, a resampling algorithm could punish a stream for having proposed such a move, even though *every* stream will eventually have to do so. In this respect the intrinsic component of SIS weight *decreases* the correlation between its current and final value.

Our aim then should be to perform resampling in such a way that we compare only the extrinsic components of the SIS weights between streams. Stopping-time resampling achieves this by running streams until their intrinsic components

Stopping scheme	Likelihood estimate	Standard error	Relative error	Resampling events
None	$3.83 \times 10^{-7}$	$1.52 \times 10^{-9}$	$3.40 \times 10^{-3}$	0
C	$3.71 \times 10^{-7}$	$8.61 \times 10^{-9}$	$-2.90 \times 10^{-2}$	2
CM	$3.85 \times 10^{-7}$	$1.06 \times 10^{-9}$	$8.26 \times 10^{-3}$	3
SCM	$3.82 \times 10^{-7}$	$0.21 \times 10^{-9}$	$4.15 \times 10^{-4}$	1
‘True’ value	$3.82 \times 10^{-7}$	$4.61 \times 10^{-11}$	-	0

Table 1: Likelihood estimates (at  $\theta = 1, \rho = 0$ ) of the dataset shown in Figure 1 under various stopping schemes, together with standard errors and number of resampling events incurred. Estimates are based on  $M = 10^4$  runs with  $B = 1$ , except for the ‘true’ value which is based on  $M = 10^7$  runs with  $B = \infty$ . The ‘true’ value is used to estimate the relative errors.

are approximately equal. The stopping-times in (4) estimate this by the number of coalescence events encountered. Crucially, when a dataset conforms to the infinite sites model and exhibits  $s$  segregating sites, we know that there will be contributions to intrinsic weight from precisely  $s$  mutation events. Defining a stopping-time by the number of coalescence events ignores how many of these  $s$  hurdles each stream has overcome. A simple example is shown in Figure 1, for which stopping-time resampling under scheme C diminishes the accuracy of the likelihood estimate by an order of magnitude, compared to no resampling (Table 1). Of the two example configurations shown in Figure 1 at  $T_1^C$ , one is clearly much closer to the MRCA than the other, though the resampling algorithm does not use this information. Indeed, if mutation events tend to decrease the SIS weight then the configuration closer to the MRCA will tend to be *removed* by resampling.

### 3.1 Infinite sites mutation

A simple correction to equation (4) is to count coalescence *or* mutation events, which will be denoted as scheme CM:

$$T_l^{\text{CM}} := \inf\{k \in \mathbb{N} : C_k^{(j)} + M_k^{(j)} \geq l\}, \quad l = 1, \dots, n + s - 2, \quad (5)$$

where  $M_k^{(j)}$  is the number of mutation events encountered by stream  $j$  after  $k$  steps. This is an improvement over C (Table 1), but (i) the total number of stopping-times is now fixed at  $n + s - 2$ , which may be large and cannot be adjusted; and (ii) more seriously, we should not expect coalescence and mutation events to have the same effect on the SIS weight. Why should a stream that has encountered three coalescence events be compared against one that has encountered, say, one

coalescence and two mutations? To solve these problems we imbue the space of configurations with a pseudometric on partially reconstructed genealogies:

$$d_{\text{SCM}}[H_{-k_i}^{(i)}, H_{-k_j}^{(j)}] := \nu \left[ |C_{k_i}^{(i)} - C_{k_j}^{(j)}| + \mu |M_{k_i}^{(i)} - M_{k_j}^{(j)}| \right], \quad (6)$$

for fixed  $\nu, \mu > 0$ , and define a new stopping-time for stream  $j$  by

$$T_l^{\text{SCM}} := \inf\{k \in \mathbb{N} : d_{\text{SCM}}[H_{-k}^{(j)}, H_{-m}] \leq d_{\text{SCM}}[H_0, H_{-m}] - l\}. \quad (7)$$

This is a scaled (SCM) version of the stopping scheme CM. The parameter  $\nu$  is for convenience, allowing us to specify the grain of the stopping-times. Throughout,  $\nu$  is set so that each stream encounters precisely  $n - 2$  stopping-times during the reconstruction, the same number as under scheme C. The parameter  $\mu$  specifies an exchange rate, in terms of the effect on the SIS weight, between coalescence and mutation events. For SCM we will use the ratio of the expected number of coalescence events to the expected number of mutation events, in a sample of size  $n$ :

$$\mu = \frac{n-1}{\theta \sum_{j=1}^{n-1} j^{-1}}. \quad (8)$$

One justification for this choice is that  $T_l^{\text{SCM}}$  enjoys the following properties of convergence in distribution:

$$\begin{aligned} T_l^{\text{SCM}} &\rightarrow \inf\{k \in \mathbb{N} : M_k^{(j)} \geq l\} \text{ as } \theta \rightarrow 0, \text{ and} \\ T_l^{\text{SCM}} &\rightarrow \inf\{k \in \mathbb{N} : C_k^{(j)} \geq l\} = T_l^{\text{C}} \text{ as } \theta \rightarrow \infty. \end{aligned}$$

Thus for example when the mutation rate is small, whether a stopping-time has been reached is largely determined simply by the number of mutation events a stream has overcome. Notice that the previous stopping-times can also be recast in this manner; scheme C corresponds to the special case  $\nu = 1, \mu = 0$ , while scheme CM corresponds to  $\nu = 1 = \mu$ . The three stopping schemes are illustrated in Figure 2. A number of other choices were explored, with less improvement (Jenkins, 2008).

### 3.2 Multiple loci and crossover recombination

The above framework can be extended to multiple loci; for clarity, exposition is restricted to a two-locus model in which recombination events always occur at position  $\frac{1}{2}$  in the unit interval representing the haplotype. Refer to the two loci as A and B. If they have separate mutation parameters  $\theta_A, \theta_B$ , then we simply introduce two stopping-time parameters  $\mu_A, \mu_B$  corresponding to (8), replacing  $\theta$  respectively

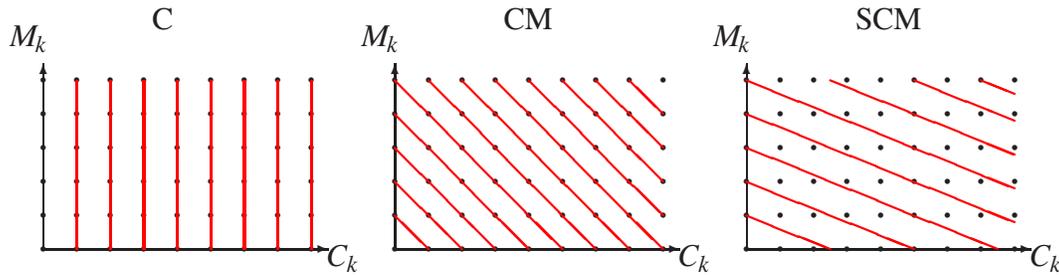


Figure 2: Illustration of each stopping scheme defined in the main text. Contours (solid red lines) represent stopping-times. The trajectory of a stream is a random walk projected onto two axes counting the number of coalescence events  $C_k$  and the number of mutation events  $M_k$ , starting at the origin. In the example shown, the rescaling under SCM corresponds to a small value of  $\theta$ ; now, more than two coalescence events are required to reach the next stopping-time compared with only one mutation event.

with  $\theta_A$  and  $\theta_B$ . We then count mutation events occurring to the two loci separately and adjust  $\nu$  so that the total number of stopping-times is unchanged.

Recombination introduces an additional complication. The state space of configurations now records the mutant alleles carried by each haplotype and also the regions over which a haplotype is ancestral to the present-day sample. For example, the two parents of a recombining haplotype ancestral over the whole of  $[0, 1]$  are themselves ancestral over only part of the interval, one over  $[0, \frac{1}{2}]$  and the other over  $(\frac{1}{2}, 1]$ . See Griffiths and Marjoram (1996) for full details of this concept and Jenkins and Griffiths (2011) for the specialization to two loci. Stopping-times are adjusted as follows. Rather than track the total number of coalescence events encountered, we should track the total length of ancestral material remaining. For stream  $j$  at step  $k$  this quantity will be denoted by  $L_k^{(j)}$ . For a fully-specified sample of  $n$  haplotypes we have  $L_0^{(j)} = n$ . This quantity then decreases, possibly taking on non-integral values, until  $L_m^{(j)} = 1$ . In each of the metrics above we replace  $C_k^{(i)}$  and  $C_k^{(j)}$  with  $L_k^{(i)}$  and  $L_k^{(j)}$  respectively (see also Larribe, 2003). Combining these observations, we arrive at a full, two-locus, definition for SCM:

$$d_{\text{SCM}}[H_{-k_i}^{(i)}, H_{-k_j}^{(j)}] = \nu \left[ |L_{k_i}^{(i)} - L_{k_j}^{(j)}| + \mu_A |M_{k_i}^{A(i)} - M_{k_j}^{A(j)}| + \mu_B |M_{k_i}^{B(i)} - M_{k_j}^{B(j)}| \right], \quad (9)$$

where  $M_k^{A(j)}$  and  $M_k^{B(j)}$  are the number of mutation events encountered by stream  $j$  at step  $k$ , at locus A and B respectively. Corresponding multi-locus definitions for schemes C and CM are obtained by setting  $\nu = 1$ ,  $\mu_A = \mu_B = 0$ , and  $\nu = \mu_A = \mu_B = 1$ , respectively. In the following section we will compare the performance of

the stopping schemes C, CM, and SCM, given by the appropriate choices for  $\nu$ ,  $\mu_A$ , and  $\mu_B$  inserted into equation (9).

### 3.3 Simulation study

To evaluate the performance of the stopping-times developed in the previous sections, we use the proposal distribution described in Jenkins and Griffiths (2011) for a two-locus, infinite sites model; recall though that stopping-time resampling is applicable to any sequential proposal distribution. I simulated a large number of datasets using `ms` (Hudson, 2002) under this model and under a variety of  $\rho$ -values; motivated by the human data examined in Jenkins and Griffiths (2011) I fixed  $\theta_A = \theta_B = 3.5$  for simulating data and for the SIS driving values. The study was restricted to a modest  $n = 20$  and  $M = 10^4$  runs, so that one is able to obtain an independent, ‘true’ estimate of the likelihood for each dataset using a much larger number of runs ( $M = 10^7$ , without resampling). The unsigned relative error of the shorter runs compared to this ‘true’ value serves as our measure of performance:

$$\left| \frac{\widehat{L}(\Psi) - L(\Psi)}{L(\Psi)} \right|, \quad (10)$$

where  $\Psi = (\theta_A, \theta_B, \rho)$ ,  $\widehat{L}(\Psi)$  is the likelihood estimate from a short run with resampling, and  $L(\Psi)$  is the likelihood estimate from the long run without resampling. The *distribution* of the unsigned relative error (10) can be estimated for a given dataset by constructing many independent estimates  $\widehat{L}(\Psi)$  on that dataset.

It is worth elaborating on this approach to obtaining a reliable measure of accuracy. Two other commonly used tools are the *standard error* (SE),  $s_e(k) = \sigma_k / \sqrt{M}$ , and the *effective sample size*,  $\text{ESS}(k) = M / (1 + \sigma_k^2 / \mu_k^2)$ , where  $\mu_k$  is the mean and  $\sigma_k$  the standard deviation of SIS weights constructed to the  $k$ th step (Kong et al., 1994). There are two issues with these tools. First, they must be estimated through the sample mean  $\widehat{\mu}_k$  and sample standard deviation  $\widehat{\sigma}_k$ , which introduces statistical difficulties:  $\widehat{\sigma}_k$  is likely to *underestimate*  $\sigma_k$  even though it is unbiased, leading to an underestimated likelihood (Stephens and Donnelly, 2000) and an overconfidence in its accuracy. As a consequence, Fearnhead and Donnelly (2001) observe that while a low ESS is indicative of a poor estimate, a large ESS does not guarantee an accurate one. Second, these tools are designed for streams propagating independently, and we should not expect them to be meaningful when we also employ a resampling algorithm. For example, resampling so much that only one genealogy is left in the sample at the final step  $m$  will result in  $\widehat{\sigma}_m = 0$  and hence  $\text{ESS}(m) = M$ , but it does not follow that the resulting likelihood estimate is optimal. Throughout the simulations presented in this article, I observed SE and ESS

to be better indicators of how often resampling had been performed than of the accuracy of the likelihood estimates, showing a roughly monotonic relationship with the number of resampling events incurred.

For these reasons it is necessary to look at the variability in likelihood estimates  $\hat{L}(\Psi)$  from independent experiments on a given dataset, and, where possible, to construct a ‘true’ estimate against which to compare these experiments, as in (10). Of course, in real applications the latter will be unavailable.

First we examine the effect of the choice of stopping scheme from among those defined above. I computed a likelihood estimate for each scheme, using simulated datasets as described above. Since we do not know in advance the appropriate choice for the resampling parameter  $B$ , this procedure was repeated over a range of values. The range  $B = 2^{-4}, 2^{-3}, \dots, 2^{14}$ , was sufficient to encompass the whole range of responses, from resampling at every step ( $B = 2^{-4}$ ) to no resampling at all ( $B = 2^{14}$ ). For each choice of  $B$ , each stopping scheme, and each dataset I computed 25 independent likelihood estimates to find the distribution of the unsigned relative error defined in (10). The results for a dataset simulated with  $\rho = 5$  are shown in detail in Figure 3. As is clear from the boxplots, the stopping scheme SCM generally exhibits the lowest relative error in the likelihood. For many choices of  $B$  the likelihood estimate resulting from scheme C was essentially zero, resulting in an unsigned relative error of 1. This error was higher than if we had not done any resampling (right-most boxplot at  $B = 2^{14}$ ). For all choices of  $B$ , resampling under scheme C offered no improvement and in fact diminished the accuracy of the likelihood estimate. This pattern was typical of other datasets simulated under these parameters (not shown), though some showed a modest improvement near  $B = 2^{12}$ . For such datasets, this confirms the observation of Chen et al. (2005) that a small number of resampling events is preferable and too many becomes inefficient. This is also true here of the schemes CM and SCM, for which the optimal number of resampling events is small and positive, roughly 2–5. Importantly, scheme SCM is highly robust to misspecification of  $B$ ; resampling too often is still preferable to no resampling at all. This is useful because in practical applications the optimal choice for  $B$  is not known: it will vary with the size and complexity of the dataset under examination and with the efficiency of the proposal distribution. On the other hand, resampling too often under scheme C or CM can be counterproductive.

To examine the effect of recombination rate on these results, this experiment was repeated for data simulated under a variety of choices of  $\rho$ . Figure 4 gives six examples. Again, our measure of performance is the distribution of the unsigned relative error (10). For clarity we focus on the median of this distribution. For the dataset simulated under a small recombination rate (panel  $\rho = 0.1$ ) the likelihood estimate is already quite accurate and resampling under any scheme offers little further improvement. Relative errors are smaller here as it is easier for the

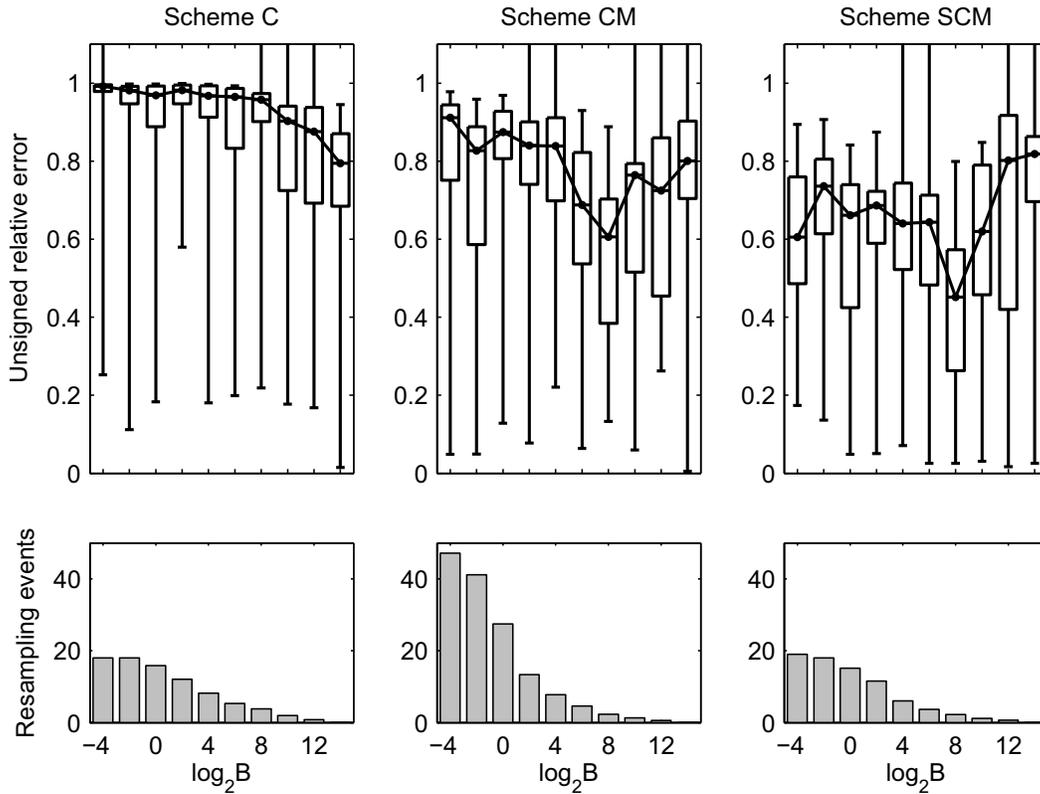


Figure 3: Performance of each stopping scheme on a simulated dataset with  $\rho = 5$ . The upper panels show boxplots of the distribution of the unsigned relative error of 25 independent likelihood estimates, repeated across a range of  $B$ . The lower panels show the number of resampling events incurred for that choice of  $B$ .

proposal distribution to explore those genealogies with few recombination events. The schemes CM and SCM perform similarly, although resampling under scheme C still diminishes the accuracy of its likelihood estimate. For intermediate recombination rates the performance of the stopping schemes on all the datasets shown here are in a clear ordering, with SCM producing the smallest relative errors and C the largest. In most cases the relative error of the likelihood estimate is minimized by using SCM and a small number of resampling events, as was the case for Figure 3. For the dataset simulated under a large recombination rate (panel  $\rho = 100$ ), with this number of runs none of the stopping schemes seems to have improved the likelihood estimate significantly, using any choice of  $B$ .

To confirm that these observations were not restricted to those datasets used in Figures 3 and 4, I repeated the procedure on a larger collection of 100 datasets for each of the same various choices of recombination parameter. For each combination

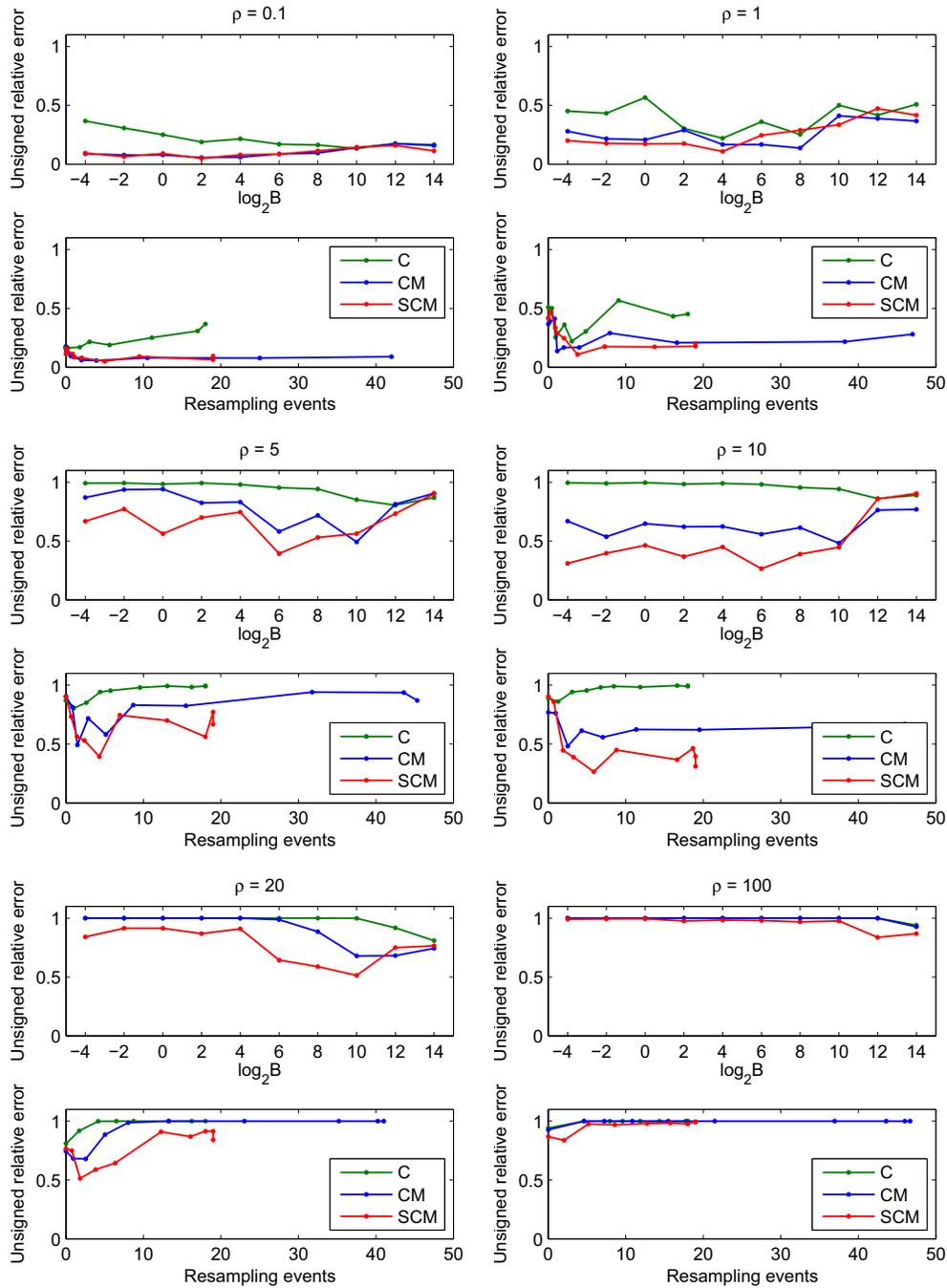


Figure 4: Performance of each stopping scheme on six datasets simulated under various choices of  $\rho$ . Each pair of plots shows the median unsigned relative error across 25 independent likelihood estimates, as a function of  $B$  (upper panels) and as a function of the number of resampling events (lower panels).

Table 2: Performance of each stopping scheme on 100 simulated datasets under various choices of  $\rho$ . Shown are the median (lower, upper quartiles) of the unsigned relative errors of these 100 likelihood estimates when  $B$  is chosen to bring the mean number of resampling events close to two in each case.

Scheme	$\rho = 0.1$		$\rho = 1$		$\rho = 5$	
None	0.14	(0.04, 0.28)	0.28	(0.10, 0.62)	0.71	(0.33, 0.88)
C	0.09	(0.04, 0.34)	0.23	(0.09, 0.46)	0.74	(0.42, 0.96)
CM	0.06	(0.03, 0.11)	0.17	(0.09, 0.34)	0.44	(0.23, 0.58)
SCM	0.07	(0.04, 0.14)	0.16	(0.07, 0.32)	0.41	(0.21, 0.73)
Scheme	$\rho = 10$		$\rho = 20$		$\rho = 100$	
None	0.79	(0.41, 0.93)	0.77	(0.48, 0.95)	0.86	(0.59, 0.98)
C	0.91	(0.60, 1.00)	0.91	(0.52, 0.99)	1.00	(0.96, 1.00)
CM	0.53	(0.27, 0.78)	0.67	(0.41, 0.94)	1.00	(0.94, 1.00)
SCM	0.40	(0.20, 0.62)	0.57	(0.31, 0.75)	0.93	(0.71, 0.99)

of  $\rho$  and stopping scheme, I used a preliminary simulation to pick the value of  $B$  that brought the mean number of resampling events close to two; other parameters remained as before. Results are shown in Table 2. As is evident from the table, the patterns seen in Figure 4 do apply more generally. The stopping scheme SCM clearly outperforms the others over a wide range of  $\rho$ , except for  $\rho = 0.1$  where there is little difference between schemes and  $\rho = 100$  where resampling under any scheme offers no improvement.

## 4 Microsatellite data

Similar reasoning to that used for infinite sites data can be used to design stopping-times under other mutation models. In this section we consider a single locus mutating under a stepwise mutation model. Under this model, the allele of a sampled haplotype is an integer, and a mutation event increments or decrements this value by one, each with probability  $\frac{1}{2}$ . This model is a simple representation for the number of repeat copies seen at a microsatellite locus.

The key difference between this and the previous mutation model is that the number of mutation events is no longer fixed. A mutation may or may not bring a configuration closer to its MRCA, and the metric used by the proposed stopping-time should account for this. The term  $M_k^{(i)}$  in (6) is no longer a single-valued function for  $H_{-m}$ , and so we replace it with the following. Let  $n_k^{(j)}(z)$  denote the

number of copies of allele  $z \in \mathbb{Z}$  in the configuration of stream  $j$  at step  $k$  of the reconstruction, and define

$$D_k^{(j)} = \max\{z \in \mathbb{Z} : n_k^{(j)}(z) > 0\} - \min\{z \in \mathbb{Z} : n_k^{(j)}(z) > 0\},$$

what will be referred to as the *diameter* of the configuration. Although the diameter of a configuration will fluctuate during the reconstruction of a genealogical history of the observed sample, it must approach zero as the reconstruction approaches its MRCA, and is thus appropriate for use in our stopping-time:

$$d_{\text{SCD}}[H_{-k}^{(i)}, H_{-k}^{(j)}] := \nu \left[ |C_{k_i}^{(i)} - C_{k_j}^{(j)}| + \mu |D_{k_i}^{(i)} - D_{k_j}^{(j)}| \right],$$

with

$$T_l^{\text{SCD}} := \inf\{k \in \mathbb{N} : d_{\text{SCD}}[H_{-k}^{(i)}, H_{-m}^{(j)}] \leq d_{\text{SCD}}[H_0, H_{-m}] - l\}, \quad (11)$$

and  $\nu$ ,  $\mu$ , and  $C_{k_i}^{(j)}$  defined as before. Incidentally, a more suitable choice for  $\mu$  than that given in equation (8) would be to replace the expected number of mutation events in its denominator with the expected diameter of a sample of size  $n$ . To my knowledge this is not known in closed-form except in the special case  $n = 2$  (Ohta and Kimura, 1973). Since the choice in (8) worked well, I do not pursue this further.

To gauge the improvement from this new stopping-time we revisit a dataset examined by Chen et al. (2005), who used a sample of  $n = 296$  allele counts sampled from locus G10M of brown bears from the Western Brooks Range of Alaska (Paetkau, Waits, Clarkson, Craighead, and Strobeck, 1997):

$$(n_0(z))_{z=98,99,\dots,117} = \{0, 0, 0, 0, 0, 24, 134, 16, 32, 81, 0, 8, 0, 1, 0, 0, 0, 0, 0, 0\}.$$

(This has initial diameter 8.) Chen et al. (2005) applied the proposal distribution of Stephens and Donnelly (2000) with driving value  $\theta = 6$  and used what I have called stopping scheme C. They found that it offered more accurate likelihood estimation compared to the case without resampling. I performed a similar analysis using the stopping scheme SCD defined in equation (11) (and its unscaled counterpart, CD). Since it is difficult to further improve on the results of Chen et al. (2005), I reverted to the simpler proposal distribution of Griffiths and Tavaré (1994), also with driving value  $\theta = 6$ . For each stopping scheme, 25 independent likelihood estimates at  $\theta = 6$  were obtained from  $M = 10^4$  runs (setting  $B = 10$ ). As in previous sections, accuracy was measured by computing the distribution of the unsigned relative error (10). For this purpose we require one longer simulation of  $M = 10^7$  runs without resampling; here I used the more efficient proposal distribution of De Iorio et al. (2005). Results are shown in Figure 5, which confirms that stopping scheme C improves on the implementation without resampling. We also find that stopping scheme SCD offers significant further improvement.

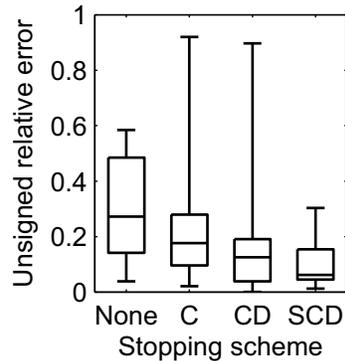


Figure 5: Performance of each stopping scheme on the brown bear dataset given in the main text. For each stopping scheme a boxplot was constructed from 25 independent likelihood estimates.

## 5 Discussion

In this article we have overcome some limitations of the stopping scheme suggested by Chen and Liu (2000) and Chen et al. (2005) for improving the efficiency of SIS methods of likelihood estimation under the coalescent model. In particular, the schemes denoted SCM and SCD are recommended, since they incorporate a number of important features giving a marked improvement in the accuracy of likelihood estimation. First, they impose no computational burden above that of scheme C (whose own burden over a regular SIS algorithm is very small). Second, they combine information from the number of both coalescence and mutation events in an appropriate way. Third, they are robust to the choice of tuning parameter  $B$ . In particular, resampling too often is seldom worse than not resampling at all.

There is obvious scope for further improvements to the stopping schemes proposed here, and for the design of stopping schemes under other models. We have recast the problem to one of designing suitable (pseudo-)metrics on the space of partially reconstructed genealogies of the data, which will further help with intuition in new situations. For example, failures of the four-gamete test visible in infinite sites data require recombination events that are another obvious source of the ‘intrinsic’ contribution to SIS weight. One could add a third dimension to the plots in Figure 2 by measuring the minimum remaining number of recombination events required for a stream to reach its MRCA, although this quantity is not a simple function of the configuration and must be approximated or computed algorithmically (e.g. Lyngsø, Song, and Hein, 2005).

The challenges encountered in this article demonstrate that ideas from SMC are not wholly transferable to all situations in which SIS can be used. Here it was

important to adjust the resampling algorithm to account for the fact that the current SIS weight is a poor predictor of the final SIS weight, and it has been useful to think about a conceptual separation between ‘intrinsic’ and ‘extrinsic’ components of the SIS weight. These complications are not restricted to coalescent models. Modifications of SMC ideas to account for related problems also arise in, for example, solving partial differential equations (Chen et al., 2005), polypeptide folding (Zhang and Liu, 2002), and sampling paths of diffusion bridges (Lin, Chen, and Mykland, 2010). In addition to stopping-time resampling, another strategy to account for the fact that current and future SIS weight may be in poor correlation is to estimate the future SIS weight of a stream by pilot exploration (Zhang and Liu, 2002, Jenkins, 2008, Lin et al., 2010), and to adjust the resampling probabilities  $\{a^{(j)} : j = 1, \dots, M\}$  accordingly. That is, at a resampling checkpoint one resamples a stream based on its expected SIS weight a few steps into the future, as estimated by a further small Monte Carlo sample of its possible future configurations. One allows a pilot ‘team’ to explore either forwards from the current position of the stream (Zhang and Liu, 2002) or backwards from the endpoint (Lin et al., 2010). Although some of these methods were developed with a specific application in mind, each is in effect dealing with a closely related underlying problem. It would be of great interest both to better characterize models for which these issues arise and to unify the optimal approach to their solution.

## References

- Beaumont, M. A., W. Zhang, and D. J. Balding (2002): “Approximate Bayesian computation in population genetics,” *Genetics*, 162, 2025–2035.
- Chen, Y. and J. S. Liu (2000): “Discussion on “Inference in molecular population genetics” by M. Stephens and P. Donnelly,” *Journal of the Royal Statistical Society: Series B*, 62, 644–645.
- Chen, Y., J. Xie, and J. S. Liu (2005): “Stopping-time resampling for sequential Monte Carlo methods,” *Journal of the Royal Statistical Society: Series B*, 67, 199–217.
- De Iorio, M. and R. C. Griffiths (2004): “Importance sampling on coalescent histories I,” *Advances in Applied Probability*, 36, 417–433.
- De Iorio, M., R. C. Griffiths, R. Leblois, and F. Rousset (2005): “Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models,” *Theoretical Population Biology*, 68, 41–53.
- Doucet, A. and A. M. Johansen (2011): “A tutorial on particle filtering and smoothing: fifteen years later,” in D. Crisan and B. Rozovskii, eds., *The Oxford handbook of nonlinear filtering*, Oxford University Press.

- Fearnhead, P. (2008): “Computational methods for complex stochastic systems: a review of some alternatives to MCMC,” *Statistics and Computing*, 18, 151–171.
- Fearnhead, P. and P. Donnelly (2001): “Estimating recombination rates from population genetic data,” *Genetics*, 159, 1299–1318.
- Griffiths, R. C., P. A. Jenkins, and Y. S. Song (2008): “Importance sampling and the two-locus model with subdivided population structure,” *Advances in Applied Probability*, 40, 473–500.
- Griffiths, R. C. and P. Marjoram (1996): “Ancestral inference from samples of DNA sequences with recombination,” *Journal of Computational Biology*, 3, 479–502.
- Griffiths, R. C. and S. Tavaré (1994): “Simulating probability distributions in the coalescent,” *Theoretical Population Biology*, 46, 131–159.
- Hey, J. and R. Nielsen (2004): “Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*,” *Genetics*, 167, 747–760.
- Hobolth, A., M. Uyenoyama, and C. Wiuf (2008): “Importance sampling for the infinite sites model,” *Statistical Applications in Genetics and Molecular Biology*, 7, Article 32.
- Hudson, R. R. (2002): “Generating samples under a Wright-Fisher neutral model of genetic variation,” *Bioinformatics*, 18, 337–338.
- Jenkins, P. A. (2008): *Importance sampling on the coalescent with recombination*, Ph.D. thesis, University of Oxford.
- Jenkins, P. A. and R. C. Griffiths (2011): “Inference from samples of DNA sequences using a two-locus model,” *Journal of Computational Biology*, 18, 109–127.
- Kong, A., J. S. Liu, and W. H. Hong (1994): “Sequential imputations and Bayesian missing data problems,” *Journal of the American Statistical Association*, 89, 278–288.
- Kuhner, M. K., J. Yamato, and J. Felsenstein (2000): “Maximum likelihood estimation of recombination rates from population data,” *Genetics*, 156, 1393–1401.
- Larribe, F. (2003): *Cartographie génétique fine par le graphe de recombinaison ancestral*, Ph.D. thesis, University of Montréal, (in French).
- Lin, M., R. Chen, and P. Mykland (2010): “On generating Monte Carlo samples of continuous diffusion bridges,” *Journal of the American Statistical Association*, 105, 820–838.
- Liu, J. S. and R. Chen (1998): “Sequential Monte Carlo methods for dynamic systems,” *Journal of the American Statistical Association*, 98, 1032–1044.
- Lyngsø, R. B., Y. S. Song, and J. Hein (2005): “Minimum recombination histories by branch and bound,” in R. Casadio and G. Myers, eds., *Algorithms in Bioinformatics*, Springer Berlin/Heidelberg, 239–250.

- Ohta, T. and M. Kimura (1973): “A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a genetic population,” *Genetical Research*, 22, 201–204.
- Paetkau, D., L. P. Waits, P. L. Clarkson, L. Craighead, and C. Strobeck (1997): “An empirical evaluation of genetic distance statistics using microsatellite data from bear (Ursidae) populations,” *Genetics*, 147, 1943–1957.
- Paul, J. S. and Y. S. Song (2010): “A principled approach to deriving approximate conditional sampling distributions in population genetics models with recombination,” *Genetics*, 186, 321–338.
- Stephens, M. and P. Donnelly (2000): “Inference in molecular population genetics,” *Journal of the Royal Statistical Society: Series B*, 62, 605–655.
- Tavaré, S., D. J. Balding, R. C. Griffiths, and P. Donnelly (1997): “Inferring coalescence times from DNA sequence data,” *Genetics*, 145, 505–518.
- Wang, Y. and B. Rannala (2008): “Bayesian inference of fine-scale recombination rates using population genomic data,” *Philosophical Transactions of the Royal Society B*, 363, 3921–3930.
- Wilson, I. J., M. E. Weale, and D. J. Balding (2003): “Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities,” 166, 155–201.
- Zhang, J. L. and J. S. Liu (2002): “A new sequential importance sampling method and its application to the two-dimensional hydrophobic-hydrophilic model,” *Journal of Chemical Physics*, 117, 3492–3498.