DOI: 10.1534/genetics.110.125534

An Accurate Sequentially Markov Conditional Sampling Distribution for the Coalescent With Recombination

Joshua S. Paul,* Matthias Steinrücken† and Yun S. Song*,†,1

**Computer Science Division and* †*Department of Statistics, University of California, Berkeley, California 94720*

Manuscript received November 30, 2010
Accepted for publication January 21, 2011

ABSTRACT

The sequentially Markov coalescent is a simplified genealogical process that aims to capture the essential features of the full coalescent model with recombination, while being scalable in the number of loci. In this article, the sequentially Markov framework is applied to the conditional sampling distribution (CSD), which is at the core of many statistical tools for population genetic analyses. Briefly, the CSD describes the probability that an additionally sampled DNA sequence is of a certain type, given that a collection of sequences has already been observed. A hidden Markov model (HMM) formulation of the sequentially Markov CSD is developed here, yielding an algorithm with time complexity linear in both the number of loci and the number of haplotypes. This work provides a highly accurate, practical approximation to a recently introduced CSD derived from the diffusion process associated with the coalescent with recombination. It is empirically demonstrated that the improvement in accuracy of the new CSD over previously proposed HMM-based CSDs increases substantially with the number of loci. The framework presented here can be adopted in a wide range of applications in population genetics, including imputing missing sequence data, estimating recombination rates, and inferring human colonization history.

THE conditional sampling distribution (CSD) describes the probability, under a particular population genetic model, that an additionally sampled DNA sequence is of a certain type, given that a collection of sequences has already been observed. In many important settings, the relevant population genetic model is the coalescent with recombination, for which the true CSD, denoted by π, does not have a known analytic formula. Nevertheless, the CSD π and, in particular, approximations thereof have found a wide range of applications in population genetics.

One important problem in which the CSD plays a fundamental role is describing the posterior distribution of genealogies under the coalescent process. Stephens and Donnelly (2000) showed that the true posterior distribution can be written in terms of π and can be approximated by using an approximate CSD, denoted $\hat{\pi}$. This observation has been used (Stephens and Donnelly 2000; Fearnhead and Donnelly 2001; De Iorio and Griffiths 2004a,b; Fearnhead and Smith 2005; Griffiths *et al.* 2008) to construct importance sampling schemes for likelihood computation and ancestral inference under the coalescent, including extensions such as recombination and population structure. In conjunction with composite-likelihood frameworks (Hudson 2001; Fearnhead and Donnelly 2002), these importance sampling methods have been used, for example, to estimate fine-scale recombination rates (McVean *et al.* 2004; Fearnhead and Smith 2005; Johnson and Slatkin 2009).

Li and Stephens (2003) introduced a different application of the CSD, observing that the probability of sampling a set of haplotypes can be decomposed into a product of CSDs and therefore can be approximated by a product of approximate CSDs $\hat{\pi}$. Similar applications of the CSD have yielded methods for estimating recombination rates (Li and Stephens 2003; Crawford *et al.* 2004; Stephens and Scheet 2005) and gene conversion parameters (Gay *et al.* 2007; Yin *et al.* 2009), for phasing genotype data into haplotype data (Stephens and Scheet 2005), for imputing missing data to improve power in association studies (Stephens and Scheet 2005; Li and Abecasis 2006; Scheet and Stephens 2006; Marchini *et al.* 2007; Howie *et al.* 2009), for inferring ancestry in admixed populations (Price *et al.* 2009), and for inferring demography (Hellenthal *et al.* 2008; Davison *et al.* 2009).

In all applications, the fidelity with which the surrogate CSD $\hat{\pi}$ approximates the true CSD π is critical to the quality of the result. Furthermore, the time required to compute probabilities under the CSD is important, as many of the above methods are now routinely applied to genome-scale data sets. As a result, many approximate CSDs have been proposed,

Supporting information is available online at http://www.genetics.org/cgi/content/full/genetics.110.125534/DC1.

[1] *Corresponding author:* Department of EECS, 683 Soda Hall, no. 1776, University of California, Berkeley, CA 94720-1776. E-mail: yss@cs.berkeley.edu

particularly for the coalescent with recombination. Fearnhead and Donnelly (2001) introduced an approximation in which an additionally sampled haplotype is constructed as an imperfect mosaic of previously sampled haplotypes, with mosaic breakpoints caused by recombination events and imperfections corresponding to mutation events. The resulting CSD, which we denote by $\hat{\pi}_{\mathrm{FD}}$, can be cast as a hidden Markov model (HMM), and the associated conditional sampling probability (CSP) can be computed with time complexity linear in both the number of previously sampled haplotypes and the number of loci. Li and Stephens (2003) proposed a related model that can be viewed as a modification to $\hat{\pi}_{\mathrm{FD}}$ limiting the state space of the HMM, hence providing a constant factor improvement in the time complexity; we denote the corresponding CSD by $\hat{\pi}_{\mathrm{LS}}$.

Following the theoretical work of De Iorio and Griffiths (2004a), Griffiths *et al.* (2008) *derived* an approximate CSD from the Wright–Fisher diffusion process associated with the two-locus coalescent with recombination. More recently, Paul and Song (2010) generalized this work to an arbitrary number of loci and demonstrated that the resulting CSD, which we denote by $\hat{\pi}_{\mathrm{PS}}$, can also be described by a genealogical process. Though it is more accurate than both $\hat{\pi}_{\mathrm{LS}}$ and $\hat{\pi}_{\mathrm{FD}}$, computing the CSP under $\hat{\pi}_{\mathrm{PS}}$ has time complexity superexponential in the number of loci. To ameliorate this limitation, Paul and Song introduced the approximate CSD $\hat{\pi}_{\mathrm{PS},1}$, which follows from prohibiting coalescence events in the genealogical process associated with $\hat{\pi}_{\mathrm{PS}}$. Computing the CSP under $\hat{\pi}_{\mathrm{PS},1}$ has time complexity exponential in the number of loci. Although this is an improvement over the superexponential complexity associated with $\hat{\pi}_{\mathrm{PS}}$, it is still impracticable to use $\hat{\pi}_{\mathrm{PS},1}$ for >20 loci.

In this article, we introduce an alternate approximation that is scalable in the number of loci, while maintaining the key features of $\hat{\pi}_{\mathrm{PS}}$ that lead to high accuracy. Specifically, motivated by the sequentially Markov coalescent (SMC) introduced by McVean and Cardin (2005), we derive a sequentially Markov approximation to $\hat{\pi}_{\mathrm{PS}}$. The key idea is to consider the marginal genealogies at each locus sequentially, using the genealogical description of $\hat{\pi}_{\mathrm{PS}}$. In general, the sequence of marginal genealogies is not Markov, but, as in McVean and Cardin (2005), we make approximations to provide a Markov construction for the sequence. We denote the resulting approximation of $\hat{\pi}_{\mathrm{PS}}$ by $\hat{\pi}_{\mathrm{SMC}}$. The CSD $\hat{\pi}_{\mathrm{SMC}}$ can also be obtained from $\hat{\pi}_{\mathrm{PS}}$ by prohibiting a certain class of coalescence events, a fact that mirrors the relation between the SMC and the coalescent with recombination (McVean and Cardin 2005). We formalize this relation by proving that $\hat{\pi}_{\mathrm{SMC}}$ is, in fact, equal to $\hat{\pi}_{\mathrm{PS},1}$.

Due to its sequentially Markov construction, $\hat{\pi}_{\mathrm{SMC}}$ can be cast as an HMM. Unfortunately, the state space of the HMM is continuous, and so efficient algorithms for CSP computation and posterior inference are not known. Our solution is to discretize the state space. The discretization procedure we develop is related, though not identical, to the Gaussian quadrature method employed by Stephens and Donnelly (2000) and Fearnhead and Donnelly (2001). Although we focus on the CSD problem here, we believe that our general approach has the potential to foster applications of the SMC in other settings as well (see Hobolth *et al.* 2007; Dutheil *et al.* 2009).

Having discretized the continuous state space, we apply standard HMM theory to obtain an efficient dynamic program for computing the CSP under the discretized approximation of $\hat{\pi}_{\mathrm{SMC}}$. The resulting time complexity is linear in both the number of previously sampled haplotypes and the number of loci. This time complexity is the same as that for $\hat{\pi}_{\mathrm{FD}}$ and $\hat{\pi}_{\mathrm{LS}}$ and hence is a substantial improvement over $\hat{\pi}_{\mathrm{PS},1}$. In summary, the work presented here provides a practical approximation to $\hat{\pi}_{\mathrm{PS}}$, which was derived from the diffusion process associated with the coalescent with recombination. Furthermore, as detailed later, the improvement in accuracy of our new CSD over $\hat{\pi}_{\mathrm{FD}}$ and $\hat{\pi}_{\mathrm{LS}}$ increases substantially with the number of loci.

The remainder of this article is organized as follows. In model, we present the necessary notation and background and describe our new CSD $\hat{\pi}_{\mathrm{SMC}}$. We also give an overview of the proof that $\hat{\pi}_{\mathrm{SMC}}$ is equivalent to $\hat{\pi}_{\mathrm{PS},1}$ and demonstrate several other useful properties. In discretization of the hmm, we describe the discretization of $\hat{\pi}_{\mathrm{SMC}}$, and in empirical results, we provide empirical evidence that the discretized approximation performs well, with regard to both accuracy and run time. Finally, in discussion we mention some connections to existing models and describe possible applications and extensions, in particular conditionally sampling more than one haplotype.

MODEL

In this section, we describe the key transition and emission distributions for the HMM underlying $\hat{\pi}_{\mathrm{SMC}}$. Further, we demonstrate that $\hat{\pi}_{\mathrm{SMC}}$ is equivalent to $\hat{\pi}_{\mathrm{PS},1}$, the variant of $\hat{\pi}_{\mathrm{PS}}$ with coalescence disallowed, and also show that the transition density satisfies several useful properties.

Notation: We consider haplotypes in the finite-sites finite-alleles setting. Denote the set of loci by $L = \{1, \ldots, k\}$ and the set of alleles at locus $\ell \in L$ by E_ℓ. Mutations occur at locus $\ell \in L$ at rate $\theta_\ell/2$ and according to the stochastic matrix $\mathbf{P}^{(\ell)} = \left(P^{(\ell)}_{a,a'}\right)_{a,a' \in E_\ell}$. Denote the set of breakpoints by $B = \{(1, 2), \ldots, (k-1, k)\}$, where recombination occurs at breakpoint $b \in B$ at rate $\rho_b/2$.

The space of k-locus haplotypes is denoted by $\mathcal{H} = E_1 \times \ldots \times E_k$. Given a haplotype $\alpha \in \mathcal{H}$, we denote by $\alpha[\ell] \in E_\ell$ the allele at locus $\ell \in L$ and by $\alpha[1:\ell]$ the partial haplotype $(\alpha[1], \ldots, \alpha[\ell])$. A sample configuration of haplotypes is specified by a vector $\mathbf{n} = (n_\alpha)_{\alpha \in \mathcal{H}}$, with n_α being the number of haplotypes of type α in the sample. The total number of haplotypes in the sample is denoted by $|\mathbf{n}| = n$. Finally, we use $\mathbf{e}_\alpha$ to denote the singleton configuration comprising a single α haplotype.

A brief review of the CSD $\hat{\pi}_{\mathrm{PS}}$: The approximate CSD $\hat{\pi}_{\mathrm{PS}}$ is described by a genealogical process closely related to the coalescent with recombination. We provide below a brief description of the framework and refer the reader to Paul and Song (2010) for further details.

Suppose that, conditioned on having already observed a haplotype configuration $\mathbf{n}$, we wish to sample a new haplotype α. Define $\mathcal{A}^*(\mathbf{n})$ to be the nonrandom *trunk* ancestry for $\mathbf{n}$, in which lineages associated with the haplotypes do not mutate, recombine, or coalesce with one another, but rather extend infinitely into the past. We assume that the unknown ancestry associated with $\mathbf{n}$ is $\mathcal{A}^*(\mathbf{n})$ and sample a *conditional ancestry* C associated with α. Within the conditional ancestry, lineages evolve backward in time with the following rates:

Mutation: Each lineage mutates at locus $\ell \in L$ with rate $\theta_\ell/2$, according to $\mathbf{P}^{(\ell)}$.
Recombination: Each lineage undergoes recombination at breakpoint $b \in B$ with rate $\rho_b/2$.
Coalescence: Each pair of lineages coalesces with rate 1.
Absorption: Each lineage is absorbed into each lineage of $\mathcal{A}^*(\mathbf{n})$ at rate 1/2.

When every lineage has been absorbed into $\mathcal{A}^*(\mathbf{n})$, the process terminates. The type of every lineage in C can now be inferred, and a sample for α is generated. An illustration of this process is presented in Figure 1A.

Although a recursion for computing the CSP $\hat{\pi}_{\mathrm{PS}}(\alpha|\mathbf{n})$ is known (Paul and Song 2010, Equation 7), it is computationally intractable, and Paul and Song approximate the genealogical process by disallowing coalescence within the conditional genealogy, denoting the resulting CSD by $\hat{\pi}_{\mathrm{PS},1}$. The recursion for $\hat{\pi}_{\mathrm{PS}}(\alpha|\mathbf{n})$ (Paul and Song 2010, Equation 12) is amenable to dynamic programming, though it still has time complexity exponential in the number k of loci.

The sequentially Markov coalescent: The sequential interpretation of the coalescent with recombination was introduced by Wiuf and Hein (1999). They observed that an ancestral recombination graph (ARG) may be simulated *sequentially* along the chromosome. In particular, the marginal coalescent tree at a given locus can be sampled conditional on the marginal ARG for all previous loci. The full ARG is then sampled by first sampling a coalescent tree at the leftmost locus and then proceeding to the right.

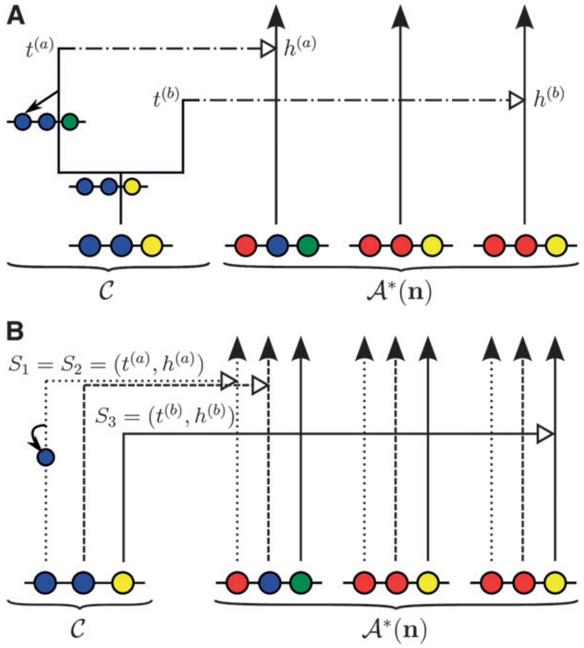
A
$t^{(a)}$
$h^{(a)}$
$t^{(b)}$
$h^{(b)}$
$\mathcal{C}$
$\mathcal{A}^*(\mathbf{n})$
B
$S_1 = S_2 = (t^{(a)}, h^{(a)})$
$S_3 = (t^{(b)}, h^{(b)})$
$\mathcal{C}$
$\mathcal{A}^*(\mathbf{n})$

Figure 1.—Illustration of the corresponding genealogical and sequential interpretations for a realization of $\hat{\pi}_{\mathrm{PS}}(\cdot|\mathbf{n})$. The three loci of each haplotype are each represented by a solid circle, with the color indicating the allelic type at that locus. The trunk genealogy $\mathcal{A}^*(\mathbf{n})$ and conditional genealogy C are indicated. Time is represented vertically, with the present (time 0) at the bottom of the illustration. (A) The genealogical interpretation: Mutation events, along with the locus and resulting haplotype, are indicated by small arrows. Recombination events, and the resulting haplotype, are indicated by branching events in C. Absorption events, and the corresponding absorption time [$t^{(a)}$ and $t^{(b)}$] and haplotype [$h^{(a)}$ and $h^{(b)}$, respectively], are indicated by dotted-dashed horizontal lines. (B) The corresponding sequential interpretation: The marginal genealogies at the first, second, and third locus (S_1, S_2, and S_3) are emphasized as dotted, dashed, and solid lines, respectively. Mutation events at each locus, along with resulting allele, are indicated by small arrows. Absorption events at each locus are indicated by horizontal lines.

McVean and Cardin (2005) proposed a simplification of this process. Though McVean and Cardin presented their work for the infinite-sites model, we state (but do not derive) the analogous results for a finite-sites, finite-alleles model. In their approach, the marginal coalescent tree at locus ℓ is sampled conditional only on the marginal coalescent tree at locus $\ell - 1$. In particular, setting $b = (\ell - 1, \ell) \in B$, (1) recombination breakpoints are realized as a Poisson process with rate $\rho_b/2$ on the marginal coalescent tree at locus $\ell - 1$, (2) the lineage branching from each recombination breakpoint associated with locus $\ell - 1$ is removed, and (3) the lineage branching from each recombination breakpoint associated with locus ℓ is subject to coalescence with other lineages at rate 1. The resulting tree is the marginal genealogy at locus ℓ. This approximation is called the sequentially Markov coalescent (SMC) and is equivalent to a variant of the coalescent with recombination that disallows coalescence between lineages ancestral to disjoint regions of the sequence (McVean and Cardin 2005).

The sequentially Markov CSD $\hat{\pi}_{\mathrm{SMC}}$: We now describe a sequentially Markov approximation to the genealogical process underlying $\hat{\pi}_{\mathrm{PS}}$. Our construction is similar to that given by McVean and Cardin (2005), described above, though the resulting dynamics are less involved since the conditional genealogy is constructed for a *single* haplotype. First, observe that under $\hat{\pi}_{\mathrm{PS}}(\cdot|\,\mathbf{n})$, the marginal conditional genealogy at a given locus $\ell \in L$ is entirely determined by two random variables: the absorption time, which we denote T_ℓ, and the absorption haplotype, which we denote H_ℓ. The present corresponds to time 0 and $T_\ell \in [0, \infty]$. See Figure 1B for an illustration. For convenience, we write $S_\ell = (T_\ell, H_\ell)$ for the random marginal conditional genealogy at locus $\ell \in L$ and $s_\ell = (t_\ell, h_\ell)$ for a realization.

Within the marginal conditional genealogy at locus $\ell \in L$, note that T_ℓ and H_ℓ are independent, with T_ℓ distributed exponentially with parameter $n/2$ and H_ℓ distributed uniformly over the n haplotypes of $\mathbf{n}$. Thus, the marginal conditional genealogy S_ℓ at locus ℓ is distributed with density $\zeta^{(\mathbf{n})}$, where

$$\zeta^{(\mathbf{n})}(s_\ell) = \frac{n_{h_\ell}}{2}\, e^{-(n/2)t_\ell}. \qquad (1)$$

Conditioning on $S_{\ell-1}{=}\,s_{\ell-1}{=}(t_{\ell-1}, h_{\ell-1})$, the marginal conditional genealogy S_ℓ, for $\ell \geq 2$, is sampled by a process analogous to that described above for the SMC. Setting $b = (\ell - 1, \ell) \in B$, the sampling procedure is as follows (see Figure 2 for an accompanying illustration): (1) Recombination breakpoints are realized as a Poisson process with rate $\rho_b/2$ on the marginal conditional genealogy $s_{\ell-1}$; (2) going backward in time, the lineage associated with locus $\ell-1$ branching from each recombination breakpoint is removed, so that only the lineage more recent than the first (*i.e.*, the most recent) breakpoint remains; and (3) the lineage associated with locus ℓ branching from the first recombination breakpoint is absorbed into a particular lineage of $\mathcal{A}^*(\mathbf{n})$ at rate 1/2.

From the above description, we deduce that there is no recombination between loci $\ell-1$ and ℓ with probability $\exp(-(\rho_b/2)t_{\ell-1})$, and in this case the marginal conditional genealogy is unchanged; that is, $S_\ell = s_{\ell-1}$. Otherwise, the time T_{r} of the first recombination breakpoint is distributed exponentially with parameter $\rho_b/2$, truncated at time $t_{\ell-1}$, and the additional time T_{a} until absorption is distributed exponentially with parameter $n/2$. Thus we have $S_\ell = (T_{\mathrm{r}} + T_{\mathrm{a}}, H_\ell)$, where H_ℓ is chosen uniformly at random from the sample $\mathbf{n}$. Taking a convolution of T_{r} and T_{a}, the transition density $\phi^{(\mathbf{n})}_{\rho_b}(\cdot|s_{\ell-1})$ is given by

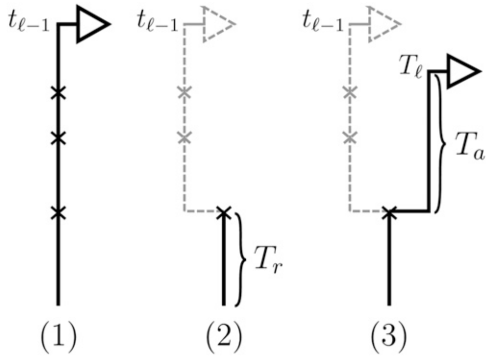

Figure 2.—Illustration of the (Markov) process for sampling the absorption time T_ℓ given the absorption time $T_{\ell-1} = t_{\ell-1}$. In step 1, recombination breakpoints are realized as a Poisson process with rate $\rho_b/2$ on the marginal conditional genealogy with absorption time $t_{\ell-1}$. In step 2, the lineage branching from each breakpoint associated with locus $_{\ell-1}$ is removed, so that only the lineage more recent than the first breakpoint, at time T_{r}, remains. In step 3, the lineage branching from the first recombination breakpoint associated with locus ℓ is absorbed after time T_{a} distributed exponentially with rate $n/2$. Thus, $T_\ell = T_{\mathrm{r}} + T_{\mathrm{a}}$.

$$\phi^{(\mathbf{n})}_{\rho_b}(s_\ell|s_{\ell-1}) = e^{-(\rho_b/2)t_{\ell-1}} \cdot \delta_{s_{\ell-1},s_\ell} + \frac{n_{h_\ell}}{n}\int_0^{t_{\ell-1}\wedge t_\ell}\left(\frac{\rho_b}{2}\,e^{-(\rho_b/2)t}\right)\frac{n}{2}\,e^{-(n/2)(t_\ell-t)}\,dt, \qquad (2)$$

where $t_{\ell-1}\wedge t_\ell$ denotes the minimum of $\mathrm{t}_{\ell-1}$ and t_ℓ.

Finally, conditioning on $S_\ell = s_\ell$, recall that mutations are realized as a Poisson process (*cf.* Stephens and Donnelly 2000) with rate $\theta_\ell/2$. Therefore, a particular allele $a \in E_\ell$ is observed with probability

$$\xi^{(\mathbf{n})}_{\theta_\ell}(a|s_\ell) = e^{-(\theta_\ell/2)t_\ell}\sum_{m=0}^{\infty}\frac{1}{m!}\left(\frac{\theta_\ell}{2}\,t_\ell\right)^m \left[(\mathbf{P}^{(\ell)})^m\right]_{h_\ell[\ell],a}. \qquad (3)$$

Hereafter, we omit the superscript $(\mathbf{n})$ and the subscripts θ_ℓ and ρ_b from these densities, whenever the context is unambiguous.

The sequentially Markov approximation to $\hat{\pi}_{\mathrm{PS}}$ can be cast as a continuous-state HMM. In generating a haplotype α, the observed state, the hidden state, and initial, transition, and emission densities are given by the following:

Observed state: At locus $\ell \in L$, the observed state is the allele $\alpha[\ell]$.
Hidden state: At locus $\ell \in L$, the hidden state is the marginal genealogy $S_\ell = (T_\ell, H_\ell)$.
Initial density: ζ is defined in (1).
Transition density: ϕ is defined in (2).
Emission density: ξ is defined in (3).

Writing $\hat{\pi}_{\mathrm{SMC}}$ for the sequentially Markov approximation to $\hat{\pi}_{\mathrm{PS}}$, we can use the forward recursion (see, *e.g.*, Doucet and Johansen 2008) to get

$$\hat{\pi}_{\mathrm{SMC}}(\alpha \mid \mathbf{n}) = \int f_{\mathrm{SMC}}(\alpha[1:k], s_k)\, ds_k, \tag{4}$$

where $f_{\mathrm{SMC}}(\cdot, \cdot)$ is defined by

$$f_{\mathrm{SMC}}(\alpha[1:\ell], s_\ell) = \xi(\alpha[\ell] \mid s_\ell) \times \int \phi(s_\ell \mid s_{\ell-1}) f_{\mathrm{SMC}}(\alpha[1:\ell-1], s_{\ell-1})\, ds_{\ell-1}, \tag{5}$$

with base case

$$f_{\mathrm{SMC}}(\alpha[1], s_1) = \xi(\alpha[1] | s_1) \cdot \zeta(s_1). \tag{6}$$

Though we cannot analytically solve the above recursion for $\hat{\pi}_{\mathrm{SMC}}$, in the next section we derive a discretized approximation with time complexity linear in both the number of loci k and the number of haplotypes n. Before doing so, we briefly discuss some appealing properties of $\hat{\pi}_{\mathrm{SMC}}$.

Properties of $\hat{\pi}_{\mathrm{SMC}}$: Recall that the SMC approximation of McVean and Cardin (2005) is equivalent to a variant of the coalescent with recombination disallowing coalescence events between lineages ancestral to disjoint regions. Similarly, the CSD $\hat{\pi}_{\mathrm{PS},1}$, when used to sample a single haplotype, is a variant of $\hat{\pi}_{\mathrm{PS}}$ disallowing the same class of coalescence events. We might therefore expect that the sequentially Markov approximation of $\hat{\pi}_{\mathrm{PS}}$ described above is equivalent to $\hat{\pi}_{\mathrm{PS},1}$, and in fact we can show that this is true.

Proposition 1. *For an arbitrary single haplotype* $\alpha \in \mathcal{H}$ *and haplotype configuration* $\mathbf{n}$, $\hat{\pi}_{\mathrm{SMC}}(\alpha \mid \mathbf{n}) = \hat{\pi}_{\mathrm{PS},1}(\alpha \mid \mathbf{n})$.

We present a sketch of the proof here and refer the reader to supporting information, File S1, for further details.

Sketch of Proof. The key idea of the proof is to introduce a *genealogical* recursion for $f(\alpha, s_k)$, the joint density function associated with sampling haplotype α (under $\hat{\pi}_{\mathrm{PS},1}$) and the marginal genealogy at the last locus s_k. This recursion can be constructed following the lines of Griffiths and Tavaré (1994) to explicitly incorporate coalescent time into a genealogical recursion.

By partitioning with respect to the most recent event occurring at the last locus k, it is possible to inductively show that $f_{\mathrm{SMC}}(\alpha, s_k) = f(\alpha, s_k)$. Furthermore, the equality $\int f(\alpha, s_k) ds_k = \hat{\pi}_{\mathrm{PS},1}(\alpha|\mathbf{n})$ can be verified, and thus we conclude that

$$\hat{\pi}_{\mathrm{PS},1}(\alpha \mid \mathbf{n}) = \int f(\alpha, s_k)\, ds_k = \int f_{\mathrm{SMC}}(\alpha, s_k)\, ds_k = \hat{\pi}_{\mathrm{SMC}}(\alpha \mid \mathbf{n}). \ \square$$

We now describe other intuitively appealing properties of $\hat{\pi}_{\mathrm{SMC}}$. In particular, it can be verified that the *detailed-balance condition*

$$\phi(s' \mid s)\zeta(s) = \phi(s \mid s')\zeta(s') \tag{7}$$

holds for the initial and transition densities, ζ and ϕ, respectively. This immediately implies that the initial distribution ζ is stationary under the given transition dynamics; *i.e.*, the *invariance condition*

$$\zeta(s) = \int \phi(s \mid s')\zeta(s')\, ds'$$

is satisfied. Thus, S_ℓ is marginally distributed according to ζ for all loci $\ell \in L$, and in particular the marginal distribution of T_ℓ is exponential with rate $n/2$. This parallels the fact that the marginal genealogies under the SMC (and the coalescent with recombination) are distributed according to Kingman's coalescent.

Similarly, the transition density exhibits a consistency property, which we call the *locus-skipping property*. Intuitively, this property states that transitioning directly from locus $\ell - 1$ to $\ell + 1$ can be accomplished by using the transition density parameterized with the sum of the recombination rates. Formally, the following equality holds for all $\rho_1, \rho_2 \geq 0$:

$$\int \phi_{\rho_2}(s_{\ell+1} \mid s_\ell)\phi_{\rho_1}(s_\ell \mid s_{\ell-1})\, ds_\ell = \phi_{\rho_1+\rho_2}(s_{\ell+1} \mid s_{\ell-1}). \tag{8}$$

This property, in conjunction with recursion (5), is computationally useful, as it enables loci $\ell \in L$ for which $\alpha[\ell]$ is unobserved to be skipped in computing the CSP $\hat{\pi}_{\mathrm{SMC}}(\alpha \mid \mathbf{n})$.

Finally, the conditional expectation of T_ℓ given $T_{\ell-1} = t_{\ell-1}$ is

$$\mathbb{E}[T_\ell | t_{\ell-1}] = \left(\frac{2}{\rho_b} + \frac{2}{n}\right)(1 - e^{-(\rho_b/2)t_{\ell-1}}), \tag{9}$$

where $b = (\ell - 1, \ell) \in B$. Asymptotically, this expression provides several intuitive results. As $\rho_b \to \infty$, $\mathbb{E}[T_\ell | t_{\ell-1}] \to 2/n$; that is, recombination happens immediately, and $2/n$ is the expectation of the additional absorption time T_{a}. As $\rho_b \to 0$, we get $\mathbb{E}[T_\ell | t_{\ell-1}] \to t_{\ell-1}$. In this case there is no recombination, and the absorption time does not change. Further, $\mathbb{E}[T_\ell | t_{\ell-1}] \to 2/\rho_b + 2/n$ holds as $t_{\ell-1} \to \infty$. Here, recombination must occur, and the exponentially distributed time is not truncated, so the expectation is the sum of the expectations of two exponentials. Finally, as $t_{\ell-1} \to 0$ we have $\mathbb{E}[T_\ell | t_{\ell-1}] \to 0$. No recombination can occur, and so the absorption time is unchanged.

DISCRETIZATION OF THE HMM

In the previous section we described a sequentially Markov approximation of the CSD $\hat{\pi}_{\mathrm{PS}}$ and showed that it can be cast as an HMM. Because the absorption time component of the hidden state is continuous, the dynamic program associated with the classical HMM forward recursion is not applicable. However, by discretizing the continuous component, we are once again able to obtain a dynamic programming algorithm, resulting in an approximate CSP computation linear in both the number of loci and the number of haplotypes.

Rescaling time: Recall from the previous section that the marginal absorption time at each locus is exponentially distributed with parameter $n/2$. To use the same discretization for all n, we follow STEPHENS and DONNELLY (2000) and FEARNHEAD and DONNELLY (2001) and transform the absorption time to a more natural scale in which the marginal absorption time is independent of n. In particular, define the transformed state $\Sigma = (\mathcal{T}, H)$ where $\mathcal{T} = (n/2)T$. We denote a realization of Σ by $\sigma = (\tau, h)$. In the APPENDIX, we provide expressions for the transformed quantities $\tilde{\zeta}(\cdot)$, $\tilde{\phi}(\cdot|\cdot)$, $\tilde{\xi}(\cdot|\cdot)$ and $\tilde{f}_{\mathrm{SMC}}(\cdot,\cdot)$ derived from (1), (2), (3), and (5), respectively.

Using this time-rescaled model, the marginal absorption time at each locus is exponentially distributed with parameter 1. Because this distribution is independent of n and the coalescent model parameters ρ and θ, we expect that a single discretization of the transformed absorption time is appropriate for a wide range of haplotype configurations and parameter values.

Discretizing absorption time: Our next objective is to discretize the absorption time $\mathcal{T} \in \mathbb{R}_{\geq 0}$. Let $0 = x_0 < x_1 < \cdots < x_d = \infty$ be a finite strictly increasing sequence in $\mathbb{R}_{\geq 0} \cup \{\infty\}$ so that $D = \{D_j = [x_{j-1}, x_j)\}_{j=1,\ldots,d}$ is a d-partition of $\mathbb{R}_{\geq 0}$.

Toward formulating a D-discretized version of the dynamics exhibited by the transformed HMM, we define the following D-discretized version of the density $\tilde{f}_{\mathrm{SMC}}$:

$$\tilde{f}_{\mathrm{SMC}}(\alpha[1:\ell], (D_j, h_\ell)) := \int_{D_j} \tilde{f}_{\mathrm{SMC}}(\alpha[1:\ell], (\tau_\ell, h_\ell))\, d\tau_\ell, \tag{10}$$

for all $\ell \in L$. Unfortunately, we cannot obtain a recursion for $\tilde{f}_{\mathrm{SMC}}\big(\alpha[1:\ell], (D_j, h_\ell)\big)$ via the definition of $\tilde{f}_{\mathrm{SMC}}$. Therefore, we make an additional approximation, namely that the transition and emission densities are conditionally dependent on the absorption time $\mathcal{T}$ only through the event $\{D_j \ni \mathcal{T}\}$; *i.e.*, the densities depend on the interval D_j to which $\mathcal{T}$ belongs but not on the actual value of $\mathcal{T}$. Abusing notation, define $\tilde{\phi}\big(\cdot|(D_j, h)\big)$ and $\tilde{\xi}\big(\cdot|(D_j, h)\big)$ as the transition and emission densities, respectively, conditioned on the event $\{D_j \ni \mathcal{T}\}$. Formally, we make the following approximations:

$$\text{Approximation 1: For all } \tau \in D_j,\ \tilde{\phi}(\cdot|(\tau, h)) \approx \tilde{\phi}(\cdot|(D_j, h)). \tag{11}$$

$$\text{Approximation 2: For all } \tau \in D_j, \tilde{\xi}(\cdot|(\tau, h)) \approx \tilde{\xi}(\cdot|(D_j, h)). \tag{12}$$

Together with the building blocks of the time-rescaled HMM, these assumptions provide a recursive approximation of $\tilde{f}_{\mathrm{SMC}}\big(\alpha[1:\ell], (D_j, h_\ell)\big)$, which we denote by $F_\ell^\alpha\big(D_j, h_\ell\big)$. Specifically, assumptions (11) and (12) imply that the integral recursion for $\tilde{f}_{\mathrm{SMC}}$ reduces to the discrete recursion

$$\begin{aligned} F_\ell^\alpha(D_j, h_\ell) &= \tilde{\xi}(\alpha[\ell] \,|\, (D_j, h_\ell)) \\ &\qquad \times \sum_{h_{\ell-1}} \sum_{i=1}^{d} \tilde{\phi}((D_j, h_\ell)|(D_i, h_{\ell-1})) F_{\ell-1}^\alpha(D_i, h_{\ell-1}), \end{aligned} \tag{13}$$

with base case

$$F_1^\alpha(D_j, h_1) = \tilde{\xi}(\alpha[1] \,|\, (D_j, h_1)) \cdot \tilde{\zeta}((D_j, h_1)), \tag{14}$$

where we have defined distributions $\tilde{\phi}\big((D_j, h_\ell) \,|\, (D_i, h_{\ell-1})\big) := \int_{D_j} \tilde{\phi}((\tau_\ell, h_\ell) \,|\, (D_i, h_{\ell-1}))\, d\tau_\ell$ and $\tilde{\zeta}\big((D_j, h_\ell)\big) := \int_{D_j} \tilde{\zeta}((\tau_\ell, h_\ell))\, d\tau_\ell$. Setting $w^{(i)} = \int_{D_i} e^{-\tau}\, d\tau$, we get

$$\tilde{\zeta}(D_i, h_\ell) = \frac{n_{h_\ell}}{n} \cdot w^{(i)}. \tag{15}$$

Turning to the transition density $\tilde{\phi}(\cdot|(D_i, h))$, which is conditioned on the event $\{D_j \ni \mathcal{T}\}$, and recalling that $\mathcal{T}$ is marginally exponentially distributed with parameter 1, we obtain

$$\begin{aligned} &\tilde{\phi}((D_j, h_\ell)|(D_i, h_{\ell-1})) \\ &\quad = \frac{1}{w^{(i)}} \int_{D_j} \int_{D_i} \tilde{\phi}((\tau_\ell, h_\ell) \,|\, (\tau_{\ell-1}, h_{\ell-1})) e^{-\tau_{\ell-1}}\, d\tau_{\ell-1}\, d\tau_\ell \\ &\quad = y^{(i)} \cdot \delta_{i,j} \delta_{h_{\ell-1}, h_\ell} + z^{(i,j)} \cdot \frac{n_{h_\ell}}{n}, \end{aligned} \tag{16}$$

with analytic expressions for $y^{(i)}$ and $z^{(i,j)}$ provided in the APPENDIX. Note that assumption (11) is not used here; rather, the formula follows from using the time-rescaled version of the transition density (2) in the double integral. An expression for the emission density $\tilde{\xi}\big(\cdot|(D_j, h)\big)$ can be similarly obtained,

$$\begin{aligned} \tilde{\xi}(\alpha[\ell] \,|\, (D_i, h_\ell)) &= \frac{1}{w^{(i)}} \int_{D_i} \tilde{\xi}(\alpha[\ell] \,|\, (\tau_\ell, h_\ell)) e^{-\tau_\ell}\, d\tau_\ell \\ &= \sum_{k=0}^{\infty} \upsilon^{(i)}(k) \cdot \left(\mathbf{P}^{(\ell)}\right)^k_{h_\ell[\ell], \alpha[\ell]}, \end{aligned} \tag{17}$$

with an analytic expression for $\upsilon^{(i)}(k)$ also given in the APPENDIX. Again, assumption (12) is not used here; the second equality of (17) follows from using the time-rescaled version of the emission probability (3) in the integral. In summary, $F_\ell^\alpha\big(D_j, h_\ell\big)$ can be computed efficiently using (13), and $\hat{\pi}_{\mathrm{SMC}}(\alpha)$ can be approximated by

$$\hat{\pi}_{\mathrm{SMC}}(\alpha) \approx \sum_{h_k} \sum_{j=1}^{d} F_k^\alpha(D_j, h_k). \tag{18}$$

Equations 13–18 provide the requisite D-discretized versions of the transformed densities. Note that these equations characterize an HMM; that the Markov property holds on the discretized state space D follows from assumptions (11) and (12) (Rosenblatt 1959). In fact, (13–18) may alternatively be obtained by *assuming* that the Markov property holds on D and writing down the relevant transition and emission probabilities with the interpretations given above. In the remainder of this section, we examine some general properties of the discretized dynamics and also provide one method for choosing a discretization D.

Computational complexity of the discretized recursion: We first consider the asymptotic complexity of computing the CSP under the D-discretized approximation for $\hat{\pi}_{\mathrm{SMC}}$. Substituting Equation 16 into the key recursion (13) gives

$$\begin{aligned} F_\ell^\alpha(D_j, h_\ell) &= \tilde{\xi}(\alpha[\ell]\,|\,(D_j, h_\ell)) \\ &\quad \times \left[y^{(j)} F_{\ell-1}^\alpha(D_j, h_\ell) + \frac{n_{h_\ell}}{n} \sum_{i=1}^{d} z^{(i,j)} \sum_{h_{\ell-1}} F_{\ell-1}^\alpha(D_i, h_{\ell-1}) \right] \end{aligned} \tag{19}$$

for $\ell \geq 2$. For a fixed discretization D, the expressions $\tilde{\xi}\big(\cdot|(D_j, h)\big)$, $y^{(i)}$, and $z^{(i,j)}$ depend only on the total sample size n, the mutation and recombination rates (θ_ℓ and ρ_ℓ), and the boundary points $x_0, \ldots, x_d$ of D; these may be precomputed and cached for relevant ranges of values. In conjunction with the base case (14), there is a dynamic program (see the appendix for details) for computing the CSP under the D-discretized approximation (18) for $\hat{\pi}_{\mathrm{SMC}}$ with time complexity $O(k \cdot (nd + d^2))$, where k is the number of loci. As in Fearnhead and Donnelly (2001), this time complexity is better than $O(k \cdot (nd)^2)$, the result that would be obtained by naive use of the HMM forward algorithm.

Properties of the discretization: Recall the *detailed-balance condition* (7) associated with $\hat{\pi}_{\mathrm{SMC}}$. Using expressions (15) and (16), together with Bayes' rule, we find that

$$\begin{aligned} &\tilde{\phi}((D_j, h_\ell)\,|\,(D_i, h_{\ell-1})) \cdot \tilde{\zeta}(D_i, h_{\ell-1}) \\ &\quad = \tilde{\phi}((D_i, h_{\ell-1})\,|\,(D_j, h_\ell)) \cdot \tilde{\zeta}(D_j, h_\ell) \end{aligned} \tag{20}$$

holds (the details are provided in the appendix). Thus, the discretized approximation of $\hat{\pi}_{\mathrm{SMC}}$ satisfies an analogous detailed balance condition. As a result, the marginal distribution at each locus of the discretized Markov chain is (again) given by $\tilde{\zeta}$ and the approximation exhibits the expected symmetries; for example, equal CSPs are computed whether starting at the leftmost locus and proceeding right or starting at the rightmost locus and proceeding left.

Furthermore, recall the *locus-skipping property* (8) associated with $\hat{\pi}_{\mathrm{SMC}}$. The first equality in (16) and assumption (11) imply the relation

$$\begin{aligned} &\tilde{\phi}_{\rho_1+\rho_2}((D_j, h_{\ell+1})\,|\,(D_i, h_{\ell-1})) \\ &\quad \approx \sum_{h_\ell} \sum_{m=1}^{d} \tilde{\phi}_{\rho_2}((D_j, h_{\ell+1})\,|\,(D_m, h_\ell)) \cdot \tilde{\phi}_{\rho_1}((D_m, h_\ell)\,|\,(D_i, h_{\ell-1})) \end{aligned} \tag{21}$$

for all $\rho_1, \rho_2 \geq 0$ (see the appendix for details). Thus, the discretized approximation of $\hat{\pi}_{\mathrm{SMC}}$ approximately satisfies an analogous locus-skipping condition, up to the error introduced via approximation (11). This approximation is particularly useful in scenarios when data are missing (*i.e.*, $\alpha[\ell]$ is unknown for one or more $\ell \in L$), since this property reduces the time complexity of the dynamic program given above. In particular, when m of the k loci are missing, the time complexity is reduced to $O((k-m) \cdot (nd + d^2))$. This is relevant, for example, in importance sampling applications (Fearnhead and Donnelly 2001).

Discretization choice and the definition of $\hat{\pi}_{\mathrm{SMC}(d)}$: Finally, we discuss a method for choosing a discretization D of the absorption time. Recalling that marginally the transformed absorption time is exponentially distributed with parameter 1, let $\{(w^{(j)}, \tau^{(j)})\}_{j=1,\ldots,d}$ be the d-point Gaussian quadrature associated with the function $f(\tau) = e^{-\tau}$ (Abramowitz and Stegun 1972, Section 25.4.45). Set $x_0 = 0$, and set x_j such that $\int_{x_{j-1}}^{x_j} e^{-\tau} d\tau = w^{(j)}$. Since $\sum_{j=1}^{d} w^{(j)} = 1$, the points $0 = x_0 < \cdots < x_d = \infty$ determine a partition $D = \{D_j = [x_{j-1}, x_j)\}_{j=1,\ldots,d}$ of $\mathbb{R}_{\geq 0}$.

The use of Gaussian quadrature evokes the work of Stephens and Donnelly (2000) and Fearnhead and Donnelly (2001). Although the method we employ is related, it is different in that we do not use the quadrature directly [for example, the values of the quadrature points $\{\tau^{(j)}\}$ are never used explicitly]; rather, we use the Gaussian quadrature as a reasonable way of choosing a discretization D. We henceforth write $\hat{\pi}_{\mathrm{SMC}(d)}$ for the d-point Gaussian quadrature-discretized version of $\hat{\pi}_{\mathrm{SMC}}$.

EMPIRICAL RESULTS

In the previous section, we defined a discretized approximation $\hat{\pi}_{\mathrm{SMC}(d)}$ of the CSD $\hat{\pi}_{\mathrm{SMC}}$. In this section, we examine the accuracy of this approximation and also compare it to the widely used CSDs $\hat{\pi}_{\mathrm{FD}}$ and $\hat{\pi}_{\mathrm{LS}}$, thereby providing evidence that $\hat{\pi}_{\mathrm{SMC}(d)}$ is a more accurate and computationally tractable CSD.

Data simulation: For simplicity, we consider a two-allele model with $\mathbf{P}^{(\ell)} = \mathbf{P} = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, $\theta_\ell = \theta$ for $\ell \in L$ and $\rho_b = \rho$ for $b \in B$. We sample a k-locus haplotype configuration $\mathbf{n}$ by (i) using a coalescent with recombination simulator, with $\rho = \rho_0$ and $\theta = \theta_0$, to sample a k_0-locus (with $k_0 \gg k$) n-haplotype configuration $\mathbf{n}_0$, and (ii) restricting attention to the central k *segregating* loci in $\mathbf{n}_0$. This procedure corresponds to the usage of the CSD on typical genomic data, in which only segregating sites are considered.

Given a k-locus n-haplotype configuration $\mathbf{n}$, we obtain a k-locus n-haplotype conditional configuration $C = (\alpha, \mathbf{n} - \mathbf{e}_\alpha)$ by withholding a single haplotype α from $\mathbf{n}$ uniformly at random. For notational simplicity, we define π on such a conditional configuration in the natural way: $\pi(C) = \pi(\alpha \mid \mathbf{n} - \mathbf{e}_\alpha)$.

CSD accuracy: We evaluate the accuracy of a CSD $\hat{\pi}$ relative to a reference CSD π_0 using the expected absolute log-ratio (ALR) error,

$$\mathrm{ALRErr}_{k,n}(\hat{\pi} \mid \pi_0) \approx \frac{1}{N} \sum_{i=1}^{N} \left| \log_{10} \left(\frac{\hat{\pi}(C^{(i)})}{\hat{\pi}_0(C^{(i)})} \right) \right|, \qquad (22)$$

where N denotes the number of simulated data sets and $C^{(i)}$ is a k-locus n-haplotype conditional configuration sampled as indicated above, and both $\hat{\pi}$ and π_0 are evaluated using the true parameter values $\theta = \theta_0$ and $\rho = \rho_0$. For example, if $\mathrm{ALRErr}_{k,n}(\hat{\pi} \mid \pi_0) = 1$, the CSP obtained using $\hat{\pi}$ differs from that obtained by π_0 by a factor of 10, on average, for a randomly sampled k-locus n-haplotype conditional configuration.

Using the ALR error, we evaluate the accuracy of several CSDs: $\hat{\pi}_{\mathrm{FD}}$ (Fearnhead and Donnelly 2001); $\hat{\pi}_{\mathrm{LS}}$ (Li and Stephens 2003); $\hat{\pi}_{\mathrm{SMC}}$, evaluated using the recursion for $\hat{\pi}_{\mathrm{PS},1}$ (Paul and Song 2010); and $\hat{\pi}_{\mathrm{SMC}(d)}$, the d-point quadrature-discretized version of $\hat{\pi}_{\mathrm{SMC}}$, for $d \in \{4, 18, 16\}$. We also evaluate $\hat{\pi}_{\mathrm{SMC\text{-}R}}$, a variant of $\hat{\pi}_{\mathrm{PS},2}$ introduced in Paul and Song (2010) with computational time complexity $O(k^3 \cdot n)$; the CSD $\hat{\pi}_{\mathrm{SMC\text{-}R}}$ is described in more detail in the appendix.

In what follows, we set $\theta_0 = 0.01$ and $\rho_0 = 0.05$ and fix $n = 10$. For $k \leq 10$, it is possible to obtain a very good approximation to the true CSD π using computationally intensive importance sampling. The resulting values of $\mathrm{ALRErr}_{k,n}(\cdot \mid \pi)$ are plotted in Figure 3A, as a function of k. Supporting the conclusion of Paul and Song (2010), $\hat{\pi}_{\mathrm{SMC}}$ is more accurate than both $\hat{\pi}_{\mathrm{LS}}$ and $\hat{\pi}_{\mathrm{FD}}$, with the disparity increasing as k increases. Moreover, the CSD $\hat{\pi}_{\mathrm{SMC}(8)}$ is nearly as accurate as $\hat{\pi}_{\mathrm{SMC}}$, suggesting that the discretization is fairly accurate even for modest values of d. Finally, the CSD $\hat{\pi}_{\mathrm{SMC\text{-}R}}$ has accuracy that is indistinguishable from $\hat{\pi}_{\mathrm{SMC}}$.

To investigate these results as k increases, we consider the ALR error relative to $\hat{\pi}_{\mathrm{SMC}}$, which can be evaluated exactly for $k \leq 20$; the resulting values of $\mathrm{ALRErr}_{k,n}(\cdot|\hat{\pi}_{\mathrm{SMC}})$ are plotted in Figure 3B, as a function of k. As k increases, both $\hat{\pi}_{\mathrm{LS}}$ and $\hat{\pi}_{\mathrm{FD}}$ continue to diverge from $\hat{\pi}_{\mathrm{SMC}}$, suggesting that the increasing disparity in accuracy, directly observable in Figure 3A, continues for larger values of k. As expected, the discretized approximation $\hat{\pi}_{\mathrm{SMC}(d)}$ shows increased fidelity to $\hat{\pi}_{\mathrm{SMC}}$ for larger values of d, and even $\hat{\pi}_{\mathrm{SMC}(4)}$ is substantially more accurate, relative to $\hat{\pi}_{\mathrm{SMC}}$, than are $\hat{\pi}_{\mathrm{LS}}$ and $\hat{\pi}_{\mathrm{FD}}$.

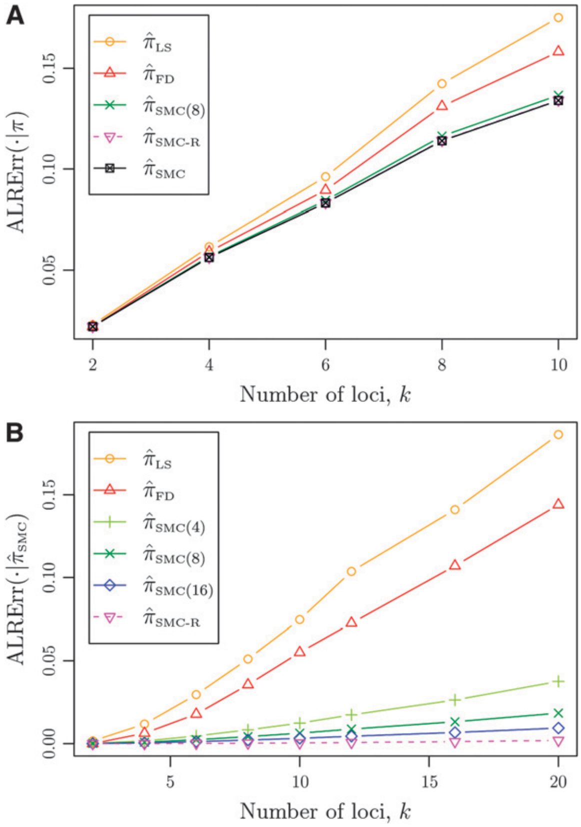

Figure 3.—Absolute log-ratio error (ALRErr) of various conditional sampling distributions. See (22) for a formal definition of $\mathrm{ALRErr}_{k,n}(\cdot \mid \cdot)$. The accuracy of $\hat{\pi}_{\mathrm{SMC\text{-}R}}$ is almost indistinguishable from that of $\hat{\pi}_{\mathrm{SMC}}$, the most accurate of all approximate CSDs considered here. As expected, discretization reduces the accuracy somewhat, but even $\hat{\pi}_{\mathrm{SMC}(4)}$ is substantially more accurate than $\hat{\pi}_{\mathrm{LS}}$ and $\hat{\pi}_{\mathrm{FD}}$. With $\theta_0 = 0.01$ and $\rho_0 = 0.05$, we used the methodology described in the text to sample 250 conditional configurations, each with $n = 10$ haplotypes and k loci. (A) Error is measured relative to the true CSD π, estimated using computationally intensive importance sampling. (B) Error is measured relative to $\hat{\pi}_{\mathrm{SMC}}$, computed by numerically solving a recursion for the equivalent CSD $\hat{\pi}_{\mathrm{PS},1}$.

It is too computationally expensive to compute $\hat{\pi}_{\mathrm{SMC}}$ for $k > 20$. However, Figure 3B suggests that the CSD $\hat{\pi}_{\mathrm{SMC\text{-}R}}$ is nearly indistinguishable from $\hat{\pi}_{\mathrm{SMC}}$. Motivated by this observation, we consider the error relative to $\hat{\pi}_{\mathrm{SMC\text{-}R}}$ for $k > 20$. The values of $\mathrm{ALRErr}_{k,n}(\cdot|\hat{\pi}_{\mathrm{SMC\text{-}R}})$ and the analogously defined signed log-ratio (SLR) error $\mathrm{SLRErr}_{k,n}(\cdot|\hat{\pi}_{\mathrm{SMC\text{-}R}})$ are plotted as a function of k in Figure 4, A and B, respectively. The trends observed in Figure 3 are recapitulated in Figure 4A, suggesting that they continue to hold for substantially larger values of k. Interestingly, Figure 4B shows that $\hat{\pi}_{\mathrm{LS}}$ and $\hat{\pi}_{\mathrm{FD}}$ produce

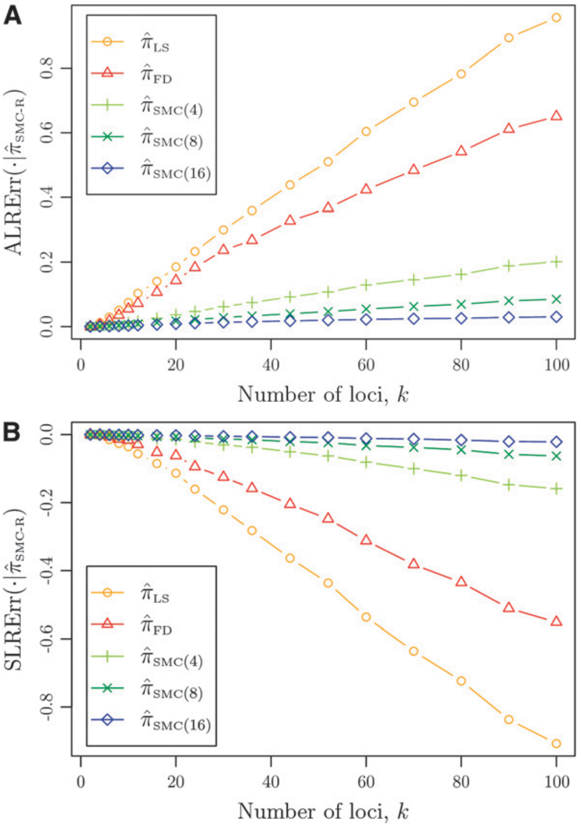

Figure 4.—Comparison of the accuracy of various conditional sampling distributions relative to $\hat{\pi}_{\text{SMC-R}}$ (see Figure 3 for the accuracy of $\hat{\pi}_{\text{SMC-R}}$). A and B illustrate that the improvement in accuracy of $\hat{\pi}_{\mathrm{SMC}(d)}$ over $\hat{\pi}_{\mathrm{LS}}$ and $\hat{\pi}_{\mathrm{FD}}$ is amplified as the number of loci k increases and that both $\hat{\pi}_{\mathrm{LS}}$ and $\hat{\pi}_{\mathrm{FD}}$ produce significantly smaller values than $\hat{\pi}_{\text{SMC-R}}$ (and $\hat{\pi}_{\mathrm{SMC}}$). For $\theta_0 = 0.01$ and $\rho_0 = 0.05$, we used the methodology described in the text to sample 250 conditional configurations with $n = 10$ haplotypes and k loci. (A) Absolute log-ratio error. (B) Signed log-ratio error.

values significantly smaller than $\hat{\pi}_{\text{SMC-R}}$ (and $\hat{\pi}_{\mathrm{SMC}}$); for example, $\hat{\pi}_{\mathrm{LS}}$ takes values that are, on average, a factor of 10 smaller than $\hat{\pi}_{\text{SMC-R}}$ for $k = 100$. In conjunction with our conclusion that $\hat{\pi}_{\mathrm{SMC}}$ is more accurate than $\hat{\pi}_{\mathrm{LS}}$ and $\hat{\pi}_{\mathrm{FD}}$, this suggests a similar systematic error with respect to the true CSD.

For a discussion of CSD accuracy in the context of the product of approximate conditionals (PAC) method (Li and Stephens 2003), we refer the reader to Paul and Song (2010). Since $\hat{\pi}_{\mathrm{SMC}(d)}$ is very close to $\hat{\pi}_{\mathrm{SMC}} = \hat{\pi}_{\mathrm{PS},1}$ (as demonstrated in the present paper), we anticipate that using it produces similar results for PAC likelihood estimation and recombination rate inference.

Running time comparison: We next consider the empirically observed running time required to compute each CSP. The results, obtained using the conditional configurations with $n = 10$ and $k \in \{1, \ldots, 100\}$ simulated as previously described, are presented in Table 1. Looking across each row, it is evident that the running time under $\hat{\pi}_{\mathrm{SMC}(d)}$, $\hat{\pi}_{\mathrm{FD}}$, and $\hat{\pi}_{\mathrm{LS}}$ depends linearly on the number of loci k, matching the asymptotic time complexity. Similarly, the running time under $\hat{\pi}_{\text{SMC-R}}$ is well matched by the theoretical cubic dependence on k.

Next, comparing $\hat{\pi}_{\mathrm{SMC}(d)}$, $\hat{\pi}_{\mathrm{FD}}$, and $\hat{\pi}_{\mathrm{LS}}$, observe that the running time for $\hat{\pi}_{\mathrm{SMC}(4)}$ is approximately a factor of 10 slower than $\hat{\pi}_{\mathrm{LS}}$ and approximately a factor of 2 slower than $\hat{\pi}_{\mathrm{FD}}$. Similarly, $\hat{\pi}_{\mathrm{SMC}(8)}$ is approximately a factor of 20 and of 4 slower than $\hat{\pi}_{\mathrm{LS}}$ and $\hat{\pi}_{\mathrm{FD}}$, respectively; and $\hat{\pi}_{\mathrm{SMC}(16)}$ is approximately a factor of 40 and of 8 slower than $\hat{\pi}_{\mathrm{LS}}$ and $\hat{\pi}_{\mathrm{FD}}$, respectively. Importantly, these factors are *constant*, depending on neither the number of loci k nor the number of haplotypes n. Also note that the time required to compute the CSD for $\hat{\pi}_{\mathrm{SMC}(d)}$ appears to depend linearly, rather than quadratically, on d for the modest (but relevant) values considered.

DISCUSSION

We have formulated a sequentially Markov approximation of $\hat{\pi}_{\mathrm{PS}}$, which we call $\hat{\pi}_{\mathrm{SMC}}$. The relationship between the genealogical process underlying $\hat{\pi}_{\mathrm{PS}}$ and $\hat{\pi}_{\mathrm{SMC}}$ is analogous to the relationship between the coalescent with recombination and the SMC. In particular, $\hat{\pi}_{\mathrm{SMC}}$ is equivalent to $\hat{\pi}_{\mathrm{PS}}$ with a certain class of coalescence events disallowed. In the case of sampling one additional haplotype, this corresponds to disallowing all coalescence events, the same approximation used to obtain $\hat{\pi}_{\mathrm{PS},1}$, and so we find that $\hat{\pi}_{\mathrm{SMC}} = \hat{\pi}_{\mathrm{PS},1}$.

Though the CSD $\hat{\pi}_{\mathrm{SMC}}$ can be cast as an HMM, the associated CSP cannot be evaluated using typical HMM methodology because of the continuous state space; to our knowledge, exact evaluation is possible only via the known recursion for $\hat{\pi}_{\mathrm{PS},1}$, which has time complexity exponential in the number of loci. By discretizing the continuous state space into d intervals, obtained using Gaussian quadrature, we obtain the discretized approximation $\hat{\pi}_{\mathrm{SMC}(d)}$ for which computing the CSP has time complexity linear in both the number of loci and the number of haplotypes. We find that, even for modest values of d, $\hat{\pi}_{\mathrm{SMC}(d)}$ is a very good approximation of $\hat{\pi}_{\mathrm{SMC}}$. Importantly, $\hat{\pi}_{\mathrm{SMC}(d)}$ is more accurate than $\hat{\pi}_{\mathrm{FD}}$ and $\hat{\pi}_{\mathrm{LS}}$ with only a (small) constant factor penalty in run time. We remark that we investigated alternative methods for discretizing the CSP computation (*e.g.*, point-based rather than interval-based methods), but settled on the described approach as it exhibited desirable properties and is theoretically well motivated.

We attribute the observed increase in accuracy of $\hat{\pi}_{\mathrm{SMC}}$ to the incorporation of two key features of the coalescent with recombination that are not integrated into either $\hat{\pi}_{\mathrm{FD}}$ or $\hat{\pi}_{\mathrm{LS}}$. Consider the genealogy associated with two particular haplotypes within an ARG. First, observe that the times to the most recent common ancestor (MRCA) at two neighboring loci are dependent, even if ancestral lineages at the two loci are

TABLE 1

Asymptotic time complexity and empirically observed average running time

		No. of loci			
Method	Complexity	$k = 10$	$k = 20$	$k = 60$	$k = 100$
$\hat{\pi}_{\mathrm{SMC}} = \hat{\pi}_{\mathrm{PS},1}$	$O(c^k \cdot n)$	6.4×10^0	4.8×10^4	NA	NA
$\hat{\pi}_{\mathrm{SMC\text{-}R}}$	$O(k^3 \cdot n)$	2.9×10^0	2.3×10^1	5.6×10^2	2.5×10^3
$\hat{\pi}_{\mathrm{SMC(16)}}$	$O(k \cdot (nd + d^2))$	1.0×10^{-1}	2.1×10^{-1}	6.1×10^{-1}	1.0×10^0
$\hat{\pi}_{\mathrm{SMC(8)}}$	$O(k \cdot (nd + d^2))$	4.6×10^{-2}	9.6×10^{-2}	3.0×10^{-1}	4.7×10^{-1}
$\hat{\pi}_{\mathrm{SMC(4)}}$	$O(k \cdot (nd + d^2))$	2.3×10^{-2}	5.1×10^{-2}	1.6×10^{-1}	2.8×10^{-1}
$\hat{\pi}_{\mathrm{FD}}$	$O(k \cdot n)$	1.1×10^{-2}	2.7×10^{-2}	7.7×10^{-2}	1.3×10^{-1}
$\hat{\pi}_{\mathrm{LS}}$	$O(k \cdot n)$	2.1×10^{-3}	4.6×10^{-3}	1.5×10^{-2}	2.5×10^{-2}

The second column shows asymptotic time complexity (with the value c indicating an unknown constant) and the last four columns show empirically observed average running time (in milliseconds) required to compute the CSP under various CSDs, for $n = 10$ and the number of loci k as specified. "NA" indicates that the computation could not be completed within a reasonable amount of time. Results were obtained on a single core of a MacPro with dual quad-core 3.0-GHz Xeon CPUs.

separated by a recombination event. $\hat{\pi}_{\mathrm{SMC}}$ explicitly models a Markov approximation to the analogous absorption-time dependence across breakpoints, whereas both $\hat{\pi}_{\mathrm{FD}}$ and $\hat{\pi}_{\mathrm{LS}}$ assume independence. Second, if the time to the MRCA at a locus is small, the probability of recombination between this locus and neighboring loci is small, since it would have had to occur prior to the MRCA. While $\hat{\pi}_{\mathrm{SMC}}$ models this property by diminishing the probability of recombination between neighboring loci if the absorption time at the first locus is small, $\hat{\pi}_{\mathrm{FD}}$ and $\hat{\pi}_{\mathrm{LS}}$ assume that recombination is independent of absorption time. We believe that $\hat{\pi}_{\mathrm{FD}}$ and $\hat{\pi}_{\mathrm{LS}}$ tend to underestimate, on average, the true CSP (as suggested in Figure 4B) due to the omission of these key features. The relationship between several CSDs, including $\hat{\pi}_{\mathrm{SMC}}$ and $\hat{\pi}_{\mathrm{FD}}$, is illustrated in Figure 5.

Toward future research, recall that the CSD can be extended to sampling more than one additional haplotype (Paul and Song 2010). Of particular importance to population genetics tools (Stephens and Scheet 2005; Marchini *et al.* 2007; Howie *et al.* 2009) for diploid organisms is sampling two additional haplotypes. Though we focused on conditionally sampling a single additional haplotype in the present work, we note that the sequentially Markov approximation to $\hat{\pi}_{\mathrm{PS}}$ is, in principle, applicable to sampling multiple haplotypes. However, the state space of the resulting HMM description increases exponentially with the number of haplotypes. In this domain, we anticipate that randomized techniques for CSP computation, such as importance sampling and Markov chain Monte Carlo, will exhibit high accuracy and the efficiency required for modern data sets. We pursue this line of research in a forthcoming article.

We believe that it is possible to extend the ideas presented here to different demographic scenarios, for example, spatial structure or models of population subdivision (Davison *et al.* 2009). It should be possible to extend the principled approach of Paul and Song (2010) toward the CSD via the diffusion generator to these scenarios, as in De Iorio and Griffiths (2004b) and Griffiths *et al.* (2008). In other scenarios, for example varying population size, the principled approach might not be applicable, so one would have to modify the genealogical interpretation heuristically, *e.g.*, varying coalescence rates. As in the present article, prohibiting certain coalescence events in the conditional genealogy should then allow for an efficient implementation of the resulting CSDs as HMMs.

Though the SMC has been used for simulating population genetic samples (Marjoram and Wall

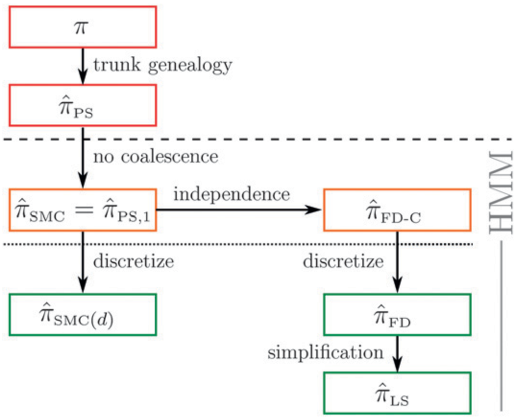
π
trunk genealogy
$\hat{\pi}_{\mathrm{PS}}$
no coalescence
$\hat{\pi}_{\mathrm{SMC}} = \hat{\pi}_{\mathrm{PS},1}$
independence
$\hat{\pi}_{\mathrm{FD\text{-}C}}$
HMM
discretize
discretize
$\hat{\pi}_{\mathrm{SMC}(d)}$
$\hat{\pi}_{\mathrm{FD}}$
simplification
$\hat{\pi}_{\mathrm{LS}}$

Figure 5.—Illustration of the relationship between various CSDs. The CSD at the head of each arrow can be seen as an approximation to the CSD at the tail. Each arrow is also annotated with a (short) description of this approximation. The CSDs below the dashed line can be cast as an HMM: Those above the dotted line (including a continuous-state version of $\hat{\pi}_{\mathrm{FD}}$, which we denote $\hat{\pi}_{\mathrm{FD\text{-}C}}$) have a continuous and infinite state space, while those below have a finite and discrete state space and are therefore amenable to simple dynamic programming algorithms. For more thorough descriptions of each approximation, see the main text and also Paul and Song (2010). Recall in particular that the equality $\hat{\pi}_{\mathrm{SMC}} = \hat{\pi}_{\mathrm{PS},1}$ holds only for conditionally sampling a single haplotype.

2006; Chen *et al.* 2009), it can also be cast as an HMM and used for inference in scenarios in which using the full coalescent with recombination is cumbersome. As described above, the state space of the HMM increases exponentially with the number of haplotypes, making exact computation intractable for large numbers of haplotypes. Nevertheless, research (Hobolth *et al.* 2007; Dutheil *et al.* 2009) is in progress for modest numbers of haplotypes. We believe that choosing a discretization using Gaussian quadrature, as described in discretization of the hmm, and the forthcoming randomized techniques alluded to above, will foster progress in this area.

We conclude by recalling that a broad range of population genetic tools have been developed, and will continue to be developed, on the basis of the CSD. These tools typically employ $\hat{\pi}_{\mathrm{LS}}$, $\hat{\pi}_{\mathrm{FD}}$, or a similar variant, because the underlying HMM structure admits simple and fast recursions for the relevant calculations (*e.g.*, the CSP). We have introduced a new CSD $\hat{\pi}_{\mathrm{SMC}}$ and a discretized approximation $\hat{\pi}_{\mathrm{SMC}(d)}$, which also have simple underlying HMM structures and substantially improve upon the accuracy of $\hat{\pi}_{\mathrm{LS}}$ and $\hat{\pi}_{\mathrm{FD}}$. We believe that $\hat{\pi}_{\mathrm{SMC}(d)}$, when used in the same contexts as $\hat{\pi}_{\mathrm{LS}}$ and $\hat{\pi}_{\mathrm{FD}}$, has the potential to produce more accurate results, with only a small constant factor penalty in run time.

This research is supported in part by National Institutes of Health (NIH) grant R01-GM094402, an Alfred P. Sloan Research Fellowship, and a Packard Fellowship for Science and Engineering. The research of J.S.P was funded in part by an NIH National Research Service Award Trainee appointment on T32-HG00047.

LITERATURE CITED

Abramowitz, M., and I. A. Stegun (Editors), 1972 *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables.* Dover, New York.

Chen, G. K., P. Marjoram and J. D. Wall, 2009 Fast and flexible simulation of DNA sequence data. Genome Res. **19:** 136–142.

Crawford, D. C., T. Bhangale, N. Li, G. Hellenthal, M. J. Rieder *et al.*, 2004 Evidence for substantial fine-scale variation in recombination rates across the human genome. Nat. Genet. **36:** 700–706.

Davison, D., J. K. Pritchard and G. Coop, 2009 An approximate likelihood for genetic data under a model with recombination and population splitting. Theor. Popul. Biol. **75**(4): 331–345.

De Iorio, M., and R. C. Griffiths, 2004a Importance sampling on coalescent histories. I. Adv. Appl. Probab. **36**(2): 417–433.

De Iorio, M., and R. C. Griffiths, 2004b Importance sampling on coalescent histories. II: Subdivided population models. Adv. Appl. Probab. **36**(2): 434–454.

Doucet, A., and A. M. Johansen, 2011 A tutorial on particle filtering and smoothing: fifteen years later, in *Handbook of Nonlinear Filtering,* edited by D. Crisan and B. Rozovsky. Oxford University Press, Oxford (in press).

Dutheil, J. Y., G. Ganapathy, A. Hobolth, T. Mailund, M. K. Uoyenoyama *et al.*, 2009 Ancestral population genomics: the coalescent hidden Markov model approach. Genetics **183:** 259–274.

Fearnhead, P., and P. Donnelly, 2001 Estimating recombination rates from population genetic data. Genetics **159:** 1299–1318.

Fearnhead, P., and P. Donnelly, 2002 Approximate likelihood methods for estimating local recombination rates. J. R. Stat. Soc. B **64:** 657–680.

Fearnhead, P., and N. G. Smith, 2005 A novel method with improved power to detect recombination hotspots from polymorphism data reveals multiple hotspots in human genes. Am. J. Hum. Genet. **77:** 781–794.

Gay, J., S. R. Myers and G. A. T. McVean, 2007 Estimating meiotic gene conversion rates from population genetic data. Genetics **177:** 881–894.

Griffiths, R. C., and S. Tavaré, 1994 Sampling theory for neutral alleles in a varying environment. Philos. Trans. R. Soc. Lond. B Biol. Sci. **344:** 403–410.

Griffiths, R. C., P. A. Jenkins and Y. S. Song, 2008 Importance sampling and the two-locus model with subdivided population structure. Adv. Appl. Probab. **40**(2): 473–500.

Hellenthal, G., A. Auton and D. Falush, 2008 Inferring human colonization history using a copying model. PLoS Genet. **4**(5): e1000078.

Hobolth, A., O. F. Christensen, T. Mailund and M. H. Schierup, 2007 Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. PLoS Genet. **3**(2): e7.

Howie, B. N., P. Donnelly and J. Marchini, 2009 A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. **5**(6): e1000529.

Hudson, R. R., 2001 Two-locus sampling distributions and their application. Genetics **159:** 1805–1817.

Johnson, P. L. F., and M. Slatkin, 2009 Inference of microbial recombination rates from metagenomic data. PLoS Genet. **5**(10): e1000674.

Li, N., and M. Stephens, 2003 Modeling linkage disequilibrium, and identifying recombination hotspots using SNP data. Genetics **165:** 2213–2233.

Li, Y., and G. R. Abecasis, 2006 Mach 1.0: rapid haplotype reconstruction and missing genotype inference. Am. J. Hum. Genet. **S79:** 2290.

Marchini, J., B. Howie, S. R. Myers, G. A. T. McVean and P. Donnelly, 2007 A new multipoint method for genome-wide association studies by imputation of genotypes. Nat. Genet. **39**(7): 906–913.

Marjoram, P., and J. D. Wall, 2006 Fast "coalescent" simulation. BMC Genet. **7:** 16.

McVean, G. A. T., and N. J. Cardin, 2005 Approximating the coalescent with recombination. Philos. Trans. R. Soc. Lond. B Biol. Sci. **360:** 1387–1393.

McVean, G. A. T., S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley *et al.*, 2004 The fine-scale structure of recombination rate variation in the human genome. Science **304:** 581–584.

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.

Price, A. L., A. Tandon, N. Patterson, K. C. Barnes, N. Rafaels *et al.*, 2009 Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet. **5**(6): e1000519.

Rosenblatt, M., 1959 Functions of a Markov process that are Markovian. J. Math. Mech. **8:** 585–596.

Scheet, P., and M. Stephens, 2006 A fast and flexible method for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am. J. Hum. Genet. **78**(4): 629–644.

Stephens, M., and P. Donnelly, 2000 Inference in molecular population genetics. J. R. Stat. Soc. Ser. B Stat. Methodol. **62**(4): 605–655.

Stephens, M., and P. Scheet, 2005 Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am. J. Hum. Genet. **76**(3): 449–462.

Wiuf, C., and J. Hein, 1999 Recombination as a point process along sequences. Theor. Popul. Biol. **55:** 248–259.

Yin, J., M. I. Jordan and Y. S. Song, 2009 Joint estimation of gene conversion rates and mean conversion tract lengths from population SNP data. Bioinformatics **25**(12): i231–i239.

Communicating editor: N. A. Rosenberg

APPENDIX

Time-transformed model: Rewriting the HMM Equations 1–6 in terms of the transformed state $\Sigma = (\mathcal{T}, H)$ introduced in DISCRETIZATION OF THE HMM yields

$$\hat{\pi}_{\mathrm{SMC}}(\alpha[1:\ell]) = \int \tilde{f}_{\mathrm{SMC}}(\alpha[1:\ell], \sigma_\ell)\, d\sigma_\ell, \tag{A1}$$

where the transformed density $\tilde{f}_{\mathrm{SMC}}$ is given by

$$\tilde{f}_{\mathrm{SMC}}(\alpha[1:\ell], \sigma_\ell) = \tilde{\xi}(\alpha[\ell] \mid \sigma_\ell) \cdot \int \tilde{\phi}(\sigma_\ell \mid \sigma_{\ell-1}) \tilde{f}_{\mathrm{SMC}}(\alpha[1:\ell-1], \sigma_{\ell-1})\, d\sigma_{\ell-1}, \tag{A2}$$

with the base case

$$\tilde{f}_{\mathrm{SMC}}(\alpha[1], \sigma_1) = \tilde{\xi}(\alpha[1] \mid \sigma_1) \cdot \tilde{\zeta}(\sigma_1). \tag{A3}$$

The transformed initial, transition, and emission densities are given by

$$\tilde{\zeta}(\sigma_\ell) = \frac{n_{h_\ell}}{n} e^{-\tau_\ell}, \tag{A4}$$

$$\tilde{\phi}(\sigma_\ell \mid \sigma_{\ell-1}) = e^{-(\rho_b/n)\tau_{\ell-1}} \delta_{\sigma_{\ell-1},\sigma_\ell} + \frac{n_{h_\ell}}{n} \int_0^{\tau_{\ell-1}\wedge\tau_\ell} \frac{\rho_b}{n} e^{-(\rho_b/n)\tau_p} e^{-(\tau_\ell - \tau_p)}\, d\tau_p, \text{ and} \tag{A5}$$

$$\tilde{\xi}(\alpha[\ell] \mid \sigma_\ell) = \sum_{k=0}^{\infty} e^{-(\theta_\ell/n)\tau_\ell} \frac{((\theta_\ell/n)\tau_\ell)^k}{k!} (P^{(\ell)})^k_{h_\ell[\ell],\alpha[\ell]}. \tag{A6}$$

Note that care must be taken upon transforming the Dirac-δ in the expression for $\tilde{\phi}(\cdot|\cdot)$.

Analytic expressions for emission and transition probabilities: We now provide analytic expressions for the quantities $y^{(i)}$, $z^{(i,j)}$, and $\upsilon^{(i)}(k)$ introduced for the transition probability (16) and the emission probability (17). Recalling that $D_i = [x_{i-1}, x_i)$ and $D_j = [x_{j-1}, x_j)$ and evaluating the associated integrals, we get

$$y^{(i)} = \frac{1}{w^{(i)}} \frac{n}{\rho_b + n} \left(e^{-((\rho_b+n)/n)x_{i-1}} - e^{-((\rho_b+n)/n)x_i}\right), \tag{A7}$$

$$z^{(i,j)} = \frac{1}{w^{(i)}} \frac{\rho_b}{\rho_b - n} \cdot \begin{cases} w^{(i)}\left(w^{(j)} - \frac{n}{\rho_b}\left(e^{-(\rho_b/n)x_{j-1}} - e^{-(\rho_b/n)x_j}\right)\right), & \text{if } j < i, \\ w^{(j)}\left(w^{(i)} - \frac{n}{\rho_b}\left(e^{-(\rho_b/n)x_{i-1}} - e^{-(\rho_b/n)x_i}\right)\right), & \text{if } j > i, \\ w^{(i)}\left(w^{(i)} - \frac{n}{\rho_b}\left(e^{-(\rho_b/n)x_{i-1}} - e^{-(\rho_b/n)x_i}\right)\right) & \\ \quad - \frac{\rho_b - n}{\rho_b} \frac{n}{\rho_b + n}\left(e^{-((\rho_b+n)/n)x_{i-1}} - e^{-((\rho_b+n)/n)x_i}\right) & \\ \quad - \frac{n}{\rho_b}\left(e^{-x_{i-1}} e^{-(\rho_b/n)x_i} - e^{-x_i} e^{-(\rho_b/n)x_{i-1}}\right), & \text{if } j = i, \end{cases} \tag{A8}$$

for $\rho_b \neq n$,

$$z^{(i,j)} = \frac{1}{w^{(i)}} \cdot \begin{cases} w^{(i)}\left(w^{(j)} + \left(x_{j-1}e^{-x_{j-1}} - x_j e^{-x_j}\right)\right), & \text{if } j < i, \\ w^{(j)}\left(w^{(i)} + \left(x_{i-1}e^{-x_{i-1}} - x_i e^{-x_i}\right)\right), & \text{if } j > i, \\ w^{(i)}\left(w^{(i)} + \left(x_{i-1}e^{-x_{i-1}} - x_i e^{-x_i}\right)\right) & \\ \quad - (x_{i-1} - x_i)e^{-(x_{i-1}+x_i)} - \frac{1}{2}\left(e^{-2x_{i-1}} - e^{-2x_i}\right), & \text{if } j = i, \end{cases} \tag{A9}$$

for $\rho_b = n$, and

$$\upsilon^{(i)}(k) = \frac{1}{w^{(i)}} \frac{((\theta_\ell/n)\tau_\ell)^k}{k!} \sum_{j=0}^{k} \left(\frac{n}{\theta_\ell + n}\right)^{j+1} \frac{k!}{(k-j)!} \left[e^{-((\theta_\ell + n)/n)x_{i-1}} x_{i-1}^{k-j} - e^{-((\theta_\ell + n)/n)x_i} x_i^{k-j} \right]. \tag{A10}$$

Note that the recursive structure of $\upsilon^{(i)}(k)$ [together with $\left(P^{(\ell)}\right)^k$ and the sum in Equation 17] suggests an efficient implementation.

Description of the dynamic program for *D*-discretized $\hat{\pi}_{\mathrm{SMC}}$: Let $D = \{D_1, \ldots, D_{\mathrm{d}}\}$ be a finite partition of $\mathbb{R}_{\geq 0}$ as described in the text. Recalling the recursion for $F_\ell^\alpha\left(D_j, h_\ell\right)$ given in Equation 19, consider the following dynamic programming algorithm for computing the *D*-discretized approximation of $\hat{\pi}_{\mathrm{SMC}}(\alpha|\mathbf{n})$:

1. For each $D_j \in D$ and $h \in \mathcal{H}$ such that $n_h > 0$, compute $F_1^\alpha\left(D_j, h\right)$ using (14), and set $Q_1\left(D_j\right) = \sum_h F_1^\alpha\left(D_j, h\right)$.
2. For each $\ell \in \{2, \ldots, k\}$,
   (a) For each $D_j \in D$, compute $R_\ell\left(D_j\right) = \sum_{i=1}^{d} z^{(i,j)} Q_{\ell-1}(D_i)$.

   (b) For each $D_j \in D$ and $h \in \mathcal{H}$ such that $n_h > 0$, compute

$$F_\ell^\alpha(D_j, h) = \tilde{\xi}(\alpha[\ell]\,|(D_j, h))\left[y^{(j)} F_{\ell-1}^\alpha(D_j, h) + \frac{n_h}{n} R_\ell(D_j)\right],$$

   and set $Q_\ell\left(D_j\right) = \sum_h F_1^\alpha\left(D_j, h\right)$.

3. Compute *D*-discretized approximation $\hat{\pi}_{\mathrm{SMC}}(\alpha\,|\,\mathbf{n}) \approx \sum_h \sum_{j=1}^{d} F_k^\alpha(D_j, h)$.

The time complexities of steps 2a and 2b are $O(d^2)$ and $O(nd)$, respectively. The time complexities of steps 1 and 3 are both $O(nd)$. We can therefore conclude that the time complexity of the dynamic program is $O(nd + (k-1)\cdot(d^2 + nd) + nd) = O(k\cdot(nd + d^2))$.

Detailed balance and locus skipping: The *detailed-balance condition* (20) for the discretized model $\hat{\pi}_{\mathrm{SMC}(d)}$ can be shown using expressions (15) and (16). Together with Bayes' rule, we find that the following holds:

$$\begin{aligned}
&\tilde{\phi}((D_j, h_\ell)|(D_i, h_{\ell-1}))\cdot\tilde{\zeta}(D_i, h_{\ell-1}) \\
&= \left[\frac{1}{w^{(i)}} \int_{D_j}\int_{D_i} \tilde{\phi}((\tau_\ell, h_\ell)|(\tau_{\ell-1}, h_{\ell-1})) e^{-\tau_{\ell-1}}\, d\tau_{\ell-1}\, d\tau_\ell\right]\cdot\frac{n_{h_{\ell-1}}}{n} w^{(i)} \\
&= \left[\int_{D_i}\int_{D_j} \frac{\tilde{\phi}((\tau_{\ell-1}, h_{\ell-1})|(\tau_\ell, h_\ell))\tilde{\zeta}(\tau_\ell, h_\ell)}{\tilde{\zeta}(\tau_{\ell-1}, h_{\ell-1})} e^{-\tau_{\ell-1}}\, d\tau_\ell\, d\tau_{\ell-1}\right]\cdot\frac{n_{h_{\ell-1}}}{n} \\
&= \left[\frac{1}{w^{(j)}} \int_{D_i}\int_{D_j} \tilde{\phi}((\tau_{\ell-1}, h_{\ell-1})\,|\,(\tau_\ell, h_\ell)) e^{-\tau_\ell}\, d\tau_\ell d\tau_{\ell-1}\right]\cdot\frac{n_{h_\ell}}{n} w^{(j)} \\
&= \tilde{\phi}((D_i, h_{\ell-1})\,|\,(D_j, h_\ell))\cdot\tilde{\zeta}(D_j, h_\ell).
\end{aligned} \tag{A11}$$

Using expression (16) and assumption (11) we can show that

$$\begin{aligned}
&\tilde{\phi}_{\rho_1+\rho_2}((D_k, h_{\ell+1})|(D_i, h_{\ell-1})) \\
&= \frac{1}{w^{(i)}} \int_{D_k}\int_{D_i} \tilde{\phi}_{\rho_1+\rho_2}((\tau_{\ell+1}, h_{\ell+1})\,|\,(\tau_{\ell-1}, h_{\ell-1})) e^{-\tau_{\ell-1}}\, d\tau_{\ell-1}\, d\tau_{\ell+1} \\
&= \frac{1}{w^{(i)}} \int_{D_k}\int_{D_i} \left[\int \tilde{\phi}_{\rho_2}((\tau_{\ell+1}, h_{\ell+1})|\sigma_\ell)\cdot\tilde{\phi}_{\rho_1}(\sigma_\ell\,|\,(\tau_{\ell-1}, h_{\ell-1}))\, d\sigma_\ell\right] e^{-\tau_{\ell-1}}\, d\tau_{\ell-1}\, d\tau_{\ell+1} \\
&\approx \sum_{h_\ell}\sum_{j=1}^{d}\left[\int_{D_k} \tilde{\phi}_{\rho_2}((\tau_{\ell+1}, h_{\ell+1})|(D_j, h_\ell))\, d\tau_{\ell+1}\right]\left[\int_{D_j} \tilde{\phi}_{\rho_1}((\tau_\ell, h_\ell)|(D_i, h_{\ell-1}))\, d\tau_\ell\right]
\end{aligned}$$

$$= \sum_{h_\ell} \sum_{j=1}^{d} \tilde{\phi}_{\rho_2}((D_k, h_{\ell+1})|(D_j, h_\ell)) \cdot \tilde{\phi}_{\rho_1}((D_j, h_\ell) \,|\, (D_i, h_{\ell-1})) \tag{A12}$$

holds; thus the *locus-skipping property* (21) for the discretized model $\hat{\pi}_{\mathrm{SMC}(d)}$ holds only approximately. Here we make explicit that the error is introduced by approximation (11) in the third step. Thus it is possible to explicitly assess the error and it goes to zero as the number of intervals used for the discretization becomes large.

A description of $\hat{\pi}_{\mathrm{SMC\text{-}R}}$: Computing the CSP for $\hat{\pi}_{\mathrm{PS},1}$ can be done via a genealogical recursion (Paul and Song 2010, Equation 12), but has time complexity exponential in the number of loci, k. To improve upon this result, Paul and Song suggest using the genealogical recursion until the *first* mutation, and thereafter using a fast alternative CSD $\hat{\pi}_{\mathrm{Alt}}$ (Paul and Song 2010, Equation 13). In particular, choosing $\hat{\pi}_{\mathrm{Alt}}=\hat{\pi}_{\mathrm{FD}}$ yields $\hat{\pi}_{\mathrm{PS},2}$, for which CSP computation has asymptotic time complexity $O(k^3 \cdot n)$.

Similarly, choosing $\hat{\pi}_{\mathrm{Alt}} = \hat{\pi}_{\mathrm{SMC}(16)}$ yields $\hat{\pi}_{\mathrm{SMC\text{-}R}}$, for which CSP computation has the same asymptotic time complexity $O(k^3 \cdot n)$. Importantly, $\hat{\pi}_{\mathrm{SMC}(16)}$ is more accurate than $\hat{\pi}_{\mathrm{FD}}$, and so the resulting CSD $\hat{\pi}_{\mathrm{SMC\text{-}R}}$ is more accurate than $\hat{\pi}_{\mathrm{PS},2}$.

GENETICS

Supporting Information
http://www.genetics.org/cgi/content/full/genetics.110.125534/DC1

An Accurate Sequentially Markov Conditional Sampling Distribution for the Coalescent With Recombination

Joshua S. Paul, Matthias Steinrücken and Yun S. Song

DOI: 10.1534/genetics.110.125534

File S1
Supporting Information

PROOF OF EQUIVALENCE OF $\hat{\pi}_{\text{SMC}}$ AND $\hat{\pi}_{\text{PS},1}$

We now want to give a more detailed proof of Proposition 1 from the main text. With the notation as in the paper we have:

Proposition 1. *For an arbitrary single haplotype $\alpha \in \mathcal{H}$ and haplotype configuration $\mathbf{n}$, $\hat{\pi}_{\text{SMC}}(\alpha|\mathbf{n}) = \hat{\pi}_{\text{PS},1}(\alpha|\mathbf{n})$.*

Recall that the initial (stationary) density was given by

$$\zeta^{(\mathbf{n})}(S_\ell = (t_\ell, h_\ell)) = \frac{n_{h_\ell}}{2} e^{-\frac{n}{2}t_\ell}, \tag{S.1}$$

the transtition density by

$$\phi^{(\mathbf{n})}_{\rho_b}(s_\ell \mid s_{\ell-1}) = e^{-\frac{\rho_b}{2}t_{\ell-1}}\delta_{s_{\ell-1},s_\ell} + \frac{n_{\alpha_\ell}}{n}\int_{t_b=0}^{t_{\ell-1}\wedge t_\ell} \frac{\rho_b}{2}e^{-\frac{\rho_b}{2}t_b}\frac{n}{2}e^{-\frac{n}{2}(t_\ell - t_b)}, \tag{S.2}$$

and the emission probability by

$$\xi^{(\mathbf{n})}_{\theta_\ell}(\alpha[\ell] \mid s_\ell) = \left[e^{\frac{\theta_\ell}{2}t_\ell\cdot(P^{(\ell)}-I)}\right]_{h_\ell[\ell],\alpha[\ell]}. \tag{S.3}$$

Since the configuration $\mathbf{n}$ is fixed, we will drop the superscript $(\mathbf{n})$ in the sequel. As in the main text we will also omit the recombination and mutation rate when unambiguous. Further, we will omit the $d\cdot$ whenever we write down integrals. If not specified differently, equation-references refer to equations from the main paper.

Proof of Proposition 1. We start by showing inductively that the joint density $f_{\text{SMC}}(\alpha[\ell' : \ell], (t_\ell, h_\ell))$ of observing the partial haplotype $\alpha[\ell' : \ell]$ and being in the hidden state (t_ℓ, h_ℓ) (basically) introduced in equations (4)-(6) satisfies a genealogical recursion f, defined as follows [c.f., Griffiths and Tavaré (1994)]:

$$\begin{aligned} f(\alpha[\ell' : \ell], (t_\ell, h_\ell)) = \int_{t_p=0}^{t_\ell} e^{-\frac{n+\sum_u \theta_u + \sum_u \rho_u}{2}t_p} \Bigg[& \frac{n_{h_\ell}\delta_{\alpha[\ell':\ell],h_\ell[\ell':\ell]}}{2}\delta_{t_p,t_\ell} \\ & + \sum_{u\in L(\ell':\ell)} \frac{\theta_u}{2} \sum_{a\in E_u} P^{(u)}_{a,\alpha[u]} f(\mathcal{S}^a_u(\alpha)[\ell' : \ell], (t_\ell - t_p, h_\ell)) \\ & + \sum_{u\in B(\ell':\ell)} \frac{\rho_u}{2}\left(\int_{s_{u_1}} f(\alpha[\ell' : u_1], s_{u_1})\right) f(\alpha[u_2, \ell], (t_\ell - t_p, h_\ell))\Bigg], \end{aligned} \tag{S.4}$$

where the sum $\sum_u$ is over $L(\ell' : \ell)$, all loci between (and including) ℓ' and ℓ, and $\sum_u$ is over $B(\ell' : \ell)$, the breakpoints between ℓ' and ℓ. Here $\mathcal{S}^a_u(\alpha)$ denotes the haplotype obtained by substituting the allele a at locus u of α, and $u = (u_1, u_2)$. For $\ell' = \ell$ (so that $\ell - \ell' = 0$),

$$f(\alpha[\ell], s_\ell) = \int_{t_p=0}^{t_\ell} e^{-\frac{n+\theta_\ell}{2}t_p}\left[\frac{n_{h_\ell}\delta_{\alpha[\ell],h_\ell[\ell]}}{2}\delta_{t_p,t_\ell} + \frac{\theta}{2}\sum_{a\in E_\ell} P^{(\ell)}_{a,\alpha[\ell]} f(a, (t_\ell - t_p, h_\ell))\right].$$

Substituting $f = f_{\text{SMC}}$ on the right-hand side,

$$\begin{aligned}
&\int_{t_p=0}^{t_\ell} e^{-\frac{n+\theta_\ell}{2}t_p}\left[\frac{n_{h_\ell}\delta_{\alpha[\ell],h_\ell[\ell]}}{2}\delta_{t_p,t_\ell} + \frac{\theta_\ell}{2}\sum_{a\in E_\ell} P^{(\ell)}_{a,\alpha[\ell]} f_{\text{SMC}}(a,(t_\ell - t_p, h_\ell))\right] \\
&\quad = e^{-\frac{n+\theta_\ell}{2}t_\ell}\frac{n_{h_\ell}\delta_{\alpha[\ell],h_\ell[\ell]}}{2} + \int_{t_p=0}^{t_\ell} e^{-\frac{n+\theta}{2}t_p}\frac{\theta_\ell}{2}\sum_{a\in E_\ell} P^{(\ell)}_{a,\alpha[\ell]} p(a \mid (t_\ell - t_p, h_\ell)) \cdot q_1(t_u - t_p, h_u) \\
&\quad = \frac{n_{h_\ell}}{2}e^{-\frac{n+\theta_\ell}{2}t_\ell}\left(\delta_{\alpha[\ell],h_\ell[\ell]} + \sum_{m=0}^{\infty}\left(\sum_{a\in E_\ell} P^{(\ell)}_{a,h[\ell]}\big[(P^{(\ell)})^m\big]_{h_\ell[\ell],a}\right)\int_{t_p=0}^{t_\ell}\frac{\theta_\ell}{2}\frac{\big(\frac{\theta_\ell}{2}(t_\ell - t_p)\big)^m}{m!}\right) \\
&\quad = \frac{n_{h_\ell}}{2}e^{-\frac{n+\theta_\ell}{2}t_\ell}\left(\delta_{\alpha[\ell],h_\ell[\ell]} + \sum_{m=0}^{\infty}\big[(P^{(\ell)})^{m+1}\big]_{h_\ell[\ell],\alpha[\ell]}\frac{\big(\frac{\theta_\ell}{2}(t_\ell)\big)^{m+1}}{(m+1)!}\right) \\
&\quad = \frac{n_{h_\ell}}{2}e^{-\frac{n}{2}t_\ell}\cdot\left[e^{\frac{\theta_\ell}{2}t_\ell\cdot(P^{(\ell)}-I)}\right]_{h_\ell[\ell],\alpha[\ell]},
\end{aligned}$$

with the final result equal to $\xi(\alpha[\ell] \mid s_\ell)\zeta(s_\ell) = f_{\text{SMC}}(\alpha[\ell], s_\ell)$. Now, inductively assuming that $f_{\text{SMC}}(\alpha[\ell' : \ell], s_\ell) = f(\alpha[\ell' : \ell], s_\ell)$ for $0 \le \ell - \ell' < m$ and all values of s_ℓ, let $\ell' < \ell$ such that $\ell - \ell' = m$. Substituting $f = f_{\text{SMC}}$ on the right-hand side of (S.4), we obtain

$$\begin{aligned}
&\int_{t_p=0}^{t_\ell} e^{-\frac{n+\sum_u\theta_u+\sum_u\rho_u}{2}t_p}\Bigg[\frac{n_{h_\ell}\delta_{\alpha[\ell':\ell],h_\ell[\ell':\ell]}}{2}\delta_{t_p,t_\ell} \\
&\qquad + \sum_{u\in L(\ell':\ell)}\frac{\theta_u}{2}\sum_{a\in E_u} P^{(u)}_{a,\alpha[u]} f_{\text{SMC}}(\mathcal{S}^a_u(\alpha)[\ell' : \ell], (t_\ell - t_p, h_\ell)) \\
&\qquad + \sum_{u\in B(\ell':\ell)}\frac{\rho_u}{2}\left(\int_{s_{u_1}} f_{\text{SMC}}(\alpha[\ell' : u_1], s_{u_1})\right) f_{\text{SMC}}(\alpha[u_2, \ell], (t_\ell - t_p, h_\ell))\Bigg].
\end{aligned} \tag{S.5}$$

We consider this expression one term at a time. Beginning with the first term:

$$\begin{aligned}
&\int_{t_p=0}^{t_\ell} e^{-\frac{n+\sum_u\theta_u+\sum_u\rho_u}{2}t_p}\frac{n_{h_\ell}\delta_{\alpha[\ell':\ell],h_\ell[\ell':\ell]}}{2}\delta_{t_p,t_\ell} \\
&\quad = \int_{s_{\ell-1}}\int_{t_p=0}^{t_{\ell-1}} e^{-\frac{n+\sum'_u\theta_u+\sum'_u\rho_u}{2}t_p}\frac{n_{h_{\ell-1}}\delta_{\alpha[\ell':\ell-1],h_{\ell-1}[\ell':\ell-1]}}{2}\delta_{t_p,t_{\ell-1}}\left[e^{-\frac{\theta_\ell+\rho_b}{2}t_p}\delta_{\alpha[\ell],h_\ell[\ell]}\delta_{s_{\ell-1},s_\ell}\right],
\end{aligned} \tag{S.6}$$

where $\sum'_u$ is over $L(\ell' : \ell - 1)$ and $\sum'_u$ is over $B(\ell' : \ell - 1)$. Moving on to the second term, expand using the definition (5) of f_{SMC}, and then use the inductive hypothesis to replace the resulting f_{SMC} terms with the corresponding f terms:

$$\begin{aligned}
&\int_{t_p=0}^{t_\ell} e^{-\frac{n+\sum_u\theta_u+\sum_u\rho_u}{2}t_p}\sum_{u\in L(\ell':\ell)}\frac{\theta_u}{2}\sum_{a\in E_u} P^{(u)}_{a,\alpha[u]} f_{\text{SMC}}(\mathcal{S}^a_u(\alpha)[\ell' : \ell], (t_\ell - t_p, h_\ell)) \\
&\quad = \int_{t_p=0}^{t_\ell} e^{-\frac{n+\sum_u\theta_u+\sum_u\rho_u}{2}t_p}\sum_{u\in L(\ell':\ell-1)}\frac{\theta_u}{2}\sum_{a\in E_u} P^{(u)}_{a,\alpha[u]} \\
&\qquad \times \xi(\alpha[\ell] \mid (t_\ell - t_p, h_\ell))\int_{s_{\ell-1}}\phi((t_\ell - t_p, h_\ell) \mid s_{\ell-1}) f(\mathcal{S}^a_u(\alpha)[\ell' : \ell - 1], s_{\ell-1}) \\
&\quad + \int_{t_p=0}^{t_\ell} e^{-\frac{n+\sum_u\theta_u+\sum_u\rho_u}{2}t_p}\frac{\theta_\ell}{2}\sum_{a\in E_u} P_{a,\alpha[\ell]} \\
&\qquad \times \xi(a \mid (t_\ell - t_p, h_\ell))\int_{s_{\ell-1}}\phi((t_\ell - t_p, h_\ell) \mid s_{\ell-1}) f(\alpha[\ell' : \ell - 1], s_{\ell-1}).
\end{aligned}$$

Concentrating on the first sub-term, making the substitution $t_{\ell-1} \to t_{\ell-1} + t_p$, and changing the order of integration, we obtain

$$\begin{aligned}\int_{s_{\ell-1}} \int_{t_p=0}^{t_\ell \wedge t_{\ell-1}} & e^{-\frac{n+\sum'_u \theta_u + \sum'_u \rho_u}{2} t_p} \sum_{u \in L(\ell':\ell-1)} \frac{\theta_u}{2} \sum_{a \in E_u} P^{(u)}_{a,\alpha[u]} f(\mathcal{S}^a_u(\alpha)[\ell' : \ell-1], (t_{\ell-1} - t_p, h_{\ell-1})) \\ & \times \left[e^{-\frac{\theta_\ell}{2} t_p} p(\alpha[\ell] \mid (t_\ell - t_p, h_\ell)) \cdot e^{-\frac{\rho_b}{2} t_p} q((t_\ell - t_p, h_\ell) \mid (t_{\ell-1} - t_p, h_{\ell-1})) \right]. \qquad \text{(S.7)}\end{aligned}$$

Now concentrating on the second sub-term and expanding using definition (S.4) of f:

$$\begin{aligned}\int_{t_p=0}^{t_\ell} & e^{-\frac{n+\sum_u \theta_u + \sum_u \rho_u}{2} t_p} \frac{\theta_\ell}{2} \sum_{a \in E_u} P_{a,\alpha[\ell]} \xi(a \mid (t_\ell - t_p, h_\ell)) \int_{s_{\ell-1}} \phi((t_\ell - t_p, h_\ell) \mid s_{\ell-1}) \\ & \times \int_{t_q=0}^{t_{\ell-1}} e^{-\frac{n+\sum'_u \theta_u + \sum'_u \rho_u}{2} t_q} \Bigg[\frac{n_{h_{\ell-1}} \delta_{\alpha[\ell':\ell-1], h_{\ell-1}[\ell':\ell-1]}}{2} \delta_{t_q, t_{\ell-1}} \\ & \quad + \sum_{u \in L(\ell':\ell-1)} \frac{\theta_u}{2} \sum_{a \in E_u} P^{(u)}_{a,\alpha[u]} f(\mathcal{S}^a_u(\alpha)[\ell' : \ell-1], (t_{\ell-1} - t_q, h_{\ell-1})) \\ & \quad + \sum_{u \in B(\ell':\ell-1)} \frac{\rho_u}{2} \left(\int_{s_{u_1}} f(\alpha[\ell' : u_1], s_{u_1}) \right) f(\alpha[u_2, \ell-1], (t_{\ell-1} - t_q, h_{\ell-1})) \Bigg] \\ = \int_{s_{\ell-1}} \int_{t_q=0}^{t_{\ell-1}} & e^{-\frac{n+\sum'_u \theta_u + \sum'_u \rho_u}{2} t_q} \Bigg[\frac{n_{h_{\ell-1}} \delta_{\alpha[\ell':\ell-1], h_{\ell-1}[\ell':\ell-1]}}{2} \delta_{t_q, t_{\ell-1}} \\ & \quad + \sum_{u \in L(\ell':\ell-1)} \frac{\theta_u}{2} \sum_{a \in E_u} P^{(u)}_{a,\alpha[u]} f(\mathcal{S}^a_u(\alpha)[\ell' : \ell-1], (t_{\ell-1} - t_q, h_{\ell-1})) \\ & \quad + \sum_{u \in B(\ell':\ell-1)} \frac{\rho_u}{2} \left(\int_{s_{u_1}} f(\alpha[\ell' : u_1], s_{u_1}) \right) f(\alpha[u_2, \ell-1], (t_{\ell-1} - t_q, h_{\ell-1})) \Bigg] \\ & \times \left[\int_{t_p=0}^{t_q \wedge t_\ell} e^{-\frac{\theta_\ell}{2} t_p} \frac{\theta_\ell}{2} \sum_{a \in E_u} P_{a,\alpha[\ell]} \xi(a \mid (t_\ell - t_p, h_\ell)) \cdot e^{-\frac{\rho_b}{2} t_p} \phi((t_\ell - t_p, h_\ell) \mid (t_{\ell-1} - t_p, h_{\ell-1})) \right], \\ & \qquad \text{(S.8)}\end{aligned}$$

with the equality obtained by making the substitutions $t_{\ell-1} \to t_{\ell-1} + t_p$ and $t_q \to t_q + t_p$ and then changing the order of integration. Finally, moving onto the third term, expand using the definition (5) of f_{SMC}, and then use the inductive hypothesis to replace the resulting f_{SMC} terms with the corresponding f terms:

$$\begin{aligned}\int_{t_p=0}^{t_\ell} & e^{-\frac{n+\sum_u \theta_u + \sum_u \rho_u}{2} t_p} \sum_{u \in B(\ell':\ell)} \frac{\rho_u}{2} \left(\int_{s_{u_l}} f_{\text{SMC}}(\alpha[\ell' : u_l], s_{u_l}) \right) f_{\text{SMC}}(\alpha[u_r, \ell], (t_\ell - t_p, h_\ell)) \\ & = \int_{t_p=0}^{t_\ell} e^{-\frac{n+\sum_u \theta_u + \sum_u \rho_u}{2} t_p} \sum_{u \in B(\ell':\ell-1)} \frac{\rho_u}{2} \left(\int_{s_{u_l}} f(\alpha[\ell' : u_l], s_{u_l}) \right) \\ & \quad \times \xi(\alpha[\ell] \mid (t_\ell - t_p, h_\ell)) \int_{s_{\ell-1}} \phi((t_\ell - t_p, h_\ell) \mid s_{\ell-1}) f(\alpha[u_r : \ell-1], s_{\ell-1}) \\ & \quad + \int_{t_p=0}^{t_\ell} e^{-\frac{n+\sum_u \theta_u + \sum_u \rho_u}{2} t_p} \frac{\rho_b}{2} \left(\int_{s_{\ell-1}} f(\alpha[\ell' : \ell-1], s_{\ell-1}) \right) \cdot f(\alpha[\ell], (t_\ell - t_p, h_\ell)).\end{aligned}$$

Concentrating on the first sub-term, making the substitution $t_{\ell-1} \to t_{\ell-1} + t_p$, and changing the order of integration, we obtain:

$$\int_{s_{\ell-1}} \int_{t_p=0}^{t_\ell \wedge t_{\ell-1}} e^{-\frac{n+\sum'_u \theta_u + \sum'_u \rho_u}{2} t_p} \sum_{u \in B(\ell':\ell-1)} \frac{\rho_u}{2} \bigg(\int_{s_{u_1}} f(\alpha[\ell' : u_1], s_{u_1}) \bigg) f(\alpha[u_2 : \ell-1], (t_{\ell-1} - t_p, h_{\ell-1})) \\ \times \bigg[e^{-\frac{\theta_\ell}{2} t_p} \xi(\alpha[\ell] \mid (t_\ell - t_p, h_\ell)) \cdot e^{-\frac{\rho_b}{2} t_p} \phi((t_\ell - t_p, h_\ell) \mid (t_{\ell-1} - t_p, h_{\ell-1})) \bigg]. \tag{S.9}$$

Now concentrating on the second sub-term and expanding using definition (S.4) of f:

$$\begin{aligned} &\int_{t_p=0}^{t_\ell} e^{-\frac{n+\sum_u \theta_u + \sum_u \rho_u}{2} t_p} \frac{\rho_b}{2} f(\alpha[\ell], (t_\ell - t_p, h_\ell)) \\ &\qquad \times \int_{s_{\ell-1}} \int_{t_q=0}^{t_{\ell-1}} e^{-\frac{n+\sum'_u \theta_u + \sum'_u \rho_u}{2} t_q} \bigg[\frac{n_{h_{\ell-1}} \delta_{\alpha[\ell':\ell-1], h_{\ell-1}[\ell':\ell-1]}}{2} \delta_{t_q, t_{\ell-1}} \\ &\qquad\qquad + \sum_{u \in L(\ell':\ell-1)} \frac{\theta_u}{2} \sum_{a \in E_u} P^{(u)}_{a,\alpha[u]} f(\mathcal{S}^a_u(\alpha)[\ell' : \ell-1], (t_{\ell-1} - t_q, h_{\ell-1})) \\ &\qquad\qquad + \sum_{u \in B(\ell':\ell-1)} \frac{\rho_u}{2} \bigg(\int_{s_{u_1}} f(\alpha[\ell' : u_1], s_{u_1}) \bigg) f(\alpha[u_2, \ell-1], (t_{\ell-1} - t_q, h_{\ell-1})) \bigg] \\ &= \int_{s_{\ell-1}} \int_{t_q=0}^{t_{\ell-1}} e^{-\frac{n+\sum'_u \theta_u + \sum'_u \rho_u}{2} t_q} \bigg[\frac{n_{h_{\ell-1}} \delta_{\alpha[\ell':\ell-1], h_{\ell-1}[\ell':\ell-1]}}{2} \delta_{t_q, t_{\ell-1}} \\ &\qquad\qquad + \sum_{u \in L(\ell':\ell-1)} \frac{\theta_u}{2} \sum_{a \in E_u} P^{(u)}_{a,\alpha[u]} f(\mathcal{S}^a_u(\alpha)[\ell' : \ell-1], (t_{\ell-1} - t_q, h_{\ell-1})) \\ &\qquad\qquad + \sum_{u \in B(\ell':\ell-1)} \frac{\rho_u}{2} \bigg(\int_{s_{u_1}} f(\alpha[\ell' : u_1], s_{u_1}) \bigg) f(\alpha[u_2, \ell-1], (t_{\ell-1} - t_q, h_{\ell-1})) \bigg] \\ &\qquad \times \bigg[\int_{t_p=0}^{t_q \wedge t_\ell} e^{-\frac{\theta_\ell}{2} t_p} \xi(\alpha[\ell] \mid (t_\ell - t_p, h_\ell)) \cdot e^{-\frac{\rho_b}{2} t_p} \frac{\rho_b}{2} \frac{n_{h_\ell}}{2} e^{-\frac{n}{2}(t_\ell - t_p)} \bigg], \end{aligned} \tag{S.10}$$

with the equality obtained by using the (one-locus) definition (6) for $f_{\text{SMC}}(\alpha[\ell], (t_\ell - t_p, h_\ell))$, making the substitutions $t_{\ell-1} \to t_{\ell-1} + t_p$ and $t_q \to t_q + t_p$, and changing the order of integration.

Having appropriately expanded each term of our key expression (S.5), we aggregate common terms across the resulting sub-expressions. Collecting the $n_{h_{\ell-1}} \delta_{\alpha[\ell':\ell-1], h_{\ell-1}[\ell':\ell-1]}$ terms from (S.6),(S.8), and (S.10),

$$
\begin{aligned}
&\int_{s_{\ell-1}} \int_{t_p=0}^{t_{\ell-1}} e^{-\frac{n+\sum'_u \theta_u+\sum'_u \rho_u}{2} t_p} \frac{n_{h_{\ell-1}} \delta_{\alpha[\ell':\ell-1],h_{\ell-1}[\ell':\ell-1]}}{2} \delta_{t_p,t_{\ell-1}} \\
&\quad \times \Bigg[e^{-\frac{\theta_\ell+\rho_b}{2} t_p} \delta_{\alpha[\ell],h_\ell[\ell]} \delta_{s_{\ell-1},s_\ell} \\
&\qquad + \int_{t_q=0}^{t_p \wedge t_\ell} e^{-\frac{\theta_\ell}{2} t_q} \frac{\theta_\ell}{2} \sum_{a \in E_u} P_{a,\alpha[\ell]} \xi(a \mid (t_\ell - t_q, h_\ell)) \cdot e^{-\frac{\rho_b}{2} t_q} \phi((t_\ell - t_q, h_\ell) \mid (t_{\ell-1} - t_q, h_{\ell-1})) \\
&\qquad + \int_{t_q=0}^{t_p \wedge t_\ell} e^{-\frac{\theta_\ell}{2} t_q} \xi(\alpha[\ell] \mid (t_\ell - t_q, h_\ell)) \cdot e^{-\frac{\rho_b}{2} t_q} \frac{\rho_b}{2} \frac{n_{h_\ell}}{2} e^{-\frac{n}{2}(t_\ell - t_q)} \Bigg] \\
&= \int_{s_{\ell-1}} \int_{t_p=0}^{t_{\ell-1}} e^{-\frac{n+\sum'_u \theta_u+\sum'_u \rho_u}{2} t_p} \frac{n_{h_{\ell-1}} \delta_{\alpha[\ell':\ell-1],h_{\ell-1}[\ell':\ell-1]}}{2} \delta_{t_p,t_{\ell-1}} \\
&\quad \times \Bigg[e^{-\frac{\rho_b}{2} t_{\ell-1}} \delta_{s_{\ell-1},s_\ell} \cdot \left(e^{-\frac{\theta_\ell}{2} t_\ell} \delta_{\alpha[\ell],h_\ell[\ell]} \right) \\
&\qquad + e^{-\frac{\rho_b}{2} t_{\ell-1}} \delta_{s_{\ell-1},s_\ell} \left(\int_{t_z=0}^{t_\ell} e^{-\frac{\theta_\ell}{2} t_z} \frac{\theta_\ell}{2} \sum_{a \in E_u} P_{a,\alpha[\ell]} \xi(a \mid (t_\ell - t_z, h_\ell)) \right) \\
&\qquad + \int_{t_q=0}^{t_{\ell-1} \wedge t_\ell} \frac{\rho_b}{2} e^{-\frac{\rho_b}{2} t_q} \frac{n_{h_\ell}}{2} e^{-\frac{n}{2}(t_\ell - t_q)} \left(\int_{t_z=0}^{t_q} e^{-\frac{\theta_\ell}{2} t_z} \frac{\theta_\ell}{2} \sum_{a \in E_u} P_{a,\alpha[\ell]} \xi(allele \mid (t_\ell - t_z, h_\ell)) \right) \\
&\qquad + \int_{t_q=0}^{t_{\ell-1} \wedge t_\ell} \frac{\rho_b}{2} e^{-\frac{\rho_b}{2} t_q} \frac{n_{h_\ell}}{2} e^{-\frac{n}{2}(t_\ell - t_q)} \left(e^{-\frac{\theta_\ell}{2} t_q} \xi(\alpha[\ell] \mid (t_\ell - t_q, h_\ell)) \right) \Bigg] \\
&= \int_{s_{\ell-1}} \int_{t_p=0}^{t_{\ell-1}} e^{-\frac{n+\sum'_u \theta_u+\sum'_u \rho_u}{2} t_p} \frac{n_{h_{\ell-1}} \delta_{\alpha[\ell':\ell-1],h_{\ell-1}[\ell':\ell-1]}}{2} \delta_{t_p,t_{\ell-1}} \\
&\quad \times \xi(\alpha[\ell] \mid s_\ell) \Bigg[e^{-\frac{\rho_b}{2} t_{\ell-1}} \delta_{s_{\ell-1},s_\ell} + \int_{t_q=0}^{t_{\ell-1} \wedge t_\ell} \frac{\rho_b}{2} e^{-\frac{\rho_b}{2} t_q} \frac{n_{h_\ell}}{2} e^{-\frac{n}{2}(t_\ell - t_q)} \Bigg] \\
&= \int_{s_{\ell-1}} \int_{t_p=0}^{t_{\ell-1}} e^{-\frac{n+\sum'_u \theta_u+\sum'_u \rho_u}{2} t_p} \frac{n_{h_{\ell-1}} \delta_{\alpha[\ell':\ell-1],h_{\ell-1}[\ell':\ell-1]}}{2} \delta_{t_p,t_{\ell-1}} \times \Bigg[\xi(\alpha[\ell] \mid s_\ell) \phi(s_\ell \mid s_{\ell-1}) \Bigg], \qquad \text{(S.11)}
\end{aligned}
$$

where the first equality is obtained by making use of the $\delta_{t_p,t_{\ell-1}}$ and $\delta_{s_{\ell-1},s_\ell}$ expressions and expanding the q term using equation (S.2) and exchanging integrals, the second equality is obtained by combining the first/second and third/fourth term along with the definition (S.3) of p, and final equality by again making use of the equation (S.2).

Similarly, collecting the $f(\mathcal{S}_u^a(\alpha)[\ell' : \ell-1], (t_{\ell-1} - t_q, h_{\ell-1}))$ terms from the resulting sub-expressions

(S.7),(S.8), and (S.10),

$$
\begin{aligned}
&\int_{s_{\ell-1}} \int_{t_p=0}^{t_{\ell-1}} e^{-\frac{n+\sum'_u \theta_u+\sum'_u \rho_u}{2} t_p} \sum_{u\in L(\ell':\ell-1)} \frac{\theta_u}{2} \sum_{a\in E_u} P^{(u)}_{a,\alpha[u]} f(\mathcal{S}^a_u(\alpha)[\ell' : \ell-1], (t_{\ell-1}-t_p, h_{\ell-1}))\\
&\quad\times \Bigg[\mathbb{I}_{(t_p\le t_\ell)} e^{-\frac{\theta_\ell}{2}t_p} \xi(\alpha[\ell] \mid (t_\ell-t_p, h_\ell)) \cdot e^{-\frac{\rho_b}{2}t_p} \phi((t_\ell-t_p,h_\ell)\mid(t_{\ell-1}-t_p,h_{\ell-1}))\\
&\qquad + \int_{t_q=0}^{t_p\wedge t_\ell} e^{-\frac{\theta_\ell}{2}t_q}\frac{\theta_\ell}{2}\sum_{a\in E_u} P_{a,\alpha[\ell]}\xi(a\mid(t_\ell-t_q,h_\ell))\cdot e^{-\frac{\rho_b}{2}t_q}\phi((t_\ell-t_q,h_\ell)\mid(t_{\ell-1}-t_q,h_{\ell-1}))\\
&\qquad + \int_{t_q=0}^{t_p\wedge t_\ell} e^{-\frac{\theta_\ell}{2}t_q}\xi(\alpha[\ell]\mid(t_\ell-t_q,h_\ell))\cdot e^{-\frac{\rho_b}{2}t_q}\frac{\rho_b}{2}\frac{n_{h_\ell}}{2}e^{-\frac{n}{2}(t_\ell-t_q)}\Bigg]\\
&= \int_{s_{\ell-1}} \int_{t_p=0}^{t_{\ell-1}} e^{-\frac{n+\sum'_u \theta_u+\sum'_u \rho_u}{2} t_p} \sum_{u\in L(\ell':\ell-1)} \frac{\theta_u}{2} \sum_{a\in E_u} P^{(u)}_{a,\alpha[u]} f(\mathcal{S}^a_u(\alpha)[\ell' : \ell-1], (t_{\ell-1}-t_p, h_{\ell-1}))\\
&\quad\times \Bigg[\mathbb{I}_{(t_p\le t_\ell)} e^{-\frac{\rho_b}{2}t_p}\phi((t_\ell-t_p,h_\ell)\mid(t_{\ell-1}-t_p,h_{\ell-1}))\Big(e^{-\frac{\theta_\ell}{2}t_p}\xi(\alpha[\ell]\mid(t_\ell-t_p,h_\ell))\Big)\\
&\qquad + \mathbb{I}_{(t_p\le t_\ell)} e^{-\frac{\rho_b}{2}t_p}\phi((t_\ell-t_p,h_\ell)\mid(t_{\ell-1}-t_p,h_{\ell-1}))\bigg(\int_{t_z=0}^{t_p} e^{-\frac{\theta_\ell}{2}t_z}\frac{\theta_\ell}{2}\sum_{a\in E_u}P_{a,\alpha[\ell]}\xi(a\mid(t_\ell-t_z,h_\ell))\bigg)\\
&\qquad + \int_{t_q=0}^{t_p\wedge t_\ell}\frac{\rho_b}{2}e^{-\frac{\rho_b}{2}t_q}\frac{n_{h_\ell}}{2}e^{-\frac{n}{2}(t_\ell-t_q)}\bigg(\int_{t_z=0}^{t_q} e^{-\frac{\theta_\ell}{2}t_z}\frac{\theta_\ell}{2}\sum_{a\in E_u}P_{a,\alpha[\ell]}\xi(a\mid(t_\ell-t_z,h_\ell))\bigg)\\
&\qquad + \int_{t_q=0}^{t_p\wedge t_\ell}\frac{\rho_b}{2}e^{-\frac{\rho_b}{2}t_q}\frac{n_{h_\ell}}{2}e^{-\frac{n}{2}(t_\ell-t_q)}\Big(e^{-\frac{\theta_\ell}{2}t_q}\xi(\alpha[\ell]\mid(t_\ell-t_q,h_\ell))\Big)\Bigg]\\
&= \int_{s_{\ell-1}} \int_{t_p=0}^{t_{\ell-1}} e^{-\frac{n+\sum'_u \theta_u+\sum'_u \rho_u}{2} t_p} \sum_{u\in L(\ell':\ell-1)} \frac{\theta_u}{2} \sum_{a\in E_u} P^{(u)}_{a,\alpha[u]} f(\mathcal{S}^a_u(\alpha)[\ell' : \ell-1], (t_{\ell-1}-t_p, h_{\ell-1}))\\
&\quad\times \xi(\alpha[\ell]\mid s_\ell)\bigg[\mathbb{I}_{(t_p\le t_\ell)}e^{-\frac{\rho_b}{2}t_p}\phi((t_\ell-t_p,h_\ell)\mid(t_{\ell-1}-t_p,h_{\ell-1})) + \int_{t_q=0}^{t_p\wedge t_\ell}\frac{\rho_b}{2}e^{-\frac{\rho_b}{2}t_q}\frac{n_{h_\ell}}{2}e^{-\frac{n}{2}(t_\ell-t_q)}\bigg]\\
&= \int_{s_{\ell-1}} \int_{t_p=0}^{t_{\ell-1}} e^{-\frac{n+\sum'_u \theta_u+\sum'_u \rho_u}{2} t_p} \sum_{u\in L(\ell':\ell-1)} \frac{\theta_u}{2} \sum_{a\in E_u} P^{(u)}_{a,\alpha[u]} f(\mathcal{S}^a_u(\alpha)[\ell' : \ell-1], (t_{\ell-1}-t_p, h_{\ell-1}))\\
&\quad\times \Big[\xi(\alpha[\ell]\mid s_\ell)\phi(s_\ell\mid s_{\ell-1})\Big],
\end{aligned}
\tag{S.12}
$$

where the first equality is obtained by expanding the ϕ term[1] in the second term using equation (S.2), the second equality is obtained by combining the first/second and third/fourth term along with the definition (S.3) of ξ, and final equality by again making use of the equation (S.2) and considering separately the case when $t_p \le t_\ell$ and $t_p > t_\ell$.

The situation is identical when collecting terms with $f(\alpha[u_2, \ell-1], (t_{\ell-1}-t_q, h_{\ell-1}))$ from (S.9),

[1] We use the following expansion for ϕ, which can be verified in the present context, namely that $t_q \le t_p \le t_{\ell-1}$ and $t_q \le t_\ell$:

$$
\begin{aligned}
\phi((t_\ell-t_q,h_\ell)\mid(t_{\ell-1}-t_q,h_{\ell-1})) &= \mathbb{I}_{(t_p\le t_\ell)}e^{-\frac{\rho_b}{2}(t_p-t_q)}\cdot\phi((t_\ell-t_p,h_\ell)\mid(t_{\ell-1}-t_p,h_{\ell-1}))\\
&\quad+\int_{t_z=0}^{(t_p\wedge t_\ell)-t_q}\frac{\rho_b}{2}e^{-\frac{\rho_b}{2}t_z}\frac{n_{h_\ell}}{2}e^{-\frac{n}{2}(t_\ell-t_q-t_z)}
\end{aligned}
$$

(S.8), and (S.10):

$$\begin{aligned}
&\int_{s_{\ell-1}} \int_{t_p=0}^{t_{\ell-1}} e^{-\frac{n+\sum_u' \theta_u + \sum_u' \rho_u}{2} t_p} \sum_{u \in B(\ell':\ell-1)} \frac{\rho_u}{2} \left(\int_{s_{u_1}} f(\alpha[\ell' : u_1], s_{u_1}) \right) f(\alpha[u_2 : \ell-1], (t_{\ell-1} - t_p, h_{\ell-1})) \\
&\quad \times \Bigg[\mathbb{I}_{(t_p \leq t_\ell)} e^{-\frac{\theta_\ell}{2} t_p} \xi(\alpha[\ell] \mid (t_\ell - t_p, h_\ell)) \cdot e^{-\frac{\rho_b}{2} t_p} \phi((t_\ell - t_p, h_\ell) \mid (t_{\ell-1} - t_p, h_{\ell-1})) \\
&\qquad + \int_{t_q=0}^{t_p \wedge t_\ell} e^{-\frac{\theta_\ell}{2} t_q} \frac{\theta_\ell}{2} \sum_{a \in E_u} P_{a,\alpha[\ell]} \xi(a \mid (t_\ell - t_q, h_\ell)) \cdot e^{-\frac{\rho_b}{2} t_q} \phi((t_\ell - t_q, h_\ell) \mid (t_{\ell-1} - t_q, h_{\ell-1})) \\
&\qquad + \int_{t_q=0}^{t_p \wedge t_\ell} e^{-\frac{\theta_\ell}{2} t_q} \xi(\alpha[\ell] \mid (t_\ell - t_q, h_\ell)) \cdot e^{-\frac{\rho_b}{2} t_q} \frac{\rho_b}{2} \frac{n_{h_\ell}}{2} e^{-\frac{n}{2}(t_\ell - t_q)} \Bigg] \\
&= \int_{s_{\ell-1}} \int_{t_p=0}^{t_{\ell-1}} e^{-\frac{n+\sum_u' \theta_u + \sum_u' \rho_u}{2} t_p} \sum_{u \in B(\ell':\ell-1)} \frac{\rho_u}{2} \left(\int_{s_{u_1}} f(\alpha[\ell' : u_1], s_{u_1}) \right) f(\alpha[u_2 : \ell-1], (t_{\ell-1} - t_p, h_{\ell-1})) \\
&\quad \times \Big[\xi(\alpha[\ell] \mid s_\ell) \phi(s_\ell \mid s_{\ell-1}) \Big]. \qquad \text{(S.13)}
\end{aligned}$$

Thus, combining equations (S.11),(S.12), and (S.13), we may re-write (S.5):

$$\begin{aligned}
&\xi(\alpha[\ell] \mid s_\ell) \int_{s_{\ell-1}} \phi(s_\ell \mid s_{\ell-1}) \cdot \int_{t_p=0}^{t_{\ell-1}} e^{-\frac{n+\sum_u' \theta_u + \sum_u' \rho_u}{2} t_p} \Bigg[\frac{n_{h_{\ell-1}} \delta_{\alpha[\ell':\ell-1], h_{\ell-1}[\ell':\ell-1]}}{2} \delta_{t_p, t_{\ell-1}} \\
&\qquad + \sum_{u \in L(\ell':\ell-1)} \frac{\theta_u}{2} \sum_{a \in E_u} P^{(u)}_{a,\alpha[u]} f(\mathcal{S}^a_u(\alpha)[\ell' : \ell-1], (t_{\ell-1} - t_p, h_{\ell-1})) \\
&\qquad + \sum_{u \in B(\ell':\ell-1)} \frac{\rho_u}{2} \left(\int_{s_{u_1}} f(\alpha[\ell' : u_1], s_{u_1}) \right) f(\alpha[u_2 : \ell-1], (t_{\ell-1} - t_p, h_{\ell-1})) \Bigg] \\
&= \xi(\alpha[\ell] \mid s_\ell) \int_{s_{\ell-1}} \phi(s_\ell \mid s_{\ell-1}) f(\alpha[\ell' : \ell-1], s_{\ell-1}) \\
&= f_{\text{SMC}}(\alpha[\ell' : \ell], s_\ell),
\end{aligned}$$

where the first equality is obtained by definition (S.4) for f, and the second equality by using the inductive hypothesis and the definition (5). Therefore, f_{SMC} satisfies the recursion for f, and we conclude that $f_{\text{SMC}} = f$. Moreover,

$$\begin{aligned}
\int_{s_\ell} f(\alpha[\ell' : \ell], s_\ell) &= \int_{s_\ell} \int_{t_p=0}^{t_\ell} e^{-\frac{n+\sum_u \theta_u+\sum_u \rho_u}{2} t_p} \Bigg[\frac{n_{h_\ell} \delta_{\alpha[\ell':\ell], h_\ell[\ell':\ell]}}{2} \delta_{t_p, t_\ell} \\
&\qquad + \sum_{u \in L(\ell':\ell)} \frac{\theta_u}{2} \sum_{a \in E_u} P^{(u)}_{a, \alpha[u]} f(\mathcal{S}^a_u(\alpha)[\ell' : \ell], (t_\ell - t_p, h_\ell)) \\
&\qquad + \sum_{u \in B(\ell':\ell)} \frac{\rho_u}{2} \left(\int_{s_{u_1}} f(\alpha[\ell' : u_1], s_{u_1}) \right) f(\alpha[u_2, \ell], (t_\ell - t_p, h_\ell)) \Bigg] \\
&= \frac{1}{n + \sum_{u \in L(\alpha[\ell':\ell])} \theta_u + \sum_{u \in B(\alpha[\ell':\ell])} \rho_u} \Bigg[\sum_{\substack{\alpha' \in \mathcal{H}: \\ \alpha'[\ell':\ell] = \alpha[\ell':\ell]}} n_{\alpha'} \\
&\qquad + \sum_{u \in L(\alpha[\ell':\ell])} \theta_u \sum_{a \in E_u} P^{(u)}_{a, \alpha[u]} \int_{s_\ell} f(\mathcal{S}^a_u(\alpha)[\ell' : \ell], s_\ell) \\
&\qquad + \sum_{u \in B(\alpha[\ell':\ell])} \rho_u \int_{s_{u_1}} f(\alpha[\ell' : u_1], s_{u_1}) \int_{s_\ell} f(\alpha[u_2, \ell], s_\ell) \Bigg],
\end{aligned}$$

where the first equality is by definition (S.4), and the second equality obtained by exchanging the integrals and making the substitution $t_\ell \to t_\ell - t_p$. Thus, $\int_{s_\ell} f(\alpha[\ell' : \ell], s_\ell)$ satisfies the recursion for $\hat{\pi}_{\text{PS},1}$ (Paul and Song, 2010, Equation (12)) and we conclude that $\int_{s_\ell} f(\alpha[\ell' : \ell], s_\ell) = \hat{\pi}_{\text{PS},1}(\alpha[\ell' : \ell])$. Thus,

$$\hat{\pi}_{\text{SMC}}(\alpha[\ell' : \ell]) = \int_{s_\ell} f_{\text{SMC}}(\alpha[\ell' : \ell], s_\ell) = \int_{s_\ell} f(\alpha[\ell' : \ell], s_\ell) = \hat{\pi}_{\text{PS},1}(\alpha[\ell' : \ell]),$$

thereby establishing the desired identity. □

LITERATURE CITED

Griffiths, R. C. and Tavaré, S. 1994. Sampling theory for neutral alleles in a varying environment. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **344,** 403–410.

Paul, J. S. and Song, Y. S. 2010. A principled approach to deriving approximate conditional sampling distributions in population genetics models with recombination. *Genetics*, **186,** 321–338.