Published in final edited form as:
Theor Popul Biol. 2013 August ; 87: 51–61. doi:10.1016/j.tpb.2012.08.004.

A sequentially Markov conditional sampling distribution for structured populations with migration and recombination

Matthias Steinrücken[a], **Joshua S. Paul**[b], and **Yun S. Song**[a,b,*]

[a]Department of Statistics, University of California, Berkeley, CA 94720, USA

[b]Computer Science Division, University of California, Berkeley, CA 94720, USA

Abstract

Conditional sampling distributions (CSDs), sometimes referred to as copying models, underlie numerous practical tools in population genomic analyses. Though an important application that has received much attention is the inference of population structure, the explicit exchange of migrants at specified rates has not hitherto been incorporated into the CSD in a principled framework. Recently, in the case of a single panmictic population, a sequentially Markov CSD has been developed as an accurate, efficient approximation to a principled CSD derived from the diffusion process dual to the coalescent with recombination. In this paper, the sequentially Markov CSD framework is extended to incorporate subdivided population structure, thus providing an efficiently computable CSD that admits a genealogical interpretation related to the structured coalescent with migration and recombination. As a concrete application, it is demonstrated empirically that the CSD developed here can be employed to yield accurate estimation of a wide range of migration rates.

Keywords

structured coalescent; recombination; migration; conditional sampling distribution; hidden Markov model; sequentially Markov coalescent

1. Introduction

Under a given population genetic model, the conditional sampling distribution (CSD), also called a *copying model* by some authors, describes the probability that an additionally sampled haplotype is of a certain type, given that a collection of haplotypes has already been observed. As described below, various applications in population genomics make use of the CSD. Although the CSD is of much importance, no exact closed-form expressions are known in the situations to which it has been applied, and so a number of approximations have been proposed.

Following the seminal work of Stephens and Donnelly (2000) and Fearnhead and Donnelly (2001), Li and Stephens (2003) proposed a widely used CSD, denoted $\widehat{\pi}_{\mathrm{LS}}$, which models the

*Corresponding author steinrue@stat.berkeley.edu (Matthias Steinrücken), jspaul@cs.berkeley.edu (Joshua S. Paul), yss@stat.berkeley.edu (Yun S. Song).

additionally observed haplotype as an imperfect mosaic of the haplotypes already observed. The model underlying $\widehat{\pi}_{\mathrm{LS}}$ can be cast as a hidden Markov model (HMM), thus admitting e cient implementation. In their paper, Li and Stephens used the CSD $\widehat{\pi}_{\mathrm{LS}}$ in a pseudo-likelihood framework to estimate fine-scale recombination rates, and subsequently $\widehat{\pi}_{\mathrm{LS}}$ and its extensions have been used in numerous other population genetic applications, including estimating gene-conversion parameters (Gay et al., 2007; Yin et al., 2009), and phasing genotype sequence data into haplotype sequence data and imputing missing data (Stephens and Scheet, 2005; Li and Abecasis, 2006; Li et al., 2010; Marchini et al., 2007; Howie et al., 2009).

Another important application of the CSD that has received much attention is the inference of population structure and demography. Hellenthal et al. (2008) employed $\widehat{\pi}_{\mathrm{LS}}$ to model human colonization history as a sequence of founder events and estimated the order of the founding events, as well as the relative contribution of different founding populations during the events. To estimate the splitting time of two populations Davison et al. (2009) modified $\widehat{\pi}_{\mathrm{LS}}$ to incorporate the split into the copying model, and used the same pseudo-likelihood framework as Li and Stephens (2003) to estimate the time of splitting. In a more recent study, Lawson et al. (2012) applied $\widehat{\pi}_{\mathrm{LS}}$ to a sample of DNA sequences and used properties of the inferred mosaic pattern to reveal structure in the underlying population.

To handle admixture, a modification to $\widehat{\pi}_{\mathrm{LS}}$ was introduced by Price et al. (2009), who assumed that the previously observed haplotypes in the CSD are from two distinct ancestral populations (e.g., African and European). In modeling the mosaic pattern for a haplotype sampled from the admixed population (e.g., African American), it is then assumed more likely that adjacent segments originate from the same ancestral population, rather than from two different ancestral populations. Price et al. applied this modified copying model to detect chromosomal segments of distinct ancestry in admixed individuals and estimated admixture fractions in recently admixed populations. The same model was applied by Wegmann et al. (2011), who used the inferred ancestry switch-points to estimate relative recombination rates between different populations.

As discussed above, $\widehat{\pi}_{\mathrm{LS}}$ is a very useful CSD with a variety of applications, but it was not *derived* from, though was certainly motivated by, principles underlying the coalescent process. To derive CSDs in a principled way, De Iorio and Griffiths (2004a) introduced a general approximation technique based on the diffusion process dual to the coalescent; this work was first presented in the case of a single locus and a panmictic population, but in a companion paper (De Iorio and Griffiths, 2004b) the authors applied the method to the case of a subdivided population with migration. Griffiths et al. (2008) extended the diffusion approximation technique to handle recombination in the special case of two loci with parent-independent mutation at each locus, and Paul and Song (2010) later generalized the framework to an arbitrary number of loci and an arbitrary finite-alleles mutation model.

Although more accurate than the CSDs developed by Fearnhead and Donnelly (2001) and by Li and Stephens (2003), the CSD $\widehat{\pi}_{\mathrm{PS}}$ proposed by Paul and Song (2010) is not amenable to efficient evaluation. More precisely, $\widehat{\pi}_{\mathrm{PS}}$ can be computed by solving a recursion that becomes intractable for a large number of loci. However, utilizing ideas related to the sequentially Markov coalescent (SMC) (Wiuf and Hein, 1999; McVean and Cardin, 2005; Marjoram and Wall, 2006), which is a simplified genealogical process that captures the essential features of the full coalescent model with recombination, we (Paul et al., 2011) recently developed an approximation to $\widehat{\pi}_{\mathrm{PS}}$ that could be cast as an HMM with continuous hidden state space. Furthermore, upon discretizing this continuous state space, we obtained

an accurate approximation with computational efficiency comparable to the CSDs of Fearnhead and Donnelly (2001) and Li and Stephens (2003).

In this paper, we extend our previous work on the sequentially Markov CSD to incorporate subdivided population structure with migration. Following Paul and Song (2010), we describe a genealogical process for an additionally sampled haplotype conditioned on the genealogy of already observed haplotypes. We present a recursion that can be used to compute the probability of the additionally sampled haplotype, but, as in Paul and Song (2010), solving this recursion is tractable only for a small number of loci. As in Paul et al. (2011), we apply the sequentially Markov framework to the conditional genealogical process with migration and recombination, and obtain an accurate approximation that facilitates computation for a large number of loci. As a concrete application, we demonstrate empirically that our new CSD can be employed in various pseudo-likelihoods to produce accurate estimation of a wide range of migration rates.

The remainder of this paper is organized as follows: In Section 2, we introduce the notation adopted throughout the paper and describe the relevant population genetic model, the coalescent with recombination and migration. We then describe the genealogical interpretation of our CSD in Section 3 and introduce several approximations in Section 4 to obtain a CSD for which computation is tractable. In Section 5, we demonstrate the applicability of our CSD by employing it to the estimation of migration rates from simulated data. Finally, we conclude in Section 6 with a discussion of further applications and extensions of the CSD developed herein to estimate demographic parameters in more complex scenarios.

2. Background

In this section, we briefly describe how migration is integrated into the coalescent with recombination, and recall the CSD $\widehat{\pi}_{\mathrm{PS}}$ proposed by Paul and Song (2010), which we extend to incorporate migration in the following section. We begin by defining some general notation that will be used throughout.

2.1. 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; recombination may occur between any consecutive pair of loci, and we denote the set of potential recombination breakpoints by $B = \{(1, 2), \ldots, (k-1, k)\}$. The space of k-locus haplotypes is denoted by $\mathscr{H} = E_1 \times \cdots \times E_k$. Given a haplotype $h \in \mathscr{H}$, we denote by $h[\ell] \in E_\ell$ the allele at locus $\ell \in L$, and by $h[\ell : \ell']$ (for $\ell \leq \ell'$) the partial haplotype $(h[\ell], \ldots, h[\ell'])$.

We consider an island model of population structure with a finite set of demes denoted by $\Gamma = \{1, \ldots, g\}$. At a given time, each individual belongs to a single deme, and the ancestors and descendants of the individual may belong to different demes by means of a migration process, detailed in Section 2.2.

A *structured sample configuration* of haplotypes is specified by $\mathbf{n} = \left(n_{\gamma,h}\right)_{\gamma \in \Gamma, h \in \mathscr{H}}$ where $n_{\gamma,h}$ denotes the number of haplotypes of type h within deme γ in the sample. The configuration of haplotypes within deme $\gamma \in \Gamma$ is denoted $\mathbf{n}_\gamma$, and the total number of haplotypes in the deme by $|\mathbf{n}_\gamma| = \mathbf{n}_\gamma$. The total number of haplotypes in $\mathbf{n}$ is denoted by $n = |\mathbf{n}| = \Sigma_{\gamma \in \Gamma} n_\gamma$. Finally, we use $\mathbf{e}_{\gamma,h}$ to denote the singleton configuration comprising a single haplotype h within deme γ.

2.2. The coalescent with recombination and migration

The stochastic process underlying our work is the coalescent with recombination and migration (Griffiths and Marjoram, 1997; Herbots, 1997). Consider a structured population with a finite set Γ of demes. We denote the relative size of deme $\gamma \in \Gamma$ by κ_γ, where $0 \leq \kappa_\gamma \leq 1$ and $\sigma_{\gamma \in \Gamma}\, \kappa_\gamma = 1$. Note that two individuals within a deme find a common ancestor at rate inversely proportional to the relative size of the deme; in the coalescent limit, coalescence within deme γ occurs at rate κ_γ^{-1}.

To allow for migration of ancestral lineages between demes, define $u_{\gamma\gamma'}$ to be the probability that an individual in deme γ has a parent in deme γ'. In the coalescent limit, as the population size N tends to infinity, an ancestral lineage in deme γ migrates, backwards in time, to deme γ' at rate $m_{\gamma\gamma'}/2$, where $m_{\gamma\gamma'} = 4Nu_{\gamma\gamma'}$ is the *scaled* migration rate. Henceforward, we consider a continuous-time Markov migration process with transition rate matrix $M = (m_{\gamma\gamma'}/2)_{\gamma,\gamma' \in \Gamma}$, where $m_{\gamma\gamma} = -\sigma_{\gamma' \neq \gamma}\, m_{\gamma\gamma'}$. For ease of notation, we define $m_\gamma = \sigma_{\gamma' \neq \gamma}\, m_{\gamma\gamma'}$.

An ancestral lineage undergoes mutation at locus $\ell \in L$ at rate $\theta/2$, where θ_ℓ is the scaled mutation rate, and according to the stochastic mutation transition matrix $\mathbf{P}^{(\ell)}$. Further, as in the ordinary coalescent with recombination, an ancestral lineage undergoes recombination, backwards in time, at breakpoint $b \in B$ at rate $\rho_b/2$, where ρ_b is the scaled recombination rate, giving rise to two lineages (in the same deme).

A structured configuration $\mathbf{n}$ with n_γ individuals in each deme γ can be sampled as follows. The process starts at present with $\mathbf{n}$ with n_γ untyped lineages in each deme γ, and lineages in each deme γ evolve backwards in time subject to the following possible events:

Mutation: Each lineage undergoes mutation at locus $\ell \in L$ with rate $\theta_\ell/2$ according to the mutation transition matrix $\mathbf{P}^{(\ell)}$.

Recombination: Each lineage undergoes recombination at breakpoint $b \in B$ with rate $\rho_b/2$.

Migration: Each lineage migrates to deme γ' with rate $m_{\gamma\gamma'}/2$.

Coalescence: Each pair of lineages coalesce with rate κ_γ^{-1}.

When a single lineage remains, the type at each locus ℓ is chosen according to the stationary distribution of the mutation matrix $\mathbf{P}^{(\ell)}$, and this type is propagated toward the present, producing a realization for the sample $\mathbf{n}$.

2.3. The CSD $\widehat{\pi}_{\mathrm{PS}}$ for a single panmictic population

The approximate CSD $\widehat{\pi}_{\mathrm{PS}}$ (Paul and Song, 2010) for a single panmictic population is described by a genealogical process closely related to the coalescent with recombination. Suppose that, conditioned on having already observed a haplotype configuration $\mathbf{n}$, we wish to sample c additional haplotypes. As described in Paul and Song (2010), given the true fully-specified genealogy $\boldsymbol{A}_\mathbf{n}$ for the conditional configuration $\mathbf{n}$, it is possible to sample a *conditional genealogy* C for the c additional haplotypes.

The conditional genealogy C comprises the following: mutation, recombination, and coalescence within C, occurring at rates given in Section 2.2; and *absorption* of lineages into the known genealogy $\boldsymbol{A}_\mathbf{n}$, occurring at rate 1 for each pair. Because the types of the lineages of $\boldsymbol{A}_\mathbf{n}$ are known, the type of an absorbed lineage is determined. Thus, when all lineages of C have been absorbed, the type may be propagated forward, thereby producing a realization for sample configuration $\mathbf{c}$ with $|\mathbf{c}| = c$.

Unfortunately, we do not generally have access to the true genealogy $\boldsymbol{A}_{\mathbf{n}}$. Making use of the diffusion-generator approximation (De Iorio and Griffiths, 2004a,b; Griffiths et al., 2008), Paul and Song (2010) propose the following: *Assume* that $\boldsymbol{A}_{\mathbf{n}} = \boldsymbol{A}_0(\mathbf{n})$, where $\boldsymbol{A}_0(\mathbf{n})$ is called the *trunk* genealogy in which lineages do not mutate, recombine, or coalesce with one another, but instead form a non-random "trunk" extending infinitely into the past. Note that $\boldsymbol{A}_0(\mathbf{n})$ does not have a most recent common ancestor, and for this reason is improper; nonetheless, it remains possible to sample a well-defined conditional genealogy C, and thus to generate an additional sample $\mathbf{c}$, in much the same way as described above. In particular, lineages within C evolve backwards in time subject to the following events:

Mutation: Each lineage undergoes mutation at locus $\ell \in L$ with rate θ_ℓ according to $\mathbf{P}^{(\ell)}$.

Recombination: Each lineage undergoes recombination at breakpoint $b \in B$ with rate ρ_b

Coalescence: Each pair of lineages coalesce with rate 2.

Absorption: Each lineage is absorbed into a lineage of $\mathscr{A}_{\mathbf{n}} = \mathscr{A}_0(\mathbf{n})$ with rate 1.

Observe that the rate of absorption is the same as in the case where $\mathscr{A}_{\mathbf{n}}$ is known. The rates for mutation, recombination, and coalescence, on the other hand, are each a factor of two larger than those given in Section 2.2; intuitively, this adjustment accounts for using the (incorrect) trunk genealogy $\mathscr{A}_0(\mathbf{n})$, and notably the absence of events therein. Importantly, the CSD $\widehat{\pi}_{\mathrm{PS}}$ has been shown to be *correct* for a one-locus model with parent independent mutation (Stephens and Donnelly, 2000; De Iorio and Griffiths, 2004a; Paul and Song, 2010), a strong argument in favor of the given rate adjustment. The CSD $\widehat{\pi}_{\mathrm{PS}}$ is completely characterized by the above genealogical process.

Remark. The rates given for the genealogical process governing $\widehat{\pi}_{\mathrm{PS}}$ are *double* those given by Paul and Song (2010) and Paul et al. (2011). Importantly, the genealogical process is *time-homogeneous*, and so for the purposes of computing the conditional sampling probability (CSP) $\widehat{\pi}_{\mathrm{PS}}(\mathbf{c}|\mathbf{n})$, this modification has no effect (indeed, any constant multiple of the rates will yield the same CSP). However, we believe that the scaling adopted here admits a natural interpretation of the absorption time as a true coalescence time. For example, consider sampling a single haplotype conditional on a configuration $\mathbf{n}$ with $|\mathbf{n}| = 1$; analogous to coalescence of two lines in Kingman's coalescent, absorption in the genealogical process associated with $\widehat{\pi}_{\mathrm{PS}}$ occurs at rate 1.

3. A new CSD $\widehat{\pi}_{\mathrm{Mig}}$ for structured populations with recombination and migration

We now introduce an approximate CSD $\widehat{\pi}_{\mathrm{Mig}}$ by extending the genealogical process of Section 2.3 to a general structured population with $|\Gamma| \geq 1$. Suppose that conditioned on having already observed a structured sample configuration $\mathbf{n}$, we wish to sample c additional haplotypes with c_γ of them in each deme γ. As before, given the true fully-specified genealogy $\mathscr{A}_{\mathbf{n}}$ for the conditional configuration $\mathbf{n}$, including migration events, it is possible to sample a conditional genealogy $\mathscr{C}$ for the c additional haplotypes. The conditional genealogy $\mathscr{C}$ comprises the events and corresponding rates of Section 2.2, this time including migration, and the *absorption* of lineages in each deme γ into lineages of $\mathscr{A}_{\mathbf{n}}$ in deme γ. These latter absorption events occur at rate κ_γ^{-1}.

In practice, we do not have access to the true genealogy $\mathscr{A}_{\mathbf{n}}$, but the diffusion-generator technique (De Iorio and Griffiths, 2004a,b; Griffiths et al., 2008; Paul and Song, 2010) again

implies the following approximation: *Assume* that $\mathscr{A}_{\mathbf{n}}=\mathscr{A}_0(\mathbf{n})=\{\mathscr{A}_0(\mathbf{n}_\gamma)\}_{\gamma\in\Gamma}$, where $\mathscr{A}_0(\mathbf{n}_\gamma)$ is the non-random sub-trunk genealogy associated with deme γ, within which lineages do not mutate, recombine, migrate, or coalesce with one another. As in Section 2.3, given this assumption it remains possible to sample a well-defined conditional genealogy $\mathscr{C}$, and thus to generate the additional structured sample **c** Specifically, lineages within each deme γ of $\mathscr{C}$ evolve backwards in time subject to the following events:

Mutation: Each lineage undergoes mutation at locus $\ell\in L$ with rate $\theta\ell$according to the mutation transition matrix $\mathbf{P}^{(\ell)}$.

Recombination: Each lineage undergoes recombination at breakpoint $b\in B$ with rate ρb

Migration: Each lineage migrates to deme γ' with rate $m_{\gamma\gamma'}$.

Coalescence: Each pair of lineages coalesces with rate $2\kappa_\gamma^{-1}$.

Absorption: Each lineage is absorbed into a lineage of $\mathscr{A}_0(\mathbf{n}_\gamma)$ with rate κ_γ^{-1}.

Observe that the rates of mutation, recombination, migration, and coalescence are a factor of two larger than when the true genealogy $\mathscr{A}_{\mathbf{n}}$ is known. Intuitively, this again accounts for using the (incorrect) trunk genealogy $\mathscr{A}_0(\mathbf{n})$, and particularly the absence of events therein; see the remark at the end of Section 2.3. The approximate CSD $\widehat{\pi}_{\text{Mig}}$ is completely characterized by this genealogical process. See Figure 1(a) for an illustration.

Remark. For strongly asymmetric migration rates, the approximate CSD $\widehat{\pi}_{\text{Mig}}$, and in particular the assumed trunk genealogy $\mathscr{A}_0(\mathbf{n})$, may be very inaccurate. Consider for example the case of two demes and $m_{12}\gg m_{21}$. The expected time for an additionally sampled haplotype in deme 2 to be absorbed into the trunk in deme 1 will be very large, since the lineages in the trunk genealogy $\mathscr{A}_0(\mathbf{n}_1)$ are confined to stay in deme 1. In case of the true genealogy An, however, one would expect the lineages of the haplotypes in the observed configuration in deme 1 to cross over to deme 2 quickly and coalesce more recently with the additional lineage.

We now consider computing the CSP $\widehat{\pi}_{\text{Mig}}(\mathbf{c}|\mathbf{n})$. It is possible to derive the following result directly using the diffusion-generator approximation, but we defer this work to Appendix A. Below, we obtain the result through the genealogical process detailed above; using typical forward-backward genealogical arguments in coalescent theory, we deduce that $\widehat{\pi}_{\text{Mig}}(\mathbf{c}|\mathbf{n})$ satisfies the following equation:

$$\begin{aligned}\widehat{\pi}_{\text{Mig}}(\mathbf{c}|\mathbf{n})=\frac{1}{\mathscr{N}}\cdot\sum_{\substack{\gamma\in\Gamma,\\ h\in\mathscr{H}}}c_{\gamma,h}\Big[&\left(n_{\gamma,h}+c_{\gamma,h}-1\right)\kappa_\gamma^{-1}\widehat{\pi}_{\text{Mig}}\left(\mathbf{c}-\mathbf{e}_{\gamma,h}|\mathbf{n}\right)\\&+\sum_{\ell\in L}\theta_\ell\sum_{a\in E_\ell}\mathbf{P}^{(\ell)}_{a,h[\ell]}\widehat{\pi}_{\text{Mig}}\left(\mathbf{c}-\mathbf{e}_{\gamma,h}+\mathbf{e}_{\gamma,\mathscr{S}_\ell^a(h)}|\mathbf{n}\right)\\&+\sum_{b\in B}\rho_b\sum_{h'\in\mathscr{H}}\widehat{\pi}_{\text{Mig}}\left(\mathbf{c}-\mathbf{e}_{\gamma,h}+\mathbf{e}_{\gamma,\mathscr{R}_b(h,h')}+\mathbf{e}_{\gamma,\mathscr{R}_b(h',h)}|\mathbf{n}\right)\\&+\sum_{\gamma'\neq\gamma}m_{\gamma\gamma'}\widehat{\pi}_{\text{Mig}}\left(\mathbf{c}-\mathbf{e}_{\gamma,h}+\mathbf{e}_{\gamma',h}|\mathbf{n}\right)\Big],\end{aligned}\tag{3.1}$$

where $\mathscr{S}_\ell^a(h)$ denotes the haplotype obtained by substituting the allele at locus $\ell$of h with allele a, and $\mathscr{R}_b(h,h')$ denotes the haplotype obtained, via recombination about breakpoint b

= ($\ell\ell$+1), by joining the (partial) haplotypes $h[1:\ell]$ and $h'[\ell+1, k]$. The first term on the right hand side of this equation corresponds to coalescence and absorption of haplotype h in deme γ, and the subsequent terms correspond to mutation, recombination, and migration, respectively. The normalizing constant $\mathscr{N}$ is given by

$$\mathscr{N} = \sum_{\gamma\in\Gamma} c_\gamma \left[\left(n_\gamma + c_\gamma - 1\right)\kappa_\gamma^{-1} + \sum_{\ell\in L}\theta_\ell + \sum_{b\in B}\rho_b + \sum_{\gamma'\neq\gamma} m_{\gamma\gamma'} \right].$$

Equation (3.1) is for the "full" (conditional) genealogical process, and, because of the recombination terms, it cannot be directly computed by solving a set of linear equations. However, as in Paul and Song (2010), it is possible to derive a "reduced" recursion related to (3.1) that can be computed by solving a finite set of linear equations. Unfortunately, the number of variables in the set of equations grows super-exponentially with both the number of loci and the number of haplotypes in the sample configuration **n**, making it computationally intractable for all but the smallest problems. In the following section, we propose accurate approximations that allow for efficient computation.

4. An efficiently computable CSD as an approximation of $\widehat{\pi}_{\text{Mig}}$

As described above, the recursion for $\widehat{\pi}_{\text{Mig}}(\mathbf{c}|\mathbf{n})$ becomes computationally intractable for even modest datasets. In what follows, we adopt a set of approximations to obtain a CSD that admits efficient implementation, while retaining the accuracy of $\widehat{\pi}_{\text{Mig}}$.

4.1. The CSD $\widehat{\pi}_{\text{MigSMC}}$: Sequentially Markov approximation of $\mathscr{C}$

We follow Paul et al. (2011) and use ideas underlying the SMC (Wiuf and Hein, 1999; McVean and Cardin, 2005; Marjoram and Wall, 2006) to approximate $\widehat{\pi}_{\text{Mig}}$. Briefly, observe that a given conditional genealogy induces a *marginal conditional genealogy* (MCG) at each locus, where each MCG comprises a series of mutation and migration events, and the eventual absorption into a lineage of the sub-trunk in a certain deme. See Figure 1(b) for an illustration. The key insight, initially provided by Wiuf and Hein (1999), is that we can generate the conditional genealogy as a *sequence* of MCGs, rather than backwards in time. Moreover, though the sequence is not formally Markov, it is well approximated (McVean and Cardin, 2005; Marjoram and Wall, 2006; Paul et al., 2011) by a Markov process using a two-locus transition density. Applying this approximation to $\widehat{\pi}_{\text{Mig}}$ yields the *sequentially Markov* CSD $\widehat{\pi}_{\text{MigSMC}}$. For ease of exposition, we restrict attention to the case of sampling a single additional haplotype, denoted $\mathfrak{h}$, but the ideas generalize, in principle, to sampling two or more additional haplotypes.

Since mutations can be superimposed onto the conditional genealogy, we first consider generating a sequence of MCGs without mutations according to a Markov process. The genealogical process underlying $\widehat{\pi}_{\text{Mig}}$ yields the following sampling procedure for the MCG at an arbitrary locus: The ancestral lineage of the additionally sampled haplotype initially resides in deme α, where the additional haplotype is sampled, and proceeds backwards in time, subject to the migration process, until being absorbed into a lineage of the sub-trunk $\mathscr{A}_0(\mathbf{n}_\gamma)$ within the current deme γ. The associated marginal distribution is used as the initial distribution at the first locus.

Conditional on the marginal genealogy at locus $\ell-1$, the marginal genealogy at locus ℓ can be sampled by first placing recombination events onto the MCG at locus $\ell-1$ according to a

Poisson process with rate $\rho_{\ell-1,\ell}$. If no recombination occurs, the marginal genealogy at locus ℓ is identical to the one at locus $\ell-1$. If recombination does occur, the MCG at locus ℓ is identical to the MCG at locus $\ell-1$ up to the time t_b of the most recent recombination event. At this point, the lineage resides in the same deme in which the ancestral lineage at locus $\ell-1$ resided at the time of the recombination event, and, independently of the lineage at locus $\ell-1$, proceeds backwards in time, subject to the migration process, until being absorbed into a lineage of the sub-trunk $\mathscr{A}_0\left(\mathbf{n}_\gamma\right)$ within the current deme γ. Figure 2 illustrates this transition mechanism for the Markov process.

Conditional on the MCG at locus ℓ, mutations are superimposed onto the MCG according to a Poisson process with rate θ_ℓ. The MCG is absorbed into a trunk lineage corresponding to some haplotype h, thereby specifying an "ancestral" allele $h[\ell]$. This allele is then propagated to the present according to the mutations and the mutation transition matrix $\mathbf{P}^{(\ell)}$, thereby generating an allele at locus ℓ of the additional haplotype. We refer to the associated distribution of alleles as the emission distribution.

It is possible to write down explicit expressions for the initial, transition, and emission distributions for $\widehat{\pi}_{\mathrm{MigSMC}}$. However, as the state space for the MCG at each locus includes the entire migrational history, an efficient algorithm for computing the CSP is not known. In the next subsections, we introduce further approximations to this model in order to admit an efficient implementation.

Although we do not prove it here, we note that, analogous to Paul et al. (2011), the sequentially Markov version of the CSD can be obtained from the genealogical process introduced in Section 3 by prohibiting coalescence events in the conditional genealogy between lineages not ancestral to any overlapping parts of the additionally sampled haplotypes. In the case of sampling one additional haplotype, this corresponds to prohibiting all coalescence events in the conditional genealogy. This observation allows one to write down a recursive formula to compute probabilities under $\widehat{\pi}_{\mathrm{MigSMC}}$, but this again does not lead to an efficient implementation.

4.2. The CSD $\widehat{\pi}_{\mathrm{MigSMC-AO}}$: Keeping track of the absorption time only

As noted in the previous subsection, if we keep track of all demes in which the additional ancestral lineage at a given locus resides at any given time in the past, then the MCG is a complicated object. To remedy this, we approximate the full marginal genealogy by just recording the time until absorption, as well as the deme in which the ancestral lineage resides at the time of absorption and also the absorbing haplotype. The reduced MCG at locus ℓ is thus given by a triplet of random variables $\left(G_\ell^A, T_\ell^A, H_\ell^A\right) \in \Gamma \times \mathbb{R}_{\geq 0} \times \mathscr{H}$, that the deme absorption G_ℓ^A, the absorption time T_ℓ^A, and the absorbing haplotype H_ℓ^A. Henceforward, of S to denote $\Gamma \times \mathbb{R}_{\geq 0} \times \mathscr{H}$.

Now, observe that the marginal migration dynamics of the ancestral lineage at a single locus can be described by a continuous-time Markov chain with a finite state space. The states can be divided in two groups: one state for each deme denoting *residence* in that deme before being absorbed into the trunk, and another one for each deme to represent being *absorbed* into a lineage of the trunk in the given deme at some previous time. We denote the set of states by $\{1, \ldots, g, a_1, \ldots, a_g\}$, where, for $1 \leq i \leq g$, state i denotes residence in deme i, and state a_i denotes absorption in deme i. The dynamics of the Markov chain is given by the $(2g \times 2g)$-dimensional block-specified rate matrix

$$Z=\begin{pmatrix} M & - & A & A \\ & 0 & & 0 \end{pmatrix},$$

where **0** is a $(g \times g)$-dimensional matrix of zeros, M is the $(g \times g)$-dimensional matrix of migration rates which govern the transitions between the first group of states (the non-absorbed states), and A is the $(g \times g)$-dimensional diagonal matrix

$$A=\begin{pmatrix} \kappa_1^{-1} n_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \kappa_g^{-1} n_g \end{pmatrix}$$

which governs the transition into the second group (the absorbed states). The diagonal form ensures that the absorbed state a_i can be reached only if the ancestral lineage currently resides in deme i. Also, note that absorption is proportional to the inverse of relative size k_i^{-1} of deme i, as well as the number of trunk-lineages n_i in deme i. Because the absorbing states are also absorbing in the context of the Markov chain, the rows of Z corresponding to these states are set to zero.

The process generating the conditional genealogy for the whole additional haplotype proceeds sequentially along the haplotype, and thus admits a natural interpretation in an HMM framework, where the MCG at a given locus is the hidden state and the allele of the additionally sampled haplotype at this locus is the emitted symbol. We now describe the initial density, the transition density, and the emission probability.

4.2.1. Initial density—Standard theory of Markov chains yields that the probabilities of interest for the initial density can be found in the respective entries of the transition semigroup. If the additional haplotype is sampled at present in deme α, the probability of residing in deme i and not being absorbed more recently than time t into the past is $(e^{tZ})_{\alpha,i}$. On the other hand, the cumulative probability of being absorbed in deme i more recently than time t is given by $(e^{tZ})_{\alpha,a_i}$. Thus, the initial density of state $s = (\omega, h, t)$, that is the density of being absorbed in deme ω at time t into the trunk-lineage of haplotype h, is given by the derivative of the latter matrix exponential:

$$\begin{aligned} \zeta^{(\mathbf{n})}(s) &= \frac{d}{dt}\mathbb{P}\left\{G^A=\omega, T^A \le t, H^A=h\right\} \\ &= \frac{d}{dt}\left\{\frac{n_{\omega,h}}{n_\omega}\left(e^{tZ}\right)_{\alpha,a_\omega}\right\} \\ &= \frac{n_{\omega,h}}{n_\omega}\left(Ze^{tZ}\right)_{\alpha,a_\omega}. \end{aligned} \tag{4.1}$$

The factor $n_{\omega,h}/n_\omega$ comes from the fact that absorption into a specific lineage of the trunk is uniform amongst those present in deme ω.

4.2.2. Transition density—The density for transition from locus $\ell - 1$ to ℓ using the full MCG, described in Section 4.1, conditions on the full migration history of the lineage of the additionally sampled haplotype at locus $\ell - 1$. Thus, at the time of a possible recombination event, all demes up to this event, including the deme where the event takes place are determined. If only the time and the deme of absorption are recorded, then the deme in which the ancestral lineage resides at the time of the recombination event is a random variable with a distribution determined by the dynamics of the Markov chain. Let $G^B_{t_b}$ denote

the random deme in which the ancestral lineage of the additional haplotype resides at the time t_b of the recombination event. Then, for $\gamma \in \Gamma$, the distribution is given by

$$\mathbb{P}\left\{G^B_{t_b}=\gamma | T^A_{\ell-1}=t_{\ell-1}, G^A_{\ell-1}=\omega_{\ell-1}\right\}=\frac{\left[e^{t_b Z}\right]_{\alpha,\gamma}\left[Ze^{(t_{\ell-1}-t_b)Z}\right]_{\gamma,a_{\omega_{\ell-1}}}}{\left[Ze^{t_{\ell-1}Z}\right]_{\alpha,a_{\omega_{\ell-1}}}}. \quad (4.2)$$

The transition from $s_{\ell-1}$ to s_ℓ is now given as follows: The time t_b of the potential recombination event is chosen according to an exponential distribution with rate $\rho_{(\ell-1,\ell)}$. If $t_b \geq t_{\ell-1}$, no recombination occurs and the MCG s_ℓ is identical to $s_{\ell-1}$. If $t_b < t_{\ell-1}$, then recombination occurs and we use (4.2) to determine the probability that the ancestral lineage resided in a certain deme at the time of the recombination event. The ancestral lineage at locus ℓ proceeds from time t_b in this deme and is again subject to migration according to the dynamics of the Markov chain governed by rate matrix Z, until it is absorbed. Integrating over the possible times of the recombination event and summing over the different possible demes yields the following transition density of the hidden state s_ℓ at locus ℓ, given $s_{\ell-1}$ at the previous locus:

$$\begin{aligned}\phi^{(\mathbf{n})}_{\rho_b}(s_\ell|s_{\ell-1}) &= \mathbb{P}\left\{G^A_\ell=\omega_\ell, T^A_\ell \in dt_\ell, H^A_\ell=h_\ell | G^A_{\ell-1}=\omega_{\ell-1}, T^A_{\ell-1}=t_{\ell-1}, H^A_{\ell-1}=h_{\ell-1}\right\} \\ &= e^{-\rho_b t_{\ell-1}}\delta_{s_\ell,s_{\ell-1}} + \frac{n_{\omega_\ell,h_\ell}}{n_{\omega_\ell}}\int_{t_b=0}^{t_{\ell-1}\wedge t_\ell} e^{-\rho_b t_b}\sum_{\gamma\in\Gamma}\frac{\left[e^{t_b Z}\right]_{\alpha,\gamma}\left[Ze^{(t_{\ell-1}-t_b)Z}\right]_{\gamma,a_{\omega_{\ell-1}}}}{\left[Ze^{t_{\ell-1}Z}\right]_{\alpha,a_{\omega_{\ell-1}}}}\left[Ze^{(t_\ell-t_b)Z}\right]_{\gamma,a_{\omega_\ell}} dt_b, \end{aligned} \quad (4.3)$$

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

4.2.3. Emission probability—Since the mutation rate does not depend on the deme in which the ancestral lineage of the additional haplotype resides, the emission probability at locus ℓ only depends on the absorption time t_ℓ and the allele of the absorbing haplotype at that locus $h_{\ell}[\ell]$. As described above, a Poisson number (with mean $t_\ell\theta/2$) of mutation events is placed onto the MCG s_ℓ and the "ancestral" allele is propagated to the present according to the mutation transition matrix $\mathbf{P}^{(\theta)}$. Thus, the probability that the allele of the additional haplotype is $\mathfrak{h}[\ell]$, given the hidden state s_ℓ can be written as the following matrix exponential:

$$\begin{aligned}\xi^{(\mathbf{n})}_{\theta_\ell}(\mathfrak{h}[\ell]|s_\ell) &= \mathbb{P}\left\{\mathfrak{H}_\ell=\mathfrak{h}[\ell] | G^A_\ell=\omega_\ell, T^A_\ell=t_\ell, H^A_\ell=h_\ell\right\} \\ &= \left[e^{t_\ell\theta_\ell(\mathbf{P}^{(\ell)}-\mathbf{I})}\right]_{h_\ell[\ell],\mathfrak{h}[\ell]'} \end{aligned} \quad (4.4)$$

where $\mathfrak{H}_\ell$ denotes the random allele emitted at locus ℓ.

4.2.4. A hidden Markov model formulation to compute the CSP—Using the quantities introduced in the previous Sections, we can now employ the forward algorithm for this HMM with continuous hidden state space (Cappé et al., 2005) by defining the quantity $f^{\mathfrak{h}}_\ell(\cdot)$ recursively. The base case for the first locus is

$$f^{\mathfrak{h}}_1(s_1) = \xi^{(\mathbf{n})}_{\theta_1}(\mathfrak{h}[1]|s_1)\cdot\zeta^{(\mathbf{n})}(s_1),$$

and the recursive step for the transition from locus $\ell-1$ to ℓ is

$$f^{\mathfrak{h}}_\ell(s_\ell) = \xi^{(\mathbf{n})}_{\theta_\ell}(\mathfrak{h}[\ell]+s_\ell)\cdot\int\phi^{(\mathbf{n})}_{\rho_b}(s_\ell|s_{\ell-1})\, f^{\mathfrak{h}}_{\ell-1}(s_{\ell-1})\, ds_{\ell-1},$$

where $b = (\ell - 1, \mathfrak{h})$. Finally, the CSP $\widehat{\pi}_{\mathrm{MigSMC-AO}}$ of an additional haplotype $\mathfrak{h}$ given the already observed sample configuration $\mathbf{n}$ is given by:

$$\widehat{\pi}_{\mathrm{MigSMC-AO}}(\mathfrak{h}|\mathbf{n}) = \int f_k^{\mathfrak{h}}(s_k)\, ds_k.$$

The sequentially Markov genealogical process corresponding to the CSD $\widehat{\pi}_{\mathrm{MigSMC-AO}}$ is illustrated in Figure 1(c).

Note that the dynamics of the Markov chain on the hidden states is reversible with respect to the initial density, i.e.,

$$\phi_\rho^{(\mathbf{n})}\left(s'|s\right)\zeta^{(\mathbf{n})}(s) = \phi_\rho^{(\mathbf{n})}\left(s|s'\right)\zeta^{(\mathbf{n})}\left(s'\right)$$

holds for all recombination rates $\rho \in \mathbb{R}_{\geq 0}$ and hidden states $s, s' \in S$. Thus, the initial density $\zeta^{(\mathbf{n})}(\cdot)$ is in fact the stationary distribution of the Markov chain on the hidden state space. Reversibility also ensures that the CSP computed starting at the first locus and proceeding forward is the same as the CSP computed when starting at the final locus and proceeding backward.

4.3. Discretizing time

The reduced hidden state space of the HMM introduced in the previous subsection yields an approximation to the full sequentially Markov conditional sampling distribution. However, the hidden state space (in particular the absorption time) is continuous, making implementation with standard (discrete) HMM methodology impossible. Thus, as in Paul et al. (2011), we propose a further approximation, by discretizing the positive real line into a finite number of intervals and recording the interval that the absorption time falls into. Formally, this corresponds to the approximation that the transition density and emission probability are equal for times that belong to the same interval.

To this end, assume that $0 = x_0 < x_1 < \cdots < x_d = \infty$ is a finite, strictly increasing sequence in $\mathbb{R}_{\geq 0} \cup \{\infty\}$ such that $\mathscr{D} = \left\{\mathscr{D}_j = \left[x_{j-1}, x_j\right)\right\}_{j=1,\ldots,d}$ is a partition of $\mathbb{R}_{\geq 0}$ into d intervals. We denote the discretized hidden state space $\Gamma \times \{1, \ldots, d\} \times \mathscr{H}$ by Σ and the hidden states by $\sigma = (\omega, i, h) \in \Sigma$, where i is the index of the interval of absorption. As before, ω denotes the deme and h the trunk lineage of absorption. Based on the partition $\mathscr{D}$, denote the discretized version of the initial density as

$$\begin{aligned}\tilde{\zeta}^{(\mathbf{n})}(\sigma) &= \mathbb{P}\left\{G^A = \omega, T^A \in D_i, H^A = h\right\} \\ &= \int_{D_i} \zeta^{(\mathbf{n})}(\omega, t, h)\, dt \\ &= u(\omega, i) \cdot \frac{n_{\omega,h}}{n_\omega},\end{aligned}$$

where

$$u(\omega, i) = \int_{t \in D_i} \left(Ze^{tZ}\right)_{\alpha, a_\omega} dt = \left(e^{x_i Z}\right)_{\alpha, a_\omega} - \left(e^{x_{i-1} Z}\right)_{\alpha, a_\omega}.$$

Note that the event $\left\{T^A \in D_i\right\}$ encodes that we only record the time interval in which absorption happens.

Similarly, we can derive the discretized version of the transition density as

$$\begin{aligned}\tilde{\phi}^{(\mathbf{n})}_{\rho_b}(\sigma_\ell|\sigma_{\ell-1}) &= \mathbb{P}\left\{G^A_\ell = \omega_\ell, T^A_\ell \in D_{i_\ell}, H^A_\ell = h_\ell | G^A_{\ell-1} = \omega_{\ell-1}, T^A_{\ell-1} \in D_{i_{\ell-1}}, H^A_{\ell-1} = h_{\ell-1}\right\} \\ &= \frac{1}{\tilde{\zeta}^{(\mathbf{n})}(\sigma_{\ell-1})} \int_{D_{i_\ell}} \int_{D_{i_{\ell-1}}} \phi^{(\mathbf{n})}_{\rho_b}(\omega_\ell, t_\ell, h_\ell | \omega_{\ell-1}, t_{\ell-1}, h_{\ell-1})\, \zeta^{(\mathbf{n})}(\omega_{\ell-1}, t_{\ell-1}, h_{\ell-1})\, dt_{\ell-1} dt_\ell \\ &= y_{\rho_b}(\omega_{\ell-1}, i_{\ell-1}) \cdot \delta_{s_{\ell-1}, s_\ell} + z_{\rho_b}(\omega_\ell, i_\ell | \omega_{\ell-1}, i_{\ell-1}) \cdot \frac{n_{\omega_\ell, h_\ell}}{n_{\omega_\ell}},\end{aligned} \tag{4.5}$$

where explicit expressions of $y_{\rho b}(\omega_{\ell-1}, i_{\ell-1})$ and $z_{\rho b}(\omega_\ell, i_\ell | \omega_{\ell-1}, i_{\ell-1})$ are shown in Appendix C. Note that we again only record the intervals containing the absorption times at locus $\ell - 1$ and ℓ.

Finally, the emission probabilities in the discretized HMM can be obtained via

$$\begin{aligned}\tilde{\phi}^{(\mathbf{n})}_{\theta_\ell}(\mathfrak{h}[\ell]|\sigma_\ell) &= \mathbb{P}\left\{\mathfrak{H}_\ell = \mathfrak{h}[\ell] | G^A_\ell = \omega_\ell, T^A_\ell \in D_{i_\ell}, H^A_\ell = h_\ell\right\} \\ &= \frac{1}{\tilde{\zeta}^{(\mathbf{n})}(\sigma_\ell)} \int_{D_{i_\ell}} \xi^{(\mathbf{n})}_{\theta_\ell}(\mathfrak{h}[\ell]|\omega_\ell, t_\ell, h_\ell)\, \zeta^{(\mathbf{n})}(\omega_\ell, t_\ell, h_\ell)\, dt_\ell,\end{aligned} \tag{4.6}$$

and we again provide a more explicit form of this quantity in Appendix C. Note that the emission probability (4.4) in the continuous case is only dependent on the time of absorption and the allele that the absorbing haplotype bears at the given locus. The discretized analog (4.6) on the other hand also depends on the deme that the absorbing haplotype resides in. This is due to the fact that the latter conditions on being absorbed at any point in a given time interval, and since the rate of absorption during that interval depends on the deme, this dependence enters expression (4.6).

With the state space discretized, the CSP can be computed via the standard forward algorithm for HMMs (Cappé et al., 2005). Thus, we define the quantity $F^{\mathfrak{h}}_\ell(\sigma_\ell)$ recursively along loci. At the first locus, we have

$$F^{\mathfrak{h}}_1(\sigma_1) = \tilde{\xi}^{(\mathbf{n})}_{\theta_1}(\mathfrak{h}[1]|\sigma_1) \cdot \tilde{\zeta}^{(\mathbf{n})}(\sigma_1).$$

The transition from locus $\ell - 1$ to locus ℓ is given by

$$F^{\mathfrak{h}}_\ell(\sigma_\ell) = \tilde{\xi}^{(\mathbf{n})}_{\theta_\ell}(\mathfrak{h}[\ell]|\sigma_\ell) \cdot \sum_{\sigma_{\ell-1} \in \Sigma} \tilde{\phi}^{(\mathbf{n})}_{\rho_b}(\sigma_\ell|\sigma_{\ell-1})\, F^{\mathfrak{h}}_{\ell-1}(\sigma_{\ell-1}),$$

and the probability of observing haplotype $\mathfrak{h}$ under the discretized HMM is given by

$$\widehat{\pi}_{\text{MigSMC-AOD}}(\mathfrak{h}|\mathbf{n}) := \sum_{\sigma_k \in \Sigma} F^{\mathfrak{h}}_k(\sigma_k),$$

which provides an approximation to $\widehat{\pi}_{\text{MigSMC-AO}}(\mathfrak{h}|\mathbf{n})$.

Remark.

1. 1. In Paul et al. (2011), the authors advocate using a discretization based on points obtained from Gaussian quadrature. However, we obtained numerically more stable results when using a logarithmic discretization, that is $x_i = -(1/r)\log((d-i)/i)$, where r is the harmonic mean of the absorption rates in each deme.

2. 2. The runtime of the standard implementation of the forward algorithm for HMMs described in the previous paragraph is quadratic in the number of hidden states. In Paul et al. (2011), the authors describe a straightforward implementation of their model that leads to a better bound on the runtime. Since our transition density is of similar form, a similar improvement can be applied here.

5. Application: Estimating migration rates

To demonstrate the utility of our approximate CSD $\widehat{\pi}_{\mathrm{MigSMC-AOD}}$, we considered estimating migration rates for data simulated under the full coalescent with recombination and migration. In particular, we simulated data for $k = 10^4$ bi-allelic loci. For simplicity, we set $\theta_\ell = 5 \times 10^{-2}$ and $\mathbf{P}^{(\ell)} = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}$ for all $\ell \in L$, and $\rho_b = 5 \times 10^{-2}$ for all $b \in B$. We considered a structured population with two demes (i.e., $\Gamma = \{1,2\}$), and set $\kappa_1 = \kappa_2 = 0.5$ and $m_{12} = m_{21} = m$. For each value of $m \in \{0.001, 0.10, 1.00, 10.0\}$, we generated 100 datasets with $n_1 = n_2 = 10$ individuals in each of the two demes.

Observe that the per-individual mutation and recombination rates are thus both approximately $10^4 \cdot 5 \times 10^{-2} = 5 \times 10^2$. In humans, for which average per-base mutation and recombination rates are on the order of 10^{-3}, these values correspond to a genomic sequence on the order of 500 kb. We thus reason that the haplotypes we simulated are representative of a relatively longer genomic sequence that has been "compressed", for reasons of computational efficiency, into 10^4 loci. Further, we chose the range of migration rates to be compliant with recent estimates in humans (Gutenkunst et al., 2009; Gravel et al., 2011), as well as Drosophila (Wang and Hey, 2010).

We considered three approximate/composite likelihood formulations that make use of the CSD. Let $\mathbf{n}$ be a particular sample configuration of n haplotypes, and write $\mathbf{n} = \sum_{i=1}^{n} \mathbf{e}_{\gamma_i, h_i}$.

LCL: In the *Leave-one-out Composite Likelihood*, the likelihood is approximated as a product of CSPs with each the result of removing a single haplotype from the sample configuration:

$$\mathrm{LCL}(\mathbf{n}) = \left[\prod_{i=1}^{n} \widehat{\pi}_{\mathrm{MigSMC-AOD}} \left(\mathbf{e}_{\gamma_i, h_i} | \mathbf{n} - \mathbf{e}_{\gamma_i, h_i} \right) \right]^{1/n}.$$

.

PAC: In the popular *Product of Approximate Conditionals* framework (Li and Stephens, 2003), the likelihood is approximated by a conditional decomposition, averaged over 20 random permutations $\{\sigma_j\}$ of the haplotypes (this number of permutations is as suggested by Li and Stephens). Defining $\sigma_j(\mathbf{e}_{\gamma i, hi}) = \mathbf{e}_{\gamma \sigma j(i), h \sigma j(i)}$:

$$\mathrm{PAC}(\mathbf{n}) = \frac{1}{20} \sum_{j=1}^{20} \prod_{i=1}^{n} \widehat{\pi}_{\mathrm{MigSMC-AOD}} \left(\sigma_j \left(\mathbf{e}_{\gamma_i, h_i} \right) | \mathbf{n} - \sum_{i'=1}^{i} \sigma_j \left(\mathbf{e}_{\gamma_{i'}, h_{i'}} \right) \right).$$

.

PCL: In the *Pairwise Composite Likelihood*, the likelihood is approximated as a product of CSPs with each a single haplotype conditioned upon a single haplotype:

$$\mathrm{PCL}(\mathbf{n}) = \prod_{i=1}^{n} \prod_{i' \neq i} \widehat{\pi}_{\mathrm{MigSMC-AOD}}\left(\mathbf{e}_{\gamma_i, h_i} | \mathbf{e}_{\gamma_{i'}, h_{i'}}\right).$$

.

We set the values of θ and ρ to the (true) values used for simulation, and considered the approximate likelihood surfaces for the parameter m. Figure 3 shows the surfaces for two example configurations (generated as described above) for $m = 0.10$. Perhaps most importantly, the likelihood surfaces appear to be unimodal and otherwise well-behaved. In Figure 3(a), the likelihood curves are quite similar to one another, and the maximum likelihood occurs near the true parameter. This is not generally true, however, as evidenced by Figure 3(b), for which the likelihood surface for the LCL method is substantially different than that of PAC and PCL.

We also considered the behavior of the maximum likelihood estimate (MLE) under each of the likelihood approximations. For each simulated dataset, we computed, using golden section search, the MLE migration rate $\widehat{m}$ and computed $\log_2(\widehat{m}/m)$, where m is the (true) migration rate used to generate the dataset. In this way, results for different values of m are directly comparable; a correct estimate of the migration rate produces a value of 0, and under- and overestimation by a factor of two produce values of −1 and 1, respectively. To assess the performance of the MLEs based on the CSD developed in this paper, we also compared with estimates obtained from the widely-used test statistic F_{ST}:

F_{ST}: It can be shown that the migration rate in a symmetric island model with two sub-populations can be estimated by

$$\widehat{m}(\mathbf{n}) = \frac{1}{4}\left(\frac{1}{\mathrm{F}_{\mathrm{ST}}(\mathbf{n})} - 1\right),$$

, where $F_{ST}(\mathbf{n}) = 1 - \boldsymbol{\pi}_S(\mathbf{n})/\boldsymbol{\pi}_T(\mathbf{n})$, with $\boldsymbol{\pi}_S(\mathbf{n})$ denoting the average within-population diversity and $\boldsymbol{\pi}_T(\mathbf{n})$ the overall diversity; c.f., Charlesworth (1998, Equation (4)). Note that, although Charlesworth discusses three different estimators for F_{ST}, the corresponding migration rate estimators coincide in models where the sub-populations have equal weights.

For each true migration rate $m \in \{0.01, 0.10, 1.00, 10.0\}$, box plots for the transformed MLE under each likelihood approximation and the F_{ST}-based estimator are presented in Figure 4. Observe that the LCL-based MLE performs very poorly for $m = 0.01$ (see Figure 4(a)), consistently underestimating the true value; this may be because the final haplotype to be sampled is generally very similar to previously-sampled haplotypes within the deme, obviating the need for migration events within the conditional genealogy. Intuitively, this effect should be diminished when the data are produced using larger migration rates, which does appears to be the case (see Figures 4(b), 4(c), and 4(d)).

On the other hand, the PCL-based MLE performed poorly for $m = 10.0$, again consistently underestimating the true value. This may be because, for large migration rates, there simply is not enough information in a pairwise analysis of the haplotypes to determine the true rate; intuitively, this effect should be diminished when the data are produced using smaller migration rates, relative to the rate of recombination. This is indeed the case, and in fact, for smaller migration rates, the PCL-based MLE is well-correlated with the PAC-based MLE (data not shown).

The PAC-based MLE appears not to suffer at either of these extremes. We speculate that this is because PAC incorporates both pairwise and higher-order terms, making it less susceptible to the problems we observe with the LCL- and PCL-based MLEs. We note that Li and Stephens (2003) came to a similar conclusion.

For low migration rates, the method based on F_{ST} consistently overestimates the true rates (see Figure 4(a)), but shows a small variance. For intermediate migration rates, on the other hand, it produces underestimates (see Figures 4(b) and 4(c)), and the variance is larger than that of the PAC-based MLE. The estimates for large migration rates (see Figure 4(d)) are similarly biased, although the variance is comparable to the MLE methods in this case. Overall, the PAC-based estimation is quite accurate, demonstrating that, using the CSD $\widehat{\pi}_{\mathrm{MigSMC-AOD}}$, it is possible to attain accurate estimates of the migration rate.

6. Discussion

Numerous applications in population genomics make use of the conditional sampling distribution, so developing accurate, efficiently computable CSDs for various population genetic models is of much interest. Recently, we proposed an accurate sequentially Markov CSD that follows from approximating the diffusion process dual to the coalescent with recombination for a single panmictic population. In this paper, we have extended that approach to incorporate subdivided population structure with migration, providing a novel CSD that facilitates computation and also admits a useful genealogical interpretation closely related to the structured coalescent with migration and recombination. We believe that this extension will have several interesting applications, some of which we list below.

Recalling the applications of CSDs described in the introduction, we note that it is straightforward to apply $\widehat{\pi}_{\mathrm{MigSMC-AOD}}$ to annotate segments of distinct ancestry in individuals. As in Price et al. (2009), the already observed configuration consists of the donor individuals from different populations. For low migration rates, the model underlying $\widehat{\pi}_{\mathrm{MigSMC-AOD}}$ leads naturally to the fact that, following a recombination event, the ancestral lineage at the next locus is more likely to get absorbed in the same deme, rather than switching demes by a migration event and then getting absorbed in a different deme. Whereas the method developed by Price et al. (2009) is applicable for recently admixed individuals, we expect our model to be more accurate in situations where the mixing of the populations happened over a long time through the continuous exchange of migrants.

Recall that Wegmann et al. (2011) estimated relative recombination rate variation in different populations based on ancestry switch-points in the chromosome detected using the method of Price et al. (2009). As detailed above, our model can be extended to detect ancestry switch-points in populations that mixed over long periods of time. In such situations we expect that the segments of different ancestry detected by our method can be used in a similar fashion as in Wegmann et al. (2011) to analyze recombination rate variation in different subpopulations, when no strong recent admixture is evident.

Recently, Li and Durbin (2011) performed a related analysis of human demography. They used the SMC in the special case of a sample consisting of only two sequences, and thus were able to obtain explicit transition functions along the sequence, as we did for our CSD (Paul et al., 2011). Li and Durbin incorporated changes in the size of the population into their model, thus allowing them to use the two sequences of a diploid individual to infer population size histories of different human populations. They do not explicitly account for population structure and migration in their analysis, but we believe that the methods developed in this paper could be readily incorporated into their model. In a similar study, Mailund et al. (2011) used a pair of sequences sampled from different populations in the

SMC framework to estimate ancestral population sizes and splitting times in an isolation model. Again, we think it is possible to incorporate migration into the model using the ideas we developed in this paper.

Paul and Song (2012) have recently developed a framework to substantially increase the speed of computations involved in dealing with HMMs for next-generation sequence data, and they demonstrate their improvements in the model introduced by Paul et al. (2011). Utilizing the fact that whole genomic sequence data consists of long stretches without sequence variation in between SNPs, and that the observed variation can be described by a small number of haplotype blocks, they were able to decrease the computation time by several orders of magnitude. The same ideas can be applied to speedup our method, fostering the application of analyses like the one detailed in Section 5 or similar applications to whole genome sequence data.

The CSDs developed in this paper assume that the structure underlying the population remains unchanged throughout the whole course of evolution. Furthermore the rates at which migrants are exchanged are assumed constant. This is mirrored by the fact that the Markov chain described by the matrix Z is time homogeneous and the number of states does not change. For more realistic populations one would pose changes of the population structure and the population sizes at different points in the past, as well as varying rates of migration. The methods developed in this paper can be readily extended to scenarios where the structural parameters of the underlying demography are piecewise constant for given periods of time. This can be implemented by allowing the Markov chain, governing the absorption of the additional ancestral lineage, to be piecewise homogeneous. Except for the work of Davison et al. (2009) and Price et al. (2009), we are not aware of any other CSDs that try to incorporate explicit population structure into the copying model. Such a CSD accounting for a more general demographic model would allow one to estimate more general demographic parameters like ancient population sizes and structure, as well as migration rates, and duration of periods of migration in certain isolation-with-migration scenarios, using, for example, the framework illustrated in Section 5, importance sampling (Stephens and Donnelly, 2000; Fearnhead and Donnelly, 2001; Griffiths et al., 2008), or other frameworks detailed in the introduction. Myers et al. (2008) show that demographic studies like Gutenkunst et al. (2009) and Gravel et al. (2011), that rely exclusively on the frequency spectrum, can be limited in resolving demographic parameters, and methods, as the one developed in this paper, that explicitly incorporate linkage structure, might alleviate such problems.

Acknowledgments

We thank John Kamm for many stimulating and fruitful discussions. This research is supported in part by a DFG Research Fellowship STE 2011/1-1 to M.S; an NIH National Research Service Award Trainee appointment on T32-HG00047 to JSP; and an NIH grant R01-GM094402, and a Packard Fellowship for Science and Engineering to Y.S.S.

Appendix A. Di usion approximation

We here provide a derivation of the sampling recursion using the diffusion generator approximation (De Iorio and Griffiths, 2004a,b; Griffiths et al., 2008; Paul and Song, 2010). The diffusion associated with the coalescent including migration and recombination has state given by the vector $\mathrm{x}=\left(x_{\gamma,h}\right)_{\gamma\in\Gamma,h\in\mathcal{H}}$ where $x_{\boldsymbol{\gamma},h}$ is the frequency of haplotype h within deme $\boldsymbol{\gamma}$. The generator for the diffiusion can then be written

$$\mathscr{L} f(\mathbf{x}) = \sum_{\substack{\gamma \in \Gamma \\ h \in \mathscr{H}}} \mathscr{L}_{\gamma,h} \frac{\partial}{\partial x_{\gamma,h}} f(\mathbf{x}),$$

where

$$\begin{aligned}\mathscr{L}_{\gamma,h} f(\mathbf{x}) = \frac{1}{2} \Bigg\{ & x_{\gamma,h} \sum_{h' \in \mathscr{H}} \left(\delta_{h,h'} - x_{\gamma,h'}\right) \kappa_\gamma^{-1} \frac{\partial}{\partial x_{\gamma,h'}} f(\mathbf{x}) \\ & + \sum_{\ell \in L} \theta_\ell \sum_{a \in E_\ell} x_{\gamma,\mathscr{S}_\ell^a(h)} \left(\mathbf{P}^{(\ell)}_{a,h[\ell]} - \delta_{\mathscr{S}_\ell^a(h),h}\right) f(\mathbf{x}) \\ & + \sum_{b \in B} \rho_b \left(\sum_{h' \in \mathscr{H}} x_{\gamma,\mathscr{R}_b(h,h')} x_{\gamma,\mathscr{R}_b(h',h)} - x_{\gamma,h} \right) f(\mathbf{x}) \\ & + \sum_{\gamma' \neq \gamma} \left(m_{\gamma\gamma'} x_{\gamma',h} - m_\gamma x_\gamma, h\right) f(\mathbf{x}) \Bigg\},\end{aligned}$$

and f is an arbitrary, bounded, twice-differentiable function with continuous second derivatives.

Denote the probability of obtaining (at stationarity) an ordered sample configuration $\mathbf{n}$ by $q(\mathbf{n})$. Then $q(\mathbf{n}) = \mathbb{E}[q(\mathbf{n}|\mathbf{X})]$, where $\mathbb{E}$ denotes expectation with respect to the stationary distribution of the diffiusion, $\mathbf{X}$ denotes the random vector of frequencies, and $q(\mathbf{n}|\mathbf{x}) = \Pi_{\gamma \in \Gamma} \Pi_{h \in \mathscr{H}} x_{\gamma,h}^{n_{\gamma,h}}$. Finally, given an additional haplotype configuration $\mathbf{c}$, the true conditional sampling probability is, by definition, $\boldsymbol{\pi}(\mathbf{c}|\mathbf{n}) = \mathbf{q}(\mathbf{c} + \mathbf{n})/q(\mathbf{n})$.

By general diffusion theory, $\mathbb{E}[\mathscr{L} f(\mathbf{X})] = 0$. The diffusion generator approximation assumes the existence of a distribution, with associated expectation $\mathbb{E}$, such that the previous condition holds *component-wise*; that is for each $\boldsymbol{\gamma} \in \Gamma$ and $h \in \mathscr{H}$,

$$\widehat{\mathbb{E}}\left[\mathscr{L}_{\gamma,h} \frac{\partial}{\partial X_{\gamma,h}} f(\mathbf{X})\right] = 0.$$

By analogy to the exact case, we assume that the sampling distribution is approximated by $\widehat{q}(\mathbf{n}) = \widehat{\mathbb{E}}[q(\mathbf{n}|\mathbf{X})]$, and define the approximate CSD $\widehat{\pi}_{\text{Mig}}(\mathbf{c}|\mathbf{n}) = \widehat{q}(\mathbf{n}+\mathbf{c}) / \widehat{q}(\mathbf{n})$. Using the component-wise approximation above,

$$\sum_{\gamma \in \Gamma} \sum_{h \in \mathscr{H}} \frac{c_{\gamma,h}}{c_{\gamma,h} + n_{\gamma,h}} \widehat{\mathbb{E}}\left[\mathscr{L}_{\gamma,h} \frac{\partial}{\partial X_{\gamma,h}} q(\mathbf{n}+\mathbf{c}|\mathbf{X})\right] = 0.$$

Using the expressions for $\mathscr{L}_{\gamma,h}$ and $q(\mathbf{n}+\mathbf{c}|\mathbf{X})$, together with the definition $\widehat{q}(\mathbf{n}) = \widehat{\mathbb{E}}[q(\mathbf{n}|\mathbf{X})]$, we obtain, after some algebra analogous to Paul and Song (2010, Appendix),

$$\widehat{q}(\mathbf{n}+\mathbf{c}) = \frac{1}{\mathscr{N}} \sum_{\substack{\gamma\in\Gamma \\ h\in\mathscr{H}}} c_{\gamma,h} \Big[\big(n_{\gamma,h}+c_{\gamma,h}-1\big)\kappa_\gamma^{-1}\widehat{q}\big(\mathbf{n}+\mathbf{c}-\mathbf{e}_{\gamma,h}\big) + \sum_{\ell\in L}\theta_\ell \sum_{a\in E_\ell} \mathbf{P}^{(\ell)}_{a,h[\ell]}\widehat{q}\Big(\mathbf{n}+\mathbf{c}-\mathbf{e}_{\gamma,h}+\mathbf{e}_{\gamma,\mathscr{S}^a_\ell(h)}\Big) + \sum_{b\in B}\rho_b \sum_{h'\in\mathscr{H}} \widehat{q}\Big(\mathbf{n}+\mathbf{c}-\mathbf{e}_{\gamma,h}+\mathbf{e}_{\gamma,\mathscr{R}_b(h,h')}+\mathbf{e}_{\gamma,\mathscr{R}_b(h',h)}\Big) + \sum_{\gamma'\neq\gamma} m_{\gamma\gamma'}\widehat{q}\big(\mathbf{n}+\mathbf{c}-\mathbf{e}_{\gamma,h}+\mathbf{e}_{\gamma',h}\big)\Big],$$

where the normalizing constant is given by

$$\mathscr{N} = \sum_{\gamma\in\Gamma} c_\gamma \left[\big(n_\gamma+c_\gamma-1\big)\kappa_\gamma^{-1} + \sum_{\ell\in L}\theta_\ell + \sum_{b\in B}\rho_b + m\gamma \right].$$

Dividing this result by $\widehat{q}(\mathbf{n})$, we thus obtain the given recursion for $\widehat{\pi}_{\text{Mig}}(\mathbf{c}|\mathbf{n})$. It is also possible to derive, in much the same way, the "reduced" version of this recursion; for details, see Paul and Song (2010).

Appendix B. Explicit transition density

We begin by assuming that the matrix Z is diagonalizable, which is true if and only if M is diagonalizable. In this case, the matrix exponentials in equations (4.1) and (4.3) admit the eigen-decomposition $\left(e^{tZ}\right)_{i,j} = \sum_{k=1}^{2g} e^{t\lambda_k}(v_k w_k)$ and $\left(Ze^{tZ}\right)_{i,j} = \sum_{k=1}^{2g}\lambda_k e^{t\lambda_k}(v_k w_k)$. Here λ_k are the eigenvalues of Z, v_k are the eigenvectors, and ω_k are the rows of the inverse of the matrix of eigenvectors $V = (v_1, \ldots, v_{2g})$. This eigen-decomposition can be used to evaluate the matrix exponential in equation (4.1), and to compute the integral in equation (4.3) analytically as

$$\phi^{(\mathbf{n})}_{\rho_b}(s_\ell|s_{\ell-1}) = e^{-\rho_b t_{\ell-1}}\delta_{s_\ell,s_{\ell-1}} + \frac{n_{\omega_\ell,h_\ell}}{n_{\omega\ell}} \frac{\rho_b}{\left(Ze^{t_{\ell-1}Z}\right)_{\alpha,a_{\omega_{\ell-1}}}} \sum_{\gamma\in\Gamma}\sum_{k=1}^{2g}\sum_{m=1}^{2g}\sum_{n=1}^{2g} \Big[(v_k w_k)_{\alpha,\gamma}(v_m w_m)_{\gamma,a_{\omega_{\ell-1}}}(v_n w_n)_{\gamma,a_{\omega_\ell}} \times \lambda_m\lambda_n e^{t_{\ell-1}\lambda_m} e^{t_\ell\lambda_n} \mathbf{I}_0^{t_{\ell-1}\wedge t_\ell}(\lambda_k-\lambda_m-\lambda_n-\rho_b)\Big],$$

where

$$\mathbf{I}_a^b(\lambda) = \int_{t=a}^b e^{\lambda t}dt = \begin{cases} \frac{1}{\lambda}\left(e^{\lambda b}-e^{\lambda a}\right), & \text{if } \lambda\neq 0, \\ b-a, & \text{if } \lambda = 0. \end{cases} \quad \text{(B.1)}$$

Note that for a non-diagonalizable matrix, a similar eigen-decomposition can be employed using generalized eigenvectors and the Jordan normal form, and similar, though more involved, explicit computations can be performed.

Appendix C. Probabilities in discretized HMM

We now give more explicit forms of the quantities involved in the probabilities of the discretized HMM, derived using the eigen-decomposition of the extended migration matrix

$Z = V \Lambda V^{-1}$. Inserting equation (4.3) and (4.1) into the second to last line in equation (4.5), combined with the eigen-decomposition $\left(Ze^{tZ}\right)_{i,j} = \sum_{k=1}^{2g} \lambda_k e^{t\lambda_k} (v_k w_k)_{i,j}$ yields

$$y_\rho(\omega_{\ell-1}, i) = \frac{1}{u(\omega, i)} \sum_{k=1}^{2g} (v_k w_k)_{\alpha, a_\omega} \lambda_k \mathbf{I}_{x_{i-1}}^{x_i}(\lambda_k - \rho),$$

for the first term in the last line of equation (4.5), with $\mathbf{I}_a^b(\lambda)$ defined as in equation (B.1). For the second term we get

$$z_\rho(\omega', j|\omega, i) = \frac{\rho}{u(\omega, i)} \sum_{\gamma \in \Gamma} \sum_{k=1}^{2g} \sum_{m=1}^{2g} \sum_{n=1}^{2g} (v_k w_k)_{\alpha,\gamma} (v_m w_m)_{\gamma, a_\omega} (v_n w_n)_{\gamma, a_{\omega'}}$$
$$\times \Big[e^{\lambda_m x_i} e^{\lambda_n x_j} \mathbf{I}_0^{x_i \wedge x_j}(\lambda_k - \lambda_m - \lambda_n - \rho) \quad - e^{\lambda_m x_i} e^{\lambda_n x_{j-1}} \mathbf{I}_0^{x_i \wedge x_{j-1}}(\lambda_k - \lambda_m - \lambda_n - \rho) \quad - e^{\lambda_m x_{i-1}} e^{\lambda_n x_j} \mathbf{I}_0^{x_{i-1} \wedge x_j}(\lambda_k - \lambda_m - \lambda_n - \rho) \quad + e^{\lambda_m \wedge}$$

Finally, using equation (4.6) one can show that

$$\tilde{\xi}_\theta^{(\mathbf{n})}(\mathfrak{h}|\omega, i, h) = \frac{1}{u(\omega, i)} \sum_{j=1}^{|E|} \sum_{k=1}^{2g} \left(q_j p_j\right)_{h[\ell], \mathfrak{h}[\ell]} (v_k w_k)_{\alpha, a_\omega} \lambda_k \mathbf{I}_{x_{i-1}}^{x_i}\left(\lambda_k + \theta\mu_j - \theta\right)$$

holds, where E is the set of alleles at the given locus, and we used the eigen-decompositions of Z and the mutation matrix $\mathbf{P} = Q\,\mathrm{diag}(\mu_1, \ldots, \mu_{|E|})\, Q^{-1}$. Here μ_j are the eigenvalues of the mutation matrix, $Q = (q_1, \ldots, q_{2g})$ is the matrix which has the eigenvectors of the mutation matrix as columns, and p_j denotes the j-th row of Q^{-1}.

References

Cappé, O.; Moulines, E.; Ryden, T. Inference in Hidden Markov Models. Springer: 2005. Springer Series in Statistics

Charlesworth B. Measures of divergence between populations and the effect of forces that reduce variability. Molecular Biology and Evolution. 1998; 15(5):538–543. [PubMed: 9580982]

Davison D, Pritchard JK, Coop G. An approximate likelihood for genetic data under a model with recombination and population splitting. Theor. Popul. Biol. 2009; 75(4):331–345. [PubMed: 19362099]

De Iorio M, Griffiths RC. Importance sampling on coalescent histories. I. Adv. in Appl. Probab. 2004a; 36(2):417–433.

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

Fearnhead P, Donnelly P. Estimating recombination rates from population genetic data. Genetics. 2001; 159:1299–1318. [PubMed: 11729171]

Gay J, Myers SR, McVean GAT. Estimating meiotic gene conversion rates from population genetic data. Genetics. 2007; 177:881–894. [PubMed: 17660532]

Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA, Project TG, Bustamante CD. Demographic history and rare allele sharing among human populations. Proceedings of the National Academy of Sciences. 2011

Griffiths, RC.; Marjoram, P. Progress in population genetics and human evolution (Minneapolis, MN, 1994), volume 87 of *IMA Vol. Math. Appl.*. Springer; New York: 1997. An ancestral recombination graph; p. 257-270.

Griffiths RC, Jenkins PA, Song YS. Importance sampling and the two-locus model with subdivided population structure. Adv. in Appl. Probab. 2008; 40(2):473–500. [PubMed: 19936262]

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. Oct.2009 5(10):e1000695. [PubMed: 19851460]

Hellenthal G, Auton A, Falush D. Inferring human colonization history using a copying model. PLoS Genet. 2008; 4(5):e1000078. [PubMed: 18497854]

Herbots, HM. Progress in population genetics and human evolution (Minneapolis, MN, 1994), volume 87 of *IMA Vol. Math. Appl.*. Springer; New York: 1997. The structured coalescent; p. 231-255.

Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009; 5(6):e1000529. [PubMed: 19543373]

Lawson D, Hellenthal G, Myers S, Falush D. Inference of population structure using dense haplotype data. PLoS Genetics. 2012; 8(1):e1002453. [PubMed: 22291602]

Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nature. 2011; 475:493–496. [PubMed: 21753753]

Li N, Stephens M. Modelling linkage disequilibrium, and identifying recombination hotspots using SNP data. Genetics. 2003; 165:2213–2233. [PubMed: 14704198]

Li Y, Abecasis GR. Mach 1.0: Rapid haplotype reconstruction and missing genotype inference. Am. J. Hum. Genet. 2006; S79:2290.

Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. Mach: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genetic Epidemiology. 2010; 34:816–834. [PubMed: 21058334]

Mailund T, Dutheil JY, Hobolth A, Lunter G, Schierup MH. Estimating divergence time and ancestral effective population size of bornean and sumatran orangutan subspecies using a coalescent hidden markov model. PLoS Genet. Mar.2011 7(3):e1001319. [PubMed: 21408205]

Marchini J, Howie B, Myers SR, McVean GAT, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat. Genet. 2007; 39(7):906–13. [PubMed: 17572673]

Marjoram P, Wall JD. Fast “coalescent” simulation. BMC Genet. 2006; 7:16. [PubMed: 16539698]

McVean GA, Cardin NJ. Approximating the coalescent with recombination. Philos. Trans. R. Soc. Lond. B Biol. Sci. 2005; 360:1387–93. [PubMed: 16048782]

Myers S, Fefferman C, Patterson N. Can one learn history from the allelic spectrum? Theoretical Population Biology. 2008; 73(3):342–348. [PubMed: 18321552]

Paul JS, Song YS. A principled approach to deriving approximate conditional sampling distributions in population genetics models with recombination. Genetics. 2010; 186:321–338. [PubMed: 20592264]

Paul JS, Song YS. Blockwise HMM computation for large-scale population genomic inference. Bioinformatics (submitted). 2012

Paul JS, Steinrücken M, Song YS. An accurate sequentially Markov conditional sampling distribution for the coalescent with recombination. Genetics. 2011; 187:1115–1128. [PubMed: 21270390]

Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, Beaty TH, Mathias R, Reich D, Myers S. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet. 2009; 5(6):e1000519. [PubMed: 19543370]

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

Stephens M, Scheet P. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am. J. Hum. Genet. 2005; 76(3):449–462. [PubMed: 15700229]

Wang Y, Hey J. Estimating divergence parameters with small samples from a large number of loci. Genetics. 2010; 184(2):363–379. [PubMed: 19917765]

Wegmann D, Kessner DE, Veeramah KR, Mathias RA, Nicolae DL, Yanek LR, Sun YV, Torgerson DG, Rafaels N, Mosley T, Becker LC, Ruczinski I, Beaty TH, Kardia SLR, Meyers DA, Barnes KC, Becker DM, Freimer NB, Novembre J. Recombination rates in admixed individuals identified by ancestry-based inference. Nat. Genet. 2011; 43:847–853. [PubMed: 21775992]

Wiuf C, Hein J. Recombination as a point process along sequences. Theor. Pop. Biol. 1999; 55:248–259. [PubMed: 10366550]

Yin J, Jordan MI, Song YS. Joint estimation of gene conversion rates and mean conversion tract lengths from population SNP data. Bioinformatics. 2009; 25(12):i231–9. [PubMed: 19477993]

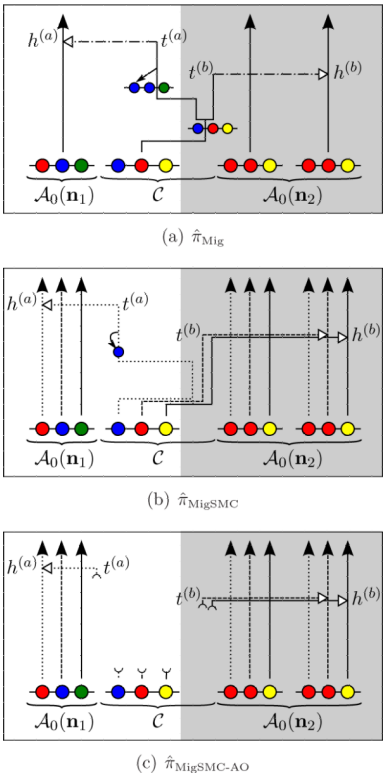

Figure 1.

Illustration of the subsequent approximations to the true conditional sampling distribution. The three loci of each haplotype are each represented by a filled circle, with the color indicating the allelic type at that locus. The trunk genealogies in deme 1 $\mathscr{A}_0(\mathbf{n}_1)$ and deme 2 $\mathscr{A}_0(\mathbf{n}_2)$, as well as the conditional genealogy $\mathscr{C}$ are indicated. The different demes are indicated by the white and the grey background. 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 $\mathscr{C}$. Migration events are indicated by switching to another deme. Absorption events, and the corresponding absorption time ($t^{(a)}$ and $t^{(b)}$) and haplotype ($H^{(a)}$ and $H^{(b)}$, respectively), are indicated by dot-dashed horizontal lines. (b) The corresponding sequential interpretation: The marginal genealogies at the first, second, and third locus 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. (c) The corresponding sequential interpretation where just the deme of absorption, the time of absorption, and the absorbing haplotype are recorded. The gap in the ancestral lineages indicates that the marginal conditional genealogy is integrated out.

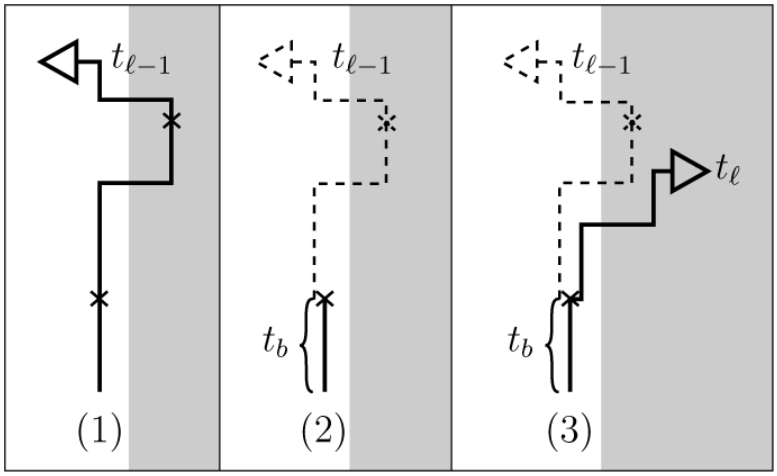

Figure 2.

The transition density from locus $\ell - 1$ to locus ℓ in the model underlying $\widehat{\pi}_{\text{MigSMC}}$ is illustrated. The white and the grey background symbolize the two different demes that the ancestral lineage can reside in. (1) A Poisson number of recombination events is placed uniformly onto the marginal conditional genealogy at locus $\ell - 1$. (2) If the time t_b of the most recent recombination event is more recent than the time of absorption $t_{\ell-1}$, then the marginal conditional genealogy up to this time is copied to locus ℓ. (3) The ancestral lineage at locus ℓ evolves according to migration until it is absorbed at time t_ℓ into the trunk in some deme.

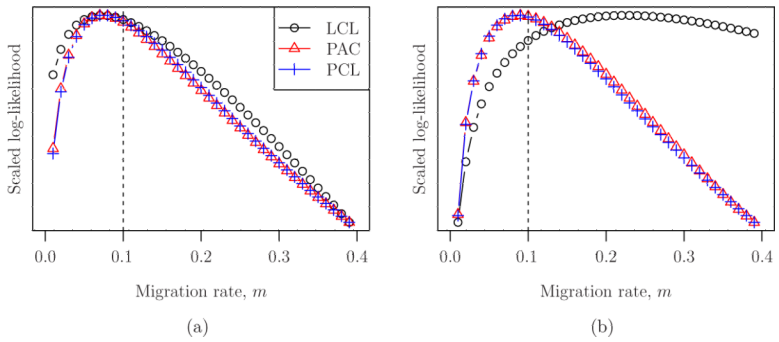

Figure 3.

Re-scaled log likelihood surfaces for two sample configurations (generated for $m = 0.10$, indicated by a vertical line in the plots), and for each of the three approximate likelihood formulations (LCL, PAC, PCL) described in the text. In both cases, the likelihoods are computed using the true values of $\theta = 5 \times 10^{-2}$ and $\rho = 5 \times 10^{-2}$. (a) A case for which all of the likelihood surfaces are similar (b) A case for which the LCL likelihood surface is substantially different than the likelihood surfaces for PAC and PCL

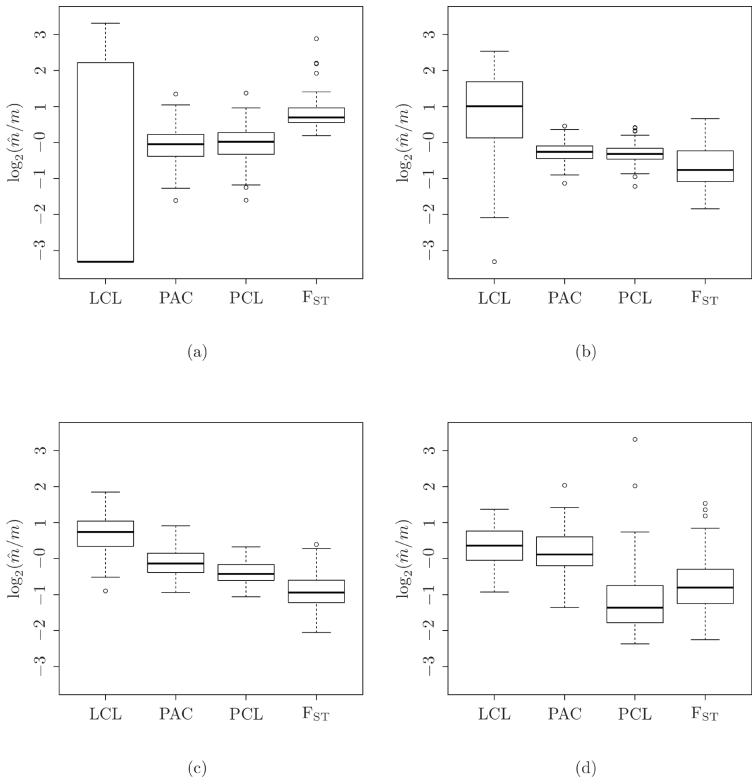

Figure 4.

Box plots (produced using the software package R, and including outliers) for the quantity $\log_2(\widehat{m}/m)$ over 100 samples, where $\widehat{m}$ is the MLE under each of the three approximate likelihood formulations (LCL, PAC, PCL) or the F_{ST}-based estimate as described in the text. The MLE values $\widehat{m}$ were found using golden section search within the interval ($m \cdot 10^{-1}$, $m \cdot 10$) (a) $m = 0.01$ (b) $m = 0.10$ (c) $m = 1.00$ (d) $m = 10.0$. Note that the median of the LCL estimator in (a) lies on the lower bound of the interval, thus at least half of the estimates reach this bound and are most likely even smaller.