# Inference of complex population histories using whole-genome sequences from multiple populations

Matthias Steinrücken[1,2,4], John A. Kamm[2], and Yun S. Song[1,2,3,5]

[1]Computer Science Division, University of California, Berkeley, USA
[2]Department of Statistics, University of California, Berkeley, USA
[3]Department of Integrative Biology, University of California, Berkeley, USA

[4]Current address: Department of Biostatistics and Epidemiology, University of Massachusetts, Amherst, MA, USA
[5]Current address: Department of Mathematics and Department of Biology, University of Pennsylvania, Philadelphia, PA, USA

Correspondence should be addressed to Y.S.S. (yss@berkeley.edu)

## Abstract

There has been much interest in analyzing genome-scale DNA sequence data to infer population histories, but the inference methods developed hitherto are limited in model complexity and computational scalability. Here, we present an efficient, flexible statistical method that can utilize whole-genome sequence data from multiple populations to infer complex demographic models involving population size changes, population splits, admixture, and migration. We demonstrate through an extensive simulation study that our method can accurately and efficiently infer demographic parameters in realistic biological scenarios. The algorithms described here are implemented in a new version of the software package diCal, which is available for download at `https://sourceforge.net/projects/dical2`.

## Introduction

Whole-genome sequences have now become available for population genetic analyses, and inference methods that can take better advantage of genome-scale data have received considerable attention over the past few years. In particular, there has been much interest in methods that can use genomic data of individuals from multiple populations to infer complex models of population histories. In addition to being of historical interest, population demography is important to study because it influences the pattern of genetic variation, and understanding the intricate interplay between demography and other evolutionary forces such as natural selection and genetic drift is a major aim in population genetics.

Inference methods (Nielsen 2000; Gutenkunst et al. 2009; Lukić and Hey 2012; Excoffier et al. 2013; Bhaskar et al. 2015) based on the sample frequency spectrum (SFS) are computationally efficient, but they ignore linkage information in the data and the minimax rate of convergence for such estimators can be quite poor (Terhorst and Song 2015). Also, their utility is limited by the fact that the number of model parameters that is theoretically possible to estimate using the SFS alone is bounded by the sample size (Bhaskar and Song 2014). Methods (Li and Durbin 2011; Mailund et al. 2011; Palamara et al. 2012; Sheehan et al. 2013; Harris and Nielsen 2013; Schiffels and Durbin 2014) that take linkage structure into account are more statistically efficient and they can be used to infer models with many parameters even when the sample size is small. This aspect is of practical importance, since an increasing number of studies now try to infer complex demographic models involving multiple populations by using a small number of individuals sampled from each population (e.g., Raghavan et al. 2015).

Full-likelihood methods based on the coalescent with recombination utilize most of the linkage information in the data. A popular demographic inference method of this kind is PSMC (Li and Durbin 2011), which uses a pair of sequences to infer piecewise-constant population size histories. Its recent extension, MSMC (Schiffels and Durbin 2014), can use sequences sampled from a pair of populations to infer a genetic separation history, in addition to population size changes.

Parallel to this development, a new inference method called diCal (Demographic Inference using Composite Approximate Likelihood) (Sheehan et al. 2013) was introduced to infer piecewise-constant effective population size histories using multiple sequences, thereby providing improved inference about the recent past. The key mathematical component of diCal is the conditional sampling distribution (CSD) $\pi_\Theta$, which describes the conditional probability of observing a new haplotype given a collection of already observed haplotypes, under a given population genetic model with parameters $\Theta$. The corresponding genealogical process, which can be formulated as a hidden Markov model (HMM), is illustrated in Figure 1. In this paper, we extend diCal in several ways, including subdivided population structure with migration (Steinrücken et al. 2013), to develop a scalable inference tool for population genomic analysis under general demographic models. In contrast to MSMC, which does not explicitly model population structure, we consider fully parametric demographic models that are easier to interpret. Specifically, our method is flexible enough to handle:

1. An arbitrary number of populations specified by the user.
2. An arbitrary pattern of population splits and mergers.
3. More general population size changes (e.g., piecewise-exponential).
4. Arbitrary migration patterns with time-varying continuous migration rates or pulse migration events.
5. An arbitrary multi-allelic mutation model at each site (including bi- or quadra-allelic).

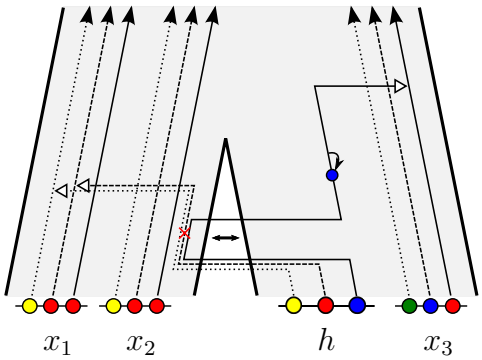



Figure 1: Example realization of the CSD, which approximates the coalescent with recombination. The demography describes an ancestral population that splits into two, with subsequent gene flow. Haplotypes $x_1$ and $x_2$ are observed in the first population and $x_3$ in the second. The lineages corresponding to $x_1, x_2, x_3$ are called "trunk" lineages. The additional haplotype $h$ is sampled in the second population. The dotted, dashed, and solid lines represent the lineages at loci 1, 2, and 3, respectively. A recombination event, indicated by the red cross, separates the lineages at loci 2 and 3. The lineage for locus 3 migrates back to the second population from where $h$ is sampled.

In addition to these features, we introduce major computational improvements which enable the use of whole-genome data. The mathematical details of our method and the computational extensions are provided in Methods and the Supplemental Material. Below we briefly highlight the key technical advances.

In PSMC and the earlier version of diCal, the demographic epochs and HMM discretization intervals are both fixed, and the discretization intervals form a strict refinement of the demographic epochs. In contrast, discretization intervals and demographic epochs are decoupled in our improved version of diCal. For example, a population size change-point or a population split time can freely vary and do not need to coincide with any discretization interval boundary. This flexibility allows for more accurate parameter estimation, especially regarding population split times.

In the original formulation of the sequentially Markov CSD (Paul et al. 2011) (see Figure 1), the trunk lineages are constant in time, extending infinitely far into the past, never coalescing or changing. However, in practice, the trunk lineages may coalesce or migrate between populations. We implement a modified version of the dynamics in the trunk, so that the absorption rate into lineages of present population $\alpha$ in ancient population $\beta$ at time $t$ is proportional to the expected number of ancestors that present population $\alpha$ has in ancient population $\beta$ at time $t$. We obtain an approximation to this number by solving a certain set of differential equations (Jewett and Rosenberg 2014).

The CSDs for different haplotypes can be combined in various ways to devise a composite likelihood that can be used in a maximum likelihood framework for parameter estimation. Our implementation of the expectation-maximization (EM) algorithm allows any composite likelihood that is composed of sums and products of CSDs. The runtime for the EM is linear in the number of haplotypes times the number of CSDs, and quadratic in the number of populations involved. Only the E-step depends linearly on the length of the haplotype, whereas the M-step is independent of this quantity.

For computational speedup, we implement the "locus-skipping" algorithm (Paul and Song 2012) in the likelihood computation, which analytically and exactly integrates over contiguous stretches of non-segregating loci. Further, we extend the method to work for the EM algorithm, leading to dramatic overall speedups. However, locus-skipping does not work well with missing data, and

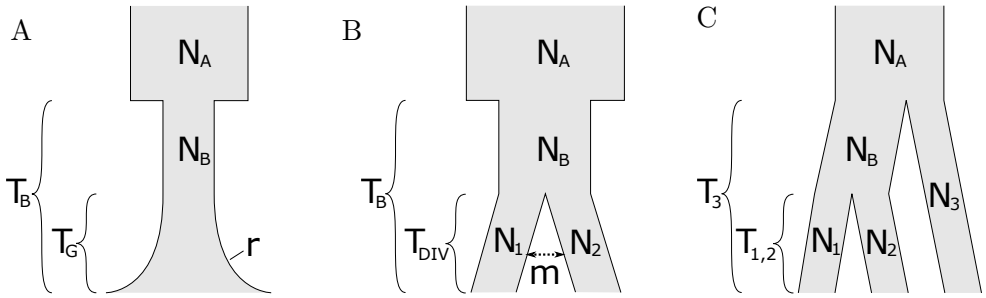



Figure 2: Demographic models used in our simulation study. **A.** Recent exponential population growth: An ancestral population of size $N_A$ undergoes a bottleneck at time $T_{\mathrm{B}}$, where its size is reduced to $N_{\mathrm{B}}$. Growth starts at time $T_{\mathrm{G}}$ at an exponential rate $r$. **B.** Demographic model of a population split: An ancestral population of size $N_{\mathrm{A}}$ undergoes a strong bottleneck that starts at time $T_{\mathrm{B}}$ in the past, and reduces the population size to $N_{\mathrm{B}}$. At time $T_{\mathrm{DIV}}$, this population then splits into two populations of size $N_1$ and $N_2$, respectively. Following the population split, migrants are exchanged at a rate $m$. **C.** Demographic model of three populations with pure splits: An ancestral population of size $N_A$ splits into two populations of size $N_B$ and $N_3$ at time $T_3$. The former then again splits into two populations of size $N_1$ and $N_2$ at time $T_{1,2}$.

missing alleles should be imputed if one wants to take full advantage of the locus-skipping algorithm. To remedy this drawback, we implement an alternative speedup by grouping loci together into larger blocks. A similar speedup was used in PSMC, but it treats the whole block as a single di-allelic site. In contrast, the blocks in our method are viewed as full non-recombining haplotypes.

For complex demographic models, the likelihood function may have local optima. To address this issue, we implement a flexible genetic algorithm and combine it with the EM procedure, thus enabling more efficient navigation of high-dimensional parameter space.

# Results

To demonstrate the flexibility, accuracy, and efficiency of our method, we performed an extensive simulation study, under a variety of biologically relevant demographic scenarios. The haplotype length was set to 250 Mbp in all cases. See Methods for further details on simulation. In each scenario, we used our method to estimate all demographic parameters of the underlying model.

## Recent exponential growth

The first model we investigated is a model of recent exponential population growth. To investigate the performance of our method in this setting, we simulated data consisting of 10 haplotypes under the demographic model depicted in Figure 2A. We fixed $T_{\mathrm{B}} = 65,000$ years ago (ya), $T_{\mathrm{G}} = 15,000$ ya, $N_{\mathrm{A}} = 15,000$, and $N_{\mathrm{B}} = 1,800$, and used three different values for the growth rate $r$, namely $0.25\%, 0.5\%$, or $1.0\%$ per generation. We simulated 100 datasets for each parameter setting.

For each simulated dataset, we used the leave-one-out (LOL) composite likelihood in our EM procedure combined with a genetic algorithm to estimate all 5 parameters of the demographic model. For the genetic algorithm, we chose 50 random starting points that were each optimized for 5 EM iterations. Then we chose the 5 best parameter values ("parents") and replaced each of them with an average of 3 "offspring" parameters sets to obtain the next "generation". These were then again optimized for 5 EM iterations. We repeated this procedure for 4 more "generations",

and reported the parameters that achieved the overall maximal likelihood value.

Shown in Figure 3 are violin plots of the accuracy results. Each violin plot shows the base-2 logarithm of the estimated parameter divided by the truth (the relative error). Thus, a value of 0 corresponds to an exact estimate, whereas +1 is a two-fold over- and −1 is a two-fold underestimate. The analysis of the simulated data shows that in these scenarios, all parameters are estimated with very little variability. However, the results indicate that the estimation of the exponential growth rate is biased upward. This bias is somewhat counterbalanced by a slight downward bias of the time when growth starts, and the population size before growth starts. In fact, the estimates lead to very accurate contemporary population sizes. We note that it is possible to empirically correct for biases in applications and using more sequence data for each individual can reduce the variability. Moreover, we stress that our method is able to get reasonably accurate estimates of the recent exponential growth using only 10 haplotypes. This is far less than the sample size (thousands to tens of thousands) required by SFS-based methods to get reasonable estimates; see Bhaskar et al. (2015) and references therein.

## Population split

We also investigated a model of a past population split, depicted in Figure 2B. This model allows for a bottleneck before the populations split, and possibly subsequent gene flow after the split. We first focused on the case with no gene flow, i.e., with migration probability $m = 0$.

We simulated datasets with two haplotypes in each of the extant populations. We simulated 100 datasets each for $T_{\mathrm{DIV}} = 10,000$ ya and $20,000$ ya, with the remaining parameters set to $T_{\mathrm{B}} = 70,000$ ya, $N_{\mathrm{A}} = 20,000$, $N_{\mathrm{B}} = 1,800$, and $N_1 = N_2 = 5,000$. This scenario has recently been introduced as part of a study of the demographic history of Native Americans (Raghavan et al. 2015). In addition, we simulated 100 datasets with $T_{\mathrm{DIV}} = 70,000$ ya, setting $N_{\mathrm{B}} = N_{\mathrm{A}} = 20,000$, thereby also removing the need for $T_{\mathrm{B}}$. This model is supposed to resemble an out-of-Africa event. For the genetic algorithm, we used 60 random starting points, and 6 "parents" for each of the following 4 "generations" for the cases $T_{\mathrm{DIV}} = 10,000$ ya and $20,000$ ya, and 40 starting points and 5 "parents" for the case $T_{\mathrm{DIV}} = 70,000$ ya. We used the leave-one-out (LOL) composite likelihood.

Figures 4 shows the accuracy results. These empirical results demonstrate that our method is able to estimate the parameters in this clean-split model with high accuracy. Most parameters show little bias and the empirical percentiles are very narrow. Only the estimates of the extant population sizes $N_1$ and $N_2$ for $T_{\mathrm{DIV}} = 10,000$ and $T_{\mathrm{DIV}} = 20,000$ show a somewhat higher variability. Since this time-frame is very recent on an evolutionary timescale, it would require either more sampled haplotypes or more sequence data to get better estimates of these parameters.

## Isolation with migration

We also investigated the demographic model shown in Figure 2B allowing for positive gene flow after the ancestral population splits into two. To this end we set the migration probability to $m = 0.00025$; that is, an individual from population 1 can have a parent from population 2, and vice versa, with a probability of 0.00025 per generation. Using this migration probability, we simulated 100 datasets each consisting of two haplotypes in each extant population, using $T_{\mathrm{DIV}} = 40,000$ ya, $T_{\mathrm{B}} = 70,000$ ya, $N_{\mathrm{A}} = 20,000$, $N_{\mathrm{B}} = 1,800$, and $N_1 = N_2 = 5,000$. We also simulated 100 datasets using $T_{\mathrm{DIV}} = 70,000$ ya, $N_{\mathrm{B}} = N_{\mathrm{A}} = 20,000$, and $N_1 = N_2 = 5,000$. In the former case we used 70 starting points and 6 "parents" for each "generation" in the genetic algorithm, whereas for the later we used 50 and 5, respectively.

Figure 5 shows the accuracy results. In both scenarios, we used the pairwise composite likelihood

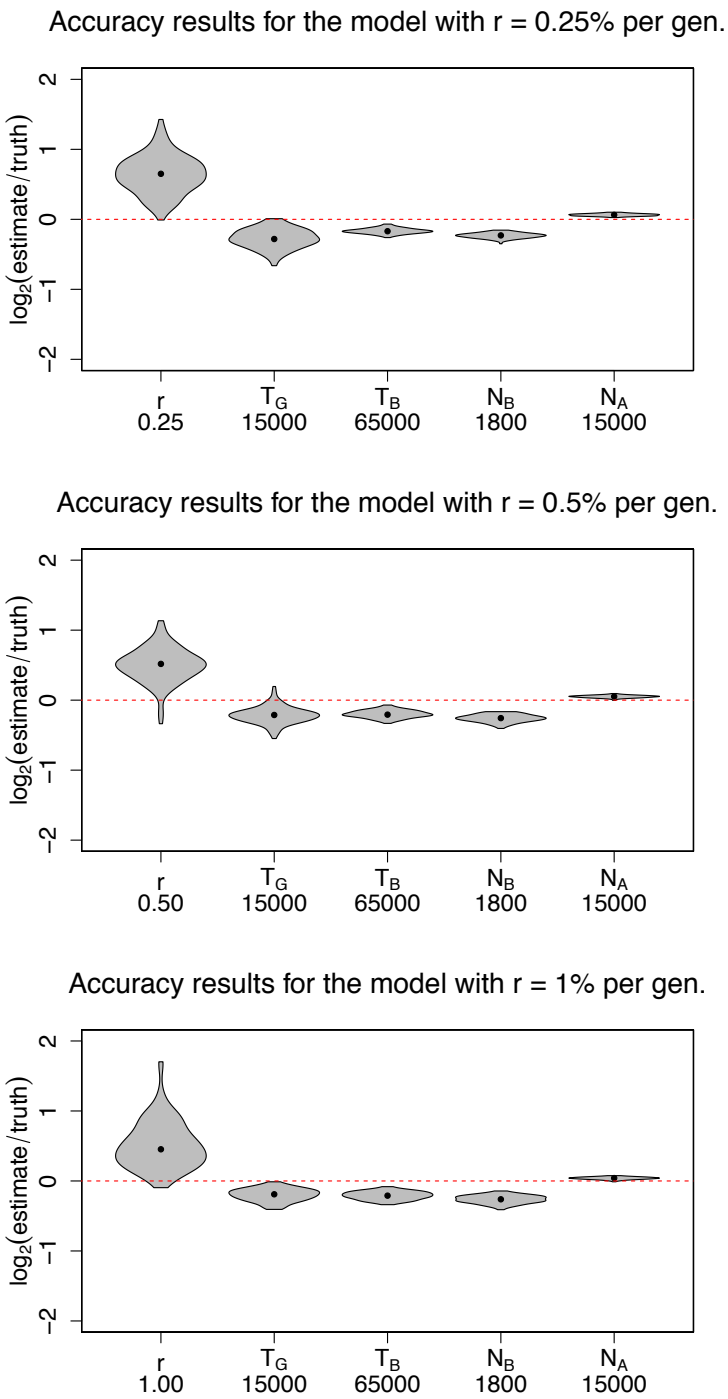



Figure 3: Accuracy results of our method, `diCal2`, for the recent exponential growth model shown in Figure 2A with expansion rate $r = 0.25\%, 0.5\%$ and $1.0\%$ per generation. Parameter estimates were obtained using only 10 haplotypes, which is much less than the sample size (thousands to tens of thousands) required by SFS-based methods to get good estimates. The violin plots show base-2 logarithm of the relative error (estimate/truth) for the analysis of 100 simulated datasets. True parameter values are shown on the $x$-axis.

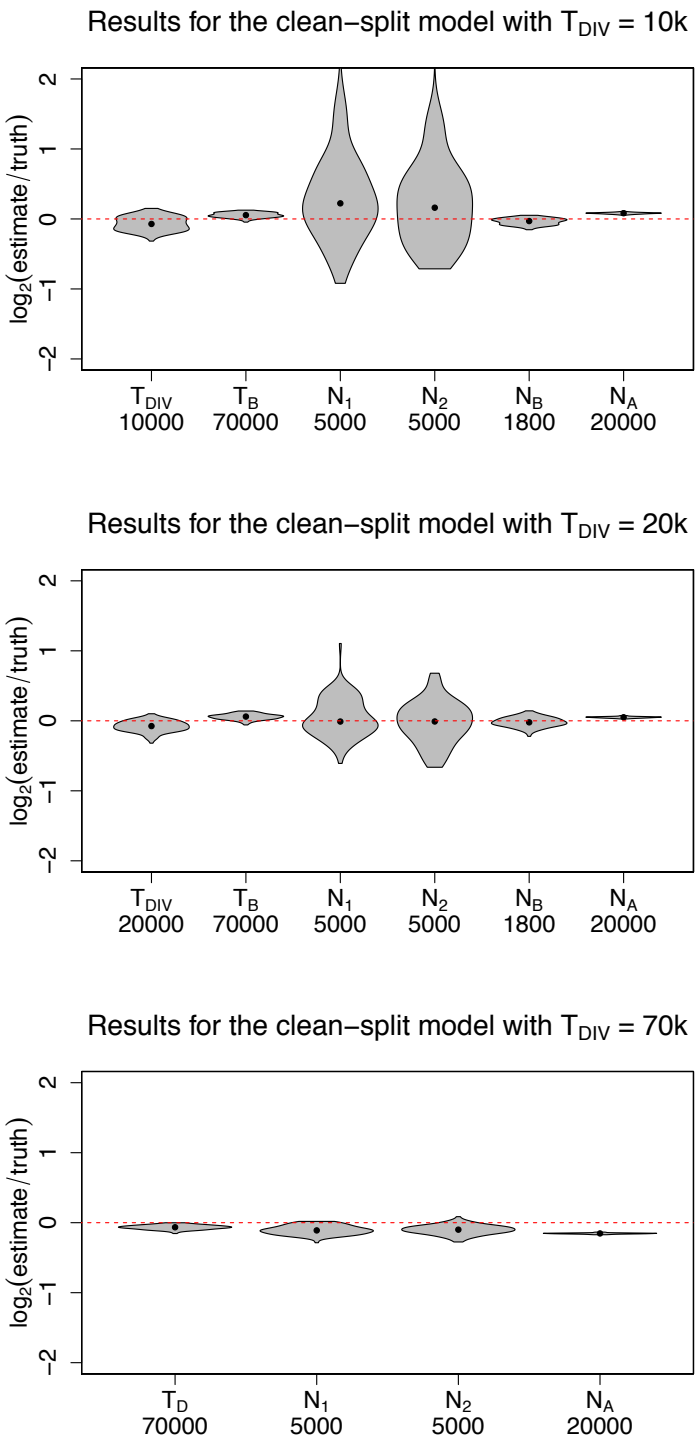



Figure 4: Accuracy results of `diCal2` for the population split model shown in Figure 2B with divergence time $T_{\mathrm{DIV}} = 10, 20$ and 70 kya, and no gene flow ($m = 0$). Using only two haplotypes in each extant population, the parameters of this clean-split model could be estimated very accurately.

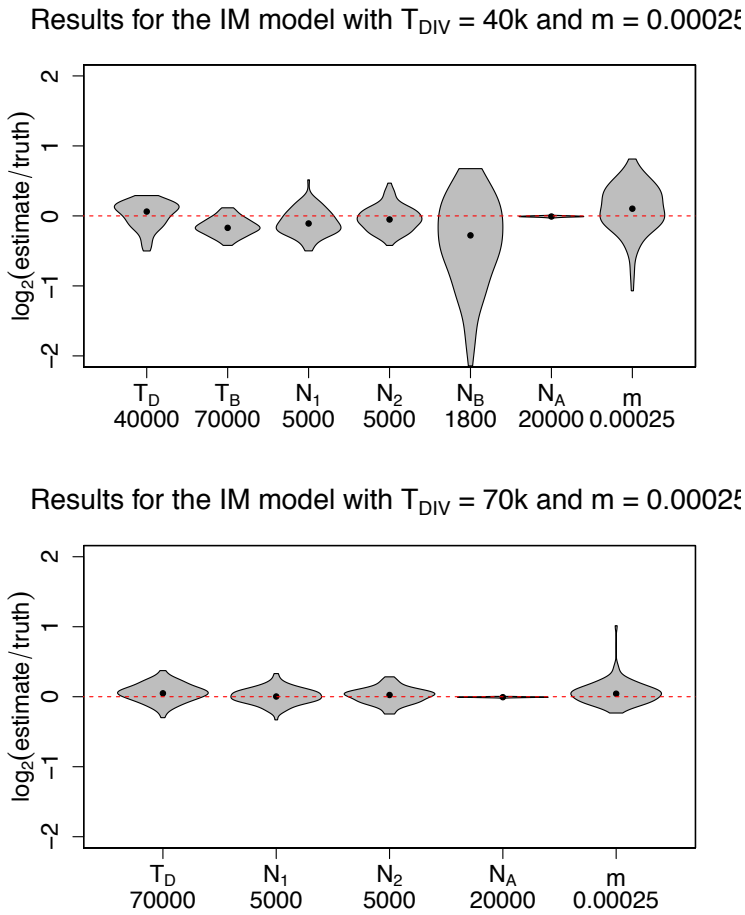



Figure 5: Accuracy results of `diCal2` for the isolation with migration (IM) model shown in Figure 2B with divergence time $T_{\text{DIV}} = 40$ and 70 kya, and migration probability $m = 0.00025$. As in the clean-split case, only two haplotypes in each extant population were used. Most parameter estimates show little bias or variability. See the text for further discussion.

(PCL). Again, most parameter estimates show little bias or variability, the exceptions being $N_B$ and $m$ in the first scenario. However, we note that the evolutionary timescales involved are rather short, and thus the number of events informative about these parameters is low. In practice, the variability could be reduced by using additional chromosomes.

## Three-population model

Lastly, we simulated data under the model depicted in Figure 2C relating three extant populations. Under this model, an ancestral population of size $N_A$ splits into two populations of size $N_B$ and $N_3$ at time $T_3$. The one of size $N_B$ then again splits into two populations of size $N_1$ and $N_2$ at time $T_{1,2}$. We simulated 100 datasets with two haplotypes in each of the extant populations. We set the parameters to $T_{1,2} = 30,000$ ya, $T_3 = 60,000$ ya, $N_A = 20,000$, $N_B = 3,000$, and $N_1 = N_2 = N_3 = 5,000$. For the genetic algorithm, we chose 70 starting points, and 6 "parents", and used the leave-one-out (LOL) composite likelihood. Figure 6 shows the accuracy of our method. Again, the empirical distribution of the estimates show little bias or variability.

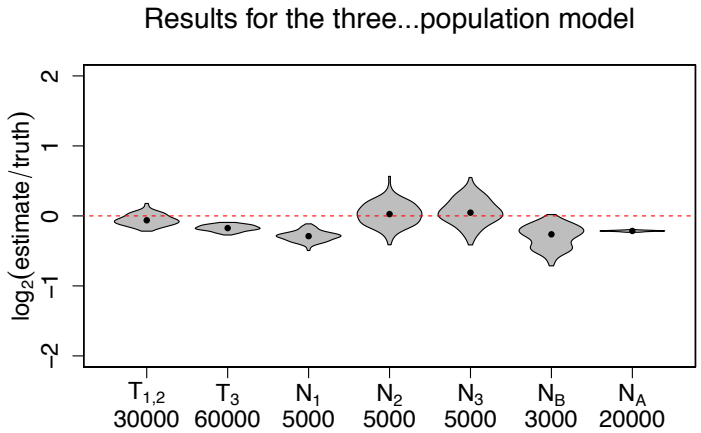



Figure 6: Accuracy results of `diCal2` for the three-population model shown in Figure 2C, using two haplotypes in each extant population.

## Discussion

The results described above demonstrate that our method can efficiently and accurately estimate biologically relevant demographic parameters. It has recently been successfully employed in a study of Native American genomes (Raghavan et al. 2015). The flexible framework underlying the method allows one to consider a wide range of population histories. Aside from demographic inference, we note that our method can be utilized in other population genetic problems of interest, for example model selection. Furthermore, the posterior decoding of latent variables in our CSD can be used in detecting admixture tracts, estimating fine-scale recombination rates in admixed individuals, distinguishing ancestral and introgressed polymorphism, and detecting incomplete lineage sorting. Also, applying our CSD in methods for phasing genotypes, imputing missing sequence data, and detecting identity-by-descent tracts would make it possible to account for demography, thus potentially improving accuracy. Lastly, it is straightforward to incorporate temporal samples (ancient DNA sequences) into our method, which leads to further interesting applications.

## Methods

Here we introduce our demography-aware CSD and describe how it can be employed in a composite likelihood framework to estimate demographic parameters using EM. Further mathematical details are provided in the Supplemental Material.

### Background and notation

We model the sampled sequences or haplotypes under the finite-sites, finite-alleles coalescent with recombination. To this end, denote the set of possible alleles at a specific site or locus by $E$. A haplotype $h$ of length $L$ carries an allele at every locus and is thus an $L$-tuple from the set $\mathcal{H} = E^L$ of possible haplotypes. Denote by $h[l]$ the allele that haplotype $h$ carries at locus $l$, and by $h[l : l']$ the vector $(h[l], \ldots, h[l'])$. At each locus, mutations can occur at a coalescent-scaled per-locus mutation rate of $\theta/2$, where $\theta = 4N_0\mu$, with $N_0$ being the reference effective population size and $\mu$ the per-locus per-generation mutation probability. Denote by $P$ the stochastic mutation

matrix, that is, if a mutation occurs, then allele $a$ mutates into allele $a'$ with probability $P_{a,a'}$, for $a, a' \in E$. A crossover recombination event occurs between each pair of consecutive loci $(l, l+1)$, for $1 \leq l < L$, at coalescent-scaled rate of $\rho/2$, where $\rho = 4N_0 r$ and $r$ denotes the per-generation recombination probability. Note that the recombination and mutation rates could, in principle, vary along the sequence. However, for notational convenience we will assume that the rates are constant.

We assume that the haplotypes are sampled in any of $g$ sub-populations, and we denote the set of possible sub-populations at present by $\Gamma = \{1, \ldots, g\}$. A sample configuration $\mathbf{n}$ can thus be described by a collection of non-negative integers $n_{\gamma,h} \geq 0$, which give the number of haplotypes of type $h \in \mathcal{H}$ sampled in sub-population $\gamma \in \Gamma$. Denote by $n$ the total number of sampled haplotypes, that is $n = \sum_{\gamma \in \Gamma} \sum_{h \in \mathcal{H}} n_{\gamma,h}$. Furthermore, $\mathbf{n}_\gamma$ denotes the configuration consisting of only those haplotypes that were sampled in sub-population $\gamma$, and $n_\gamma = \sum_{h \in \mathcal{H}} n_{\gamma,h}$ denotes the number of such haplotypes. In addition, $x \in \mathbf{n}$ holds if $n_{\gamma,h} > 0$ for any $\gamma$.

We allow for a general demographic model, where the demographic structure and the migration rates can differ at different times in the past. To this end, choose $E+1$ times $0 = t_0 \leq t_1 \leq \cdots \leq t_E = \infty$ to obtain a partition of the positive real line $[0, \infty)$ into $E$ epochs denoted by $I_\epsilon = [t_{\epsilon-1}, t_\epsilon)$. Here $t_0 = 0$ corresponds to the present and $t_E = \infty$ to an infinite time in the past. Denote the set of all epochs by $\mathcal{E} := \{1, \ldots, E\}$. Note that this notation allows for an epoch to have length zero. To allow for changes in the ancient demographic structure, define for each epoch $\epsilon \in \mathcal{E}$ a partition $\Gamma_\epsilon = \{\gamma_\epsilon^{(1)}, \ldots, \gamma_\epsilon^{(g_\epsilon)}\}$ of $\Gamma$. This notation reflects that all present sub-populations whose indices are in the set $\gamma_\epsilon^{(i)}$ derive from the $i$th ancestral population in epoch $\epsilon$, and that there were $g_\epsilon = |\Gamma_\epsilon|$ sub-populations during that epoch. We require that in the first epoch $\Gamma_1 = \{\gamma_1^{(1)}, \ldots, \gamma_1^{(g)}\}$, the equality $\gamma_1^{(i)} = \{i\}$ holds, and that for all $\epsilon \in \mathcal{E} \backslash \{E\}$ the partition $\Gamma_\epsilon$ is a refinement of $\Gamma_{\epsilon+1}$.

The size of sub-population $\gamma \in \Gamma_\epsilon$ is given by $\kappa_\gamma^{(\epsilon)} N_0$, and the coalescent rate is inversely proportional. Furthermore, during an epoch $\epsilon$ of positive length, migration (backwards in time) from sub-population $\gamma \in \Gamma_\epsilon$ into sub-population $\delta \in \Gamma_\epsilon$ occurs at a coalescent-scaled rate of $m_{\gamma,\delta}^{(\epsilon)}/2$. Here $m_{\gamma,\delta}^{(\epsilon)} = 4N_0 v_{\gamma,\delta}^{(\epsilon)}$ and $v_{\gamma,\delta}^{(\epsilon)}$ is the per-generation probability that an individual in sub-population $\gamma$ has a parent from sub-population $\delta$. To handle scenarios of population admixture we introduce a mechanism for instantaneous migration during an epoch $\epsilon$ of length zero, where $t_{\epsilon-1} = t_\epsilon$ and $I_\epsilon = \emptyset$ hold. Instantaneous migration from sub-population $\gamma$ to $\delta$ during such an epoch occurs with probability $y_{\gamma,\delta}^{(\epsilon)}$, the probability that an individual residing in sub-population $\gamma \in \Gamma_\epsilon$ at time $t_{\epsilon-1}$ has an ancestor residing in sub-population $\delta \in \Gamma_\epsilon$ at time $t_\epsilon$. We denote all the parameters necessary to describe a demographic history by $\Theta$, and present an example in Figure 7.

## Demography-aware conditional sampling distribution

The conditional sampling probability (CSP) $\pi_\Theta(h|\alpha, \mathbf{n})$ denotes the probability of observing haplotype $h$ in sub-population $\alpha$, given that the haplotype configuration $\mathbf{n}$ has already been observed and the underlying demography is described by $\Theta$. We will follow along the lines of Steinrücken et al. (2013) and extend their model to the general demographic framework introduced above. To this end we first describe a sequentially Markovian genealogical process (Wiuf and Hein 1999; McVean and Cardin 2005) that approximates the true conditional genealogical process. Subsequent approximations to this process lead to an HMM with finite hidden state space that can be used to efficiently compute approximate CSPs under this model. The CSDs presented in Steinrücken et al. (2013) and Sheehan et al. (2013) can be obtained as special cases of the model presented here.

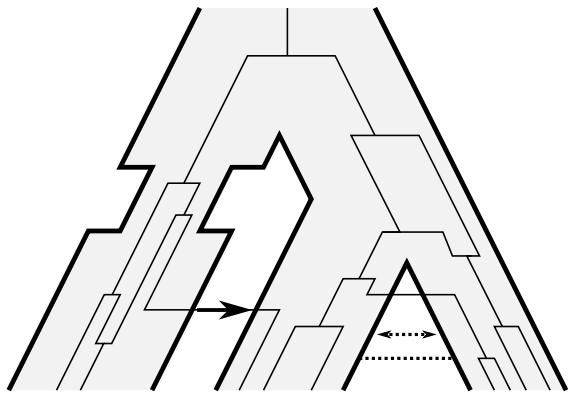


Figure 7: An example of a demographic history and a realization of the coalescent with recombination under this history. In this example there are $E = 8$ epochs. Note that epoch 3 has length zero, thus $t_2 = t_3$. The demographic structure is given by $\Gamma_1 = \Gamma_2 = \Gamma_3 = \Gamma_4 = \{\{1\}, \{2\}, \{3\}\}$, $\Gamma_5 = \Gamma_6 = \Gamma_7 = \{\{1\}, \{2, 3\}\}$, and $\Gamma_8 = \{\{1, 2, 3\}\}$. All migration rates associated with all epochs but 3 are zero, except $m^{(2)}_{2,3} = m^{(2)}_{3,2} = m^{(4)}_{2,3} = m^{(4)}_{3,2} = m$ (and the rates on the diagonal accordingly). The instantaneous migration probabilities associated with epoch 3 are all zero, but $y^{(3)}_{2,1} \geq 0$. The bottleneck indicated in the figure is implemented by setting $\kappa^{(6)}_{\gamma_{6,1}} < \kappa^{(5)}_{\gamma_{5,1}} = \kappa^{(7)}_{\gamma_{7,1}}$.

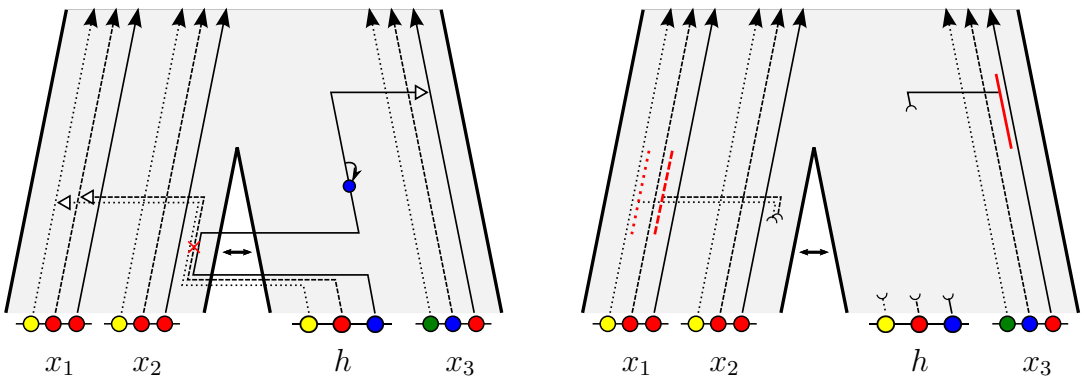
**Trunk approximation.** Computing the true CSP $\pi_\Theta(h|\alpha, \mathbf{n})$ would require to integrate over all possible genealogies relating the haplotypes in the already observed configuration $\mathbf{n}$ and the possible ways of attaching the lineage of the additional haplotype $h$ to these genealogies. To approximate this high-dimensional integration, assume that the unknown genealogy of the configuration $\mathbf{n}$ is given by an unchanging trunk of ancestral lineages for each haplotype extending infinitely into the past. If sub-populations are merged at some point in the past, then the trunk-lineage continues in the merged sub-population. Paul and Song (2010) and Steinrücken et al. (2013) motivated this trunk-approximation using an approach based on the generator of the underlying diffusion process that was developed by De Iorio and Griffiths (2004a,b), and provide an extensive analysis of its accuracy. The trunk-approximation for a given configuration $\mathbf{n}$ is depicted in Figure 8(a).

The following generative process describes the distribution of the ancestral lineage and the allelic composition of the random additional haplotype $H$ under the trunk approximation $\pi^T_\Theta(\cdot|\alpha, \mathbf{n})$. First, a sequence of marginal additional ancestral lineages is sampled that include, at each locus, a history of migration events performed by ancestors of $H$ along the ancestral lineage at this locus, the lineage of the trunk the ancestor coalesces into, and the times of these events. Under assumptions, similar to the Sequentially Markov Coalescent (Wiuf and Hein 1999; McVean and Cardin 2005), these marginal lineages can be generated sequentially starting from the first (left-most) locus in a Markovian fashion. At the first locus, an additional ancestral lineage starts at the present in sub-population $\alpha$ and extends into the past. During an epoch $\epsilon$ of positive length, if the lineage resides in sub-population $\gamma \in \Gamma_\epsilon$, then it is subject to the events:

- *Migration:* The lineage migrates to sub-population $\delta \in \Gamma_\epsilon$ with rate $m^{(\epsilon)}_{\gamma,\delta}$.
- *Absorption:* The lineage is absorbed into a uniformly chosen trunk-lineage in the sub-population it currently resides in at rate $\left(\kappa^{(\epsilon)}_\gamma\right)^{-1}$, the inverse of its size.

During an epoch $\epsilon$ of length zero, the only possible event is

- *Pulse-migration:* The additional lineage migrates to sub-population $\delta \in \Gamma_\epsilon$ with probability $y^{(\epsilon)}_{\gamma,\delta}$.




(a) $\pi^T_\Theta(h|2, \{x_1, x_2, x_3\})$ (b) $\pi^D_\Theta(h|2, \{x_1, x_2, x_3\})$

Figure 8: Two example realizations of the approximations to the true CSD. In these examples, the demography $\Theta$ describes an ancestral population that splits into two, with subsequent gene flow. The already observed configuration consists of haplotypes $x_1$ and $x_2$ in the first sub-population and $x_3$ in the second. The additional haplotype $h$ is sampled in the second sub-population. (a) The CSD $\pi^T_\cdot(\cdot|\cdot)$ approximates the true genealogy relating the observed haplotypes by an unchanging trunk. The dotted, dashed, and solid lines represent the lineages at locus 1, 2, and 3, respectively. At the first locus, the marginal additional lineage undergoes a migration event and is absorbed into the trunk-lineage of $x_1$. A recombination event, indicated by the red cross, separates the lineages at locus 2 and 3. Thus, up to the time of the breakpoint, the additional lineages are the same. At locus 3, it then undergoes migration independently and is absorbed at a different time into the trunk-lineage of $x_3$. The alleles at each locus are then propagated to the present accounting for possible mutations, indicated by the black arrow. (b) Under the approximation $\pi^D_\cdot(\cdot|\cdot)$, the absorbing trunk-lineages at each locus are as before, however, only the intervals (indicated in red) that the absorption times falls into are recorded.

If at the end of an epoch $\epsilon$ the additional lineage resides in a sub-population that merges with other sub-populations into a single ancestral population in epoch $\epsilon + 1$, then it continues in this ancestral population after time $t_\epsilon$. This Markov process backwards in time specifies the initial distribution at the first locus and also describes the marginal distribution of an additional lineage. The migration rate is two-fold higher then in the standard coalescent to balance out the non-migrating trunk.

Under the full coalescent with recombination, the ancestral lineages of two loci that are separated by a recombination distance $\rho$ evolve together into the past for an exponentially distributed amount of time with parameter $\rho/2$ until they are decoupled by a recombination event, and evolve independently beyond this event. Thus, under the approximate CSD $\pi^T_\Theta(\cdot|\alpha, \mathbf{n})$, given the marginal additional genealogy at a certain locus $l-1$, the marginal additional genealogy at locus $l$ is sampled as follows. Denote the time of absorption at locus $l-1$ by $t_{l-1}$. To determine whether and at what time an ancestral recombination event separates locus $l-1$ and locus $l$, a time $t_b$ is sampled from an exponential distribution with parameter $\rho$. If $t_b > t_{l-1}$, then the two loci are not separated by an ancestral recombination event. In this case, the complete marginal additional genealogy at locus $l-1$ is copied to the next locus, including the history of migration events, thus $t_{l-1} = t_l$. If $t_b \leq t_{l-1}$, then a recombination event separates the two loci. In this case, the marginal additional lineage at locus $l-1$ from the present up to the time of the breakpoint $t_b$ is copied to locus $l$, including the sub-population it resides in at that time. The marginal additional lineage at locus $l$ beyond the time of the breakpoint then evolves independently according to the marginal dynamics,

that is, it independently is subject to the migration dynamics until it is ultimately absorbed into a lineage of the trunk. Note that the recombination rate is two-fold higher than in the standard coalescent to again compensate for the lack of events in the trunk.

Once a sequence of marginal additional genealogies is generated, the alleles of the additional haplotype are sampled as follows. At each locus, the allele carried by the haplotype corresponding to the absorbing lineage at the respective locus is propagated along the marginal additional lineage of length $t_l$ from the time of absorption to the present. Mutation events occur at rate $\theta$ and change the current allele according to the stochastic mutation matrix $P$. Note that the rate of evolution is again multiplied by two. This generative process describes the distribution of the additional haplotype under $\pi^T_\Theta(\cdot|\alpha, \mathbf{n})$. A realization can be seen in Figure 8(a).

To compute the approximate conditional sampling probability $\pi^T_\Theta(h|\alpha, \mathbf{n})$, one would have to integrate the probability of observing $h$ given a certain sequence of marginal additional genealogies over all possible such sequences. As this infinite dimensional integration cannot be implemented efficiently, we will introduce another approximation with a discretized hidden state space.

**Discretized hidden state space.** At locus $l$, denote by $T^A_l$ the random absorption time, by $G_l$ the random sub-population where the absorption event takes place, and by $X_l$ the random trunk-lineage that the additional lineage is absorbed into. Since lineages in the trunk do not migrate, the absorbing lineage $X_l$ would be sufficient to determine the sub-population where absorption takes place. However, we keep the sub-population explicit for later convenience. For notational convenience we will also use the partition of the past into demographic epochs as the discretization of the absorption time. However, we describe in Section S6 of the Supplemental Material how this restriction can be relaxed.

The hidden state space then comprises of an epoch of absorption $i \in \mathcal{E}$, a sub-population where absorption takes place $\omega \in \Gamma_i$, and an absorbing trunk-lineage $x \in \mathbf{n}_\omega$. For arbitrary hidden states $s_l = (i_l, \omega_l, x_l)$ and $s_{l-1} = (i_{l-1}, \omega_{l-1}, x_{l-1})$, the initial probabilities

$$\nu(s_l) := \mathbb{P}\big\{T^A_l \in I_{i_l}, G_l = \omega_l, X_l = x_l\big\},$$

the transition probabilities

$$\begin{aligned}\phi(s_l|s_{l-1}) := \mathbb{P}\big\{&T^A_l \in I_{i_l}, G_l = \omega_l, X_l = x_l\big| \\ &T^A_{l-1} \in I_{i_{l-1}}, G_{l-1} = \omega_{l-1}, X_{l-1} = x_{l-1}\big\}\end{aligned}$$

and the emission probabilities

$$\xi(h[l]|s_l) := \mathbb{P}\big\{H[l] = h[l]\big|T^A_l \in I_{i_l}, G_l = \omega_l, X_l = x_l\big\},$$

can be computed using suitable combinations of matrix exponentials describing the evolution of the Markov chain that governs the dynamics of the marginal additional lineage backwards in time. We provide the details of these computations in Section S1.3 of the Supplemental Material. The requisite Markov chain becomes inhomogeneous when considering more general population size models, like exponential growth. The solution then cannot be obtained by matrix exponentials, but we rather have to resort to numerical approximations via step-wise solving of the associated differential equations. Both procedures are implemented in our software package. A realization of the discretized CSD can be seen in Figure 8(b).

These probabilities can then be used in an HMM framework to compute the CSP $\pi^D_\Theta(h|\alpha, \mathbf{n})$.

Defining the *forward variable* $F$ as

$$F_l(s_l) := \mathbb{P}\big\{H[1:l] = h[1:l], T_l^A \in I_{i_l}, G_l = \omega_l, X_l = x_l\big\},$$

the CSP is given by

$$\pi_\Theta^D(h|\alpha, \mathbf{n}) = \sum_{s_L} F_L(s_L).$$

The details of the dynamic programming algorithm used to compute $F$ and ultimately $\pi_\Theta^D$ are given in Section S2 of the Supplemental Material.

## Expectation-Maximization algorithm for demographic inference

The CSD can be used to define a composite likelihood function, which in turn enables us to perform maximum-likelihood inference of the demographic parameters $\Theta$. We can use any such function that is composed of sums and products of CSDs. Here, we describe the product of approximate conditionals (PAC) framework proposed by Li and Stephens (2003). The underlying idea is as follows. For convenience, order the $n$ haplotypes in the configuration $\mathbf{n}$ by enumerating them from 1 to $n$, that is, $x_i$ denotes the $i$th haplotype and $\alpha_i$ denotes the sub-population that it was sampled from, here $1 \leq i \leq n$. Let $\mathbb{P}_\Theta(\mathbf{n})$ denote the probability of sampling the configuration $\mathbf{n}$ under the structured coalescent with recombination, where the demographic model is given by $\Theta$. Then, for the true CSD, the following equality holds by the definition of the CSP:

$$\mathbb{P}_\Theta(\mathbf{n}) = \prod_{i=1}^{n} \pi_\Theta\Big(x_i \Big| \alpha_i, \mathbf{n} - \sum_{j=1}^{i} \mathbf{e}_{\alpha_j, x_j}\Big) \tag{1}$$

Here $\mathbf{e}_{\alpha,x}$ denotes the configuration with just one haplotype $x$ in sub-population $\alpha$. The entire configuration $\mathbf{n}$ can be sampled sequentially by sampling the first haplotype, then sampling the second haplotype given the first, followed by sampling the third given the first two, and so forth until the whole sample $\mathbf{n}$ is obtained. If the true CSP could be computed, this formula could be used to compute the exact sampling probability.

As we cannot obtain the true CSP in general, substituting the approximate CSD $\pi_\Theta^D$ introduced above into this formula yields an approximation of the sampling probability. An interesting feature of equation (1) is that, for the true CSD, the value of the product does not depend on the ordering of the haplotypes. However, if an approximate CSD is substituted into this formula, the value of the product can and will most certainly depend on the order. This fact had already been noticed by Li and Stephens (2003). To mediate the influence of the haplotype-order, they proposed to average the product over several random permutations of the ordering. Thus, the sampling probability can be approximated by

$$\mathbb{P}_\Theta(\mathbf{n}) \approx \frac{1}{K} \sum_{j=1}^{K} \prod_{i=1}^{n} \pi_\Theta^D\Big(x_{\sigma_j(i)} \Big| \alpha_{\sigma_j(i)}, \mathbf{n} - \sum_{j=1}^{i} \mathbf{e}_{\alpha_{\sigma_j(j)}, x_{\sigma_j(j)}}\Big), \tag{2}$$

where $\sigma_j$ is a random permutation of $\{1, \ldots, n\}$, and $K$ is the number of such permutations. Surprisingly, a moderate number of permutations performs well in practice.

We investigated different approaches of using the CSD in the composite likelihood framework. For example, the arithmetic mean in equation (2) can be replaced by a geometric mean. Furthermore, the CSDs can be combined in a leave-one-out fashion or using all pairs of haplotypes, rather than removing the haplotypes sequentially. Some of these composite likelihood schemes have been

successfully applied for demographic inference in the past (Sheehan et al. 2013; Steinrücken et al. 2013). We provide some details of the different composite likelihood schemes in Section S3 of the Supplemental Material.

To find the maximum likelihood estimate, we employ the composite likelihood in the standard expectation-maximization (EM) framework (Dempster et al. 1977), rather than evaluating (2) directly for different values of the model parameters. While in principle all parameters of the model can be inferred, we focus on the demographic parameters $\Theta$. Unfortunately, it is not possible to derive a closed form solution for the maximum in the maximization step in general, but numerical optimization schemes, like the Nelder-Mead simplex algorithm (Nelder and Mead 1965) can be used to efficiently determine the requisite maximum. We provide mathematical details for the implementation of the EM algorithm in Section S3 of the Supplemental Material.

In Section S4 of the Supplemental Material, we provide details on the implementation of the "locus-skipping" algorithm, and the alternative speedup that groups loci into larger blocks. Furthermore, in Section S5 of the Supplemental Material, we provide mathematical details of the modifications to the trunk genealogy to increase accuracy. Finally, we describe in Section S6 of the Supplemental Material how to employ a discretization for the HMM computations that differs from the partition induced by the demography and remains fixed throughout the optimization procedure.

### Simulation of data

To simulate DNA sequence data, we used the software **scrm** (Staab et al. 2015). We simulated 100 datasets for each demographic scenario and set the haplotype length to 250 Mbp for each dataset. We used $1.25 \times 10^{-8}$ per generation for both the per-site mutation rate and the recombination rate between two adjacent nucleotides. A crucial parameter for **scrm** is `-l`, the length of the recombination history to be used during the simulation, which we set to 100 kbp. Generation time was assumed to be 30 years.

## Software availability

The algorithms described in this article has been implemented in the new version of diCal, which is available for download at `https://sourceforge.net/projects/dical2`.

## Acknowledgments

We thank Sara Sheehan, Jeff Spence, and Geno Guerra for helpful discussions and for testing our software. This research is supported in part by an NIH Grant R01-GM094402, and a Packard Fellowship for Science and Engineering.

## References


Bhaskar, A. and Song, Y. S., 2014. Descartes' rule of signs and the identifiability of population demographic models from genomic variation data. *The Annals of Statistics*, **42**(6):2469–2493.

Bhaskar, A., Wang, Y. R., and Song, Y. S., 2015. Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. *Genome Research*, **25**(2):268–279.

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

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

Dempster, A. P., Laird, N. M., and Rubin, D. B., 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B Met.*, **39**(1):1–38.

Excoffier, L., Dupanloup, I., Huerta-Sanchez, E., Sousa, V., and Foll, M., 2013. Robust demographic inference from genomic and SNP data. *PLoS Genetics*, **9**:e1003905.

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and Bustamante, C. D., 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, **5**:e1000695.

Harris, K. and Nielsen, R., 2013. Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genetics*, **9**(6):e1003521.

Jewett, E. M. and Rosenberg, N. A., 2014. Theory and applications of a deterministic approximation to the coalescent model. *Theor. Popul. Biol.*, **93**(0):14–29.

Li, H. and Durbin, R., 2011. Inference of human population history from individual whole-genome sequences. *Nature*, **475**:493–496.

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

Lukić, S. and Hey, J., 2012. Demographic inference using spectral methods on SNP data, with an analysis of the human out-of-africa expansion. *Genetics*, **192**:619–639.

Mailund, T., Dutheil, J. Y., Hobolth, A., Lunter, G., and Schierup, M. H., 2011. Estimating divergence time and ancestral effective population size of Bornean and Sumatran orangutan subspecies using a coalescent hidden Markov model. *PLoS Genetics*, **7**(3):e1001319.

McVean, G. A. and Cardin, N. J., 2005. Approximating the coalescent with recombination. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **360**:1387–93.

Nelder, J. A. and Mead, R., 1965. A simplex method for function minimization. *Comput. J.*, **7**(4):308–313.

Nielsen, R., 2000. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics*, **154**(2):931–942.

Palamara, P. F., Lencz, T., Darvasi, A., and Pe'er, I., 2012. Length distributions of identity by descent reveal fine-scale demographic history. *The American Journal of Human Genetics*, **91**(5):809–822.

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.

Paul, J. S. and Song, Y. S., 2012. Blockwise HMM computation for large-scale population genomic inference. *Bioinformatics*, **28**:2008–2015.

Paul, J. S., Steinrücken, M., and Song, Y. S., 2011. An accurate sequentially markov conditional sampling distribution for the coalescent with recombination. *Genetics*, **187**:1115–1128.

Raghavan, M., Steinrücken, M., Harris, K., Schiffels, S., Rasmussen, S., DeGiorgio, M., Albrechtsen, A., Valdiosera, C., Ávila Arcos, M. C., Malaspinas, A.-S., *et al.*, 2015. Genomic evidence for the pleistocene and recent population history of native americans. *Science*, **349**:aab3884.

Schiffels, S. and Durbin, R., 2014. Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, **46**(8):919–25.

Sheehan, S., Harris, K., and Song, Y. S., 2013. Estimating variable effective population sizes from multiple genomes: A sequentially markov conditional sampling distribution approach. *Genetics*, **194**(3):647–662.

Staab, P. R., Zhu, S., Metzler, D., and Lunter, G., 2015. scrm: efficiently simulating long sequences using the approximated coalescent with recombination. *Bioinformatics*, **31**(10):1680–1682.

Steinrücken, M., Paul, J. S., and Song, Y. S., 2013. A sequentially markov conditional sampling distribution for structured populations with migration and recombination. *Theor. Popul. Biol.*, **87**:51–61.

Terhorst, J. and Song, Y. S., 2015. Fundamental limits on the accuracy of demographic inference based on the sample frequency spectrum. *Proc Natl Acad Sci USA*, **112**(25):7677–7682.

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

# SUPPLEMENTAL MATERIAL

## INFERENCE OF COMPLEX POPULATION HISTORIES USING WHOLE-GENOME SEQUENCES FROM MULTIPLE POPULATIONS

MATTHIAS STEINRÜCKEN, JOHN A. KAMM, AND YUN S. SONG

### S1. Hidden Markov Model formulation of the approximate CSPs

In this section, we first present some additional notation and tools to describe the marginal migration-and-absorption process for the approximate conditional sampling distributions (CSDs). Employing these, we then give a more detailed description of the Hidden Markov Model (HMM) that can be used to approximate the conditional sampling probability (CSP).

S1.1. **Markov chain governing the marginal dynamics.** We now introduce the mathematical notation to formalize the Markov chain that governs, backwards in time, the marginal migration and absorption dynamics in our CSD; see Demography-aware Conditional Sampling Distribution in METHODS. Furthermore, we provide details on how to compute the requisite transition probabilities for this Markov chain.

S1.1.1. *Migration matrices.* The migration rates for a given epoch $\epsilon$ can be subsumed in the migration matrix

$$M_\epsilon := \begin{pmatrix} -m_1^{(\epsilon)} & m_{1,2}^{(\epsilon)} & \cdots & m_{1,g_\epsilon}^{(\epsilon)} \\ m_{2,1}^{(\epsilon)} & -m_2^{(\epsilon)} & \cdots & m_{2,g_\epsilon}^{(\epsilon)} \\ \vdots & \vdots & \ddots & \vdots \\ m_{g_\epsilon,1}^{(\epsilon)} & \cdots & \cdots & -m_{g_\epsilon}^{(\epsilon)} \end{pmatrix}, \tag{S1}$$

where we denoted the elements in $\Gamma_\epsilon$ by $1, \ldots, g_\epsilon$, and $m_\gamma^{(\epsilon)} = \sum_{\delta \neq \gamma} m_{\gamma,\delta}^{(\epsilon)}$ for each $\gamma \in \Gamma_\epsilon$.

Along similar lines, for epochs of length zero with $t_{\epsilon-1} = t_\epsilon$, the matrix

$$Y_\epsilon := \begin{pmatrix} y_{1,1}^{(\epsilon)} & y_{1,2}^{(\epsilon)} & \cdots & y_{1,g_\epsilon}^{(\epsilon)} \\ y_{2,1}^{(\epsilon)} & y_{2,2}^{(\epsilon)} & \cdots & y_{2,g_\epsilon}^{(\epsilon)} \\ \vdots & \vdots & \ddots & \vdots \\ y_{g_\epsilon,1}^{(\epsilon)} & \cdots & \cdots & y_{g_\epsilon,g_\epsilon}^{(\epsilon)} \end{pmatrix} \tag{S2}$$

comprises the instantaneous migration probabilities.

S1.1.2. *Extended migration matrix.* Consider the approximate CSDs introduced in Demography-aware Conditional Sampling Distribution of METHODS. During an epoch $\epsilon$ of positive length, in addition to the migration dynamics described by the migration matrix defined in equation (S1), the marginal additional lineage can be absorbed into a lineage of the trunk in the sub-population it currently resides in.

To model this behavior, the Markov chain describing the dynamics has two states per sub-population. One state for the case when the lineage only resides in the respective sub-population, and one state for the case when the lineage is actually absorbed. The dynamics between the unabsorbed states is governed by the migration rates given in the migration matrix $M_\epsilon$. An absorbed

state can only be reached from the unabsorbed state associated with the same sub-population, since the additional lineage can only be absorbed into a trunk-lineage in the sub-population it currently resides in. While the lineage resides in sub-population $\gamma \in \Gamma_\epsilon$, it gets absorbed with rate $(\kappa_\gamma^{(\epsilon)})^{-1} n_\gamma$, proportional to the inverse population size $(\kappa_\gamma^{(\epsilon)})^{-1}$ and $n_\gamma$, the number of trunk lineages in the sub-population. The latter is given as $n_\gamma = \sum_{\delta \in \gamma} n_\delta$, the sum over the number of haplotypes in all present sub-populations that $\gamma$ is ancestral to. Furthermore, since the Markov chain cannot exit an absorbed state, the rates for leaving absorbed states are zero.

Thus, the Markov chain describing the migration and absorption dynamics for epoch $\epsilon$ backwards in time evolves according to the $2g_\epsilon \times 2g_\epsilon$ rate matrix

$$Z_\epsilon := \begin{pmatrix} M_\epsilon - A_\epsilon & A_\epsilon \\ 0 & 0, \end{pmatrix} \tag{S3}$$

where the matrix

$$A_\epsilon = \mathrm{diag}((\kappa_{\gamma_\epsilon^{(1)}}^{(\epsilon)})^{-1} n_{\gamma_1}, \ldots, (\kappa_{\gamma_\epsilon^{(g)}}^{(\epsilon)})^{-1} n_{\gamma_g}), \tag{S4}$$

for $\gamma_\epsilon^{(i)} \in \Gamma_\epsilon$ and $g = |\Gamma_\epsilon|$, governs the absorption of the additional lineage into the trunk. Further, let $a_{\gamma_\epsilon^{(i)}}$ denote the index in this matrix of the state "being absorbed in $\gamma_\epsilon^{(i)}$."

S1.1.3. *Spectral representation (Eigendecomposition).* For an epoch $\epsilon$ of positive length, the spectral representation of $Z_\epsilon$ is helpful to compute certain integrals and matrix exponentials necessary for calculating the requisite probabilities of the HMM underlying the CSD $\pi_\Theta^D$. Assume that the $2g_\epsilon \times 2g_\epsilon$ matrix $Z_\epsilon$ is diagonalizable, and denote by $\{\lambda_1^{(\epsilon)}, \ldots \lambda_{2g_\epsilon}^{(\epsilon)}\}$ the eigenvalues and by $\{v_1^{(\epsilon)}, \ldots, v_{2g_\epsilon}^{(\epsilon)}\}$ the corresponding eigenvectors. Note that $g_\epsilon$ eigenvalues are zero due to the $g_\epsilon$ absorbing states.

Now define

$$V_\epsilon := \begin{pmatrix} v_1^{(\epsilon)} & \cdots & v_{2g_\epsilon}^{(\epsilon)} \end{pmatrix} \tag{S5}$$

to be the matrix that has the eigenvectors as columns. With this definition we can write

$$Z_\epsilon = V_\epsilon \begin{pmatrix} \lambda_1^{(\epsilon)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_{2g_\epsilon}^{(\epsilon)} \end{pmatrix} V_\epsilon^{-1} = \sum_{k=1}^{2g_\epsilon} \lambda_k^{(\epsilon)} v_k^{(\epsilon)} w_k^{(\epsilon)}, \tag{S6}$$

which in turn yields

$$e^{tZ_\epsilon} = V_\epsilon \begin{pmatrix} e^{t\lambda_1^{(\epsilon)}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & e^{t\lambda_{2g_\epsilon}^{(\epsilon)}} \end{pmatrix} V_\epsilon^{-1} = \sum_{k=1}^{2g_\epsilon} e^{t\lambda_k^{(\epsilon)}} v_k^{(\epsilon)} w_k^{(\epsilon)}, \tag{S7}$$

where $w_k^{(\epsilon)}$ is the $k$-th row of $V_\epsilon^{-1}$. Then

$$\left(e^{tZ_\epsilon}\right)_{\gamma,\delta} = \sum_{k=1}^{2g_\epsilon} e^{t\lambda_k^{(\epsilon)}} (v_k^{(\epsilon)} w_k^{(\epsilon)})_{\gamma,\delta} \tag{S8}$$

holds, and furthermore,

$$\left(Z_\epsilon e^{tZ_\epsilon}\right)_{\gamma,\delta} = \sum_{k=1}^{2g_\epsilon} \lambda_k^{(\epsilon)} e^{t\lambda_k^{(\epsilon)}} (v_k^{(\epsilon)} w_k^{(\epsilon)})_{\gamma,\delta}. \tag{S9}$$

From equations (S8) and (S9) it follows that

$$\frac{d}{dt}\left(e^{tZ_\epsilon}\right)_{\gamma,\delta} = \left(Z_\epsilon e^{tZ_\epsilon}\right)_{\gamma,\delta}. \tag{S10}$$

Note that if $Z_\epsilon$ is not diagonalizable, a similar spectral decomposition could be employed, using generalized eigenvalues and the Jordan normal form. However, for ease of notation, we will only present the computations in the sequel for diagonalizable matrices.

S1.2. **Continuous HMM.** Before we introduce the initial, the transition, and the emission probability for the discrete HMM underlying $\pi^D_\Theta$, we will first introduce the corresponding quantities in a model with a continuous absorption time, to illustrate our approach and introduce some useful concepts.

S1.2.1. *Marginal/Initial density.* The transition density in this model is reversible with respect to the initial density, thus the initial and transitions densities are identical. They can be obtained as follows.

First, define

$$f^\epsilon_{\mu_\epsilon,\gamma_\epsilon} := \begin{cases} \left(e^{(t_\epsilon - t_{\epsilon-1})Z_\epsilon}\right)_{\mu_\epsilon,\zeta_\epsilon}, & \text{if } I_\epsilon \neq \emptyset, \\ \left(Y_\epsilon\right)_{\mu_\epsilon,\zeta_\epsilon}, & \text{if } I_\epsilon = \emptyset, \end{cases} \tag{S11}$$

the probability that a lineage residing in sub-population $\mu_\epsilon \in \Gamma_\epsilon$ at time $t_{\epsilon-1}$ resides in sub-population $\gamma_\epsilon \in \Gamma_\epsilon$ at time $t_\epsilon$. In an epoch of length zero ($I_\epsilon = \emptyset$), this given by the instantaneous migration probabilities, whereas in an epoch of positive length ($I_\epsilon \neq \emptyset$), the matrix exponential of the extended migration matrix accounts for the fact that the lineage is not absorbed during the interval $I_\epsilon$.

The quantity (S11) can be employed to recursively define the probability $p^{(0,\epsilon-1)}_{\alpha,\gamma_\epsilon}$ that the additional lineage resides in sub-population $\alpha \in \Gamma_1$ (where the additional haplotype is sampled) at the beginning of epoch 1 (time $t_0$) and resides in sub-population $\gamma_\epsilon \in \Gamma_\epsilon$ at $t_{\epsilon-1}$, while not having been absorbed by that time. The latter can thus be calculated by dynamic programming using the formulas $p^{0,0}_{\alpha,\gamma_1} = \delta_{\alpha,\gamma_1}$, where $\delta$ is the Kronecker-delta, and

$$p^{(0,\epsilon-1)}_{\alpha,\gamma_\epsilon} = \sum_{\mu_{\epsilon-1}\in\Gamma_{\epsilon-1}} \sum_{\substack{\zeta_{\epsilon-1}\in\Gamma_{\epsilon-1} \\ \zeta_{\epsilon-1}\subset\gamma_\epsilon}} p^{(0,\epsilon-2)}_{\alpha,\mu_{\epsilon-1}} f^{\epsilon-1}_{\mu_{\epsilon-1},\zeta_{\epsilon-1}}. \tag{S12}$$

The sum $\sum_{\substack{\zeta_{\epsilon-1}\in\Gamma_{\epsilon-1} \\ \zeta_{\epsilon-1}\subset\gamma_\epsilon}}$ is necessary, since it sums over all the sub-populations that merge into the sub-population $\gamma_\epsilon$ at time $t_{\epsilon-1}$, and thus their probabilities have to be combined.

Now, for an arbitrary locus $l$ and a time $t_l \in \mathbb{R}_{\geq 0}$, let $e = \epsilon(t_l)$ denote the epoch of positive length such that $t_l \in I_e$. With $\omega_l \in \Gamma_e$, and $x_l \in \mathbf{n}_{\omega_l}$, the marginal density is then given as

$$\begin{aligned} \mathbb{P}\{T^A_l \in dt_l, G_l = \omega_l, X_l = x_l\} &= \frac{1}{n_{\omega_l}} \sum_{\gamma_e\in\Gamma_e} p^{(0,e-1)}_{\alpha,\gamma_e} \left(Z_e e^{(t_l - t_{e-1})Z_e}\right)_{\gamma_e,a_{\omega_l}} \\ &=: \frac{1}{n_{\omega_l}} q^{(0,e)}_{\alpha,a_{\omega_l}}(t_l - t_{e-1}). \end{aligned} \tag{S13}$$

Here $\left(Z_e e^{(t_l - t_{e-1})Z_e}\right)_{\gamma_e,a_{\omega_l}}$ is the density of the event that the additional lineage is absorbed into a trunk-lineage in sub-population $\omega_l$ at time $t_l$. The factor $\frac{1}{n_{\omega_l}}$ appears, since it is absorbed into a specific trunk-lineage in this sub-population.

S1.2.2. *Transition density.* For ease of exposition, we focus on deriving the joint density first, which can then be combined with the marginal density to obtain the transition density. As detailed in METHODS, the additional lineages at two loci $l-1$ and $l$ can either be separated by a recombination event or not. The time of the recombination $T^B$ event is given by an exponential random variable with rate $\rho$.

Thus, with $t_{l-1}, t_l \in \mathbb{R}_{\geq 0}$, $\omega_{l-1} \in \Gamma_{\epsilon(t_{l-1})}$, $x_l \in \mathbf{n}_{\omega_{l-1}}$, $\omega_l \in \Gamma_{\epsilon(t_l)}$, and $x_l \in \mathbf{n}_{\omega_l}$, partitioning with respect to the time of the recombination event yields

$$\begin{aligned}
&\mathbb{P}\{T^A_{l-1} \in dt_{l-1}, G_{l-1} = \omega_{l-1}, X_{l-1} = x_{l-1}, T^A_l \in dt_l, G_l = \omega_l, X_l = x_l\} \\
&\quad = \int_{t_b=0}^{\infty} \mathbb{P}\{T^A_{l-1} \in dt_{l-1}, G_{l-1} = \omega_{l-1}, X_{l-1} = x_{l-1}, T^A_l \in dt_l, G_l = \omega_l, X_l = x_l, T^B \in dt_b\} \\
&\quad = \int_{t_b=t_{l-1}\wedge t_l}^{\infty} \mathbb{P}\{T^A_{l-1} \in dt_{l-1}, G_{l-1} = \omega_{l-1}, X_{l-1} = x_{l-1}, T^A_l \in dt_l, G_l = \omega_l, X_l = x_l, T^B \in dt_b\} \\
&\qquad + \int_{t_b=0}^{t_{l-1}\wedge t_l} \mathbb{P}\{T^A_{l-1} \in dt_{l-1}, G_{l-1} = \omega_{l-1}, X_{l-1} = x_{l-1}, T^A_l \in dt_l, G_l = \omega_l, X_l = x_l, T^B \in dt_b\}
\end{aligned} \tag{S14}$$

for the joint distribution of the hidden states at locus $l-1$ and $l$.

The first term in equation (S14) represents the case when the lineages at both loci are absorbed together before the recombination event can decouple them. It is given by

$$\begin{aligned}
&\int_{t_b=t_{l-1}\wedge t_l}^{\infty} \mathbb{P}\{T^A_{l-1} \in dt_{l-1}, G_{l-1} = \omega_{l-1}, X_{l-1} = x_{l-1}, T^A_l \in dt_l, G_l = \omega_l, X_l = x_l, T^B \in dt_b\} \\
&\qquad = \delta_{t_{l-1},t_l}\delta_{\omega_{l-1},\omega_l}\delta_{x_{l-1},x_l}\frac{1}{n_{\omega_{l-1}}} q^{(0,e)}_{\alpha,a_{\omega_l}}(t_{l-1} - t_{e-1})e^{-\rho t_{l-1}},
\end{aligned} \tag{S15}$$

with $e = \epsilon(t_{l-1} \wedge t_l)$. The second term in (S14) represents the case when recombination decouples the lineages at the two loci, and they are both absorbed independently. It yields

$$\begin{aligned}
&\int_{t_b=0}^{t_{l-1}\wedge t_l} \mathbb{P}\{T^A_{l-1} \in dt_{l-1}, G_{l-1} = \omega_{l-1}, X_{l-1} = x_{l-1}, T^A_l \in dt_l, G_l = \omega_l, X_l = x_l, T^B \in dt_b\} \\
&\qquad = \frac{1}{n_{\omega_{l-1}}}\frac{1}{n_{\omega_l}} \\
&\qquad\quad \times \Bigg( \sum_{\epsilon=1}^{e-1} \int_{t_b=t_{\epsilon-1}}^{t_\epsilon} \sum_{\eta\in\Gamma_\epsilon} \mathbb{P}\{T^A_{l-1} \in dt_{l-1}, G_{l-1} = \omega_{l-1}, T^A_l \in dt_l, G_l = \omega_l, T^B \in dt_b, G^B_{t_b} = \eta\} \\
&\qquad\qquad + \int_{t_b=t_{e-1}}^{t_{l-1}\wedge t_l} \sum_{\eta\in\Gamma_\epsilon} \mathbb{P}\{T^A_{l-1} \in dt_{l-1}, G_{l-1} = \omega_{l-1}, T^A_l \in dt_l, G_l = \omega_l, T^B \in dt_b, G^B_{t_b} = \eta\} \Bigg),
\end{aligned} \tag{S16}$$

which is partitioned with respect to the epoch $\epsilon$ during which the recombination event occurred and the random sub-population $\{G^B_{t_b} = \eta\}$ that the coupled additional lineages were residing in at the time of the event. Note that in this partitioning, only the epochs of positive length have to be considered, since the probability of recombination in an epoch of length zero is zero.

For a given epoch $\epsilon$ of positive length, partitioning with respect to the possible sub-populations at the beginning and the end of epoch $\epsilon$, the inner term of the first summand in (S16) yields

$$
\begin{aligned}
&\int_{t_b=t_{\epsilon-1}}^{t_\epsilon} \sum_{\eta\in\Gamma_\epsilon} \mathbb{P}\{T_{l-1}^A \in dt_{l-1}, G_{l-1}=\omega_{l-1}, T_l^A \in dt_l, G_l=\omega_l, T^B \in dt_b, G_{t_b}^B=\eta\} \\
&= \int_{t_b=t_{\epsilon-1}}^{t_\epsilon} \mathbb{P}\{T^B \in dt_b\} \sum_{\eta\in\Gamma_\epsilon} \mathbb{P}\{T_{l-1}^A \in dt_{l-1}, G_{l-1}=\omega_{l-1}, T_l^A \in dt_l, G_l=\omega_l, G_{t_b}^B=\eta | T^B \in dt_b\} \\
&= \int_{t_b=t_{\epsilon-1}}^{t_\epsilon} \rho e^{-\rho t_b} \sum_{\eta\in\Gamma_\epsilon} \sum_{\gamma_\epsilon\in\Gamma_\epsilon} p_{\alpha,\gamma_\epsilon}^{(0,\epsilon-1)} \left(e^{(t_b-t_{\epsilon-1})Z_\epsilon}\right)_{\gamma_\epsilon,\eta} \\
&\qquad \times \sum_{Z_{\epsilon+1}\in\Gamma_{\epsilon+1}} \sum_{\substack{\zeta_\epsilon\in\Gamma_\epsilon \\ \zeta_\epsilon\subset Z_{\epsilon+1}}} \left(e^{(t_\epsilon-t_b)Z_\epsilon}\right)_{\eta,\zeta_\epsilon} q_{Z_{\epsilon+1},a_{\omega_{l-1}}}^{(\epsilon,\epsilon(t_{l-1}))}(t_{l-1}-t_{\epsilon(t_{l-1})-1}) \\
&\qquad \times \sum_{\Xi_{\epsilon+1}\in\Gamma_{\epsilon+1}} \sum_{\substack{\xi_\epsilon\in\Gamma_\epsilon \\ \xi_\epsilon\subset \Xi_{\epsilon+1}}} \left(e^{(t_\epsilon-t_b)Z_\epsilon}\right)_{\eta,\xi_\epsilon} q_{\Xi_{\epsilon+1},a_{\omega_l}}^{(\epsilon,\epsilon(t_l))}(t_l-t_{\epsilon(t_l)-1})\, dt_b \\
&= \sum_{\gamma_\epsilon\in\Gamma_\epsilon} \sum_{Z_{\epsilon+1}\in\Gamma_{\epsilon+1}} \sum_{\substack{\zeta_\epsilon\in\Gamma_\epsilon \\ \zeta_\epsilon\subset Z_{\epsilon+1}}} \sum_{\Xi_{\epsilon+1}\in\Gamma_{\epsilon+1}} \sum_{\substack{\xi_\epsilon\in\Gamma_\epsilon \\ \xi_\epsilon\subset \Xi_{\epsilon+1}}} p_{\alpha,\gamma_\epsilon}^{(0,\epsilon-1)} \\
&\qquad \times R_{\gamma_\epsilon,(\zeta_\epsilon,\xi_\epsilon)}^{(\epsilon)} q_{Z_{\epsilon+1},a_{\omega_{l-1}}}^{(\epsilon,\epsilon(t_{l-1}))}(t_{l-1}-t_{\epsilon(t_{l-1})-1}) q_{\Xi_{\epsilon+1},a_{\omega_l}}^{(\epsilon,\epsilon(t_l))}(t_l-t_{\epsilon(t_l)-1}),
\end{aligned}
\tag{S17}
$$

where

$$
\begin{aligned}
R_{\gamma_\epsilon,(\zeta_\epsilon,\xi_\epsilon)}^{(\epsilon)} &:= \sum_{\eta\in\Gamma_\epsilon} \int_{t_b=t_{\epsilon-1}}^{t_\epsilon} \rho e^{-\rho t_b} \left(e^{(t_b-t_{\epsilon-1})Z_\epsilon}\right)_{\gamma_\epsilon,\eta} \left(e^{(t_\epsilon-t_b)Z_\epsilon}\right)_{\eta,\zeta_\epsilon} \left(e^{(t_\epsilon-t_b)Z_\epsilon}\right)_{\eta,\xi_\epsilon} dt_b \\
&= \rho \sum_{\eta\in\Gamma_\epsilon} \sum_{k=1}^{2g_\epsilon} \sum_{m=1}^{2g_\epsilon} \sum_{n=1}^{2g_\epsilon} (v_k^{(\epsilon)} w_k^{(\epsilon)})_{\gamma_\epsilon,\eta} (v_m^{(\epsilon)} w_m^{(\epsilon)})_{\eta,\zeta_\epsilon} (v_n^{(\epsilon)} w_n^{(\epsilon)})_{\eta,\xi_\epsilon} \\
&\qquad \times \int_{t_b=t_{\epsilon-1}}^{t_\epsilon} e^{-\rho t_b} e^{\lambda_k^{(\epsilon)}(t_b-t_{\epsilon-1})} e^{\lambda_m^{(\epsilon)}(t_\epsilon-t_b)} e^{\lambda_n^{(\epsilon)}(t_\epsilon-t_b)}\, dt_b \\
&= \rho \sum_{\eta\in\Gamma_\epsilon} \sum_{k=1}^{2g_\epsilon} \sum_{m=1}^{2g_\epsilon} \sum_{n=1}^{2g_\epsilon} (v_k^{(\epsilon)} w_k^{(\epsilon)})_{\gamma_\epsilon,\eta} (v_m^{(\epsilon)} w_m^{(\epsilon)})_{\eta,\zeta_\epsilon} (v_n^{(\epsilon)} w_n^{(\epsilon)})_{\eta,\xi_\epsilon} \\
&\qquad \times H_{t_{\epsilon-1}}^{t_\epsilon}((\lambda_m^{(\epsilon)}+\lambda_n^{(\epsilon)})t_\epsilon - \lambda_k^{(\epsilon)} t_{\epsilon-1}, \lambda_k^{(\epsilon)} - \lambda_m^{(\epsilon)} - \lambda_n^{(\epsilon)} - \rho),
\end{aligned}
\tag{S18}
$$

using the spectral decomposition to simplify the matrix exponentials. Note that this quantity is independent of $t_{l-1}$ and $t_l$. The definition

$$
H_a^b(u,\lambda) = \int_{t=a}^{b} e^{\lambda t+u}\, dt = \begin{cases} \frac{1}{\lambda}(e^{\lambda b+u} - e^{\lambda a+u}), & \text{if } \Re(u) \neq \pm\infty,\ b \neq \infty,\ \lambda \in \mathbb{C}\backslash\{0\}, \\ e^u(b-a), & \text{if } \Re(u) \neq \pm\infty,\ b \neq \infty,\ \lambda = 0, \\ -\frac{1}{\lambda} e^{\lambda a+u}, & \text{if } \Re(u) \neq \pm\infty,\ b = \infty,\ \Re(\lambda) < 0, \\ 0, & \text{if } \Re(u) = -\infty, \end{cases}
\tag{S19}
$$

is used for the integral term in equation (S18), with $b > a \geq 0$. This definition covers all the relevant cases, since $\Re(\lambda_n^{(\epsilon)}) \leq 0$ holds for all $\epsilon$ and $n$, and $\Re(\lambda_n^{(\epsilon)}) = 0$ implies $\lambda_n^{(\epsilon)} = 0$ for $Z_\epsilon$ considered here. Also, whenever $\Re(u) \neq \pm\infty$ and $b = \infty$, then $\Re(\lambda) < 0$, if $\rho > 0$. In addition, for

$a < \infty$ and $u \neq \infty$, the definition

$$0 \cdot H_a^{\infty}(u,0) = 0 \cdot \lim_{b\to\infty} H_a^b(u,0) = \lim_{b\to\infty} \big(0 \cdot H_a^b(u,0)\big) = \lim_{b\to\infty} 0 = 0 \tag{S20}$$

has to be used in the appropriate cases.

Focusing on the second summand in (S16), assume without loss of generality that $t_{l-1} < t_l$. For the case $\epsilon(t_{l-1}) \neq \epsilon(t_l)$,

$$\begin{aligned}
&\int_{t_b=t_{e-1}}^{t_{l-1}} \sum_{\eta\in\Gamma_e} \mathbb{P}\{T_{l-1}^A \in dt_{l-1}, G_{l-1} = \omega_{l-1}, T_l^A \in dt_l, G_l = \omega_l, T^B \in dt_b, G_{t_b}^B = \eta\} \\
&\quad = \int_{t_b=t_{e-1}}^{t_{l-1}} \mathbb{P}\{T^B \in dt_b\} \sum_{\eta\in\Gamma_e} \mathbb{P}\{T_{l-1}^A \in dt_{l-1}, G_{l-1} = \omega_{l-1}, T_l^A \in dt_l, G_l = \omega_l, G_{t_b}^B = \eta | T^B \in dt_b\} \\
&\quad = \int_{t_b=t_{e-1}}^{t_{l-1}} \rho e^{-\rho t_b} \sum_{\eta_e\in\Gamma_e} \sum_{\gamma\in\Gamma_e} p_{\alpha,\gamma_e}^{(0,e-1)} \big(e^{(t_b-t_{e-1})Z_e}\big)_{\gamma_e,\eta} \big(Z_e e^{(t_{l-1}-t_b)Z_e}\big)_{\eta,a_{\omega_{l-1}}} \\
&\qquad\qquad \times \sum_{\Xi_{e+1}\in\Gamma_{e+1}} \sum_{\substack{\xi_e\in\Gamma_e \\ \xi_e\subset\Xi_{e+1}}} \big(e^{(t_e-t_b)Z_e}\big)_{\eta,\xi_e} q_{\Xi_{e+1},a_{\omega_l}}^{(e,\epsilon(t_l))} (t_l - t_{\epsilon(t_l)-1})\, dt_b,
\end{aligned} \tag{S21}$$

holds, whereas the summand yields

$$\begin{aligned}
&\int_{t_b=t_{e-1}}^{t_{l-1}\wedge t_l} \rho e^{-\rho t_b} \sum_{\eta\in\Gamma_e} \sum_{\gamma_e\in\Gamma_e} p_{\alpha,\gamma_e}^{(0,e-1)} \big(e^{(t_b-t_{e-1})Z_e}\big)_{\gamma_e,\eta} \\
&\qquad\qquad \times \big(Z_e e^{(t_{l-1}-t_b)Z_e}\big)_{\eta,a_{\omega_{l-1}}} \big(Z_e e^{(t_l-t_b)Z_e}\big)_{\eta,a_{\omega_l}}\, dt_b
\end{aligned} \tag{S22}$$

for the case $\epsilon(t_{l-1}) = \epsilon(t_l)$. It is possible to obtain more explicit expressions for the integrals in equation (S21) and (S22), and to compute them numerically, using the spectral decompositions introduced in Section S1.1.3. However, we will not provide these computations here, but rather provide the full details for the discretized HMM underlying $\pi_{\Theta}^D$ in the next section. Finally, note that we provided the details for computing the joint density for the absorbing lineages at locus $l-1$ and $l$ here, but dividing this density by the marginal density at locus $l-1$ yields the requisite transition density for the HMM.

S1.2.3. *Emission.* Conditional on the absorption time $t_l$, the number of mutation events is Poisson distributed with parameter $\theta$, the mutation rate. Thus, the emission probability is given as

$$\begin{aligned}
&\mathbb{P}\{H[l] = a | T_l^A \in dt_l, G_l = \omega_l, X_l = x_l\} \\
&\quad = \big(e^{t_l\theta(P-\mathbb{1})}\big)_{x_l[l],a},
\end{aligned} \tag{S23}$$

where $H$ denotes the additionally sampled haplotype, $a \in E$, $x_l[l]$ is the allele that the absorbing lineage bears at locus $l$, and $P$ is the mutation matrix that governs the transitions between the alleles.

S1.3. **Discretized HMM.** In this section, we will provide details about the initial, transition and emission probabilities for the discrete HMM underlying the CSD $\pi_{\Theta}^D$, introduced in Demography-aware Conditional Sampling Distribution of METHODS. The basic idea is to integrate the respective densities introduced in the previous section over the discretization intervals. As mentioned in METHODS, we assume here that the time discretization intervals match with the partitioning into epochs induced by the demography. We detail in Section S6 how this assumption can be relaxed. Furthermore, note that the epochs of length zero do not yield valid hidden states of the discretized HMM.

S1.3.1. *Marginal/Initial probability.* The probability that the additional lineage residing in sub-population $\Xi$ at time $t_\epsilon$ is absorbed into any lineage of the trunk within the sub-population $\omega$ during the interval $I_i$ is given by

$$\begin{aligned} Q^{(\epsilon)}_{\Xi,a_\omega}(i) &:= \int_{t=t_{i-1}}^{t_i} q^{(\epsilon,i)}_{\Xi,a_\omega}(t-t_{i-1})\,dt \\ &= \int_{t=t_{i-1}}^{t_i} \sum_{\gamma_i\in\Gamma_i} p^{(\epsilon,i-1)}_{\Xi,\gamma_i}\big(Z_i e^{(t-t_{i-1})Z_i}\big)_{\gamma_i,a_\omega}\,dt \\ &= \sum_{\gamma_i\in\Gamma_i} p^{(\epsilon,i-1)}_{\Xi,\gamma_i} \sum_{k=1}^{2g_i} (v^{(i)}_k w^{(i)}_k)_{\gamma_i,a_\omega} \lambda^{(i)}_k e^{-\lambda^{(i)}_k t_{i-1}} \int_{t=t_{i-1}}^{t_i} e^{\lambda^{(i)}_k t}\,dt \\ &= \sum_{\gamma_i\in\Gamma_i} p^{(\epsilon,i-1)}_{\Xi,\gamma_i} \sum_{k=1}^{2g_i} (v^{(i)}_k w^{(i)}_k)_{\gamma,a_\omega} \lambda^{(i)}_k e^{-\lambda^{(i)}_k t_{i-1}} H^{t_i}_{t_{i-1}}(0,\lambda^{(i)}_k) \end{aligned} \tag{S24}$$

where we used $q$ as defined in equation (S13), and $H$ as in equations (S19) and (S20). The discretized initial probability of being absorbed during the interval $I_i$ in sub-population $\omega_l \in \Gamma_i$ into the lineage $x_l \in \mathbf{n}_{\omega_l}$ is then given as

$$\begin{aligned} \nu(\omega_l, i, x_l) &:= \mathbb{P}\{T^A_l \in I_i, G_l = \omega_l, X_l = x_l\} \\ &= \frac{1}{n_{\omega_l}} u(\omega_l, i) \end{aligned} \tag{S25}$$

with

$$u(\omega, i) := \int_{t=t_{i-1}}^{t_i} q^{(0,i)}_{\alpha,a_\omega}(t-t_{i-1})\,dt = Q^{(0)}_{\alpha,a_\omega}(i). \tag{S26}$$

S1.3.2. *Transition probability.* To derive the transition probability, for given $i, j \in \mathcal{E}$, we again start by focusing on the joint probability that the additional lineage at locus $l-1$ is absorbed during the interval $I_i$ into the trunk-lineage $x_{l-1} \in \mathbf{n}_{\omega_{l-1}}$ residing in sub-population $\omega_{l-1} \in \Gamma_i$, and the lineage at locus $l$ is absorbed during $I_j$ into $x_l \in \mathbf{n}_{\omega_l}$, with $\omega_l \in \Gamma_j$. This probability is given by

$$\begin{aligned} &\mathbb{P}\{T^A_{l-1} \in I_i, G_{l-1} = \omega_{l-1}, X_{l-1} = x_{l-1}, T^A_l \in I_j, G_l = \omega_l, X_l = x_l\} \\ &\quad = \int_{t_{l-1}=t_{i-1}}^{t_i} \int_{t_l=t_{j-1}}^{t_j} \mathbb{P}\{T^A_{l-1} \in dt_{l-1}, G_{l-1} = \omega_{l-1}, X_{l-1} = x_{l-1}, T^A_l \in dt_l, G_l = \omega_l, X_l = x_l\}. \end{aligned} \tag{S27}$$

Substituting equation (S14) into (S27), the *coupled* term isolated in equation (S15) yields

$$
\begin{aligned}
&\int_{t_{l-1}=t_{i-1}}^{t_i}\int_{t_l=t_{j-1}}^{t_j}\delta(t_{l-1}-t_l)\delta_{\omega_{l-1},\omega_l}\delta_{x_{l-1},x_l}e^{-\rho t_{l-1}}q^{(0,\mu)}_{\alpha,a_{\omega_{l-1}}}(t_{l-1}-t_{\mu-1})\frac{1}{n_{\omega_{l-1}}}\,dt_{l-1}\,dt_l\\
&\quad=\int_{t_{l-1}=t_{i-1}}^{t_i}\mathbb{1}_{\{t_{l-1}\in I_j\}}\delta_{\omega_{l-1},\omega_l}\delta_{x_{l-1},x_l}e^{-\rho t_{l-1}}q^{(0,\mu)}_{\alpha,a_{\omega_l}}(t_{l-1}-t_{\mu-1})\frac{1}{n_{\omega_{l-1}}}\,dt_{l-1}\\
&\quad=\delta_{i,j}\delta_{\omega_{l-1},\omega_l}\delta_{x_{l-1},x_l}\frac{1}{n_{\omega_{l-1}}}\int_{t_{l-1}=t_{i-1}}^{t_i}e^{-\rho t_{l-1}}\sum_{\gamma_\mu\in\Gamma_\mu}p^{(0,\mu-1)}_{\alpha,\gamma_\mu}\big(Z_\mu e^{(t_{l-1}-t_{\mu-1})Z_\mu}\big)_{\gamma_\mu,a_{\omega_{l-1}}}\,dt_{l-1}\\
&\quad=\delta_{i,j}\delta_{\omega_{l-1},\omega_l}\delta_{x_{l-1},x_l}\frac{1}{n_{\omega_{l-1}}}\sum_{\gamma_\mu\in\Gamma_\mu}p^{(0,\mu-1)}_{\alpha,\gamma_\mu}\\
&\qquad\qquad\times\sum_{k=1}^{2g_\mu}(v^{(\mu)}_k w^{(\mu)}_k)_{\gamma_\mu,a_{\omega_{l-1}}}\lambda^{(\mu)}_k e^{-\lambda^{(\mu)}_k t_{\mu-1}}\int_{t_{l-1}=t_{i-1}}^{t_i}e^{-\rho t_{l-1}}e^{\lambda^{(\mu)}_k t_{l-1}}\,dt_{l-1}\\
&\quad=\delta_{i,j}\delta_{\omega_{l-1},\omega_l}\delta_{x_{l-1},x_l}\frac{1}{n_{\omega_{l-1}}}\sum_{\gamma_\mu\in\Gamma_\mu}p^{(0,\mu-1)}_{\alpha,\gamma_\mu}\\
&\qquad\qquad\times\sum_{k=1}^{2g_\mu}(v^{(\mu)}_k w^{(\mu)}_k)_{\gamma_\mu,a_{\omega_{l-1}}}\lambda^{(\mu)}_k H^{t_\mu}_{t_{\mu-1}}(-\lambda^{(\mu)}_k t_{\mu-1},\lambda^{(\mu)}_k-\rho),
\end{aligned}
\tag{S28}
$$

with $\mu = i \wedge j$, and $H$ as defined in equations (S19) and (S20).

Furthermore, after substituting, the term in parentheses in the *decoupled* term isolated in (S16) yields

$$
\begin{aligned}
\int_{t_{l-1}=t_{i-1}}^{t_i}\int_{t_l=t_{j-1}}^{t_j}\Bigg(&\sum_{\epsilon=1}^{\mu-1}\int_{t_b=t_{\epsilon-1}}^{t_\epsilon}\sum_{\eta\in\Gamma_\epsilon}\mathbb{P}\{T^A_{l-1}\in dt_{l-1},G_{l-1}=\omega_{l-1},T^A_l\in dt_l,G_l=\omega_l,T^B\in dt_b,G^B_{t_b}=\eta\}\\
&+\int_{t_b=t_{\mu-1}}^{t_{l-1}\wedge t_l}\sum_{\eta\in\Gamma_\mu}\mathbb{P}\{T^A_{l-1}\in dt_{l-1},G_{l-1}=\omega_{l-1},T^A_l\in dt_l,G_l=\omega_l,T^B\in dt_b,G^B_{t_b}=\eta\}\Bigg).
\end{aligned}
\tag{S29}
$$

Again, fixing $\epsilon$ and focusing on a summand of the first sum in (S29) yields

$$
\begin{aligned}
&\int_{t_b=t_{\epsilon-1}}^{t_\epsilon} \rho e^{-\rho t_b} \sum_{\eta\in\Gamma_\epsilon}\sum_{\gamma_\epsilon\in\Gamma_\epsilon} p^{(0,\epsilon-1)}_{\alpha,\gamma_\epsilon}\big(e^{(t_b-t_{\epsilon-1})Z_\epsilon}\big)_{\gamma_\epsilon,\eta} \\
&\quad\times \sum_{Z_{\epsilon+1}\in\Gamma_{\epsilon+1}} \sum_{\substack{\zeta_\epsilon\in\Gamma_\epsilon\\ \zeta_\epsilon\subset Z_{\epsilon+1}}} \big(e^{(t_\epsilon-t_b)Z_\epsilon}\big)_{\eta,\zeta_\epsilon} \int_{t_{l-1}=t_{i-1}}^{t_i} q^{(\epsilon,i)}_{Z_{\epsilon+1},a_{\omega_{l-1}}}(t_{l-1}-t_{i-1})\,dt_{l-1} \\
&\quad\times \sum_{\Xi_{\epsilon+1}\in\Gamma_{\epsilon+1}} \sum_{\substack{\xi_\epsilon\in\Gamma_\epsilon\\ \xi_\epsilon\subset \Xi_{\epsilon+1}}} \big(e^{(t_\epsilon-t_b)Z_\epsilon}\big)_{\eta,\xi_\epsilon} \int_{t_l=t_{j-1}}^{t_j} q^{(\epsilon,j)}_{\Xi_{\epsilon+1},a_{\omega_l}}(t_l-t_{j-1})\,dt_l\,dt_b \\
&= \sum_{\gamma_\epsilon\in\Gamma_\epsilon}\sum_{Z_{\epsilon+1}\in\Gamma_{\epsilon+1}}\sum_{\Xi_{\epsilon+1}\in\Gamma_{\epsilon+1}} p^{(0,\epsilon-1)}_{\alpha,\gamma_\epsilon} Q^{(\epsilon)}_{Z_{\epsilon+1},a_{\omega_{l-1}}}(i) Q^{(\epsilon)}_{\Xi_{\epsilon+1},a_{\omega_l}}(j) \\
&\quad\times \sum_{\eta\in\Gamma_\epsilon}\sum_{\substack{\zeta_\epsilon\in\Gamma_\epsilon\\ \zeta_\epsilon\subset Z_{\epsilon+1}}}\sum_{\substack{\xi_\epsilon\in\Gamma_\epsilon\\ \xi_\epsilon\subset \Xi_{\epsilon+1}}} \int_{t_b=t_{\epsilon-1}}^{t_\epsilon} \rho e^{-\rho t_b}\big(e^{(t_b-t_{\epsilon-1})Z_\epsilon}\big)_{\gamma_\epsilon,\eta}\big(e^{(t_\epsilon-t_b)Z_\epsilon}\big)_{\eta,\zeta_\epsilon}\big(e^{(t_\epsilon-t_b)Z_\epsilon}\big)_{\eta,\xi_\epsilon}\,dt_b \\
&= \sum_{\gamma_\epsilon\in\Gamma_\epsilon}\sum_{Z_{\epsilon+1}\in\Gamma_{\epsilon+1}}\sum_{\Xi_{\epsilon+1}\in\Gamma_{\epsilon+1}} p^{(0,\epsilon-1)}_{\alpha,\gamma_\epsilon} R^{(\epsilon)}_{\gamma_\epsilon,(Z_{\epsilon+1},\Xi_{\epsilon+1})} Q^{(\epsilon)}_{Z_{\epsilon+1},a_{\omega_{l-1}}}(i) Q^{(\epsilon)}_{\Xi_{\epsilon+1},a_{\omega_l}}(j),
\end{aligned}
\tag{S30}
$$

where $Q^{(\cdot)}_{\cdot,\cdot}(\cdot)$ was defined in (S24), and $R^{(\cdot)}_{\cdot,(\cdot,\cdot)}$ was defined in (S18).

For the second summand in (S29) the cases $i \neq j$ and $i = j$ have to be distinguished Assume without loss of generality $i < j$, so $\mu = i$. Then, focusing on the case $i < j$ gives

$$
\begin{aligned}
&\int_{t_{l-1}=t_{\mu-1}}^{t_\mu}\int_{t_l=t_{j-1}}^{t_j}\int_{t_b=t_{\mu-1}}^{t_{l-1}} \rho e^{-\rho t_b}\sum_{\eta\in\Gamma_\mu}\sum_{\gamma_\mu\in\Gamma_\mu} p^{(0,\mu-1)}_{\alpha,\gamma_\mu}\big(e^{(t_b-t_{\mu-1})Z_\mu}\big)_{\gamma_\mu,\eta}\big(Z_\mu e^{(t_{l-1}-t_b)Z_\mu}\big)_{\eta,a_{\omega_{l-1}}} \\
&\quad\times \sum_{\Xi_{\mu+1}\in\Gamma_{\mu+1}}\sum_{\substack{\xi_\mu\in\Gamma_\mu\\ \xi_\mu\subset\Xi_{\mu+1}}} \big(e^{(t_\mu-t_b)Z_\mu}\big)_{\eta,\xi_\mu} q^{(\mu,j)}_{\Xi_{\mu+1},a_{\omega_l}}(t_l-t_{j-1})\,dt_b\,dt_l\,dt_{l-1} \\
&= \sum_{\eta\in\Gamma_\mu}\sum_{\gamma_\mu\in\Gamma_\mu}\sum_{\Xi_{\mu+1}\in\Gamma_{\mu+1}}\sum_{\substack{\xi_\mu\in\Gamma_\mu\\ \xi_\mu\subset\Xi_{\mu+1}}} p^{(0,\mu-1)}_{\alpha,\gamma_\mu} Q^{(\mu)}_{\Xi_{\mu+1},a_{\omega_l}}(j) \\
&\quad\times \int_{t_{l-1}=t_{\mu-1}}^{t_\mu}\int_{t_b=t_{\mu-1}}^{t_{l-1}} \rho e^{-\rho t_b}\big(e^{(t_b-t_{\mu-1})Z_\mu}\big)_{\gamma_\mu,\eta}\big(e^{(t_\mu-t_b)Z_\mu}\big)_{\eta,\xi_\mu}\big(Z_\mu e^{(t_{l-1}-t_b)Z_\mu}\big)_{\eta,a_{\omega_{l-1}}}\,dt_b\,dt_{l-1} \\
&= \sum_{\eta\in\Gamma_\mu}\sum_{\gamma_\mu\in\Gamma_\mu}\sum_{\Xi_{\mu+1}\in\Gamma_{\mu+1}}\sum_{\substack{\xi_\mu\in\Gamma_\mu\\ \xi_\mu\subset\Xi_{\mu+1}}} p^{(0,\mu-1)}_{\alpha,\gamma_\mu} Q^{(\mu)}_{\Xi_{\mu+1},a_{\omega_l}}(j) \\
&\quad\times \sum_{k=1}^{2g_\mu}\sum_{m=1}^{2g_\mu}\sum_{n=1}^{2g_\mu} (v^{(\mu)}_k w^{(\mu)}_k)_{\gamma_\mu,\eta}(v^{(\mu)}_m w^{(\mu)}_m)_{\eta,\xi_\mu}(v^{(\mu)}_n w^{(\mu)}_n)_{\eta,a_{\omega_{l-1}}} \\
&\qquad\times \rho\lambda^{(\mu)}_n e^{\lambda^{(\mu)}_m t_\mu-\lambda^{(\mu)}_k t_{\mu-1}}\int_{t_{l-1}=t_{\mu-1}}^{t_\mu} e^{\lambda^{(\mu)}_n t_{l-1}}\int_{t_b=t_{\mu-1}}^{t_{l-1}} e^{(\lambda^{(\mu)}_k-\lambda^{(\mu)}_m-\lambda^{(\mu)}_n-\rho)t_b}\,dt_b\,dt_{l-1}.
\end{aligned}
\tag{S31}
$$

Changing the order of integration in integral expression in (S31) yields

$$\begin{aligned}
\lambda_n^{(\mu)} \int_{t_b=t_{\mu-1}}^{t_\mu} e^{(\lambda_k^{(\mu)}-\lambda_m^{(\mu)}-\lambda_n^{(\mu)}-\rho)t_b} \int_{t_{l-1}=t_b}^{t_\mu} e^{\lambda_n^{(\mu)} t_{l-1}}\, dt_{l-1}\, dt_b \\
&= \int_{t_b=t_{\mu-1}}^{t_\mu} e^{(\lambda_k^{(\mu)}-\lambda_m^{(\mu)}-\lambda_n^{(\mu)}-\rho)t_b} (e^{\lambda_n^{(\mu)} t_\mu} - e^{\lambda_n^{(\mu)} t_b})\, dt_b \\
&= e^{\lambda_n^{(\mu)} t_\mu} \int_{t_b=t_{\mu-1}}^{t_\mu} e^{(\lambda_k^{(\mu)}-\lambda_m^{(\mu)}-\lambda_n^{(\mu)}-\rho)t_b} dt_b - \int_{t_b=t_{\mu-1}}^{t_\mu} e^{(\lambda_k^{(\mu)}-\lambda_m^{(\mu)}-\rho)t_b}\, dt_b.
\end{aligned} \tag{S32}$$

Note that the second term in equation (S32) does not depend on $n$ anymore. Thus, when substituting it back into expression (S31) this term vanishes, since $\sum_{n=1}^{2g_\mu} (v_n^{(\mu)} w_n^{(\mu)})_{\eta_b, a_{\omega_{l-1}}} = (\mathbb{1})_{\eta_b, a_{\omega_{l-1}}} = 0$. The first term in equation (S32) can be written as

$$H_{t_{\mu-1}}^{t_\mu}(\lambda_n^{(\mu)} t_\mu, \lambda_k^{(\mu)} - \lambda_m^{(\mu)} - \lambda_n^{(\mu)} - \rho), \tag{S33}$$

using the definitions (S19) and (S20). Note that $i < j$ and thus $t_\mu < \infty$ hold, so this quantity is well defined. Substituting it into (S31) yields

$$\begin{aligned}
&\sum_{\eta\in\Gamma_\mu} \sum_{\gamma_\mu\in\Gamma_\mu} \sum_{\Xi_{\mu+1}\in\Gamma_{\mu+1}} \sum_{\substack{\xi_\mu\in\Gamma_\mu \\ \xi_\mu\subset\Xi_{\mu+1}}} p_{\alpha,\gamma_\mu}^{(0,\mu-1)} Q_{\Xi_{\mu+1},a_{\omega_l}}^{(\mu)}(j) \\
&\quad \times \sum_{k=1}^{2g_\mu} \sum_{m=1}^{2g_\mu} \sum_{n=1}^{2g_\mu} (v_k^{(\mu)} w_k^{(\mu)})_{\gamma_\mu,\eta} (v_m^{(\mu)} w_m^{(\mu)})_{\eta,\xi_\mu} (v_n^{(\mu)} w_n^{(\mu)})_{\eta,a_{\omega_{l-1}}} \\
&\qquad \times \rho H_{t_{\mu-1}}^{t_\mu}\big((\lambda_m^{(\mu)} + \lambda_n^{(\mu)}) t_\mu - \lambda_k^{(\mu)} t_{\mu-1}, \lambda_k^{(\mu)} - \lambda_m^{(\mu)} - \lambda_n^{(\mu)} - \rho\big) \\
&= \sum_{\gamma_\mu\in\Gamma_\mu} p_{\alpha,\gamma_\mu}^{(0,\mu-1)} \sum_{\Xi_{\mu+1}\in\Gamma_{\mu+1}} \sum_{\substack{\xi_\mu\in\Gamma_\mu \\ \xi_\mu\subset\Xi_{\mu+1}}} R_{\gamma_\mu,(a_{\omega_{l-1}},\xi_\mu)}^{(\mu)} Q_{\Xi_{\mu+1},a_{\omega_l}}^{(\mu)}(j).
\end{aligned} \tag{S34}$$

In the case $i = j = \mu$, the second summand in equation (S29) gives

$$\begin{aligned}
&\int_{t_{l-1}=t_{\mu-1}}^{t_\mu} \int_{t_l=t_{\mu-1}}^{t_\mu} \int_{t_b=t_{\mu-1}}^{t_{l-1}\wedge t_l} \rho e^{-\rho t_b} \sum_{\eta\in\Gamma_\mu} \sum_{\gamma_\mu\in\Gamma_\mu} p_{\alpha,\gamma_\mu}^{(0,\mu-1)} \\
&\qquad \times \left(e^{(t_b-t_{\mu-1})Z_\mu}\right)_{\gamma_\mu,\eta_b} \left(Z_\mu e^{(t_{l-1}-t_b)Z_\mu}\right)_{\eta,a_{\omega_{l-1}}} \left(Z_\epsilon e^{(t_l-t_b)Z_\mu}\right)_{\eta,a_{\omega_l}} dt_b\, dt_l\, dt_{l-1} \\
&= \sum_{\eta\in\Gamma_\epsilon} \sum_{\gamma_\mu\in\Gamma_\epsilon} p_{\alpha,\gamma_\mu}^{(0,\mu-1)} \sum_{k=1}^{2g_\mu} \sum_{m=1}^{2g_\mu} \sum_{n=1}^{2g_\mu} (v_k^{(\mu)} w_k^{(\mu)})_{\gamma_\mu,\eta} (v_m^{(\mu)} w_m^{(\mu)})_{\eta,a_{\omega_{l-1}}} (v_n^{(\mu)} w_n^{(\mu)})_{\eta,a_{\omega_l}} \\
&\quad \times \rho \lambda_m^{(\mu)} \lambda_n^{(\mu)} e^{-\lambda_k^{(\mu)} t_{\mu-1}} \int_{t_{l-1}=t_{\mu-1}}^{t_\mu} \int_{t_l=t_{\mu-1}}^{t_\mu} \int_{t_b=t_{\mu-1}}^{t_{l-1}\wedge t_l} e^{(\lambda_k^{(\mu)}-\lambda_m^{(\mu)}-\lambda_n^{(\mu)}-\rho)t_b} e^{\lambda_m^{(\mu)} t_{l-1}} e^{\lambda_n^{(\mu)} t_l}\, dt_b\, dt_l\, dt_{l-1}.
\end{aligned} \tag{S35}$$

Again, considering only the integral part in equation (S29), and exchanging the order of integration yields

$$\begin{aligned}
&\lambda_m^{(\mu)}\lambda_n^{(\mu)}\int_{t_b=t_{\mu-1}}^{t_\mu} e^{(\lambda_k^{(\mu)}-\lambda_m^{(\mu)}-\lambda_n^{(\mu)}-\rho)t_b}\Big[\int_{t_{l-1}=t_b}^{t_\mu} e^{\lambda_m^{(\mu)}t_{l-1}}\,dt_{l-1}\Big]\Big[\int_{t_l=t_b}^{t_\mu} e^{\lambda_n^{(\mu)}t_l}\,dt_l\Big]\,dt_b.\\
&=\int_{t_b=t_{\mu-1}}^{t_\mu} e^{(\lambda_k^{(\mu)}-\lambda_m^{(\mu)}-\lambda_n^{(\mu)}-\rho)t_b}\Big[e^{\lambda_m^{(\mu)}t_\mu}-e^{\lambda_m^{(\mu)}t_b}\Big]\Big[e^{\lambda_n^{(\mu)}t_\mu}-e^{\lambda_n^{(\mu)}t_b}\Big]\,dt_b\\
&=\int_{t_b=t_{\mu-1}}^{t_\mu} e^{(\lambda_k^{(\mu)}-\lambda_m^{(\mu)}-\lambda_n^{(\mu)}-\rho)t_b}e^{(\lambda_m^{(\mu)}+\lambda_n^{(\mu)})t_\mu}\,dt_b\\
&=H_{t_{\mu-1}}^{t_\mu}\big((\lambda_m^{(\mu)}+\lambda_n^{(\mu)})t_\mu,\lambda_k^{(\mu)}-\lambda_m^{(\mu)}-\lambda_n^{(\mu)}-\rho\big),
\end{aligned}\tag{S36}$$

with $H$ as defined in equations (S19) and (S20). Here the second equality holds, since upon solving the brackets, only the first summand is dependent on $k$, $m$, and $n$. The other summands then vanish due to a similar argument as was used in deriving equation (S33). Substituting the right hand side of equation (S36) into (S35) gives

$$\begin{aligned}
&\sum_{\eta\in\Gamma_\epsilon}\sum_{\gamma_\mu\in\Gamma_\epsilon} p_{\alpha,\gamma_\mu}^{(0,\mu-1)}\sum_{k=1}^{2g_\mu}\sum_{m=1}^{2g_\mu}\sum_{n=1}^{2g_\mu}(v_k^{(\mu)}w_k^{(\mu)})_{\gamma_\mu,\eta}(v_m^{(\mu)}w_m^{(\mu)})_{\eta,a_{\omega_{l-1}}}(v_n^{(\mu)}w_n^{(\mu)})_{\eta,a_{\omega_l}}\\
&\qquad\times\rho H_{t_{\mu-1}}^{t_\mu}\big((\lambda_m^{(\mu)}+\lambda_n^{(\mu)})t_\mu-\lambda_k^{(\mu)}t_{\mu-1},\lambda_k^{(\mu)}-\lambda_m^{(\mu)}-\lambda_n^{(\mu)}-\rho\big)\\
&=\sum_{\gamma_\mu\in\Gamma_\epsilon} p_{\alpha,\gamma_\mu}^{(0,\mu-1)}R_{\gamma_\mu,(a_{\omega_{l-1}},a_{\omega_l})}^{(\mu)}.
\end{aligned}\tag{S37}$$

Combining equations (S28), (S30), (S34), and (S37) gives the joint absorption probability (S27).

The discretized transition probability can then be obtained by dividing the joint probability through the marginal probability at locus $l-1$:

$$\begin{aligned}
&\phi(i_l,\omega_l,x_l|i_{l-1},\omega_{l-1},x_{l-1})\\
&:=\mathbb{P}\{T_l^A\in I_{i_l},G_l=\omega_l,X_l=x_l|T_{l-1}^A\in I_{i_{l-1}},G_{l-1}=\omega_{l-1},X_{l-1}=x_{l-1}\}\\
&=y(i_{l-1},\omega_{l-1})\delta_{i_{l-1},i_l}\delta_{\omega_{l-1},\omega_l}\delta_{x_{l-1},x_l}+z(i_l,\omega_l|i_{l-1},\omega_{l-1})\frac{1}{n_{\omega_l}}.
\end{aligned}\tag{S38}$$

Here

$$y(i,\omega_{l-1}):=\frac{1}{u(i,\omega_{l-1})}\sum_{\gamma_i\in\Gamma_i}p_{\alpha,\gamma_i}^{(0,i-1)}\sum_{k=1}^{2g_i}(v_k^{(i)}w_k^{(i)})_{\gamma_i,a_{\omega_{l-1}}}\lambda_k^{(i)}H_{t_{i-1}}^{t_i}(-\lambda_k^{(i)}t_{i-1},\lambda_k^{(i)}-\rho),\tag{S39}$$

with $u(\cdot,\cdot)$ as defined in (S26). Furthermore, with $\mu=i\wedge j$, define

$$\begin{aligned}
z(j,\omega_l|i,\omega_{l-1}):=\frac{1}{u(i,\omega_{l-1})}\Bigg[&\sum_{\epsilon=1}^{\mu-1}\sum_{\gamma_\epsilon\in\Gamma_\epsilon}p_{\alpha,\gamma_\epsilon}^{(0,\epsilon-1)}\sum_{Z_{\epsilon+1}\in\Gamma_{\epsilon+1}}\sum_{\substack{\zeta_\epsilon\in\Gamma_\epsilon\\ \zeta_\epsilon\subset Z_{\epsilon+1}}}\sum_{\Xi_{\epsilon+1}\in\Gamma_{\epsilon+1}}\sum_{\substack{\xi_\epsilon\in\Gamma_\epsilon\\ \xi_\epsilon\subset\Xi_{\epsilon+1}}}R_{\gamma_\epsilon,(\zeta_\epsilon,\xi_\epsilon)}^{(\epsilon)}\\
&\qquad\times Q_{Z_{\epsilon+1},a_{\omega_{l-1}}}^{(\epsilon)}(i)Q_{\Xi_{\epsilon+1},a_{\omega_l}}^{(\epsilon)}(j)\\
&+\sum_{\gamma_\mu\in\Gamma_\mu}p_{\alpha,\gamma_\mu}^{(0,\mu-1)}W_{\gamma_\mu,(a_{\omega_{l-1}},a_{\omega_l})}^{(\mu)}(i,j)\Bigg],
\end{aligned}\tag{S40}$$

where

$$W^{(\mu)}_{\gamma_\mu,(a_{\omega_{l-1}},a_{\omega_l})}(i,j) := \begin{cases} \sum_{\Xi_{\mu+1}\in\Gamma_{\mu+1}} \sum_{\substack{\xi_\mu\in\Gamma_\mu \\ \xi_\mu\subset\Xi_{\mu+1}}} R^{(\mu)}_{\gamma_\mu,(a_{\omega_{l-1}},\xi_\mu)} Q^{(\mu)}_{\Xi_{\mu+1},a_{\omega_l}}(j), & \text{if } i<j, \\ R^{(\mu)}_{\gamma_\mu,(a_{\omega_{l-1}},a_{\omega_l})}, & \text{if } i=j, \\ W^{(\mu)}_{\gamma_\mu,(a_{\omega_l},a_{\omega_{l-1}})}(j,i), & \text{if } i>j. \end{cases} \tag{S41}$$

S1.3.3. *Emission probability.* Finally, the emission probability, that is, the probability that the observed haplotype $H$ carries the allele $a$ at locus $l$ given that the additional lineage at this locus is absorbed during the interval $I_i$ in sub-population $\omega_l \in \Gamma_i$ into the lineage $x_l \in \mathbf{n}_{\omega_l}$ can be computed as

$$\begin{aligned} &\mathbb{P}\{H[l]=a|T_l^A\in I_i, G_l=\omega_l, X_l=x_l\} \\ &= \frac{\mathbb{P}\{H[l]=a, T_l^A\in I_i, G_l=\omega_l, X_l=x_l\}}{\mathbb{P}\{T_l^A\in I_i, G_l=\omega_l, X_l=x_l\}} \\ &= \frac{\mathbb{P}\{H[l]=a, T_l^A\in I_i, G_l=\omega_l, X_l=x_l\}}{u(i,\omega_l)} n_{\omega_l}. \end{aligned} \tag{S42}$$

Using equations (S23) and (S13), the numerator in (S42) yields

$$\begin{aligned} &\mathbb{P}\{H[l]=a, T_l^A\in I_i, G_l=\omega_l, X_l=x_l\} n_{\omega_l} \\ &= \int_{t_l=t_{i-1}}^{t_i} \mathbb{P}\{H[l]=a, T_l^A\in dt_l, G_l=\omega_l, X_l=x_l\} n_{\omega_l} \\ &= \int_{t_l=t_{i-1}}^{t_i} \mathbb{P}\{H[l]=a|T_l^A\in dt_l, G_l=\omega_l, X_l=x_l\}\mathbb{P}\{T_l^A\in dt_l, G_l=\omega_l, X_l=x_l\} n_{\omega_l} \\ &= \int_{t_l=t_{i-1}}^{t_i} \big(e^{t_l\theta(P-\mathbb{1})}\big)_{x_l[l],a} \sum_{\gamma_i\in\Gamma_i} p^{(0,i-1)}_{\alpha,\gamma_i} \big(Z_i e^{(t_l-t_{i-1})Z_i}\big)_{\gamma_i,a_{\omega_l}} dt_l \\ &= \sum_{\gamma_i\in\Gamma_i} p^{(0,i-1)}_{\alpha,\gamma_i} \sum_{j=1}^{|E|}\sum_{k=1}^{2g_i} (\mathfrak{v}_j\mathfrak{w}_j)_{x_l[l],a} (v_k^{(i)} w_k^{(i)})_{\gamma_i,a_{\omega_l}} \lambda_k^{(i)} H^{t_i}_{t_{i-1}}\big(-\lambda_k^{(i)} t_{i-1}, \theta(\mathfrak{l}_j-1)+\lambda_k^{(i)}\big), \end{aligned} \tag{S43}$$

where $\mathfrak{l}_j$ (with $\Re(\mathfrak{l}_j)\le 0$) are the eigenvalues of $P$, $\mathfrak{v}_j$ are its eigenvectors, and $\mathfrak{w}_j$ the are row-vectors of the matrix inverse to the matrix made up of the column vectors $\mathfrak{v}_j$. Again, $H$ is as defined in equations (S19) and (S20). Combining equation (S42) with equation (S43) yields, with $u(\cdot,\cdot)$ as defined in (S26),

$$\begin{aligned} \xi(a|i,\omega_l,x_l) &:= \mathbb{P}\{H[l]=a|T_l^A\in I_i, G_l=\omega_l, X_l=x_l\} \\ &= \frac{1}{u(i,\omega_l)} \sum_{\gamma_i\in\Gamma_i} p^{(0,i-1)}_{\alpha,\gamma_i} \sum_{j=1}^{|E|}\sum_{k=1}^{2g_i} (\mathfrak{v}_j\mathfrak{w}_j)_{x_l[l],a} (v_k^{(i)} w_k^{(i)})_{\gamma_i,a_{\omega_l}} \lambda_k^{(i)} \\ &\qquad\qquad \times H^{t_i}_{t_{i-1}}\big(-\lambda_k^{(i)} t_{i-1}, \theta(\mathfrak{l}_j-1)+\lambda_k^{(i)}\big) \end{aligned} \tag{S44}$$

for the emission probability of the discretized HMM underlying the CSD $\pi^D_\Theta$.

## S2. Forward-Backward Algorithm

Given a certain demographic history $\Theta$ and an observed configuration $\mathbf{n}$, denote by $H^{\alpha,\mathbf{n}}_\Theta \in E^L$ the random haplotype additionally sampled in sub-population $\alpha$ which is distributed according to the CSD $\pi^D_\Theta$, that is, $H^{\alpha,\mathbf{n}}_\Theta \sim \pi^D_\Theta(\cdot|\alpha,\mathbf{n})$. Note that the distribution implicitly depends on the recombination rate $\rho$ and the mutational model $(\theta, P)$ as well. As mentioned in Demography-aware Conditional Sampling Distribution of METHODS, the probability $\mathbb{P}\{H^{\alpha,\mathbf{n}}_\Theta = h\}$ of observing a certain

additional haplotype $h \in E^L$ can be computed under the HMM defined by the probabilities $\nu$, $\phi$, and $\xi$ given in Section S1.3 using the forward algorithm. To this end denote by $\mathcal{S} := \big\{(i,\omega,x)\big| i \in \mathcal{E}, \omega \in \Gamma_i, x \in \mathbf{n}_\omega\big\}$ the set of hidden states, so a hidden state comprises of an interval $i$ during which the additional lineage is absorbed, a sub-population $\omega$ in which absorption happens, and a trunk-lineage $x$ that the lineage is absorbed into. Further, for $1 \le l \le L$, denote by $S_l \in \mathcal{S}$ the random hidden state at locus $l$, and by $S_\Theta^{\alpha,\mathbf{n}} := (S_1, \ldots, S_L)$ the full sequence of hidden states.

S2.1. **Forward Algorithm.** Given the hidden state $s_l = (i_l, \omega_l, x_l) \in \mathcal{S}$, the forward probability

$$\begin{aligned} F_l(s_l) &:= \mathbb{P}\big\{H_\Theta^{\alpha,\mathbf{n}}[1:l] = h[1:l], S_l = s_l\big\} \\ &= \mathbb{P}\big\{H_\Theta^{\alpha,\mathbf{n}}[1:l] = h[1:l], T_l^A \in I_{i_l}, G_l = \omega_l, X_l = x_l\big\} \end{aligned} \tag{S45}$$

is the joint probability of observing the partial haplotype $h[1:l]$ up to locus $l$, and the additional lineage being absorbed into haplotype $x_l$ in sub-population $\omega_l$ during interval $i_l$ at locus $l$. Dynamic programming can be used to compute $F_l(s_l)$ via the dynamic program:

$$\begin{aligned} F_l(s_l) &= \xi(h_l|s_l) \sum_{s_{l-1}\in\mathcal{S}} F_{l-1}(s_{l-1})\phi(s_l|s_{l-1}) \\ &= \xi(h_l|i_l,\omega_l,x_l)\Bigg[y(i_l,\omega_l)F_{l-1}(i_l,\omega_l,x_l) \\ &\qquad + \frac{1}{n_{\omega_l}} \sum_{\substack{i_{l-1}\in\mathcal{E},\\ \omega_{l-1}\in\Gamma_{i_{l-1}}}} z(i_l,\omega_l|i_{l-1},\omega_{l-1}) \sum_{x_{l-1}\in\mathbf{n}_{\omega_{l-1}}} F_{l-1}(i_{l-1},\omega_{l-1},x_{l-1})\Bigg]. \end{aligned} \tag{S46}$$

The initial value for this dynamic program is given by

$$F_1(i_1,\omega_1,x_1) = \xi_\theta(h_1|i_1,\omega_1,x_1,)\frac{1}{n_{\omega_1}}u(i_1,\omega_1). \tag{S47}$$

Note that if the haplotypes associated with lineages $x$ and $x'$ from the trunk are identical, then we have $F_l(i,\omega,x) = F_l(i,\omega,x')$ for all $l,i,\omega$. Finally, the probability of observing the additional haplotype is given as

$$\mathbb{P}\{H_\Theta^{\alpha,\mathbf{n}} = h\} = \sum_{s_L\in\mathcal{S}} \mathbb{P}\big\{H_\Theta^{\alpha,\mathbf{n}}[1:L] = h_{[1:L]}, S_L = s_L\big\} = \sum_{s_L\in\mathcal{S}} F_L(s_L). \tag{S48}$$

A naïve implementation of equation (S46) would, for all $s_l \in \mathcal{S}$, iterate over every $s_{l-1} \in \mathcal{S}$. This would result in a quadratic dependence of the runtime on the size of the hidden state space, implying a quadratic dependence on the number of haplotypes in the trunk. To this end, define

$$Q[i_{l-1},\omega_{l-1}] := \sum_{x_{l-1}\in\mathbf{n}_{\omega_{l-1}}} F_{l-1}(i_{l-1},\omega_{l-1},x_{l-1}), \tag{S49}$$

and

$$R[i_l,\omega_l] := \sum_{\substack{i_{l-1}\in\mathcal{E},\\ \omega_{l-1}\in\Gamma_{i_{l-1}}}} z(i_l,\omega_l|i_{l-1},\omega_{l-1})Q[i_{l-1},\omega_{l-1}]. \tag{S50}$$

Pre-computing $R[i_l,\omega_l]$ and re-using it in equation (S46) allows for an implementation whose runtime only depends linearly on the number of haplotypes in the trunk. Thus, the algorithm to compute the forward probabilities and ultimately the likelihood has runtime complexity $O(Lnd^2)$, where $d = Eg$. Recall that $L$ denotes the number of loci, $n$ the number of haplotypes, $E$ the number of discretization intervals, and $g$ the number of sub-populations at present.

S2.2. **Backward algorithm.** The backward probability

$$B_l(s_l) = \mathbb{P}\{H_\Theta^{\alpha,\mathbf{n}}[l+1:L] = h[l+1:L] | T_l^A \in I_{i_l}, G_l = \omega_l, X_l = x_l) \tag{S51}$$

is the probability of observing the alleles $h[l+1:L]$ following locus $l$, conditional on the hidden state at locus $l$. This quantity can again be used to compute the observation probability, but it is also necessary for the expectation-maximization procedure that will be introduced in Section S3. It is possible to write down an explicit backward algorithm for the computation, however, here we proceed along a different route.

To this end define

$$\begin{aligned} F_l^*(s_l) &:= \mathbb{P}\{H_\Theta^{\alpha,\mathbf{n}}[1:(L-l+1)] = h_{[L:l]}, S_{L-l+1} = s_l\} \\ &= \mathbb{P}\{H_\Theta^{\alpha,\mathbf{n}}[l:L] = h_{[L:l]}, S_L = s_l\} \\ &= \mathbb{P}\{H_\Theta^{\alpha,\mathbf{n}}[l:L] = h_{[l:L]}, S_l = s_l\}, \end{aligned} \tag{S52}$$

where $h[L:l]$ denotes the reversed vector $(h[L], \ldots, h[l])$, and equality in (S52) holds since the transition probability is reversible with respect to the initial distribution. Note that (S46) can also be used to compute $F_l^*$, if $F_l$ is replaced by $F_{l-1}^*$, $F_{l-1}$ replaced by $F_l^*$, and the observed alleles are adjusted accordingly to the reversed haplotype $h[L:l]$. Using the modified forward probability $F_l^*$, the backward probability can be obtained via

$$B_l(s_l) = \frac{F_l^*(s_l)}{\xi_\theta(h_l|s_l)\frac{1}{n_{\omega_l}}u(i_l,\omega_l)}. \tag{S53}$$

## S3. Parametric Inference via EM

We now present several ways of combining the CSDs introduced in the previous sections in suitable composite likelihood frameworks. We then detail the application of the Expectation Maximization (EM) algorithm to infer demographic parameters in each of these frameworks.

S3.1. **Composite Likelihoods.** As in Expectation-Maximization Algorithm for Demographic Inference of METHODS, assume that the haplotypes in a given sample configuration $\mathbf{n}$ are ordered by enumerating them from 1 to $n$. Thus, $x_i$ denotes the $i$-th haplotype and $\alpha_i$ denotes the sub-population that the $i$-th haplotype resides in at the time the sample is taken, with $1 \leq i \leq n$. Furthermore, for a given permutation $\sigma$ of $\{1, \ldots, n\}$, define

$$\sigma(i, \mathbf{n}) := \sum_{j=1}^{i} \mathbf{e}_{\alpha_{\sigma(j)}, x_{\sigma(j)}} \tag{S54}$$

to be the configuration induced by $\sigma$ and a given index $i$, where $\mathbf{e}_{\alpha,x}$ again denotes the configuration with a single haplotype $x$ in sub-population $\alpha$. Further, let

$$\mathbf{n}_{-i} := \mathbf{n} - \mathbf{e}_{\alpha_i, x_i} \tag{S55}$$

denote the configuration where haplotype $i$ is removed. As before, denote by $H_\Theta^{\alpha,\mathbf{n}}$ the random additionally sampled haplotype distributed according to the CSD $\pi_\Theta^D(\cdot|\alpha, \mathbf{n})$.

With this notation the product of approximate conditionals (PAC) composite likelihood introduced in METHODS is given by

$$\mathrm{PAC}_\Theta(\mathbf{n}) := \frac{1}{K} \sum_{\sigma \in \Pi} \prod_{i=1}^{n} \mathbb{P}\big\{H_\Theta^{\alpha_{\sigma(i)}, \sigma(i-1,\mathbf{n})} = x_{\sigma(i)}\big\}, \tag{S56}$$

where $\Pi := \{\sigma_1, \ldots, \sigma_K\}$ are $K$ random permutations of $\{1, \ldots, n\}$. Replacing the arithmetic mean in definition (S56) by a geometric mean yields

$$\mathrm{SuperPAC}_\Theta(\mathbf{n}) := \sqrt[K]{\prod_{\sigma\in\Pi}\prod_{i=1}^{n} \mathbb{P}\big\{H_\Theta^{\alpha_{\sigma(i)},\sigma(i-1,\mathbf{n})} = x_{\sigma(i)}\big\}}, \tag{S57}$$

another approximation to the sampling probability which we term SuperPAC. Note that the latter also yields the true likelihood if the true CSD would be used instead of an approximation.

As mentioned in METHODS, the approximate CSD $\pi_\Theta^D(\cdot|\cdot)$ can also be employed in a leave-one-out composite likelihood (LCL)

$$\mathrm{LCL}_\Theta(\mathbf{n}) := \prod_{i=1}^{n} \mathbb{P}\big\{H_\Theta^{\alpha_i,\mathbf{n}_{-i}} = x_i\big\}, \tag{S58}$$

evaluating the product of all CSDs obtained by leaving each haplotype in turn out of the trunk, or a pairwise composite likelihood (PCL)

$$\mathrm{PCL}_\Theta(\mathbf{n}) := \prod_{i\neq j} \mathbb{P}\big\{H_\Theta^{\alpha_i,\mathbf{e}_{\alpha_j,x_j}} = x_i\big\}, \tag{S59}$$

consisting of the product of CSDs between all pairs of haplotypes.

S3.2. **Objective Functions.** Since the composite likelihoods introduced in the previous paragraphs are combinations of the HMMs underlying the different CSDs, they all comprise of observed random variables $H_\Theta^{\alpha,\mathbf{n}}$, the additionally sampled haplotypes, and latent random variables $S_\Theta^{\alpha,\mathbf{n}}$, the associated sequences of hidden states. To obtain a maximum likelihood estimate (MLE) of the demographic parameters $\Theta$ that best describe a given sample of haplotypes $\mathbf{n}$ under a certain composite likelihood, we apply the standard expectation-maximization (EM) framework [1].

The general outline of the EM algorithm is as follows. Suppose we have parameters $\Theta$, and random variables $\mathbb{X}_\Theta, \mathbb{S}_\Theta$, where $\mathbb{X}_\Theta = \mathbb{X}$ is observed, and $\mathbb{S}_\Theta$ is unobserved (hidden). We would like to find the value of $\Theta$ that maximizes the likelihood $\mathbb{L}(\Theta) = \mathbb{P}(\mathbb{X}_\Theta = \mathbb{X})$. To do so, first choose initial parameters $\Theta^{(0)}$, and then update them iteratively. At step $k+1$, the parameters $\Theta^{(k+1)}$ are obtained by maximizing a certain objective function $Q$ based on $\Theta^{(k)}$, that is

$$\Theta^{(k+1)} = \operatorname*{argmax}_\Theta Q(\Theta|\Theta^{(k)}). \tag{S60}$$

where $Q(\Theta|\Theta^{(k)}) = \mathbb{E}_{\mathbb{S}_{\Theta^{(k)}}}[\log \mathbb{P}(\mathbb{X}_\Theta = \mathbb{X}, \mathbb{S}_\Theta = \mathbb{S}_{\Theta^{(k)}}) \mid \mathbb{X}_{\Theta^{(k)}} = \mathbb{X}]$, with the expectation is taken over $\mathbb{S}_{\Theta^{(k)}}$, as indicated by the subscript. Then $\Theta^{(k)}$ is guaranteed to converge to a local maximum of the likelihood surface $\mathbb{L}(\Theta)$.

We can apply EM to find local maxima of our composite likelihoods $\mathrm{PAC}_\Theta(\mathbf{n})$, $\mathrm{SuperPAC}_\Theta(\mathbf{n})$, $\mathrm{LCL}_\Theta(\mathbf{n})$, $\mathrm{PCL}_\Theta(\mathbf{n})$. To do so, for each composite likelihood, we construct a generative model and random variables $\mathbb{X}_\Theta, \mathbb{S}_\Theta$, such that the composite likelihood is equal to $\mathbb{P}(\mathbb{X}_\Theta = \mathbb{X})$. We then derive $Q_\cdot(\Theta|\Theta^{(k)})$ for each such model.

Note that it is in general not possible to solve the maximization problem in equation (S60) analytically. Thus, in the remainder of this section, we will describe how to evaluate the objective functions for given $\Theta$ and $\Theta^{(k)}$, and employ it in a numerical framework, like the Nelder-Mead simplex algorithm [2], to find the requisite maximum. The EM framework guarantees that the overall likelihood of the data increases with each parameter update.

**PAC.** Fixing the set of random permutations $\Pi$, definition (S56) can be interpreted as a mixture model: First, pick a permutation $\Psi$ uniformly at random from the pool $\Pi$. Then, conditional on $\Psi = \sigma$, generate a random sample $\mathfrak{N}_\Theta^\sigma$: First, sample a haplotype in sub-population $\alpha_{\sigma(1)}$ given an empty trunk. Each allele at each locus is sampled from the stationary distribution of the mutation matrix $P$. Then, sample a second haplotype in sub-population $\alpha_{\sigma(2)}$ given the first haplotype as

the already observed trunk; a third haplotype in sub-population $\alpha_{\sigma(3)}$ given the first two in the trunk; and so forth, until a sample of size $n$ is generated. The event that the sample $\mathbf{n}$ is generated in this way is given by

$$\{\mathfrak{N}_\Theta^\sigma = \mathbf{n}\} = \bigcap_{i=1}^{n} \{H_\Theta^{\alpha_{\sigma(i)},\sigma(i-1,\mathbf{n})} = x_{\sigma(i)}\}, \tag{S61}$$

Finally, $\mathrm{PAC}_\Theta(\mathbf{n}) = \mathbb{P}\{\mathfrak{N}_\Theta^\Psi = \mathbf{n}\}$ gives the likelihood of observing the configuration $\mathbf{n}$ under this mixture model, and is equal to equation (S56). Using our previous notation, we have the observed variable $\mathbb{X}_\Theta = \mathfrak{N}_\Theta^\Psi$, and the hidden latent variable $\mathbb{S}_\Theta = \{\Psi, S_\Theta^{\cdot,\cdot}\}$, where $S_\Theta^{\cdot,\cdot}$ is the sequence of hidden states for every CSD.

Let $\Upsilon$ be the random permutation associated with $\Theta^{(k)}$, so $\mathbb{S}_{\Theta^{(k)}} = \{\Upsilon, S_{\Theta^{(k)}}^{\cdot,\cdot}\}$. Then we have

$$\begin{aligned}
&Q_{\mathrm{PAC}}(\Theta|\Theta^{(k)}) \\
&= \mathbb{E}\Bigg[\log\Bigg(\mathbb{P}\{\Psi=\Upsilon\}\prod_{i=1}^{n}\mathbb{P}\Big\{S_\Theta^{\alpha_{\Psi(i)},\Psi(i-1,\mathbf{n})}=S_{\Theta^{(k)}}^{\alpha_{\Upsilon(i)},\Upsilon(i-1,\mathbf{n})}, H_\Theta^{\alpha_{\Psi(i)},\Psi(i-1,\mathbf{n})}=x_{\Upsilon(i)}\Big|\Psi=\Upsilon\Big\}\Bigg)\Bigg|\mathfrak{N}_{\Theta^{(k)}}^\Upsilon=\mathbf{n}\Bigg] \\
&= -\log(K) + \sum_{\sigma\in\Pi}\mathbb{P}\{\Upsilon=\sigma|\mathfrak{N}_{\Theta^{(k)}}^\Upsilon=\mathbf{n}\} \\
&\qquad \times \sum_{i=1}^{n}\mathbb{E}\Bigg[\log\mathbb{P}\Big\{S_\Theta^{\alpha_{\sigma(i)},\sigma(i-1,\mathbf{n})} = S_{\Theta^{(k)}}^{\alpha_{\sigma(i)},\sigma(i-1,\mathbf{n})}, H_\Theta^{\alpha_{\sigma(i)},\sigma(i-1,\mathbf{n})} = x_{\sigma(i)}\Big\}\Bigg|H_{\Theta^{(k)}}^{\alpha_{\sigma(i)},\sigma(i-1,\mathbf{n})} = x_{\sigma(i)}\Bigg] \\
&= -\log(K) + \sum_{\sigma\in\Pi}\frac{\mathbb{P}\big\{\mathfrak{N}_{\Theta^{(k)}}^\sigma = \mathbf{n}\big|\Upsilon=\sigma\big\}}{\sum_{\tau\in\Pi}\mathbb{P}\big\{\mathfrak{N}_{\Theta^{(k)}}^\tau = \mathbf{n}\big|\Upsilon=\tau\big\}}\sum_{i=1}^{n}Q_{x_{\sigma(i)}}^{\alpha_{\sigma(i)},\sigma(i-1,\mathbf{n})}(\Theta|\Theta^{(k)}).
\end{aligned} \tag{S62}$$

The second equality follows from partitioning the conditional expectation with respect to $\{\Upsilon = \sigma\}$ and the fact that $\mathbb{P}\{\Psi = \sigma\} = 1/K$. The third equality follows from an application of Bayes' rule and using the definition

$$Q_x^{\alpha,\mathbf{n}}(\Theta|\Theta^{(k)}) := \mathbb{E}\Big[\log\mathbb{P}\big\{S_\Theta^{\alpha,\mathbf{n}} = S_{\Theta^{(k)}}^{\alpha,\mathbf{n}}, H_\Theta^{\alpha,\mathbf{n}} = x\big\}\Big|H_{\Theta^{(k)}}^{\alpha,\mathbf{n}} = x\Big]; \tag{S63}$$

the objective function for a single HMM.

**SuperPAC.** Here the generating model is as follows. Again, fix the random set of permutations $\Pi$, but instead of sampling a dataset for a single random permutation as in the PAC mixture model, we obtain $\mathbb{X}_\Theta$ by independently sampling a dataset $\mathfrak{N}_\Theta^\sigma$ for every permutation $\sigma$. We then have $\mathrm{SuperPAC}_\Theta(\mathbf{n})^K = \mathbb{P}(\mathbb{X}_\Theta = (\mathbf{n}, \mathbf{n}, \ldots, \mathbf{n}))$, the likelihood of observing $\{\mathfrak{N}_\Theta^\sigma = \mathbf{n}\}$ for each of the $K$ permutations. The hidden latent variable $\mathbb{S}_\Theta$ is given by the sequence of hidden states for every CSD. The objective function for the SuperPAC composite likelihood (S57) is given by

$$Q_{\mathrm{SuperPAC}}(\Theta|\Theta^{(k)}) = \sum_{\sigma\in\Pi}\sum_{i=1}^{n}Q_{x_{\sigma(i)}}^{\alpha_{\sigma(i)},\sigma(i-1,\mathbf{n})}(\Theta|\Theta^{(k)}), \tag{S64}$$

where taking the root can be omitted, since it is a monotone function.

**LCL.** In the LCL (S58) case, the objective function is

$$Q_{\mathrm{LCL}}(\Theta|\Theta^{(k)}) = \sum_{i=1}^{n}Q_{x_i}^{\alpha_i,\mathbf{n}_{-i}}(\Theta|\Theta^{(k)}), \tag{S65}$$

which is obtained by constructing a generative model where we independently sample the haplotype for each leave-one-out model.

**PCL.** Lastly, the objective function for PCL (S59) is

$$Q_{\mathrm{PCL}}(\Theta|\Theta^{(k)}) = \sum_{i\neq j} Q_{x_i}^{\alpha_i,\mathbf{e}_{\alpha_j,x_j}}(\Theta|\Theta^{(k)}), \tag{S66}$$

which is obtained by a generative model where we independently sample the additional haplotype for each pair.

Equations (S62), (S64), (S65), and (S66) show that for each of the composite likelihoods considered here the objective function can be written in terms of the objective functions $Q^{\cdot,\cdot}$ for the individual HMMs involved. For a general $h$, $\alpha$, and $\mathbf{n}$, this function can be further simplified to obtain

$$\begin{aligned}
Q_h^{\alpha,\mathbf{n}}(\Theta|\Theta^{(k)}) &= \mathbb{E}\Big[\log \mathbb{P}\big\{S_\Theta^{\alpha,\mathbf{n}} = S_{\Theta^{(k)}}^{\alpha,\mathbf{n}}, H_\Theta^{\alpha,\mathbf{n}} = h\big\}\Big| H_{\Theta^{(k)}}^{\alpha,\mathbf{n}} = h\Big] \\
&= \sum_{s\in\mathcal{S}} \mathbb{E}\Big[\log\Big(\big(\nu_\Theta(s)\big)^{\mathbb{1}_{\{(S_{\Theta^{(k)}}^{\alpha,\mathbf{n}})_1 = s\}}}\Big)\Big| H_{\Theta^{(k)}}^{\alpha,\mathbf{n}} = h\Big] \\
&\quad + \sum_{s,s'\in\mathcal{S}} \mathbb{E}\Big[\log\Big(\big(\phi_\Theta(s'|s)\big)^{\#\{s\to s'\}}\Big)\Big| H_{\Theta^{(k)}}^{\alpha,\mathbf{n}} = h\Big] \\
&\quad + \sum_{s\in\mathcal{S}}\sum_{a\in E} \mathbb{E}\Big[\log\Big(\big(\xi_\Theta(a|s)\big)^{\#\{s\uparrow a\}}\Big)\Big| H_{\Theta^{(k)}}^{\alpha,\mathbf{n}} = h\Big] \\
&= \sum_{s\in\mathcal{S}} \log\big(\nu_\Theta(s)\big)\mathbb{P}\big\{(S_{\Theta^{(k)}}^{\alpha,\mathbf{n}})_1 = s\big| H_{\Theta^{(k)}}^{\alpha,\mathbf{n}} = h\big\} \\
&\quad + \sum_{s,s'\in\mathcal{S}} \log\big(\phi_\Theta(s'|s)\big)\mathbb{E}\big[\#\{s\to s'\}\big| H_{\Theta^{(k)}}^{\alpha,\mathbf{n}} = h\big] \\
&\quad + \sum_{i\in\mathcal{E},\omega\in\Gamma_i}\sum_{a,t\in E} \log\big(\xi_\Theta(a|i,\omega,t)\big)\mathbb{E}\big[\#\{(i,\omega,t)\uparrow a\}\big| H_{\Theta^{(k)}}^{\alpha,\mathbf{n}} = h\big].
\end{aligned} \tag{S67}$$

The initial $\nu$, transition $\phi$, and emission $\xi$ probabilities are given in (S25), (S38), and (S44). Here the subscripts $\Theta$ is used to emphasize their dependence on the demographic parameters. Furthermore, $\#\{s\uparrow a\}$ denotes the number of times allele $a$ is emitted from hidden state $s$ for a given realization of $S_{\Theta^{(k)}}^{\alpha,\mathbf{n}}$ and $H_{\Theta^{(k)}}^{\alpha,\mathbf{n}}$, and $\#\{s\to s'\}$ is the number of transitions from hidden state $s$ to $s'$. Note that we slightly abuse the notation by conditioning on the trunk allele instead of the trunk haplotype in the emission probability $\xi$ on the last line. We adjust the number of emissions appropriately.

The second summand on the right hand side of equation (S67) (the transition part) can be further modified to

$$\begin{aligned}
&\sum_{i\in\mathcal{E},\omega\in\Gamma_i}\sum_{i'\in\mathcal{E},\omega'\in\Gamma_{i'}}\sum_{x\in\mathbf{n}_\omega,x'\in\mathbf{n}_{\omega'}} \log\big(y_\Theta(i,\omega)\delta_{i,i'}\delta_{\omega,\omega'}\delta_{x,x'} + \frac{1}{n_{\omega'}} z_\Theta(i',\omega'|i,\omega)\big) \\
&\qquad\qquad \times \mathbb{E}\Big[\#\big\{(i,\omega,x)\to(i',\omega',x')\big\}\Big| H_{\Theta^{(k)}}^{\alpha,\mathbf{n}} = h\Big] \\
&= \sum_{i\in\mathcal{E},\omega\in\Gamma_i}\sum_{i'\in\mathcal{E},\omega'\in\Gamma_{i'}} \log\big(\frac{1}{n_{\omega'}} z_\Theta(i',\omega'|i,\omega)\big)\Big(\mathbb{E}\Big[\#\big\{(i,\omega)\to(i',\omega')\big\}\Big| H_{\Theta^{(k)}}^{\alpha,\mathbf{n}} = h\Big] \\
&\qquad\qquad - \delta_{i,i'}\delta_{\omega,\omega'}\sum_{x\in\mathbf{n}_\omega}\mathbb{E}\Big[\#\big\{(i,\omega,x)\to(i,\omega,x)\big\}\Big| H_{\Theta^{(k)}}^{\alpha,\mathbf{n}} = h\Big]\Big) \\
&\quad + \sum_{i\in\mathcal{E},\omega\in\Gamma_i} \log\big(y_\Theta(i,\omega) + \frac{1}{n_\omega} z_\Theta(i,\omega|i,\omega)\big)\sum_{x\in\mathbf{n}_\omega}\mathbb{E}\Big[\#\big\{(i,\omega,x)\to(i,\omega,x)\big\}\Big| H_{\Theta^{(k)}}^{\alpha,\mathbf{n}} = h\Big],
\end{aligned} \tag{S68}$$

with

$$\#\{(i,\omega)\to(i',\omega')\} := \sum_{x\in\mathbf{n}_\omega, x'\in\mathbf{n}_{\omega'}} \#\{(i,\omega,x)\to(i',\omega',x')\}. \tag{S69}$$

We introduce this modification, since a naïve implementation of the left hand side of equation (S68) would depend quadratically on the number of haplotypes in the trunk, whereas the right hand side only depends linearly on this number.

S3.3. **Computing the Conditional Expectations.** We now provide the details on how to compute the conditional probabilities and expectations that are required to evaluate equation (S68) and the objective function (S67), which can then be used to evaluate the objective functions for the different composite likelihoods. Assume that for all $l \in \{1,\dots,L\}$ and all $s \in \mathcal{S}$ the forward probabilities $F_l(s)$ and the backward probabilities $B_l(s)$ introduced in Section S2 haven been computed under the parameters $\Theta^{(k)}$.

The posterior probabilities for the initial hidden state are then given by

$$\begin{aligned}\mathbb{P}\big\{(S^{\alpha,\mathbf{n}}_{\Theta^{(k)}})_1 = s \big| H^{\alpha,\mathbf{n}}_{\Theta^{(k)}} = h\big\} &= \frac{\mathbb{P}\big\{(S^{\alpha,\mathbf{n}}_{\Theta^{(k)}})_1 = s, H^{\alpha,\mathbf{n}}_{\Theta^{(k)}} = h\big\}}{\mathbb{P}\big\{H^{\alpha,\mathbf{n}}_{\Theta^{(k)}} = h\big\}} \\ &= \frac{1}{\mathbb{P}\big\{H^{\alpha,\mathbf{n}}_{\Theta^{(k)}} = h\big\}} \sum_{s\in\mathcal{S}} \nu_{\Theta^{(k)}}(s)\xi_{\Theta^{(k)}}(h_1|s)B_1(s).\end{aligned} \tag{S70}$$

The conditional expectation in equation (S68) that is marginalized over the absorbing haplotypes can be evaluated using

$$\begin{aligned}&\mathbb{E}\Big[\#\{(i,\omega)\to(i',\omega')\}\Big| H^{\alpha,\mathbf{n}}_{\Theta^{(k)}} = h\Big] \\ &= \frac{1}{\mathbb{P}\big\{H^{\alpha,\mathbf{n}}_{\Theta^{(k)}} = h\big\}} \sum_{l=1}^{L} \sum_{x\in\mathbf{n}_\omega, x'\in\mathbf{n}_{\omega'}} \Big(y_{\Theta^{(k)}}(i,\omega)\delta_{i,i'}\delta_{\omega,\omega'}\delta_{x,x'} + \frac{1}{n_{\omega'}} z_{\Theta^{(k)}}(i',\omega'|i,\omega)\Big) \\ &\qquad\qquad \times F_l(i,\omega,x)\xi_{\Theta^{(k)}}(h_{l+1}|i',\omega',x')B_{l+1}(i',\omega',x') \\ &= \frac{1}{\mathbb{P}\big\{H^{\alpha,\mathbf{n}}_{\Theta^{(k)}} = h\big\}} \sum_{l=1}^{L} \Bigg[\frac{1}{n_{\omega'}} z_{\Theta^{(k)}}(i',\omega'|i,\omega)\Big(\sum_{x\in\mathbf{n}_\omega} F_l(i,\omega,x)\Big) \times \sum_{x'\in\mathbf{n}_{\omega'}} \xi_{\Theta^{(k)}}(h_{l+1}|i',\omega',x')B_{l+1}(i',\omega',x') \\ &\qquad + \delta_{i,i'}\delta_{\omega,\omega'} y_{\Theta^{(k)}}(i,\omega) \sum_{x\in\mathbf{n}_\omega} F_l(i,\omega,x)\xi_{\Theta^{(k)}}(h_{l+1}|i,\omega,x)B_{l+1}(i,\omega,x)\Bigg].\end{aligned} \tag{S71}$$

Again, the computation of right hand side only depends linearly on the number of haplotypes in the trunk. The expectation involving the transition from a certain hidden state to itself is given by

$$\begin{aligned}&\mathbb{E}\Big[\#\{(i,\omega,x)\to(i,\omega,x)\}\Big| H^{\alpha,\mathbf{n}}_{\Theta^{(k)}} = h\Big] \\ &\quad= \frac{1}{\mathbb{P}\big\{H^{\alpha,\mathbf{n}}_{\Theta^{(k)}} = h\big\}} \big(y_{\Theta^{(k)}}(i,\omega) + \frac{1}{n_{\omega'}} z_{\Theta^{(k)}}(i',\omega'|i,\omega)\big) \sum_{l=1}^{L} F_l(i,\omega,x)\xi_{\Theta^{(k)}}(h_{l+1}|i,\omega,x)B_{l+1}(i,\omega,x).\end{aligned} \tag{S72}$$

Finally,

$$\mathbb{E}\big[\#\{(i,\omega,t)\uparrow a\}\big| H^{\alpha,\mathbf{n}}_{\Theta^{(k)}} = h\big] = \frac{1}{\mathbb{P}\big\{H^{\alpha,\mathbf{n}}_{\Theta^{(k)}} = h\big\}} \sum_{l=1}^{L} \mathbb{1}_{\{h_l=a\}} \sum_{\substack{x\in\mathbf{n}_\omega \\ x_l=t}} F_l(i,\omega,x)B_l(i,\omega,x) \tag{S73}$$

gives the conditional expectation of the number of emissions of a certain type. The time complexity to evaluate the objective function (S67) is $O(nd^2)$, where $d = Eg$. The overall complexity for the EM algorithm depends on the particular composite likelihood that is used.

## S4. Improving computational efficiency

We now introduce two modifications in order to speed up the computations of the forward-backward algorithm. The runtimes will still depend linearly on the number of loci, but the number of loci that effectively have to be considered will be reduced. In what follows, assume that the demographic parameters $\Theta$, an additional haplotype $h$, a corresponding additional sub-population $\alpha$, and a configuration of trunk haplotypes $\mathbf{n}$ are given, and consider the computations for the CSD $\pi^D_\Theta(h|\alpha, \mathbf{n}) = \mathbb{P}\big\{H^{\alpha,\mathbf{n}}_\Theta = h\big\}$.

S4.1. **Locus skipping.** First we will detail a modification that decreases the number of effective loci by "skipping" over non-polymorphic loci. A similar modification has been introduced before [3] and it requires that the mutation matrix is such that every allele mutates to every other allele at the same rate. For example, this requirement is satisfied by the mutation matrix with $P_{a,a'} = \frac{1}{|E|-1}$ if $a \neq a'$, and $P_{a,a} = 0$. It follows that

$$\xi(a|i, \omega, a) = \xi(a'|i, \omega, a') \tag{S74}$$

holds for all $a, a' \in E$, $i \in \mathcal{E}$, and $\omega \in \Gamma_i$. The modified computations produce the exact result, if the requirement is met by the given mutation matrix. However, even this is not the case, the computational benefit might outweigh the approximation error. Furthermore, define the set of non-polymorphic loci by

$$\mathscr{N} := \big\{1 \leq l \leq L \big| h[l] = x[l], \forall x \in \mathbf{n}\big\}, \tag{S75}$$

that is, the set of all loci, where the additional haplotype and all the trunk haplotypes carry the same allele.

Then, given two hidden states $s = (i, \omega, x), s' = (i', \omega', x') \in \mathcal{S}$, define the $k$-step transition probability as

$$\begin{aligned}&\phi^{(k)}(s'|s)\\ &:= \mathbb{P}\big\{T^A_{l+k} \in I_{i'}, G_{l+k} = \omega', X_{l+k} = x' \big| T^A_l \in I_i, G_l = \omega, X_l = x, \{l+1, \ldots, l+k-1\} \subset \mathscr{N}\big\}.\end{aligned} \tag{S76}$$

By conditioning on $\{l+1, \ldots, l+k-1\} \subset \mathscr{N}$ we include the requirement that the intervening loci $\{l+1, \ldots, l+k-1\}$ are not polymorphic. The base case is given by

$$\begin{aligned}\phi^{(1)}(s'|s) &= y^{(1)}(i, \omega)\delta_{i',i}\delta_{\omega',\omega}\delta_{x',x} + z^{(1)}(i', \omega'|i, \omega)\frac{1}{n_{\omega'}}\\ &= y(i, \omega)\delta_{i',i}\delta_{\omega',\omega}\delta_{x',x} + z(i', \omega'|i, \omega)\frac{1}{n_{\omega'}} = \phi(s'|s),\end{aligned} \tag{S77}$$

the transition probability defined in equation (S38). Now for a given $k$, and any $k_1, k_2 \in \mathbb{N}$ with $k_1 + k_2 = k$, the recursive relation

$$\begin{aligned}\phi^{(k)}(s'|s) &= \sum_{t \in \mathcal{S}} \phi^{(k_2)}(s'|t)\xi(h_{l+k_1}|t)\phi^{(k_1)}(t|s)\\ &= \sum_{j \in \mathcal{E}, \psi \in \Gamma_j} \xi(a|j, \psi, a) \sum_{v \in \mathbf{n}_\psi} \phi^{(k_2)}(s'|j, \psi, v)\phi^{(k_1)}(j, \psi, v|s)\\ &= y^{(k)}(i, \omega)\delta_{i',i}\delta_{\omega',\omega}\delta_{x',x} + z^{(k)}(i', \omega'|i, \omega)\frac{1}{n_{\omega'}}\end{aligned} \tag{S78}$$

holds, with

$$y^{(k)}(i,\omega) := \xi(a|i,\omega,a)y^{(k_2)}(i,\omega)y^{(k_1)}(i,\omega), \tag{S79}$$

and

$$\begin{aligned} z^{(k)}(i',\omega'|i,\omega) := {} & \xi(a|i,\omega,a)z^{(k_2)}(i',\omega'|i,\omega)y^{(k_1)}(i,\omega) \\ & + \xi(a|i',\omega',a)y^{(k_2)}(i',\omega')z^{(k_1)}(i',\omega'|i,\omega) \\ & + \sum_{j\in\mathcal{E},\psi\in\Gamma_j} \xi(a|j,\psi,a)z^{(k_2)}(i',\omega'|j,\psi)z^{(k_1)}(j,\psi|i,\omega). \end{aligned} \tag{S80}$$

The $k$-step transition probabilities can be employed as follows. Denote by $\mathscr{L}' := \{1\} \cup \overline{\mathscr{N}} \cup \{L\}$ the set of polymorphic loci plus the first and the last. Further, define

$$\mathfrak{p}(l,l') := \big\{\mathfrak{n}(l,l')\big\} \cup \mathfrak{p}\big(\mathfrak{n}(l,l'),l'\big), \tag{S81}$$

with $\mathfrak{n}(l,l') := \max\{l+2^m | m \in \mathbb{N}_0, l+2^m < l'\}$ for $l+1 < l'$, and $\mathfrak{p}(l,l+1) := \emptyset$. Now

$$\mathscr{L} := \mathscr{L}' \cup \bigcup_{(l,l') \text{ consecutive in } \mathscr{L}'} \mathfrak{p}(l,l') \tag{S82}$$

is the set of polymorphic loci, plus a scaffold that guarantees that the distance between two consecutive loci in $\mathscr{L}$ is always a power of 2. Further, every locus between 1 and $L$ that is not an element of $\mathscr{L}$ is guaranteed to be non-polymorphic. Thus, the forward $F_l(s)$ and backward $B_l(s)$ probabilities can be computed for $l \in \mathscr{L}$ using only transition probabilities of the form $\phi^{(k)}$ given in equation (S76) with $k = 2^m$, where $m$ is an non-negative integer from 0 to the maximal exponent needed for the possible steps in $\mathscr{L}$. The initial and emission probabilities do not need to be modified.

Previously, since all steps along the sequence in the EM algorithm detailed in section S3 have the same size, only one term involving the transition probability occurs on the right hand side of equation (S67). To implement the possibility of different step sizes, steps of the same size have to be grouped together, and a term like the former has to be added for each group. If the set of polymorphic loci $\mathscr{L}'$ would be used directly for the computations, there would in general be a large number of different sizes, and the EM algorithm would not be very efficient. However, by using the set $\mathscr{L}$ instead, it is guaranteed that the sizes of the possible steps are all powers of two, and thus the EM algorithm can still be implemented efficiently.

S4.2. **Multi-locus HMM-step handler.** A different approach to reduce the effective number of loci is by combining neighboring loci within a window into "meta"-loci. To this end, assume that a window-size $b \in \mathbb{N}$ is given, and define the set of "meta"-loci as $\mathscr{L}^\star := \{0,\ldots,\lfloor (L-1)/b \rfloor\}$. Mathematically, this approach is equivalent to restricting the hidden states at all loci in the set $\{(l^\star \cdot b)+1,\ldots,(l^\star \cdot b)+b\}$ to be identical, and setting the recombination rates between $(l^\star \cdot b)+b$ and $((l^\star+1)\cdot b)+1$ to $b \cdot \rho$, for each $l^\star \in \mathscr{L}^\star$.

Combining the loci can be implemented as follows. Define modified forward probabilities $F^\star_{l^\star}(s)$ for all $l^\star \in \mathscr{L}^\star$, and compute them according to equation (S46), with modified transition $\phi^\star$ and emission $\xi^\star$ probabilities. The transition probabilities $\phi^\star$ are essentially given by definition (S38), only a recombination rate of $b \cdot \rho$ has to be used instead of $\rho$. At a given locus $l^\star \in \mathscr{L}^\star$, for a given hidden state $s = (i,\omega,x)$, the "allele" of the additional haplotype is $h[(l^\star \cdot b)+1 : (l^\star \cdot b)+b]$ and the "allele" of the trunk haplotype is $x[(l^\star \cdot b)+1 : (l^\star \cdot b)+b]$. Thus, the emission probability at locus $l^\star$ is given by

$$\begin{aligned} & \xi^\star\big(h[(l^\star \cdot b)+1 : (l^\star \cdot b)+b]\big|i,\omega,x[(l^\star \cdot b)+1 : (l^\star \cdot b)+b]\big) \\ & \qquad = \prod_{l\in\{(l^\star\cdot b)+1,\ldots,(l^\star\cdot b)+b\}} \xi\big(h[l]\big|i,\omega,x[l]\big). \end{aligned} \tag{S83}$$

The initial probabilities $\nu$ remain unchanged. The modified backward probabilities $B^{\star}_{l^{\star}}(s)$ and the EM algorithm can be adjusted accordingly.

## S5. Migrating trunk

The unchanging trunk previously described has some drawbacks (also mentioned in the main text). In particular, under the true coalescent, a trunk lineage may absorb with the additional lineage in a sub-population different from the one it resides in at present, due to migration of the trunk lineage. Furthermore, going backwards in time, the rate of absorption of the additional lineage decreases, due to coalescence events within the trunk.

To mitigate these drawbacks, we modify the approximate CSD. Under the model outlined in Section S1, a hidden state $s = (i, \omega, x) \in \mathcal{S}$ represents the event that the additional lineage is absorbed during interval $I_i$, in sub-population $\omega \in \Gamma_i$, into the trunk-lineage $x$. In the modified model, a hidden state $s^{\dagger} = (i, \omega^{\dagger}, x) \in \mathcal{E} \times \Gamma \times \mathbf{n}_{\omega^{\dagger}}$ represents the event that during the interval $I_i$ the lineage of the additional haplotype absorbs into the lineage of the haplotype $x$ that resides in $\omega^{\dagger}$ at the present.

Now, approximate the genealogy relating the haplotypes in the trunk under the true model as follows. First, recall that the number of absorbing lineages in the trunk determines the absorption rates in definition (S4), and they were assumed constant for each sub-population in Section S1. Under the coalescent with migration, these numbers are given by a stochastic process, the ancestral process, that evolves due to coalescence and migration events. Using the full stochastic process is prohibitive, however, Jewett and Rosenberg [4] showed that often times its expected value can be used instead without much loss in accuracy. To this end consider a given epoch $\epsilon \in \mathcal{E}$, with $I_{\epsilon} = [t_{\epsilon-1}, t_{\epsilon})$. The expected number of trunk lineages in each sub-population $\left\{n^{(\epsilon)}_{\gamma}(t_{\epsilon-1})\right\}_{\gamma \in \Gamma_{\epsilon}}$ at the beginning of epoch $\epsilon$ are given by $\{n_{\gamma}\}_{\gamma \in \Gamma}$, if $\epsilon = 1$, and

$$\Big\{ \sum_{\substack{\delta \in \Gamma_{\epsilon-1} \\ \delta \subset \gamma}} n^{(\epsilon-1)}_{\delta}(t_{\epsilon-1}) \Big\}_{\gamma \in \Gamma_{\epsilon}} \tag{S84}$$

otherwise. The dynamics of the expected number of lineages during epoch $\epsilon$ can be approximated by the system of differential equations

$$\frac{d}{dt} n^{(\epsilon)}_{\gamma}(t) = -\frac{1}{\kappa^{(\epsilon)}_{\gamma}} \binom{n^{(\epsilon)}_{\gamma}(t)}{2} \mathbb{1}_{\{n^{(\epsilon)}_{\gamma}(t) > 1\}} + \sum_{\substack{\delta=1 \\ \delta \neq \gamma}}^{|\Gamma_{\epsilon}|} \big( m^{(\epsilon)}_{\delta,\gamma} n^{(\epsilon)}_{\delta}(t) - m^{(\epsilon)}_{\gamma,\delta} n^{(\epsilon)}_{\gamma}(t) \big), \tag{S85}$$

for $\gamma \in \Gamma_{\epsilon}$, c.f. [4, Equation 26]. The additional indicator function in the first summand balances out the fact that a term involving the variance is missing in this approximation. For each epoch $\epsilon \in \mathcal{E}$, these differential equations can be solved numerically. Then, replace $n_{\gamma}$ by $\frac{1}{2}\big(n^{(\epsilon)}_{\gamma}(t_{\epsilon-1}) + n^{(\epsilon)}_{\gamma}(t_{\epsilon})\big)$ for each $\gamma \in \Gamma_{\epsilon}$ in equation (S4), and compute $\tilde{u}$, $\tilde{y}$, $\tilde{z}$, and $\tilde{\xi}$ using these modified absorption matrices in equation (S26), (S39), (S40), and (S44), respectively.

To approximate the effect of the migration dynamics in the trunk on the absorption dynamics of the lineage of the additional haplotype, define $p^{(\epsilon)}(\gamma, \delta)$ as the probability that a lineage residing in sub-population $\gamma$ at present resides in sub-population $\delta$ at the beginning of epoch $\epsilon$. Here $\gamma \in \Gamma$ and $\delta \in \Gamma_{\epsilon}$. Further, let $q^{(\epsilon)}(\gamma, \delta)$ be the corresponding probability at the end of the epoch. If $\epsilon = 1$, then $p^{(\epsilon)}(\gamma, \delta) = \delta_{\gamma,\delta}$. If $\epsilon > 1$, then

$$p^{(\epsilon)}(\gamma, \delta) = \sum_{\substack{\zeta \in \Gamma_{\epsilon-1} \\ \zeta \subset \delta}} q^{(\epsilon-1)}(\gamma, \zeta) \tag{S86}$$

holds. Furthermore,

$$q^{(\epsilon)}(\gamma,\delta) = \sum_{\mu\in\Gamma_\epsilon} p^{(\epsilon)}(\gamma,\mu) \begin{cases} \left(e^{(t_\epsilon - t_{\epsilon-1})M_\epsilon}\right)_{\mu,\delta}, & \text{if } I_\epsilon \neq \emptyset, \\ \left(Y_\epsilon\right)_{\mu,\delta}, & \text{if } I_\epsilon = \emptyset. \end{cases} \tag{S87}$$

Lastly, define the average of $p^{(\epsilon)}$ and $q^{(\epsilon)}$, weighted by the number of haplotypes in a certain sub-population, as

$$r^{(\epsilon)}(\gamma,\delta) := \frac{1}{2}\big(p^{(\epsilon)}(\gamma,\delta) + q^{(\epsilon)}(\gamma,\delta)\big)\cdot n_\gamma, \tag{S88}$$

and define

$$\tilde{r}^{(\epsilon)}(\gamma,\delta) = \frac{1}{\sum_{\mu\in\Gamma} r^{(\epsilon)}(\mu,\delta)} r^{(\epsilon)}(\gamma,\delta) \tag{S89}$$

as the fraction of lineages residing sub-population $\delta$ during epoch $\epsilon$ that reside in sub-population $\gamma$ at present.

Combining the quantities introduced in the previous paragraphs, define the modified initial probabilities as

$$\nu^\dagger(i,\omega^\dagger,x) := \frac{1}{n_{\omega^\dagger}} \underbrace{\sum_{\gamma\in\Gamma_i} \tilde{r}^{(i)}(\omega^\dagger,\gamma)\tilde{u}(i,\gamma)}_{=:\tilde{v}(i,\omega^\dagger)}, \tag{S90}$$

that is, the probability of being absorbed during epoch $i$ into lineage $x$ residing in sub-population $\omega^\dagger$ at present is given by considering the probability of being absorbed in a certain sub-population times the fraction of lineages in that sub-population that are ancestral to lineages in sub-population $\omega^\dagger$ at present, and then summing this over all sub-populations. Along similar lines, define the modified transition probabilities

$$\begin{aligned} &\phi^\dagger(i',\psi^\dagger,x'|i,\omega^\dagger,x) \\ &\quad = \frac{1}{\tilde{v}(i,\omega^\dagger)}\Big(\delta_{i',i}\delta_{\psi^\dagger,\omega^\dagger}\delta_{x',x}\sum_{\gamma\in\Gamma_i}\tilde{r}^{(i)}(\omega^\dagger,\gamma)\tilde{y}(i,\gamma)\tilde{u}(i,\gamma) \\ &\qquad + \frac{1}{n_{\psi^\dagger}}\sum_{\delta\in\Gamma_{i'}}\sum_{\gamma\in\Gamma_i}\tilde{r}^{(i')}(\psi^\dagger,\delta)\tilde{r}^{(i)}(\omega^\dagger,\gamma)\tilde{z}(i',\delta|i,\gamma)\tilde{u}(i,\gamma)\Big). \end{aligned} \tag{S91}$$

and the modified emission probabilities

$$\xi^\dagger(a|i,\omega^\dagger,t) = \frac{1}{\tilde{v}(i,\omega^\dagger)}\sum_{\gamma\in\Gamma_i}\tilde{r}^{(i)}(\omega^\dagger,\gamma)\tilde{\xi}(a|i,\gamma,t)\tilde{u}(i,\gamma). \tag{S92}$$

These probabilities can then be used in appropriately modified versions of the forward, backward, and EM algorithm.

## S6. Decoupling HMM discretization from demographic history

We now describe how to employ a discretization for the HMM computations that differs from the partition induced by the demographic history. To this end, define a partition of the real line into intervals $\{J_j\} := \big\{[t_{j-1},t_j)\big\}$ that is to be used to discretize the HMM. Note that, for convenience, we abuse the subscript notation for $t$ slightly. The hidden states of our HMM are then $(j,\omega^\dagger,x)$, where $j$ is an index of the discretization intervals $\{J_j\}$, and $\omega^\dagger$ and $x$ are as before. We define a third partition

$$\{K_k\} := \bigcup_{\{I_\epsilon\}}\bigcup_{\{J_j\}}\{I_\epsilon\cap J_j\}. \tag{S93}$$

Note that $\{K_k\}$ is a refinement of $\{I_\epsilon\}$ and $\{J_j\}$, that is for all $k$ the inclusion $K_k \subset I_\epsilon$ holds for some $\epsilon$, and $K_k \subset J_j$ for some $j$. In particular, note that the population sizes and migration rates are constant within each refined interval $K_k$. Thus we can work with the "refined" demographic history with epochs $\{K_k\}$ instead of $\{I_\epsilon\}$. Specifically, associate with interval $K_k$ the set of subpopulations $\Gamma_k := \Gamma_\epsilon$ and a migration matrix $M_k := M_\epsilon$, with $\epsilon$ such that $K_k \subset I_\epsilon$. Assign $Y_k$ to the intervals of length zero accordingly.

As in Section S5, compute $\tilde{u}$, $\tilde{y}$, $\tilde{z}$, and $\tilde{\xi}$ according to equation (S26), (S39), (S40), and (S44), respectively, using the modified absorption rates and the discretization $\{K_k\}$. Also, define $\tilde{r}^{(k)}(\gamma, \delta)$ accordingly, using this discretization in equation (S89). Since $\{K_k\}$ is a refinement of $\{J_j\}$, we can then use these quantities to compute the initial $\nu^\dagger$, transition $\phi^\dagger$, and emission $\xi^\dagger$ probabilities for the discretization $\{J_j\}$, analogously to equations (S90), (S91), and (S92). Specifically, replace $\tilde{r}^{(j)}(\omega^\dagger, \gamma)\tilde{u}(j, \gamma)$ in equation (S90) by

$$\sum_{k:K_k \subset J_j} \tilde{r}^{(k)}(\omega^\dagger, \gamma)\tilde{u}(k, \gamma). \tag{S94}$$

In equation (S92), replace $\tilde{r}^{(j)}(\omega^\dagger, \gamma)\tilde{y}(j, \gamma)\tilde{u}(j, \gamma)$ with

$$\sum_{k:K_k \subset J_j} \tilde{r}^{(k)}(\omega^\dagger, \gamma)\tilde{y}(k, \gamma)\tilde{u}(k, \gamma), \tag{S95}$$

and $\tilde{r}^{(i')}(\psi^\dagger, \delta)\tilde{r}^{(i)}(\omega^\dagger, \gamma)\tilde{z}(i', \delta|i, \gamma)\tilde{u}(i, \gamma)$ with

$$\sum_{k':K_{k'} \subset J_{j'}} \sum_{k:K_k \subset J_j} \tilde{r}^{(k')}(\psi^\dagger, \delta)\tilde{r}^{(k)}(\omega^\dagger, \gamma)\tilde{z}(k', \delta|k, \gamma)\tilde{u}(k, \gamma). \tag{S96}$$

Lastly, replace $\tilde{r}^{(j)}(\omega^\dagger, \gamma)\tilde{\xi}(a|j, \gamma, t)\tilde{u}(j, \gamma)$ in equation (S92) by

$$\sum_{k:K_k \subset J_j} \tilde{r}^{(k)}(\omega^\dagger, \gamma)\tilde{\xi}(a|k, \gamma, t)\tilde{u}(k, \gamma). \tag{S97}$$

Using these probabilities all computations for the HMM, and the algorithms based on them, can be computed using the discretization $\{J_j\}$ independent of the partition $\{I_\epsilon\}$ that is used for the demographic history.

## References


[1] Dempster AP, Laird NM, Rubin DB. *J Roy Stat Soc B Met* **39**, 1–38 (1977).
[2] Nelder JA, Mead R. *Comput J* **7**, 308–313 (1965).
[3] Paul JS, Song YS. *Bioinformatics* **28**, 2008–2015 (2012).
[4] Jewett EM, Rosenberg NA. *Theor Popul Biol* **93**, 14–29 (2014).