Running head: MSCSMC

The Sequential Multispecies Coalescent

Analisa Milkey[1], Ming-Hui Chen[2], Yu-Bo Wang[3], Aolan Li[2], and Paul O. Lewis[1]

[1] *Department of Ecology and Evolutionary Biology, University of Connecticut, 75 N. Eagleville Road, Unit 3043, Storrs, Connecticut 06269, U.S.A.*

[2] *Department of Statistics, University of Connecticut, 215 Glenbrook Road, Unit 4120, Storrs, Connecticut 06269, U.S.A.*

[3] *School of Mathematical and Statistical Sciences, Clemson University, Clemson, SC 29634, U.S.A.*

Corresponding author: Analisa Milkey, Department of Ecology and Evolutionary Biology, University of Connecticut, 75 N. Eagleville Road, Unit 3043, Storrs, Connecticut 06269, U.S.A.; Tel: +01 860 486-2069; FAX: +01 860 486-6364; E-mail: `analisa.milkey@uconn.edu`

Abstract

The multispecies coalescent (MSC) model applies coalescent theory to gene evolution within and among reproductively isolated populations ("species") to estimate a species tree in the face of gene tree conflict resulting from deep coalescence. Sequential Monte Carlo (SMC) uses particle filtering to sample a posterior distribution, providing a fully-Bayesian and easily parallelized alternative to traditional MSC tree inference approaches. The method we propose samples first from the joint posterior distribution of gene and species trees, then samples species trees conditional on gene trees sampled previously, employing SMC for both rounds. Analyses of simulated and empirical datasets yield results comparable to state-of-the-art Bayesian MCMC approaches. Sampling the multispecies coalescent using SMC retains the advantages of fully Bayesian methods and is parallelizable in ways that Bayesian MCMC methods are not but also adds unique challenges. We demonstrate the performance of SMC compared to other commonly-used species tree methods using two empirical datasets and 400 simulated datasets.

Keywords: particle filtering, Bayesian, phylogenetics, multispecies coalescent, Sequential Monte Carlo, POSET-SMC

Phylogenetic trees are essential tools in evolutionary biology, and many conclusions rest on their accuracy. Gene histories can differ from species histories due to deep coalescence, paralogy, horizontal gene transfer, and gene tree estimation error, which means many genes are necessary to accurately estimate the history of a set of species (Maddison, 1997) and, ideally, modeling should account for all sources of conflict. The multispecies coalescent model (MSC) (Rannala and Yang, 2003) accounts for deep coalescence (Pamilo and Nei, 1988), a major source of gene tree conflict, applying coalescent theory (Kingman, 1982) to explain the evolution of genes within populations constrained by the history of the reproductively-isolated units (species) to which they belong. In the most common version of the MSC, "species" are defined as completely reproductively isolated populations, so the "species tree" inferred is better termed a population structure tree (Sukumaran and Knowles, 2017); nevertheless, in this paper we will use the terms species and species tree to describe the reproductively-isolated units inferred by the MSC.

Under the basic Wright-Fisher population model (Fisher, 1930; Wright, 1931), generations are non-overlapping, mating is random (including a random amount of selfing), population size (N diploid individuals; $2N$ genes at any given locus) is constant through time, and different loci are unlinked and thus evolve independently. When multiple copies of a given gene are passed along to the next generation, the copying is termed a *coalescence event* when viewed looking backwards in time.

Deep coalescence occurs when two gene lineages sampled from a single species fail to coalesce before that species' origin. Deep coalescence can result in the topology of the gene tree differing from the topology of other gene trees (gene tree conflict) as well as the species tree (Fig.1).

While fully Bayesian implementations of the MSC have considerable advantages (for

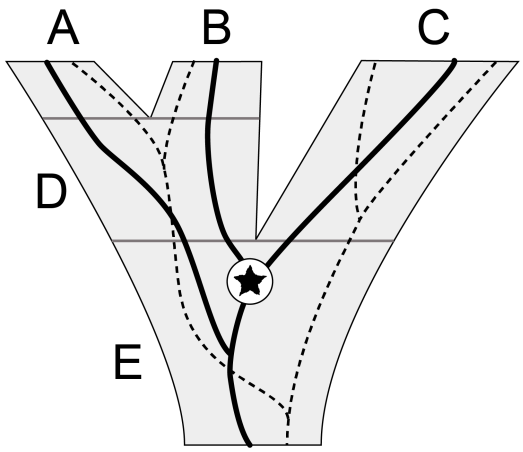


Figure 1: Gene trees for locus 1 (solid black) and locus 2 (dashed lines) embedded within a species tree (shaded). The species tree comprises 3 extant species (A, B, and C) and two ancestral species (D and E). Species boundaries are indicated by horizontal gray bars. The 2 lineages entering at the top of species D in locus 1 failed to coalesce until species E, illustrating deep coalescence (star), which results in the gene tree topology for locus 1 (A,(B,C)) conflicting with the topology ((A,B),C) of both the species tree and the gene tree for locus 2.

example, estimation of gene trees and effective population sizes), they are computationally intensive and struggle to achieve convergence to the stationary posterior distribution on large data sets. StarBEAST3 is a fully Bayesian program that uses Markov Chain Monte Carlo (MCMC) to sample a multispecies coalescent posterior distribution (Heled and Drummond, 2010; Ogilvie et al., 2022; Douglas et al., 2014). Widely used alternatives include ASTRAL (Mirarab and Warnow, 2015), a summary method that takes estimated gene tree topologies as input and estimates the species tree topology from the frequency of different quartets of taxa. SVDQuartets, like ASTRAL, estimates only the species tree topology but has the advantage of taking into account uncertainty in gene trees while still being very computationally efficient (Chifman and Kubatko, 2014). Non-parametric bootstrapping of the original data can be performed with either ASTRAL or SVDQuartets to assess clade confidence. Recently, progress has been made on estimating speciation times in addition to

topology using quartet methods (Peng, Swofford, and Kubatko, 2022), and local posteriors based on the MSC model provide support values for each edge of an ASTRAL species tree (Sayyari and Mirarab, 2016).

MCMC based on the Metropolis-Hastings algorithm is widely used in Bayesian phylogenetics and begins with a complete state (i.e., fully resolved phylogenetic tree with specified edge lengths) and proceeds by proposing local perturbations of the current tree as it runs. MCMC algorithms are challenging to parallelize because each iteration in the algorithm conditions on the previous iteration. Typically, parallelized MCMC approaches to phylogenetics involve splitting up likelihood calculations across processors, placing independent runs on different processors, or placing differently heated chains on different processors in the case of Metropolis-coupled MCMC. StarBEAST3 additionally places updates of gene trees from different loci on different processors.

Sequential Monte Carlo (SMC) or, more specifically, POSET SMC (Bouchard-Côté, 2012, 2014), is a Bayesian alternative to MCMC that uses particle filtering to sample a posterior distribution. In contrast to MCMC, SMC builds up a tree (whose state is stored in a "particle") sequentially, resampling particles at each step based on particle weights that measure improvement in likelihood relative to the previous step (Fig. 2). SMC is naturally parallelizable in several ways that are not possible using the Metropolis-Hastings algorithm.

The filtering steps involved in SMC approaches are analogous to natural selection, and selective sweeps often occur, resulting in one particle's "genome" replacing that of all (or nearly all) other particles, resulting in a condition known as particle degeneracy. Such sweeps lead to low effective sample size (ESS), analogous to the low ESS caused by autocorrelation in MCMC analyses. Thus, while SMC and MCMC both have potential failings, they are sufficiently different approaches that it is worthwhile exploring how well SMC can compete

with MCMC-based MSC.

The SMC approach to the MSC model that we describe here differs from existing non-MCMC methods (ASTRAL, SVDQuartets) in its ability to deliver a sample from the posterior distribution of species trees (rather than a single point estimate) while sharing with these methods the ability to scale to datasets with many loci. Like Metropolis-Hastings MCMC methods, SMC is an approximation whose accuracy depends on the number of particles and the nature of the proposal distributions used. While Bouchard-Côté (2012) showed that, given unlimited computational resources, Metropolis-Hastings approaches can achieve higher accuracy than SMC, SMC can deliver reasonable results on a more practical time scale.

Materials and Methods

Our approach employs SMC hierarchically. At the first level, SMC is used to obtain a sample from the joint posterior distribution of the species tree and gene trees from all loci. Species tree marginal distributions resulting from joint estimation typically suffer from particle degeneracy, and thus we employ a second level SMC to sample from the species tree posterior distribution conditional on the gene trees sampled during the lower-level SMC.

We first describe sampling from the joint posterior distribution (first level SMC) in the section entitled *SMC for Joint Gene and Species Tree Sampling*, then discuss the sampling from the conditional posterior (second level SMC) in the section *SMC for Species Trees Given Gene Trees*.

SMC for Joint Gene and Species Tree Sampling

The joint posterior distribution for the simplest multispecies coalescent model (i.e., constant population size and speciation rate) can be written:

$$p(\boldsymbol{\mathcal{G}}, \mathcal{S}, \boldsymbol{\theta}, \lambda | \mathbf{D}) \propto p(\mathbf{D}|\boldsymbol{\mathcal{G}})\ p(\boldsymbol{\mathcal{G}}|\mathcal{S}, \boldsymbol{\theta})\ p(\mathcal{S}|\lambda)\ p(\boldsymbol{\theta}|\bar{\theta})\ p(\lambda), \tag{1}$$

where:

$\mathbf{D}$ is a vector of aligned sequence data sets $\{\mathbf{D}_l : l = 1, \cdots, L\}$ for L loci. Data set $\mathbf{D}_l$ comprises sequences of length n_l sites from each of T sampled genes from locus l;

$\boldsymbol{\theta}$ is a vector of mutation-scaled population size parameters (Watterson, 1975) $\{\theta_b : b = 1, \cdots, 2M-1\}$, where M is the number of extant species ($M \leq T$). Each edge b of species tree $\mathcal{S}$ is associated with one element $\theta_b = 4N_b\mu_b$ of this vector, where N_b and μ_b are the effective (diploid) population size and mutation rate, respectively, for edge b;

$\boldsymbol{\mathcal{G}}$ is a vector of gene trees $\boldsymbol{\mathcal{G}} = \{\boldsymbol{\mathcal{G}}_l : l = 1, \cdots, L\}$, where each component gene tree $\boldsymbol{\mathcal{G}}_l$ comprises $T-1$ increments and joins: $\boldsymbol{\mathcal{G}}_l = \{(\delta_t, \xi_t) : t = 1, \cdots, T-1\}$;

$\mathcal{S}$ is the species tree, comprising $M-1$ increments and joins: $\mathcal{S} = \{(\Delta_m, \Xi_m) : m = 1, \cdots, M-1\}$;

$p(\mathbf{D}|\boldsymbol{\mathcal{G}})$ is the product of *Felsenstein likelihoods* (Felsenstein, 1981) computed using a substitution model (e.g., JC69; Jukes and Cantor, 1969) on each gene tree in $\boldsymbol{\mathcal{G}}$;

$p(\boldsymbol{\mathcal{G}}|\mathcal{S}, \boldsymbol{\theta})$ is the *coalescent likelihood* (Rannala and Yang, 2003), which is the probability density of the vector of gene trees $\boldsymbol{\mathcal{G}}$ conditioned on species tree $\mathcal{S}$ and $\boldsymbol{\theta}$;

$p(\mathcal{S}|\lambda)$ is the *species tree prior*, assumed in this paper to be a pure-birth Yule model (Yule, 1925), where λ is the birth rate;

$p(\boldsymbol{\theta}|\bar{\theta})$ is the prior for the value of θ associated with each individual species, conditional on the assumed mean ($\bar{\theta}$); and

$p(\lambda)$ is the prior on the speciation rate λ, i.e., the single parameter of the Yule model.

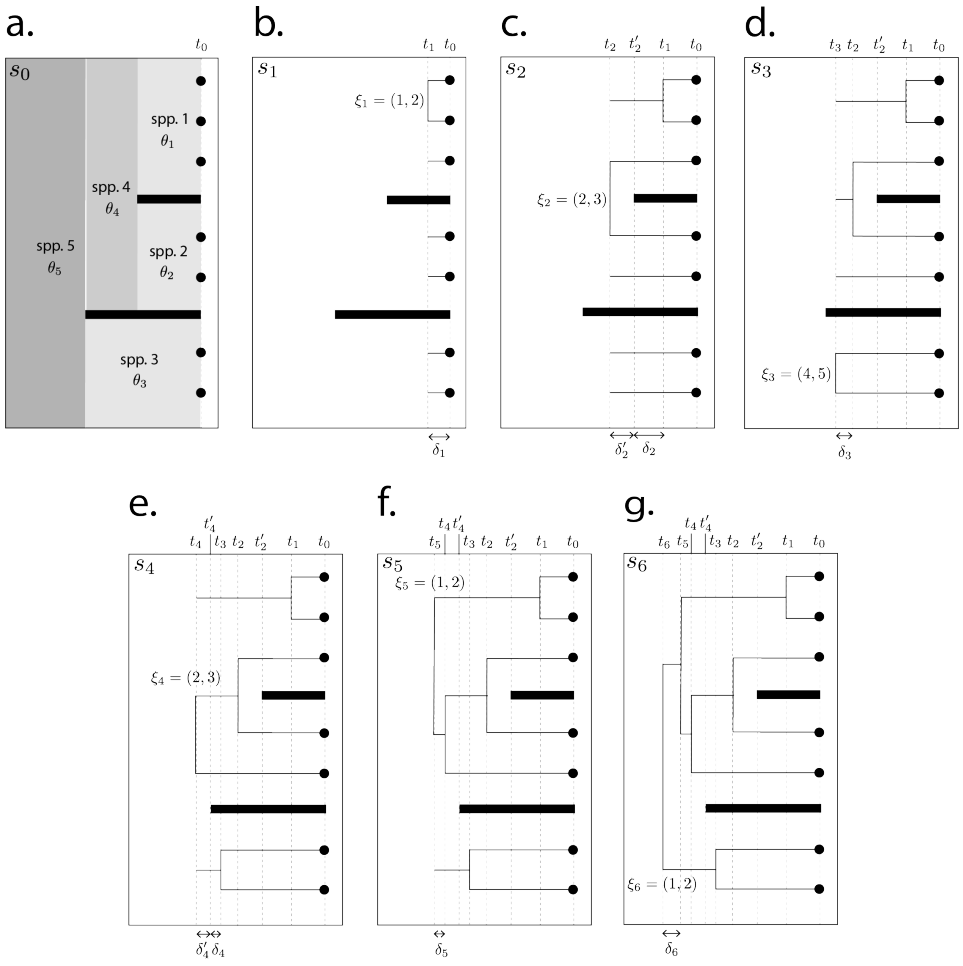


Figure 2: Species tree and growth of gene forest for one locus in one particle. Thick black lines represent species tree barriers to gene flow. Notation simplified by omitting locus and particle subscripts.

Proposing new states.— The term *forest* is used for partial states comprising sets of

disjoint, ultrametric subtrees, with each subtree having its own root vertex. The disjoint union of the leaves of the component subtrees yields the set of all sampled genes (for gene forests) or extant species (if the species forest). The term *tree* is reserved for complete states (i.e., a forest composed of just one subtree).

Each particle k ($k = 1, \cdots, K$) begins with a vector of trivial gene forests $\boldsymbol{\mathcal{G}}_k = \{\boldsymbol{\mathcal{G}}_{kl} : l = 1, \cdots, L\}$. A trivial forest (state s_0) comprises only leaf nodes and has height 0.0. A value for $\bar{\theta}_k$ is drawn from its prior distribution. For each particle k, a complete species forest $\mathcal{S}_k$ is drawn from its Yule prior. For each edge b in $\mathcal{S}_k$, a mutation-scaled population size parameter value θ_{kb} is drawn from an InverseGamma $(2, \bar{\theta}_k)$ prior (Fig. 2a).

The remainder of this section describes proposing state i given state $i-1$ in the gene forest for locus l in particle k; however, the subscript k is omitted to simplify notation. Each step i results in state i having one more coalescence event than state $i-1$. Proposing one coalescence event for all L loci in a single particle thus involves L steps. Loci are visited in randomized order in each round of L steps.

Let $t = t_{i-1}$ be the time at the start of step i ($i > 0$). A time increment δ_{li} is drawn from the coalescent prior distribution for locus l, which has rate r_{li},

$$r_{li} = \sum_{j=1}^{M(t)} \frac{n_{lj}(t)(n_{lj}(t)-1)}{\theta_j} \tag{2}$$

$$\delta_{li} \sim \text{Exp}(r_{li}), \tag{3}$$

where $n_{lj}(t)$ is the number of uncoalesced gene tree lineages in existence at time t in species j and at locus l, and $M(t)$ is the number of species at time t.

Let $\tau_1 < \cdots < \tau_{M-1}$ be the heights of nodes representing speciation events in the

species tree, and let $t'_i = \min\{\tau_m : \tau_m \geq t, m = 1, \cdots, M-1\}$. If $t + \delta_{li} \geq t'_i$ (e.g., Fig. 2c,e), then all lineages in gene forest l are advanced to time t'_i and another increment δ'_{li} is drawn from an Exponential distribution whose rate now reflects the gene forest lineages merged as a result of the speciation event at time t'_i. If $t + \delta_{li} < t'_i$ (e.g., Fig. 2b,d,f,g), all lineages in gene forest l are advanced by an amount δ_{li} and the species in which the next coalescent event occurs is determined using a draw from the multinomial probability distribution having parameters

$$p_{lij} = \frac{n_{lj}(t)(n_{lj}(t) - 1)}{\theta_j r_{li}}, \tag{4}$$

where $j = 1, \cdots, M(t)$ and r_{li} is the normalizing constant.

If species j is chosen, then two lineages ξ_{li} from species j in gene forest l are selected randomly from a Discrete Uniform distribution and joined. The new gene forest ancestral node is assigned to the same species as its two descendant lineages. The construction of state i is now complete.

Each step of this joint estimation SMC algorithm results in a new state. The number of steps S (and thus states) is thus the total number of coalescent events over all gene trees:

$$S = \sum_{l=1}^{L} (n_l - 1). \tag{5}$$

Particle weights.— Particle weights are calculated as the ratio of the product of the gene forest likelihoods after a coalescent event to the product of the gene forest likelihoods before a coalescent event (all prior terms in the numerator cancel with proposal terms in the denominator, leaving only the likelihood ratio).

The weight w_i for any given particle at state s_i $(i > 0)$ is

$$w_i = \frac{p(s_i|\mathbf{D}, s_{i-1})}{q(s_i|s_{i-1})} \tag{6}$$

$$p(s_i|D, s_{i-1}) = \frac{p(\mathbf{D}|\delta_{..i}, \xi_{..i})}{p(\mathbf{D}|\delta_{..(i-1)}, \xi_{..(i-1)})}\, p(\delta_i, \xi_i|\mathcal{S}, s_{i-1}) \tag{7}$$

$$q(s_i|s_{i-1}) = p(\delta_i, \xi_i|\mathcal{S}, s_{i-1}). \tag{8}$$

Note that we have suppressed the particle index k from the notation for readability, and use the shorthand notation $x_{..i} = x_1 \cdots x_i$ for products of similar terms. The weight can be viewed as an importance weight in the context of importance sampling, where the conditional prior $q(s_i|s_{i-1})$ represents the importance density (Bouchard-Côté, 2014).

An exception is w_1, which must take into account the prior probability of the species tree, which was simulated from the Yule prior when particles were first initialized:

$$w_1 = \frac{p(s_0, s_1|\mathbf{D})}{q(s_0, s_1)} = \frac{p(s_1|\mathbf{D})\; p(\Delta, \Xi|\lambda)\; p(\boldsymbol{\theta}|\bar{\theta})}{q(s_1)\; q(\Delta, \Xi|\lambda)\; q(\boldsymbol{\theta}|\bar{\theta})}. \tag{9}$$

Because of the cancellation of prior terms with proposal terms, the weight w_i simplifies to

$$w_i = \frac{p(\mathbf{D}|\delta_{..i}, \xi_{..i})}{p(\mathbf{D}|\delta_{..(i-1)}, \xi_{..(i-1)})}. \tag{10}$$

UPGMA gene tree completion.— The particle weight used in practice differs from (10) in one significant way. The weight defined in (10) sometimes leads to poor choices in the particle filtering stage (see below) because it lacks foresight. That is, a coalescence event proposed at step i may lead to a large improvement in likelihood at step i, but, in the

final gene tree, results in a lower likelihood than if a different join had been made at step i. Bouchard-Côté (2012) described an alternative proposal where trees in a partial state are (temporarily) completed using a fast, deterministic approach such as neighbor-joining (NJ). After the weight is calculated using this temporary complete state, the portion of the tree completed using NJ is removed.

We implement this approach using a proposal where gene forests are completed using a UPGMA algorithm (Sokal and Michener, 1958) with Jukes-Cantor distances (Jukes and Cantor, 1969). We use UPGMA rather than NJ to maintain ultrametricity. The likelihoods in both numerator and denominator of the particle weight are thus based on complete gene trees in which state s_i is embedded. This allows SMC to "look ahead" and leads to better-informed choices at each step. The weight including this modification is

$$w_i = \frac{p(\mathbf{D}|\delta_{..i}, \xi_{..i}, \delta_{i+}, \xi_{i+})}{p(\mathbf{D}|\delta_{..(i-1)}, \xi_{..(i-1)}, \delta_{(i-1)+}, \xi_{(i-1)+})}, \tag{11}$$

where δ_{i+} and ξ_{i+} are the increments and joins added to state i using the UPGMA algorithm and $\delta_{(i-1)+}$ and $\xi_{(i-1)+}$ are the increments and joins added to state $i-1$ using the UPGMA algorithm.

Filtering.— Filtering of particles takes place after a coalescent event has been proposed in each particle for a given locus. Filtering is performed using multinomial sampling with normalized weights as bin probabilities. If a particle is selected, its species tree $\mathcal{S}$ (including $\boldsymbol{\theta}$) as well as its vector of gene forests $\boldsymbol{\mathcal{G}}$ is copied to the new particle generation.

A selective sweep may result in one particle (with its associated species tree) replacing all other particles. At this point, the part of the species tree beyond the deepest coalescent event in any locus has not been influenced by the data, yet all future proposals will be

constrained by this species tree. Thus, after each locus has undergone filtering, the species tree in each particle is trimmed back to the deepest coalescent event across all loci in that particle and rebuilt from that point on, with new values of θ drawn for each new species population. This reintroduces variation in species trees across particles.

Why is filtering performed after gene forests in a single locus have been updated rather than after all loci have been considered? If loci differ in sequence length, a locus with a longer sequence will often achieve a larger difference in likelihood compared to a locus with a shorter sequence length. Filtering loci independently ensures that loci with longer sequences do not exert undo dominance in determining particle populations, especially at early stages.

SMC for Species Trees Given Gene Trees

The first-level SMC described in the previous section often results in a marginal posterior species tree sample that is dominated by very few distinct species trees. This is because the Felsenstein likelihood is only indirectly influenced by the species tree hyperparameter through its effect on gene tree edge lengths and joins. In contrast, the Felsenstein likelihood directly affects the filtering of gene forests at each stage. We thus resample the marginal species tree posterior in the second-level SMC, conditioning on the gene trees sampled in the first-level SMC.

Using SMC to estimate the distribution of species trees conditional on complete gene trees and $\bar{\theta}$ (mean mutation-scaled population size θ) is very similar to the method described in detail by Bouchard-Côté (2012) and Bouchard-Côté (2014), with the primary difference being that the likelihood function is the integrated coalescent likelihood (Jones, 2017) rather

than the Felsenstein (1981) likelihood.

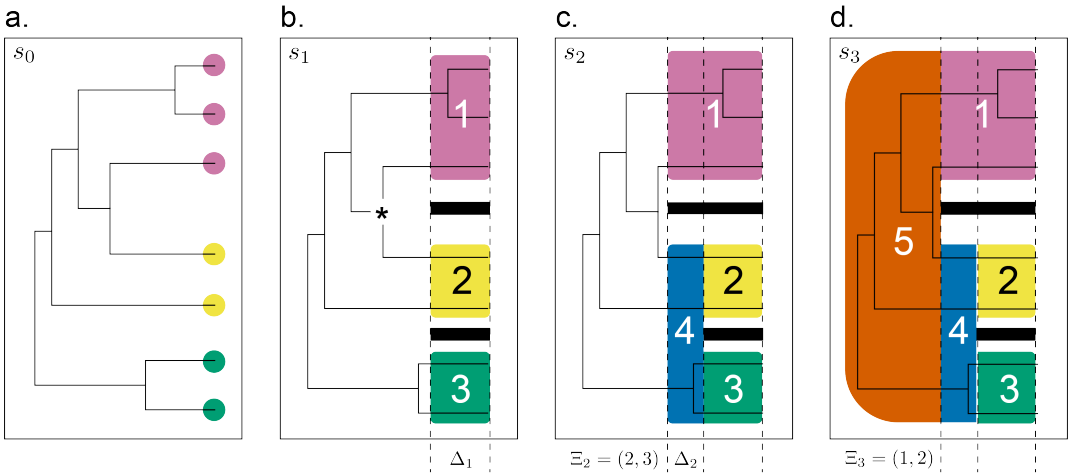


Figure 3: Growth of species forest ($M = 3$ species), constrained by a gene tree from one locus in a single particle. a. trivial state s_0. b. partial state s_1. c. partial state s_2. d. complete state s_3. Black bars represent reproductive isolation barriers separating distinct species, numbers indicate distinct species, and the asterisk (*) denotes the coalescence event that marks the maximum possible value of Δ_1 (and also Δ_2 given that Ξ_2 joined species 2 and 3).

If complete gene trees are available, the conditional posterior distribution of species trees does not require calculation of the Felsenstein likelihood,

$$p(\mathcal{S}|\boldsymbol{\mathcal{G}}, \bar{\theta}, \lambda) = \frac{p(\boldsymbol{\mathcal{G}}, \mathcal{S}|\bar{\theta}, \lambda)}{p(\boldsymbol{\mathcal{G}}|\bar{\theta}, \lambda)} \tag{12}$$

$$= \frac{\left[\int p(\boldsymbol{\mathcal{G}}|\mathcal{S}, \boldsymbol{\theta})\ p(\boldsymbol{\theta}|\bar{\theta})\ d\boldsymbol{\theta}\right] p(\mathcal{S}|\lambda)}{p(\boldsymbol{\mathcal{G}}|\bar{\theta}, \lambda)} \tag{13}$$

$$\propto p(\boldsymbol{\mathcal{G}}|\mathcal{S}, \bar{\theta})\ p(\mathcal{S}|\lambda), \tag{14}$$

where $p(\boldsymbol{\mathcal{G}}|\mathcal{S}, \bar{\theta})$ is the *integrated coalescent likelihood* (Jones, 2017). Because this distribution involves only the integrated coalescent likelihood and not the Felsenstein likelihood, sampling is much less computationally demanding.

Increments and joins.— The second-level SMC begins with K^* particles, each having L complete gene trees and a species forest in the trivial state (s_0), consisting of only the M leaf vertices, each having height 0.0 (Fig. 3a). The vector of gene trees is sampled randomly from the first-level particle population and is shared by every particle in the second-level analysis.

For example, a first-level analysis might use $K = 1000$ particles, of which 10% are sampled for use in the second-level analysis. If $K^* = 200$ particles are used for the second-level analysis, then a total of $(0.1K)K^* = (0.1 * 1000) * 200 = 20000$ species trees would constitute the second-level posterior sample, with each of $0.1K = 0.1 * 1000 = 100$ gene tree vectors retained from the first-level SMC forming the basis for an independent second-level SMC involving K^* particles.

The state of every particle is advanced from the trivial (species forest) state through a series of partial states to a complete state via a series of proposals. A weight is computed for a proposed new state and particles are filtered by drawing K^* particles with replacement from a multinomial distribution in which the bin probabilities are the normalized particle weights.

There are $j = M - i + 2$ lineages before state i is proposed ($i = 2, \cdots, M$). The transition from partial state s_{i-1} to partial state s_i involves first choosing a pair Ξ_i of existing lineages to join, with probability $\binom{j}{2}^{-1}$, and then a height increment Δ_i from the Exponential($j\lambda$) prior distribution:

$$s_i = (\Xi_i, \Delta_i), i = 1, \cdots, M \tag{15}$$

$$\Xi_1 = \emptyset \tag{16}$$

$$\Delta_M = \infty. \tag{17}$$

The first ($i = 1$) and last ($i = M$) steps represent exceptions: (1) two lineages are not joined in creating partial state s_1 because there is no reason to assume that the most recent speciation event occurred exactly at time 0 and (2) the final increment is, necessarily, ∞ and thus is not a random variable; hence, Ξ_1 is the empty set, and $\Delta_M = \infty$. In all partial states s_i ($i < M$), we implicitly assume an *ephemeral* ancestral species, which extends from $\tau_i = \sum_{j \leq i} \Delta_j$ to ∞. This is necessary because species forests are conditioned on complete gene *trees* (not partial-state gene *forests*).

To illustrate why Ξ_i is chosen before Δ_i in each step, consider state s_1 in Figure 3b. The coalescent likelihood must account for 1 coalescent event in species 1 (top) and 0 coalescent events in both species 2 and 3 (middle and bottom) during the time interval $(0, \Delta_1)$. It must also account for the remaining 5 coalescent events in the history of the 6 lineages that exist at time Δ_1. These 6 lineages and their ancestors are all members of the ephemeral ancestral species. Joining two lineages after choosing the increment Δ_1 would therefore have no effect on the coalescent likelihood because such a join would be associated with a time interval of zero length between the join and the start of the ephemeral species. The consequences of such a join would not be realized until the *next* step, at which point there is no longer any opportunity to make the particle pay for a poor join decision. Thus, joins always follow increments when constructing species forests in the second-level SMC.

In general, the number of loci is greater than 1, so the species forest constructed within each particle is conditioned on gene trees from more than one locus. The coalescent events within gene trees place constraints on the maximum value that any given species tree increment can attain. For example, the increment Δ_1 in Figure 3b must be less than or equal to the time of the coalescence event indicated by the asterisk (*). This is the first coalescence event (over all loci) where lineages from two distinct species join. Extending Δ_i further back

in time than this gene tree node would imply gene flow across species boundaries, which is not allowed under the multispecies coalescent model. Such constraints lead to sampling efficiencies even if the prior on species tree increments is vague.

Particle weights.— As for the first-level SMC, the weight w_k for particle k ($k = 1, \cdots, K$) at state s_i can be viewed as an importance weight in the context of importance sampling:

$$w_k = \frac{p(s_i|s_{i-1}, \boldsymbol{\mathcal{G}}, \bar{\theta}, \lambda)}{q(s_i|s_{i-1}, \boldsymbol{\mathcal{G}}, \lambda)} = \frac{p(s_i|\boldsymbol{\mathcal{G}}, \bar{\theta}, \lambda)}{p(s_{i-1}|\boldsymbol{\mathcal{G}}, \bar{\theta}, \lambda)\ q(s_i|s_{i-1}, \boldsymbol{\mathcal{G}}, \lambda)}, \tag{18}$$

where $p(s|\boldsymbol{\mathcal{G}}, \bar{\theta}, \lambda)$ is the posterior probability density of state s and $q(s_i|s_{i-1}, \boldsymbol{\mathcal{G}}, \lambda)$ is the importance density, which, in this case, equals the proposal density for Δ_i and Ξ_i given the previous state s_{i-1}:

$$p(s_i|\boldsymbol{\mathcal{G}}, \bar{\theta}, \lambda) = \frac{p(\boldsymbol{\mathcal{G}}|s_i, \bar{\theta})\ p(s_i|\lambda)}{p(\boldsymbol{\mathcal{G}}|\bar{\theta}, \lambda)} \tag{19}$$

$$p(s_{i-1}|\boldsymbol{\mathcal{G}}, \bar{\theta}, \lambda) = \frac{p(\boldsymbol{\mathcal{G}}|s_{i-1}, \bar{\theta})\ p(s_{i-1}|\lambda)}{p(\boldsymbol{\mathcal{G}}|\bar{\theta}, \lambda)} \tag{20}$$

$$q(s_i|s_{i-1}, \boldsymbol{\mathcal{G}}, \lambda) = p(\Delta_i|s_{i-1}, \lambda, \boldsymbol{\mathcal{G}})\ p(\Xi_i|s_{i-1}). \tag{21}$$

The multispecies coalescent likelihood can be computed piecewise by species tree edge. Figure 4 illustrates the calculation of the coalescent likelihood for locus l on edge b of the species tree. If edge b is one of the edges added to construct state i, then $\Delta_i = \sum_{j=0}^{k_{lb}} c_{lbj}$.

Jones (2017) observed that the coalescent likelihood takes the form of an InverseGamma(q_b-

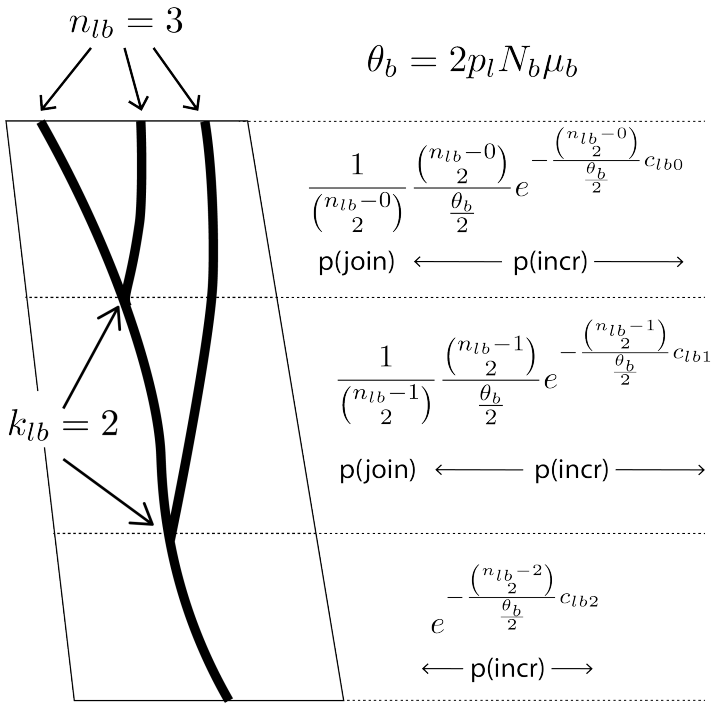


Figure 4: Computation of the multispecies coalescent likelihood for edge b of the species tree at locus l. The k_{lb} coalescent events partition the edge length into intervals of length c_{lb0}, c_{lb1}, and c_{lb2}. n_{lb} and $n_{lb} - k_{lb}$ are the number of lineages entering and leaving the edge, respectively.

1, γ_b) distribution for each edge b of the species tree:

$$p(\boldsymbol{\mathcal{G}}|\boldsymbol{\theta}) = \prod_b \frac{r_b}{\theta_b^{q_b}} e^{-\gamma_b/\theta_b}, \tag{22}$$

where

$$r_b = \prod_{l=1}^{L} \left(\frac{4}{p_l}\right)^{k_{lb}} \tag{23}$$

$$q_b = \sum_{l=1}^{L} k_{lb} \tag{24}$$

$$\gamma_b = \sum_{l=1}^{L} \sum_{j=0}^{k_{lb}} \frac{4\binom{n_{lb}-j}{2}}{p_l} c_{lbj}, \tag{25}$$

b indexes edges (i.e., species) in the species tree, l indexes loci, i indexes coalescent events within species b in gene tree $\boldsymbol{\mathcal{G}}_l$, p_l is the ploidy level of locus l (e.g., $p_l = 1$ for a plastid or mitochondrial locus and $p_l = 2$ for nuclear loci in diploid organisms), c_{lbj} is the jth time interval in gene tree $\boldsymbol{\mathcal{G}}_l$ for edge b, and n_{lb} is the number of lineages in gene tree $\boldsymbol{\mathcal{G}}_l$ at the start of edge b. Note that our formulas for r_b and γ_b contain the factor 4 because the Watterson (1975) definition of θ_b that we use ($\theta_b = 4N_b\mu_b$) differs from Jones (2017), who used $\theta_b = N_b\mu_b$, where N_b is the effective population size and μ_b the mutation rate specific to edge b.

Assuming an InverseGamma(α, β) prior distribution for θ_b allows θ_b to be analytically integrated out of the coalescent likelihood (Jones, 2017):

$$p(\boldsymbol{\mathcal{G}}|\Delta_{\cdot i}, \Xi_{\cdot i}, \bar{\theta}) = \prod_b r_b \frac{\beta^\alpha}{(\beta+\gamma_b)^{\alpha+q_b}} \frac{\Gamma(\alpha+q_b)}{\Gamma(\alpha)}. \tag{26}$$

We assume $\alpha = 2$, $\beta = \bar{\theta}$, where the parameter $\bar{\theta}$ specifies the mean value of θ among species.

For the Yule pure-birth tree model, the proposal density for Δ_i is

$$p(\Delta_i|s_{i-1}, \lambda, \boldsymbol{\mathcal{G}}) = \frac{\lambda(n_t - \mathbf{1}_{i>1})e^{-\lambda n_t \Delta_i}}{1 - e^{-\lambda(n_t - \mathbf{1}_{i>1})t_{\max}}}, \tag{27}$$

where n_t is the number of species tree lineages at time t, $t = \sum_{j<i} \Delta_j$ denotes the height of the species forest at state s_{i-1}, $t_{\max}$ is the upper bound for Δ_i determined by $\boldsymbol{\mathcal{G}}$, and $\mathbf{1}_{i>1}$ is an indicator variable that is 1 if $i > 1$ and 0 if $i = 1$.

The join proposal probability is

$$p(\Xi_i|s_{i-1}) = \frac{1}{\binom{n_t - \mathbf{1}_{i>1}}{2}}. \tag{28}$$

Cancellation of prior terms in the numerator with proposal terms in the denominator results in the following particle weight specification:

$$\begin{aligned}
w_k &= \frac{p(s_i|s_{i-1}, \boldsymbol{\mathcal{G}}, \bar{\theta}, \lambda)}{q(s_i|s_{i-1}, \boldsymbol{\mathcal{G}}, \lambda)} && (29)\\
&= \frac{p(s_i, s_{i-1}|\boldsymbol{\mathcal{G}}, \bar{\theta}, \lambda)}{p(s_{i-1}|\boldsymbol{\mathcal{G}}, \bar{\theta}, \lambda)\; q(s_i|s_{i-1}, \boldsymbol{\mathcal{G}}, \lambda)} && (30)\\
&= \frac{\frac{p(\boldsymbol{\mathcal{G}}|s_i, \bar{\theta})\; p(s_i|\lambda)}{\cancel{p(\boldsymbol{\mathcal{G}}|\theta, \lambda)}}}{\frac{p(\boldsymbol{\mathcal{G}}|s_{i-1}, \bar{\theta})\; p(s_{i-1}|\lambda)\; q(s_i|s_{i-1}, \boldsymbol{\mathcal{G}}, \lambda)}{\cancel{p(\boldsymbol{\mathcal{G}}|\theta, \lambda)}}} && (31)\\
&= \frac{p(\boldsymbol{\mathcal{G}}|s_i, \bar{\theta})\; \cancel{p(s_{i-1}|\lambda)}\; p(s_i|s_{i-1}, \lambda)}{p(\boldsymbol{\mathcal{G}}|s_{i-1}, \bar{\theta})\; \cancel{p(s_{i-1}|\lambda)}\; q(s_i|s_{i-1}, \boldsymbol{\mathcal{G}}, \lambda)} && (32)\\
&= \frac{p(\boldsymbol{\mathcal{G}}|s_i, \bar{\theta})\; \cancel{p(s_i|s_{i-1}, \lambda)}}{p(\boldsymbol{\mathcal{G}}|s_{i-1}, \bar{\theta})\; \frac{\cancel{p(s_i|s_{i-1}, \lambda)}}{1 - e^{-\lambda(n_t - \mathbf{1}_{i>1})t_{\max}}}} && (33)\\
&= \frac{p(\boldsymbol{\mathcal{G}}|s_i, \bar{\theta})}{p(\boldsymbol{\mathcal{G}}|s_{i-1}, \bar{\theta})} \left(1 - e^{-\lambda(n_t - \mathbf{1}_{i>1})t_{\max}}\right). && (34)
\end{aligned}$$

The first term is the ratio of the integrated coalescent likelihood for state s_i to the integrated coalescent likelihood for state s_{i-1}. The second term is the normalizing constant of the truncated Exponential proposal distribution for increment Δ_i.

Particle filtering.— After parameters Δ_i and Ξ_i are proposed and weights are determined for each particle, the weights are normalized and multinomial sampling is used to

draw K^* new particles using the normalized weight $\tilde{w}_k = w_k / \sum_k w_k$ as the probability for bin k. After the filtering step, the particle population represents a sample from the posterior distribution of state $s_i = \{\Delta_{\cdot i}, \Xi_{\cdot i}\}$ conditional on $\bar{\theta}$ and gene trees $\boldsymbol{\mathcal{G}}$.

Simulations

We performed simulations to compare the performance of the SMC approach to the Bayesian MCMC approach of StarBEAST3 with respect to accuracy of the species tree topology. At every point in a 20 by 20 grid, data were simulated for 10 loci, 5 species, and 2 sampled individuals per species. The 20 grid rows corresponded to evenly-spaced values of T, the species tree height, from 0.0 to 0.5. The 20 grid columns corresponded to evenly-spaced values of $\theta/2$, from 0.0 to 0.3. θ was fixed across all species within a species tree. Smaller values of T and larger values of $\theta/2$ yield greater expected deep coalescences, and thus those areas of the grid present more challenges to species tree methods.

Species trees were simulated using a pure-birth Yule (1925) model with speciation rate λ. Gene trees were simulated within the species tree from the prior. Sequences between 200 and 1000 nucleotides were simulated under the Jukes-Cantor substitution model with equal rates among sites and loci.

SMC analyses used $K = 10000$ particles for the first level and $K^* = 500$ particles for the second level. A randomly-chosen 2.5% of particles from the first-level analysis were used for the second-level analysis. A randomly-chosen 0.8% of second-level particles were saved, yielding a sample of size 1000 species trees. StarBEAST3 analyses used 15 million iterations, saving every 15000 for a total sample size of 1000 species trees. Burn-in was set to 1.5 million iterations.

Analyses using both SMC and StarBEAST3 assumed a fixed speciation rate determined from the tree length estimated by the QAGE method (Peng, Swofford, and Kubatko, 2022). Note that the simulations assumed θ was constant throughout the species tree, whereas both SMC and StarBEAST3 allowed θ to vary among species; however, both assumed constant θ within a species. The mean θ used by both SMC and StarBEAST3 analyses was also estimated by the QAGE method. The definition of θ differed between SMC and StarBEAST3. StarBEAST3 defines $\theta = N_e\mu$ whereas SMC assumes the Watterson (1975) mutation-scaled population size definition, $\theta = 4N_e\mu$. To make analyses comparable, we thus set StarBEAST3 mean θ to 1/4 that of SMC.

The Robinson-Foulds (RF; Robinson and Foulds, 1981) distances between each sampled species tree and the true species tree used for simulation were averaged to provide a measure of species tree topology accuracy.

Empirical Data

We also explored the performance of our method on two datasets, one containing gopher species and one containing snake species. The gopher dataset consists of a subset of data from Belfiore et al. (2008). The data contain 6 loci sampled from 27 individuals representing 10 species of pocket gophers. The 9 ingroup species belong to the genus *Thomomys*, and the outgroup belongs to the genus *Orthogeomys*. The snake dataset consists of a subset of data from Chifman and Kubatko (2014). The data contain 15 loci sampled from 52 individuals representing 7 species of snakes. The original data set contained 19 loci; we removed 4 loci in which one or more taxa had completely missing data. The 6 ingroup species belong to the genus *Sistrurus*, pygmy rattlesnakes, and the outgroup species belongs to the genus *Agkistrodon*. We chose these datasets because they are well-studied, making them good

measures of program performance. They are also small enough that it is feasible for multiple programs, even computationally intensive programs, to run them in a reasonable amount of time.

Gopher dataset.— We estimated a species tree using StarBEAST3 (Douglas et al., 2014) and SMC. In both programs, we used a Jukes Cantor (Jukes and Cantor, 1969) substitution model and a fixed speciation rate determined from the tree length estimated by QAGE (Peng, Swofford, and Kubatko, 2022). Mean θ was also estimated by the QAGE method. We ran StarBEAST3 for 20 million generations, with 2 million generations of burn-in, saving trees every 10000 generations for a total of 2000 species trees sampled. We allowed each locus to have a different relative rate of substitution with default prior LogNormal(1.0, 0.6). SMC analyses used $K = 10000$ particles for the first level and $K^* = 1000$ particles for the second level. A randomly-chosen 2.5% of particles from the first-level analysis were used for the second-level analysis. A randomly-chosen 0.8% of second-level particles were saved, yielding a sample of size 2000 species trees. We fixed relative rates for each locus according to estimates under a site-specific model implemented in PAUP*. We conducted 10 independent runs in both StarBEAST3 and SMC and combined the resulting species trees, yielding a sample 20,000 species trees from each program.

We also estimated species trees using SVDQuartets and ASTRAL IV (as implemented in ASTER (Zhang & Mirarab, 2022; Tabatabaee et al., 2023)). The SVDQuartets analysis comprised 1000 bootstrap replicates, and the tree was rooted at outgroup *O. heterodus*. For the ASTRAL analysis, we used maximum likelihood gene trees estimated from IQTREE (Nguyen et al., 2015), using a Jukes Cantor model for each locus. We rooted the tree at outgroup *O. heterodus*.

Snake dataset.— We estimated a species tree using StarBEAST3 (Douglas et al., 2014)

and SMC. In both programs, we used a Jukes Cantor (Jukes and Cantor, 1969) substitution model and a fixed speciation rate determined from the tree length estimated by QAGE (Peng, Swofford, and Kubatko, 2022). Mean θ was also estimated by the QAGE method. We ran StarBEAST3 for 100 million generations, with 10 million generations of burn-in, saving trees every 50000 generations for a total of 2000 species trees sampled. We allowed each locus to have a different relative rate of substitution with default prior LogNormal(1.0, 0.6). SMC analyses used $K = 15000$ particles for the first level and $K^* = 1000$ particles for the second level. A randomly-chosen 1.67% of particles from the first-level analysis were used for the second-level analysis. A randomly-chosen 0.8% of second-level particles were saved, yielding a sample of size 2000 species trees. We fixed relative rates for each locus according to estimates under a site-specific model implemented in PAUP*. We conducted 10 independent runs in both StarBEAST3 and SMC and combined the resulting species trees, yielding a sample of 20,000 species trees from each program.

We also estimated species trees using SVDQuartets and ASTRAL IV (as implemented in ASTER (Zhang & Mirarab, 2022; Tabatabaee et al., 2023)). For the SVDQuartets analysis, we used 1000 bootstrap replicates and rooted the tree at outgroup *Agkistrodon spp.* For the ASTRAL analysis, we used maximum likelihood gene trees estimated from IQTREE (Nguyen et al., 2015), using a Jukes Cantor model for each locus. We rooted the tree at outgroup *Agkistrodon spp.*

Results

Simulations

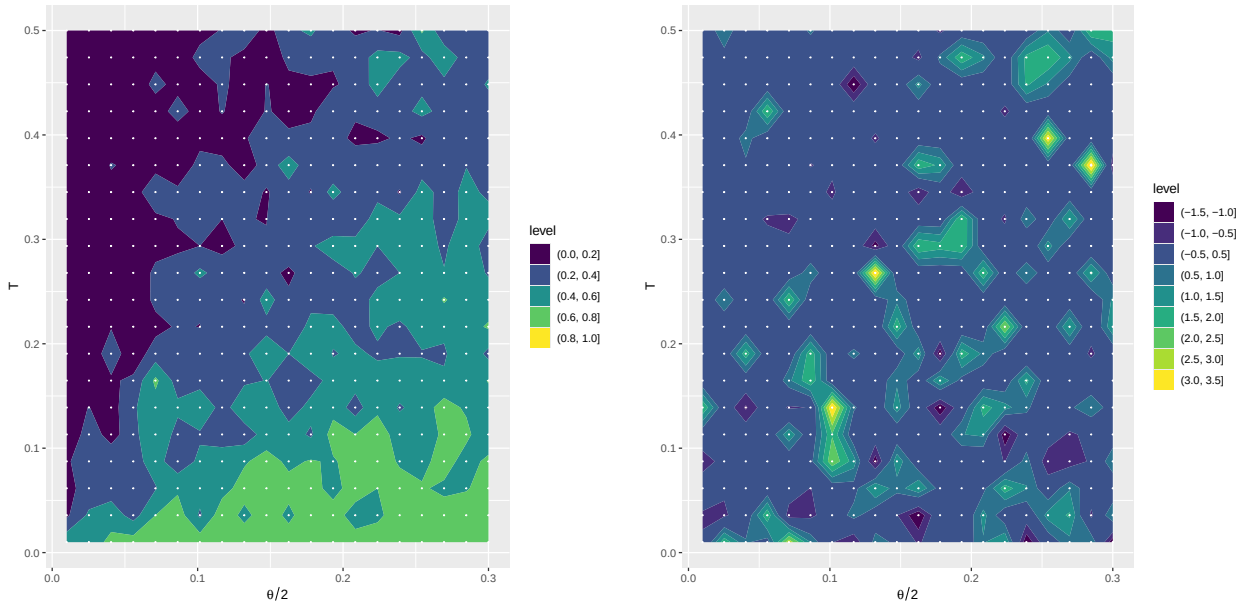
Without accounting for parallelization, SMC took 1126 seconds on average, and StarBEAST3 took 1359 seconds on average. We chose settings that gave each program roughly the same computational budget to fairly assess accuracy.

As expected, smaller values of T combined with larger values of $\theta/2$ produced difficult conditions, with more than 60% of the maximum possible number of deep coalescences being deep (Fig. 5a, lower right). Larger values of T combined with larger $\theta/2$ yielded easier conditions (Fig. 5a, upper left).

SMC and StarBEAST3 performed similarly (average difference in RF distance to true species tree between -0.5 and 0.5) for the majority (270 out of 400) of combinations of T and $\theta/2$ (Fig. 5b). There is no area of the grid where one method consistently outperforms the other. That said, StarBEAST3 performed slightly better than SMC on average, with an average RF distance of 0.399, compared to 0.638 for SMC. SVDQuartets produced an average RF distance of 1.005, and ASTRAL produced an average RF distance of 0.775.

Empirical Data

Gopher dataset.— All four methods found support for the clade containing *T. townsendii*, *T. bottae*, and *T. umbrinus* (Fig. 6). SMC, StarBEAST3, and ASTRAL recovered *T. bottae* and *T. townsendii* as sister, while SVDQuartets recovered *T. townsendii* and *T. umbrinus* as sister. All methods also found support for the clade containing *T. idahoensis*, *T. monticola*, *T. mazama*, and *T. talpoides*. ASTRAL and SVDQuartets found support for *T. idahoensis*



(a) Deep coalescences (b) Relative performance

Figure 5: Plots show simulation results for 400 combinations of expected species tree height T and $\theta/2$. a) Ratio of true number of deep coalescences to the maximum possible number of deep coalescences. (Higher values represent more difficult parameter combinations.) b) Mean SMC Robinson-Foulds (RF) distance to the true tree minus mean StarBEAST3 RF distance to the true tree. Positive values (yellow) indicate StarBEAST3 performed better for the simulation; negative values (purple) indicate SMC performed better. Points indicate $(\theta/2, T)$ combinations. Plot shows surface smoothed between sampled points.

and *T. talpoides* as sister within this clade, with *T. mazama* as the most distantly related species within this clade. SMC and StarBEAST3 found <50% support for any sister-group relationship in this clade. Neither StarBEAST3 nor SMC found support for a sister-group relationship between the putative outgroup *O. heterodus* and the ingroup species, though trees are displayed rooted at this taxon. (Neither SVDQuartets nor ASTRAL infer the root.) Without accounting for parallelization, SMC took 2298 seconds on average, and StarBEAST3 took 2629 seconds on average.

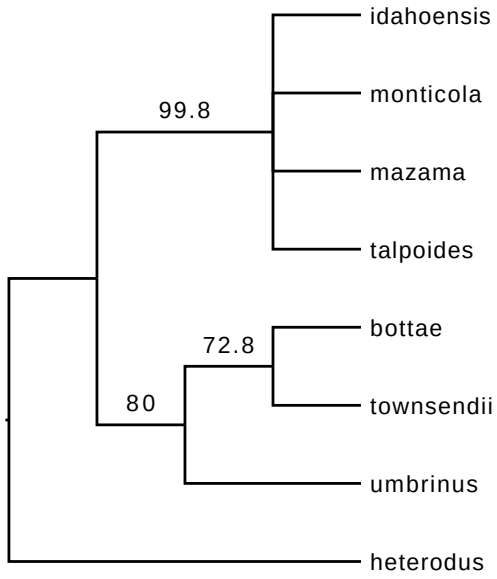
Maximum clade credibility (MCC) trees for SMC (Fig. 7a) and StarBEAST3 (Fig. 7b) show comparable branch lengths between programs. SMC found slightly longer branch

lengths for the clade containing *T. townsendii*, *T. bottae*, and *T. umbrinus*, but branch lengths on the MCC trees for this clade are no more than 0.0026 units different. Posterior probabilities for the MCC trees (Fig. 7c,d) are also comparable. Although both programs found strongest support for different sister relationships within the *T. idahoensis*, *T. monticola*, *T. mazama*, and *T. talpoides* clade, posterior probabilities were no greater than 0.5 for any relationship in this clade. The MCC trees are *not* rooted at the putative outgroup *O. heterodus*. SMC found lower support than StarBEAST3 for this taxon in the clade containing *T. townsendii*, *T. bottae*, and *T. umbrinus*, placing more support for this taxon as sister to the putative ingroup.

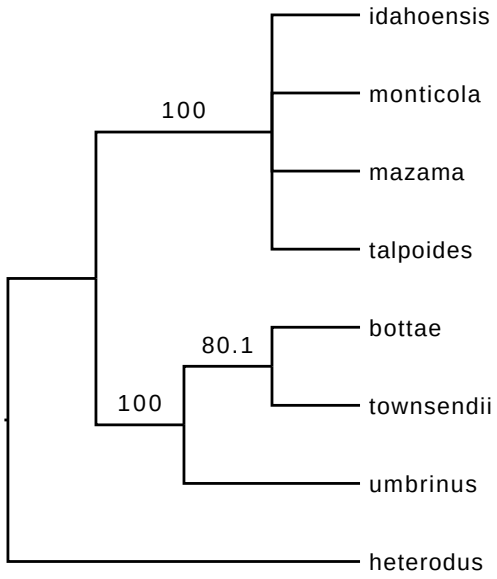
Snake dataset.— All methods recovered the same topology, with high support values for the two major clades: *S. edwardsii*, *S. tergeminus*, and *S. catenatus*; and *S. miliarius*, *S. barbouri*, and *S. streckeri*. Without accounting for parallelization, SMC took 27411 seconds on average, and StarBEAST3 took 40603 seconds on average.

Maximum clade credibility (MCC) trees (Fig. 9a,b) for SMC and StarBEAST3 show comparable branch lengths between programs. SMC found slightly longer branch lengths for the clade containing *S. barbouri*, *S. miliarius*, and *S. streckeri*, but branch lengths on the MCC trees for this clade are no more than 0.0039 units different.

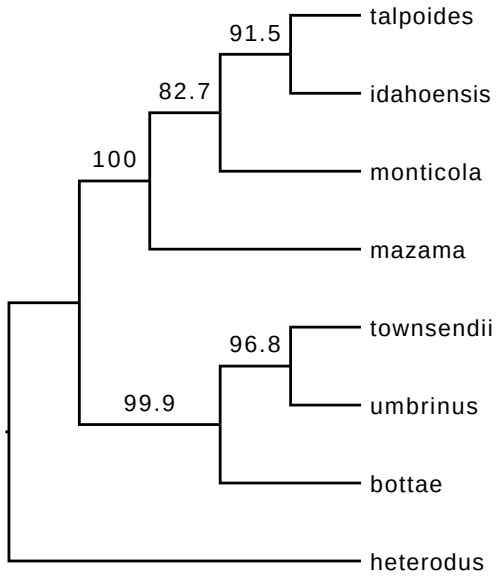
Posterior probabilities for the MCC trees (Fig. 9c,d) are also comparable. SMC found slightly lower support for *Agkistrodon spp.* as sister to the ingroup taxa and slightly lower support for the group *S. edwardsii* and *S. tergeminus*, but the topologies of both programs are identical.



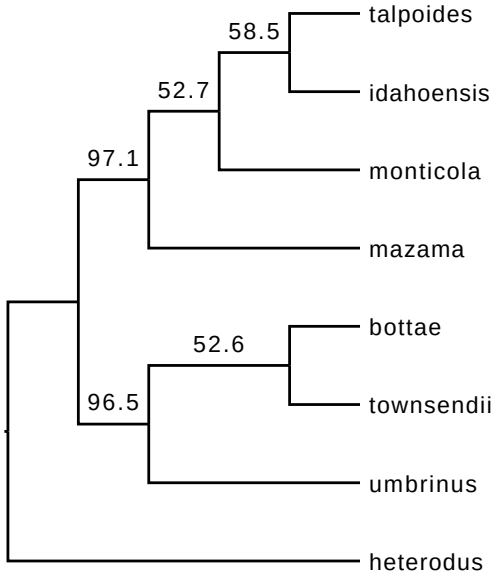
(a) SMC



(b) StarBEAST3

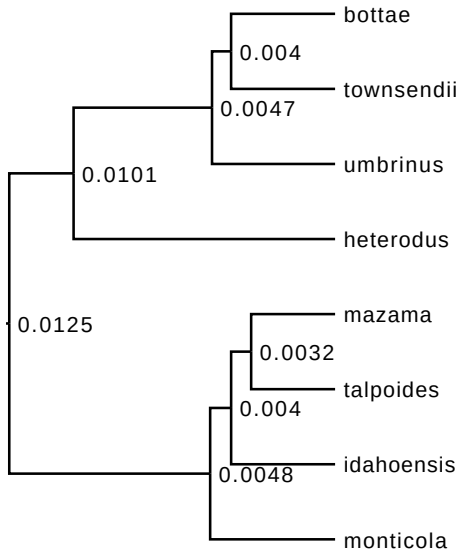


(c) SVDQuartets

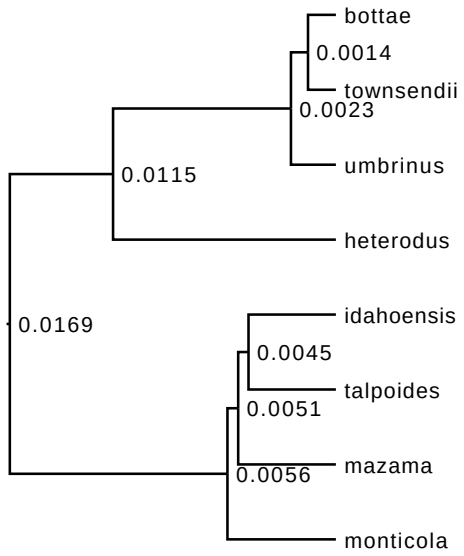


(d) ASTRAL

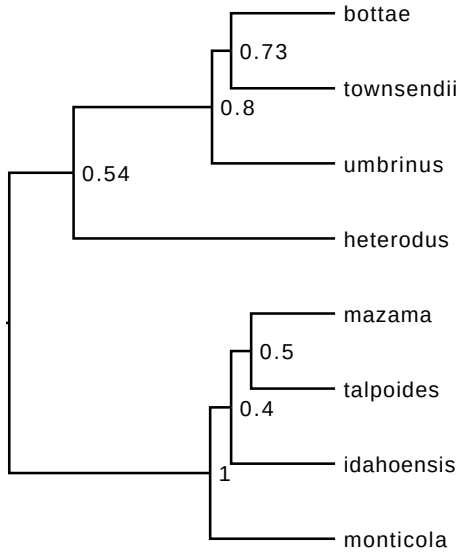
Figure 6: Gopher dataset. a) 50% majority rule consensus tree estimated by SMC. b) 50% majority rule consensus tree estimated by StarBEAST3. c) Bootstrapped consensus tree estimated by SVDQuartets. d) Tree estimated by ASTRAL with local posterior probabilities. All trees rooted at outgroup *O. heterodus*.



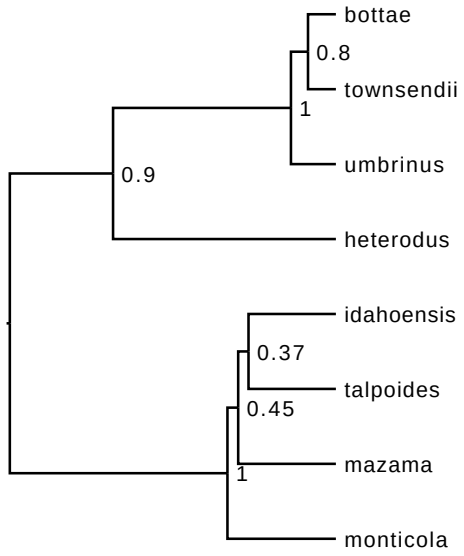
(a) SMC MCC tree with node heights



(b) StarBEAST3 MCC tree with node heights

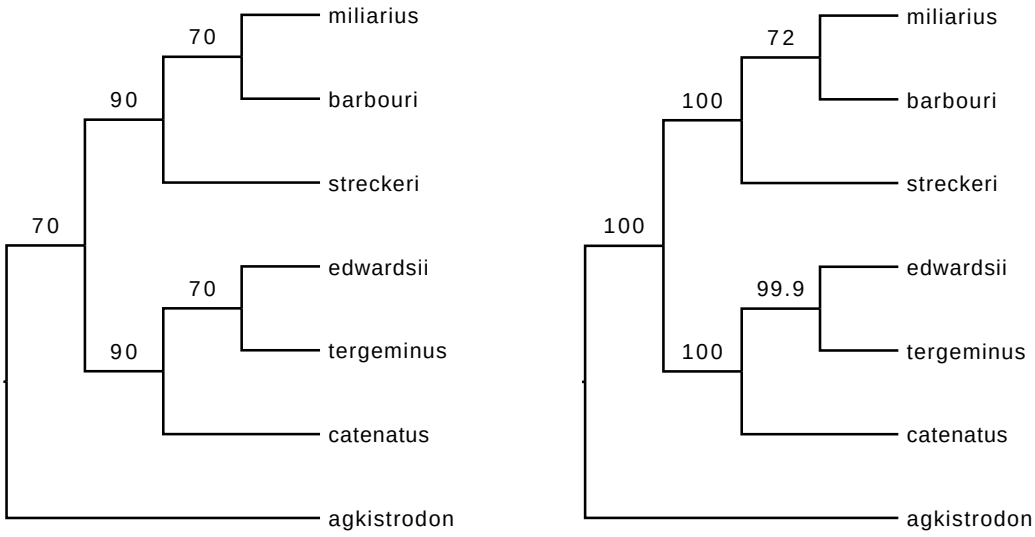


(c) SMC MCC tree with posterior probabilities



(d) StarBEAST3 MCC tree with posterior probabilities

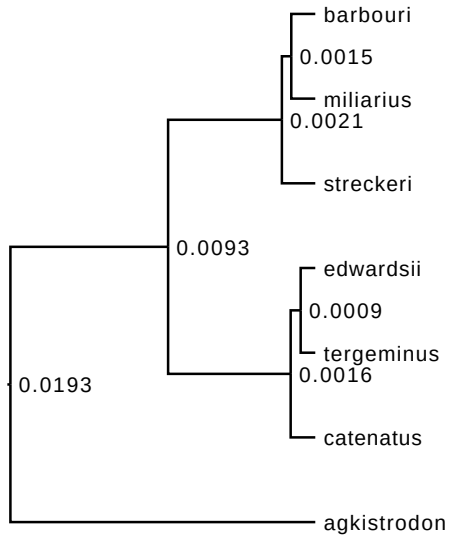
Figure 7: Gopher dataset. a) Maximum clade credibility tree estimated by SMC with node heights. b) Maximum clade credibility tree estimated by StarBEAST3 with node heights. c) Maximum clade credibility tree estimated by SMC with posterior probabilities. d) Maximum clade credibility tree estimated by StarBEAST3 with posterior probabilities. Figures created with TreeAnnotator (Drummond et al., 2007).



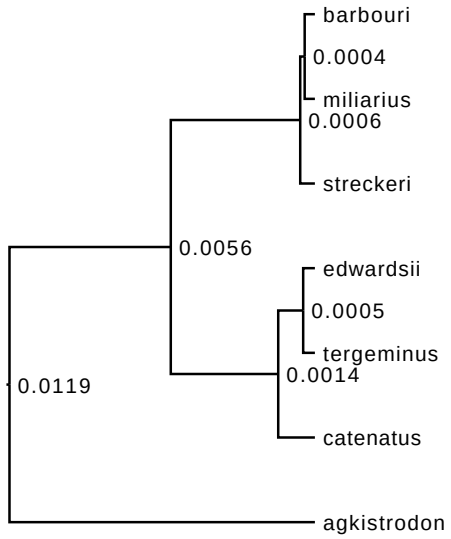
(a) SMC (b) StarBEAST3

(c) SVDQuartets (d) ASTRAL

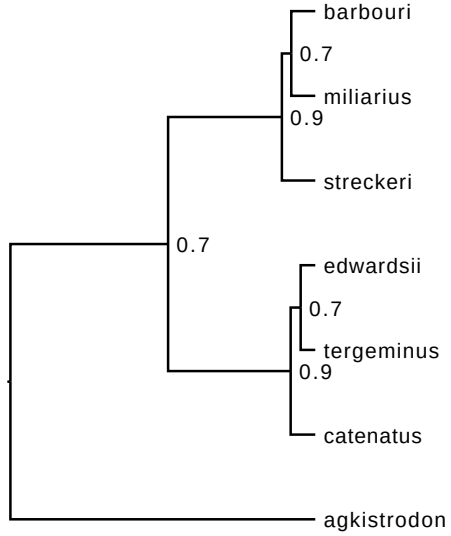
Figure 8: Snake dataset. a) 50% majority rule consensus tree estimated by SMC. b) 50% majority rule consensus tree estimated by StarBEAST3. c) Bootstrapped consensus tree estimated by SVDQuartets. d) Tree estimated by ASTRAL with local posterior probabilities. SVDQuartets and ASTRAL trees rooted at outgroup *Agkistrodon spp.*



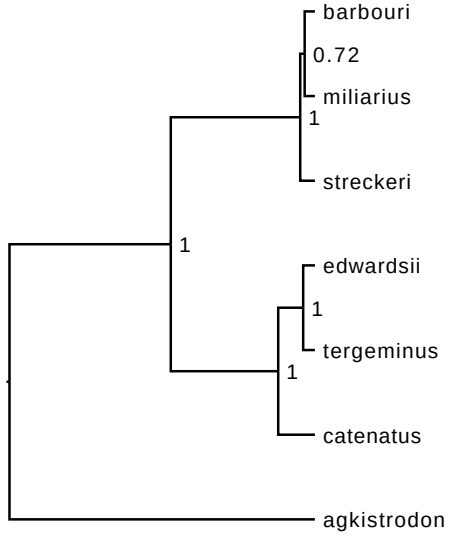
(a) SMC MCC tree with node heights



(b) StarBEAST3 MCC tree with node heights



(c) SMC MCC tree with posterior probabilities



(d) StarBEAST3 MCC tree with posterior probabilities

Figure 9: Snake dataset. a) Maximum clade credibility tree estimated by SMC with node heights. b) Maximum clade credibility tree estimated by StarBEAST3 with node heights. c) Maximum clade credibility tree estimated by SMC with posterior probabilities. d) Maximum clade credibility tree estimated by StarBEAST3 with posterior probabilities. Figures created with TreeAnnotator (Drummond et al., 2007).

Discussion

It is becoming increasingly important for phylogenetic methods to be able to accommodate datasets with hundreds of nuclear loci that conflict with one another due to incomplete lineage sorting or other factors. We have described a fully-Bayesian MSC method that is parallelizable in ways that are not possible for existing multispecies coalescent MCMC algorithms. Bayesian methods sample joint posterior distributions of continuous parameters (e.g., edge lengths, population sizes) and marginalize over discrete tree topologies, which makes them useful even as faster, non-Bayesian alternatives continue to be developed (Chifman and Kubatko, 2014; Mirarab and Warnow, 2015). Bayesian methods are also flexible with respect to model and can be used, for example, to estimate divergence times (Ogilvie et al., 2022), though our SMC implementation does not currently do this.

Speed and accuracy are important metrics for determining the utility of a program, and we have found that, while SMC cannot compete with ASTRAL and SVDQuartets in terms of speed, its accuracy and computational efficiency are comparable to that of StarBEAST3, currently recognized as the state of the art, even when parallelization is not taken into account. It is difficult to compare speed with parallelization between programs. While StarBEAST3 can be parallelized by placing loci on different processors, there is no speed advantage to using more processors than the number of loci. In contrast, SMC can parallelize across loci and particles, enabling it to take advantage of any number of processors. It is difficult to compare the computational efficiency of programs that are so divergent in their approach. Probably the best unit to use is the number of partial likelihood array calculations; however, we chose the simpler route of choosing settings for both programs that yielded runs of approximately the same total user seconds, giving StarBEAST3 slightly more time than SMC.

Particle degeneracy is a common problem with SMC-based algorithms (Truszkowski et al., 2023), and our SMC approach is no exception. If all samples from a locus have the same topology at the end of the first level SMC, the second level may produce inflated support values for clades in the species tree. We found that doing multiple independent SMC replicates and combining samples produced results comparable to StarBEAST3 for both empirical datasets. While this does increase the computational budget, multiple replicates can be run in parallel, and multiple independent StarBEAST3 runs are also recommended for assessing convergence. We found that particle degeneracy was less of an issue in our simulations, likely because the datasets were smaller and simulated from the exact model used for analysis.

Assessing convergence and the number of particles required in SMC algorithms is difficult. Bouchard-Côté (2012) suggested using effective sample size (ESS) to determine the number of particles needed; if ESS is low, it may mean more particles are needed. However, we have found that there are nearly always SMC steps for which the ESS is very low, regardless of how many particles are used. If the likelihood of one particular join in a gene tree is much greater than that of any other possible join, then any particle that gets this join correct will enjoy a selective sweep. This is as it should be; the low ESS in this case is a consequence of the high marginal gene tree clade posterior. Correctly diagnosing the causes of low ESS remains an area for future work.

We have found that increasing the number of loci in a dataset increases the number of SMC steps but does not necessarily increase the number of particles needed in the first level because only one locus is addressed during each step. Increasing the number of individuals sampled increases the number of particles required in the first-level SMC. This is because more individuals means more join possibilities, especially at early steps or when deep coa-

lescence is common. For example, sampling 100 individuals for a particular species requires 14828 particles to have a 95% chance of getting the first join correct (assuming that the first coalescence event is shallow). Finally, increasing the number of species also increases the number of particles needed because of the smaller chance that any one particle proposes a correct species tree join. This is not surprising because MCMC approaches also require more effort for larger sample sizes (loci, species, individuals per species).

Sampling the multispecies coalescent using SMC thus retains the advantages of fully Bayesian methods and is parallelizable in ways that Bayesian MCMC methods are not but also adds unique challenges. We demonstrated the performance of SMC compared to other commonly-used species tree methods using two empirical datasets and 400 simulated datasets.

Funding

This work was supported by the National Science Foundation Graduate Research Fellowship Program (Grant No. DGE 2136520 to AAM). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Acknowledgements

We would like to thank Laura Kubatko, Elizabeth Jockusch, Kent Holsinger, and Jill Wegrzyn for their constructive comments.

References

Belfiore, N., L. Liang, and C. Moritz. 2008. Multilocus phylogenetics of a rapid radiation in the genus *Thomomys* (Rodentia: Geomyidae). Systematic Biology 57:294–310. https://doi.org/10.1080/10635150802044011

Brook, D. 1964. On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbor systems. Biometrika 51:481–483. https://doi.org/10.2307/2334154

Bouchard-Côté, A., S. Sankararaman, and M. I. Jordan. 2012. Phylogenetic inference via Sequential Monte Carlo. Systematic Biology 61:579–593. https://doi.org/10.1093/sysbio/syr131

Bouchard-Côté, A. 2014. SMC (Sequential Monte Carlo) for Bayesian phylogenetics. Chapter 8, pp. 163–185, in: Chen M.-H., Kuo L., and Lewis P. O. (eds.), Bayesian phylogenetics: methods, algorithms, and applications. Chapman & Hall/CRC. https://doi.org/10.1201/b16965

Bryant, D., R. Bouckaert, J. Felsenstein, N. A. Rosenberg, and A. RoyChoudhury. 2012. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. Molecular Biology and Evolution 29:1917–1932. https://doi.org/10.1093/molbev/mss086

Chifman, J., and L. Kubatko. 2014. Quartet inference from SNP data under the coalescent model. Bioinformatics 30:3317–3324. https://doi.org/10.1093/bioinformatics/btu530

Douglas, J., C. L. Jiménez-Silva, and R. Bouckaert. 2022. StarBeast3: adaptive paral-

lelized Bayesian inference under the multispecies coalescent. Systematic Biology 71:901–916. https://doi.org/10.1093/sysbio/syac010

Drummond, A. J., and A. Rambaut. 2022. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evolutionary Biology 7:1–8. https://doi.org/10.1186/1471-2148-7-214

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of Molecular Evolution 17:368–376. https://doi.org/10.1007/BF01734359

Fisher, R. A. 1930. The genetical theory of natural selection. 1st ed., Clarendon Press. ISBN: 9780198504405

Jones, G. 2017. Algorithmic improvements to species delimitation and phylogeny estimation under the multispecies coalescent. J. Math. Biol. 74:447–467. https://doi.org/10.1007/s00285-016-1034-0

Heled, J., and A. J. Drummond. 2010. Bayesian inference of species trees from multilocus data. Molecular Biology and Evolution 27:570–580. https://doi.org/10.1093/molbev/msp274

Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Chapter 24, pp. 21–132, in: Munro, H. N. (ed.), Mammalian Protein Metabolism III, Academic Press, New York. http://dx.doi.org/10.1016/B978-1-4832-3211-9.50009-7

Kingman, J. F. C. 1982. The coalescent. Stochastic Processes and their applications 13:235–248. https://doi.org/10.1016/0304-4149(82)90011-4

Kimura, M. 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution 16:111–120. https://doi.org/10.1007/BF01731581

Kuhner, M., and J. Felsenstein. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Molecular Biology and Evolution, 11:459–468. https://doi.org/10.1093/oxfordjournals.molbev.a040126

Liu, J. S., F. Liang, and W. H. Wong. 2000. The multiple-try method and local optimization in Metropolis sampling. Journal of the American Statistical Association 95:121–134. https://www.tandfonline.com/doi/abs/10.1080/01621459.2000.10473908

Maddison, W. 1997. Gene trees in species trees. Systematic Biology, 46:523–536. https://doi.org/10.1093/sysbio/46.3.523

Mirarab, S., and T. Warnow. 2015. ASTRAL-II: Coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics 31:i44–i52. https://doi.org/10.1093/bioinformatics/btv234

Nguyen, L. T., H. A. Schmidt, A. Haeseler and B. Q. Minh. 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Molecular Biology and Evolution 32:268–274. https://doi.org/10.1093/molbev/msu300

Ogilvie, H. A., R. R. Bouckaert, and A. J. Drummond. 2017. StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. Molecular Biology and Evolution 34:L2101–2114. https://doi.org/10.1093/molbev/msx126

Ogilvie, H. A., F. K. Mendes, T. G. Vaughan, N. J. Matzke, T. Stadler, D. Welch, and A. J. Drummond. 2022. Novel integrative modeling of molecules and morphology across evolu-

tionary timescales. Systematic Biology 71:208–220. https://doi.org/10.1093/molbev/msx126

Pamilo, P., and M. Nei. 1988. Relationships between gene trees and species trees. Molecular Biology and Evolution 5:568–583. https://doi.org/10.1093/oxfordjournals.molbev.a040517

Peng, J., D. L. Swofford, and L. Kubatko. 2003. Estimation of speciation times under the multispecies coalescent. Bioinformatics 38:5182–5190. https://doi.org/10.1093/bioinformatics/btac679

Rannala, B., and Z. Yang. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164:1645–1656. https://doi.org/10.1093/genetics/164.4.1645

Robinson, D. F., and L. R. Foulds. 1981. Comparison of phylogenetic trees. Mathematical Biosciences, 53:131–147. https://doi.org/10.1016/0025-5564(81)90043-2

Sayyari, E., and S. Mirarab. 2016. Fast coalescent-based computation of local branch support from quartet frequencies. Molecular Biology and Evolution 33:1654–1668. https://doi.org/10.1093/molbev/msw079

Sokal, R. R., and C. D. Michener. 1958. A statistical method for evaluating systematic relationships. University of Kansas Science Bulletin 38:1409–1438. https://www.sid.ir/paper/549615/en

Sukumaran, J., and L. L. Knowles. 2017. Multispecies coalescent delimits structure, not species. Proceedings of the National Academy of Sciences USA 114:1607–1612. https://doi.org/10.1073/pnas.1607921114

Tabatabaee, Y., C. Zhang, T. Warnow, and S. Mirarab. 2023. Phylogenomic branch length estimation using quartets. Bioinformatics 39:i185–i193. https://doi.org/10.1093/bioinformatics/btad221

Truszkowski, J., A. Perrigo, D. Broman, F. Ronquist, and A. Antonelli. 2023. Online tree expansion could help solve the problem of scalability in Bayesian phylogenetics. Systematic Biology 72:199–1206. https://doi.org/10.1093/bioinformatics/btac679

Watterson, G. A. 1975. On the number of segregating sites in genetical models without recombination. Theoretical Population Biology 7:256–276. https://doi.org/10.1016/0040-5809(75)90020-9

Wright, S. 1931. Evolution in Mendelian populations. Genetics 16:97–159. https://doi.org/10.1093/genetics/16.2.97

Yang, Z. 2002. Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. Genetics 162:1811–1823. https://doi.org/10.1093/genetics/162.4.1811

Yule, G. U. 1925. II.—A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F. R. S. Philosophical Transactions of the Royal Society B: Biological Sciences 213:402–410. https://doi.org/10.1098/rstb.1925.0002

Zhang, C., and S. Mirarab. 2022. Weighting by gene tree uncertainty improves accuracy of quartet-based species trees. Molecular Biology and Evolution 39:msac215. https://doi.org/10.1093/molbev/msac215