Inferring Species Trees Directly from Biallelic Genetic Markers: Bypassing Gene Trees in a Full Coalescent Analysis

David Bryant,*[,1] Remco Bouckaert,[2] Joseph Felsenstein,[3] Noah A. Rosenberg,[4] and Arindam RoyChoudhury[5]

[1]Department of Mathematics and Statistics and the Allan Wilson Centre for Molecular Ecology and Evolution, University of Otago, Dunedin, New Zealand

[2]Computational Evolution Group, Department of Computer Science, University of Auckland, Auckland, New Zealand

[3]Department of Genome Sciences and Department of Biology, University of Washington

[4]Department of Biology, Stanford University

[5]Department of Biostatistics, Mailman School of Public Health, Columbia University

***Corresponding author:** E-mail: david.bryant@otago.ac.nz.

Associate editor: Rasmus Nielsen

Research article

Abstract

The multispecies coalescent provides an elegant theoretical framework for estimating species trees and species demographics from genetic markers. However, practical applications of the multispecies coalescent model are limited by the need to integrate or sample over all gene trees possible for each genetic marker. Here we describe a polynomial-time algorithm that computes the likelihood of a species tree directly from the markers under a finite-sites model of mutation effectively integrating over all possible gene trees. The method applies to independent (unlinked) biallelic markers such as well-spaced single nucleotide polymorphisms, and we have implemented it in SNAPP, a Markov chain Monte Carlo sampler for inferring species trees, divergence dates, and population sizes. We report results from simulation experiments and from an analysis of 1997 amplified fragment length polymorphism loci in 69 individuals sampled from six species of *Ourisia* (New Zealand native foxglove).

Key words: multispecies coalescent, species trees, SNP, AFLP, effective population size, SNAPP.

Introduction

Biallelic markers such as single nucleotide polymorphisms (SNPs) and amplified fragment length polymorphisms (AFLPs) are potentially rich sources of information about species radiations, species divergences, and historical demographics. However, extracting this information is not always straightforward. Patterns of genetic variation at these markers are not just a product of the relationships between the species; they also reflect inheritance patterns within each species. Any full-likelihood (or full-Bayesian) method for inferring species histories from genetic markers needs to model the random distribution of gene tree histories for each marker. To date, this task has often meant implementing massive Monte Carlo simulation-based sampling of both species trees and the gene trees at every locus (Rannala and Yang 2003; Wilson et al. 2003; Hey and Nielsen 2007; Liu and Pearl 2007; Heled and Drummond 2010).

In this paper, we describe an algorithm that allows us to bypass the gene trees and compute species tree likelihoods directly from the markers. The likelihood values, or posterior probabilities, computed by the algorithm are identical to those that would be obtained by sampling every possible gene tree topology and every possible set of gene tree branch lengths at each locus. The algorithm makes use of new formulae for lineage and allele probabilities under the coalescent and employs recently developed numerical techniques (Sidje 1998; Schmelzer and Trefethen 2007) to evaluate these formulae.

Our approach makes the following assumptions of the data:

(A1) Each marker is a single biallelic character (e.g., a biallelic SNP or AFLP banding pattern);

(A2) The genealogies for separate markers are conditionally independent given the species tree. In practice, this assumption applies to unlinked markers or linked markers that have so little linkage that they do not possess a discernible excess of linkage disequilibrium.

This latter assumption is clearly not valid for sites in a single-gene sequence. However, it is satisfied for SNPs that are well spaced along the genome. If the independence assumption (A2) is only partially violated, the effect of linkage could be investigated by subsampling sets of markers with varying degrees of independence. Even when there is linkage between sites, treating the markers as independent often still provides statistically responsible inferences (Gutenkunst et al. 2009; RoyChoudhury 2011).

In principle, the full-likelihood methods of Liu and Pearl (2007) and Heled and Drummond (2010), which are designed primarily for linked sequence data, could be applied to data satisfying assumptions (A1) and (A2) by encoding each marker as a separate locus. This strategy would quickly become computationally infeasible as the number of markers increased. Nielsen et al. (1998) demonstrated that a full-likelihood approach is tractable

for data satisfying (A1) and (A2), presenting an algorithm that uses biallelic characters directly to compute the likelihood of the species tree, though their method was computationally feasible only for small species trees. RoyChoudhury (2006) and RoyChoudhury et al. (2008) made a substantial advance on the computational problem. They took the approach of Nielsen et al. (1998) and placed it within a dynamic programming framework, thereby giving an efficient algorithm for computing the likelihood of a tree with an arbitrary number of species.

The methods developed by Nielsen et al. (1998) and RoyChoudhury et al. (2008) both make a significant and mathematically convenient assumption about mutation. Under their models, mutation can only occur within the population at the root of the species tree. It cannot occur within the populations represented by the branches of the species tree. This assumption is reasonable when comparing closely related populations for which recent mutations are sufficiently rare that they can be ignored. It is less appropriate when analyzing rapidly mutating markers or when comparing more distantly related populations or species.

Here, we extend the dynamic programming structure of RoyChoudhury et al. (2008) to allow mutations within the populations represented by the species tree. In many ways, this is a more parsimonious model: The mutation model in the root population is the same as the mutation model used for the populations along the branches. We address the algorithmic, mathematical, and computational challenges resulting from this deceptively minor change in model assumptions.

Our algorithm implements a "finite-sites" model for mutation. Nielsen (1998) derived a recursion for computing the likelihood of a tree under the "infinite-sites" model for mutation. In general, the recursive formula has too many terms to be evaluated directly, so a Markov chain Monte Carlo (MCMC) method was used instead. However, if the data satisfy (A1) and (A2), then the recursion described by Nielsen (1998) can be evaluated efficiently using algorithms similar to those described here.

One issue that arises when modeling mutation is that some of parameters might not be identifiable from data. Under the infinite-sites model and the "no-branch-mutation" model of Nielsen et al. (1998) and RoyChoudhury et al. (2008), the length of a branch in the species tree and the corresponding population size are confounded: Doubling the effective population size has the same effect on the likelihood as halving the branch length. A similar issue arises when inferring species trees from gene tree topologies without branch lengths (Degnan and Salter 2005; Wu 2011). We show that fully including mutations in the finite-sites model permit the identification of both branch lengths (times) and population sizes (θ), at least in situations where sufficiently many mutations have occurred throughout the species tree.

The coalescent process is often viewed as a dual process to the Wright–Fisher diffusion (Donnelly and Kurtz 1996), and each has some practical advantages over the other. Diffusion-based approaches for analyzing SNP data from multiple populations have been proposed by Gutenkunst et al. (2009) and Siren et al. (2010). The main difference is that coalescent-based methods such as ours' model the history of the ancestral lineages, whereas diffusion-based approaches model variation in the continuous allele frequencies. In practice, it is not clear which approach is preferable for a given data set: both require some level of approximation and each has advantages and disadvantages computationally.

We have implemented the new finite-sites model likelihood algorithm and incorporated it within a Bayesian MCMC sampler, which we call SNAPP ("SNP and AFLP Phylogenies"). SNAPP, which interfaces with the BEAST package (Drummond and Rambaut 2007), takes a range of biallelic data types as input and returns a sample of species trees with (relative) divergence times and population sizes. We have tested and validated the algorithm and software using a range of techniques, and we report results of several experiments with simulated data. The software is open source and is available for download from http://snapp.otago.ac.nz.

To illustrate the application of SNAPP, we analyze AFLP loci in 69 individuals sampled from 6 species of New Zealand *Ourisia* or native foxglove. The New Zealand *Ourisia* form a relatively recent species radiation and inference of branching patterns between these species has proven difficult (Meudt et al. 2009). Meudt et al. propose that the difficulties are due in part to "incomplete lineage sorting," which occurs when the coalescence of lineages within species predates the divergence of different species. Our Bayesian analysis, which models lineage sorting explicitly, provides a relatively clear picture of ancestral species relations in the group and, up to a scale constant, effective population sizes.

Materials and Methods

The Multispecies Coalescent

Our models are all based on the assumption that the lineage dynamics within populations (or species) are well described by the conventional Wright–Fisher model. The distribution of the gene trees within each population is approximated by the "coalescent process" (reviewed in Felsenstein 2004; Hein et al. 2005; Wakeley 2009). This process models the number of ancestral lineages of the sample from a single population as a Markov process that goes backward in time. Initially, the number of ancestral lineages equals the size of the sample. Going backward in time (upward in a branch), lineages meet at common ancestors, and the number of ancestral lineages decreases.

It is customary in coalescent theory to rescale time in terms of effective population size, so that two lineages coalesce at rate 1. This rescaling is not generally possible in the multispecies coalescent since different species can have different effective population sizes. Instead, we adopt the standard practice from phylogenetics and rescale time in terms of expected mutations (as in Rannala and Yang 2003). Hence, the expected time to a coalescence for two lineages is $\theta/2$ and the expected time to a coalescence for k lineages is $\theta/[k(k-1)]$, where θ denotes the expected number of

mutations separating two randomly chosen individuals in the population.

At the first coalescent event, two lineages are selected at random and combined, and we are left with $k - 1$ lineages. This coalescence of lineages continues until the top of the branch is reached, at which anywhere from 1 to k lineages could be present.

The nodes in the species tree represent species divergences or population splits. The individuals in each of the child populations are descendants of individuals in the parent population. In terms of the coalescent process, the lineages coming upward from the child population become lineages at the base of the parent population. This process continues upward in the species tree until the species tree root is reached. At that point, any remaining lineages coalesce according to the standard single-population coalescent model.

See Felsenstein (2004), Degnan and Rosenberg (2009), and Heled and Drummond (2010) for general introductions to the multispecies coalescent. Early contributions to the development of multispecies models built on the branches of a species tree were made by Hudson (1983), Tajima (1983), Takahata and Nei (1985), Nei (1987), Pamilo and Nei (1988), and Takahata (1989).

The multispecies coalescent determines a distribution for gene trees and their branch lengths, conditional on a species tree. The parameters of the distribution are the shape of the species tree, the divergence times within the species tree, and the population sizes along the branches of the species tree (one parameter for each branch). We bundle these parameters into the single composite parameter S, so that the probability of a gene tree G given the species tree is $P(G|S)$. We treat this quantity as a density rather than a discrete probability because of the continuous branch lengths of G.

Let X denote the alignment of sequences for a locus. Conventional phylogenetic models (e.g., Felsenstein 2004) give us the probability that X evolved along a specified gene tree G. These models provide the distribution of states at the root and the mutation probabilities down the edges of the tree. Accordingly, they determine $P(X|G)$, the probability of the data (alignment) given the gene tree. Note that once the gene tree is chosen, the species tree has no further influence on the probability of the data.

Putting $P(G|S)$ and $P(X|G)$ together, we obtain the "joint" probability (or density) of the alignment X and the gene tree G:

$$P(X, G|S) = P(X|G)P(G|S). \tag{1}$$

The gene tree G is not observed directly and it can be difficult to estimate. Since our focus is on the species tree and the features of the species tree, we work with the "marginal" probability of the data. Let Ψ denote the set of all possible genealogies for the individuals incorporating both the topologies and branch lengths. The marginal probability for the data is then found by integrating over Ψ:

$$P(X|S) = \int_{\Psi} P(X|G)P(G|S)\mathrm{d}G. \tag{2}$$

Equation (2) is sometimes called the "Felsenstein equation" (Felsenstein 1988; Rosenberg and Nordborg 2002; Hey and Nielsen 2007).

Generally, we consider multiple genetic markers. We assume that the gene trees for separate markers are independent (given the species tree). Let X_i be the alignment for the ith gene and let G_i be a corresponding gene tree. Under the independence assumption, the total probability of the m alignments at m genes is a product over all the genes:

$$\begin{aligned} P(X_1, X_2, \ldots, X_m|S) &= \prod_{i=1}^{m} P(X_i|S) \\ &= \prod_{i=1}^{m} \int_{\Psi} P(X_i|G_i)P(G_i|S)\mathrm{d}G_i. \end{aligned} \tag{3}$$

If we were to plug this formula into a Bayesian analysis, we would specify a prior distribution $P(S)$ on the species trees and then sample from the posterior distribution

$$P(S|X_1, \ldots, X_m) \propto \left(\prod_{i=1}^{m} \int_{\Psi} P(X_i|G_i)P(G_i|S)\mathrm{d}G_i \right) P(S). \tag{4}$$

Sampling from $P(S|X_1, \ldots, X_m)$ is equivalent to sampling from the joint posterior distribution

$$P(S, G_1, \ldots, G_m|X_1, \ldots, X_m) \propto \left(\prod_{i=1}^{m} P(X_i|G_i)P(G_i|S) \right) P(S) \tag{5}$$

and only considering the marginal distribution of the species trees S. This is the approach taken by BATWING (Wilson et al. 2003), BEST (Liu and Pearl 2007), and STAR-BEAST (Heled and Drummond 2010), among others. Note that if the actual gene trees G_i are provided or if they can be inferred with high accuracy, they can be treated as data and the species tree can be inferred directly (Degnan and Salter 2005; Kubatko et al. 2009).

At this point, it is appropriate to reflect on what exactly is required when applying equations (3) or (4) to large numbers of unlinked biallelic markers. To evaluate the likelihood exactly, we would need to sum (or integrate) over all possible gene trees of all loci. In a Bayesian setting, we would need to sample over a space containing not only every possible choice of species tree but also every possible choice of gene tree for every locus. Furthermore, the marginal probabilities for the gene trees depend not only on the data but also on the species tree, and so the analyses for the separate genes are all interdependent. An analysis of 1,000 independent loci then amounts to 1,001 interlinked Bayesian analyses (1,000 gene trees and one species tree). Even with modern Monte Carlo algorithms, this scale of this analysis is computationally daunting.

Overview of the Likelihood Algorithm

We circumvent these computational difficulties by calculating the integral in equation (3) analytically. In the following sections, we describe a pruning algorithm that we use to compute the likelihood of a species tree given genotype data

at unlinked biallelic markers. The algorithm works in a similar manner to Felsenstein's pruning algorithm (Felsenstein 1981) for computing the likelihood of a gene tree: we define partial likelihoods that focus only on a specific subtree; the partial likelihoods are then computed starting at the leaves (of the species tree), working upward to the root.

There are two major differences. In Felsenstein's pruning algorithm, one partial likelihood is defined for every node and every state (i.e., amino acid or nucleotide). In our algorithm, we have separate partial likelihoods for the top and bottom of each branch in the species tree, for every possible number of ancestral lineages at each point, and for every possible count of the number among these lineages carrying each allele.

Second, we need to deal with the complication that the coalescent process works backward in time (and is not reversible), whereas the mutation process works forward in time. We were not able to define a simple transition process taking numbers of ancestral lineages to numbers of descendant lineages. Instead, we first compute probability distributions for the numbers of ancestral lineages at each node in the species tree. We then define partial likelihoods for subtrees in the species tree and derive the equations required to compute them efficiently. Finally, we show how to handle the probabilities at the root of the species tree when computing the full probability of the genotype data of a marker.

We orient trees so that the ancestral nodes are at the top and time travels downward. Thus, the base of a branch in the species tree corresponds to the population at the time nearest to the present, whereas the top of a branch corresponds to the population just after it has diverged from its ancestral population. In a similar fashion, the genotypic state in a gene tree evolves from the top of the gene tree (the common ancestor) downward to the leaves.

Red and Green Alleles

The multispecies coalescent model for the evolution of markers (SNPs, AFLPs etc.) has two components: the model for the gene trees in the species tree and the model for the markers evolving down the gene tree (i.e., forward in time). The model for gene trees uses a coalescent process that works backward in time, whereas the mutation model for genetic markers (SNPs, AFLPs, etc.) typically works forward in time.

Given a gene tree with branch lengths specified, we model the evolution of a genetic marker using standard phylogenetic machinery. Suppose that there are two alleles, which for ease of illustration we label "red" and "green." Let u be the rate of mutation from the red allele to the green allele per unit time (forward in time), and let v be the corresponding rate of mutating from green to red. We say that a lineage is a red lineage if it has the red allele and a green lineage otherwise.

The allele of the most recent common ancestor at the root of the gene tree is red with stationary probability $v/(u+v)$ and green with probability $u/(u+v)$. The marker evolves down the gene tree as a continuous-time Markov chain whose instantaneous rate matrix has rate u of mutating from red to green and rate v for mutating from green to red. The alleles at the leaves of the gene tree are then the observed alleles. The probability of the allele frequencies at a marker, given the species tree, is therefore the probability of the site given a gene tree multiplied by the probability of the gene tree given the species tree, summed over all possible gene tree topologies and integrated over all possible gene tree branch lengths (eq. [3]).

Ancestral Lineage Counts and the Likelihood

The multispecies coalescent can be used to generate a random gene tree conditional on a species tree. If we take any node or point in the species tree, we can count the number of lineages in the gene tree in that species at that point in time. We say that at a specified time point, this quantity is the number of "ancestral lineages." The count of ancestral lineages is a random variable with distribution determined by the multispecies coalescent process and its resulting distribution of gene trees. The first step in our likelihood algorithm is the calculation of these lineage count distributions. See RoyChoudhury et al. (2008) and Efromovich and Kubatko (2008) for similar computations.

Let x be a branch (i.e., ancestral species) in the species tree. Let $\mathbf{n}_x^{\mathrm{B}}$ denote the number of gene tree lineages at the base of the branch x. Let $\mathbf{n}_x^{\mathrm{T}}$ denote the number of ancestral lineages at the top of the branch and let t be the length of the branch, measured in units of expected number of mutations (see fig. 1). The minimum possible value for $\mathbf{n}_x^{\mathrm{B}}$ and $\mathbf{n}_x^{\mathrm{T}}$ is 1, whereas the maximum possible value is the total number of individuals sampled in populations at or below x, a quantity that we denote by m_x. The distribution of $\mathbf{n}_x^{\mathrm{T}}$ given $\mathbf{n}_x^{\mathrm{B}}$ is given by the probability in the standard coalescent model of going from n ancestors to k ancestors over time t (measured in units of expected mutations):

$$\Pr[\mathbf{n}_x^{\mathrm{T}}=k|\mathbf{n}_x^{\mathrm{B}}=n] = \sum_{r=k}^{n} e^{\frac{-r(r-1)t}{\theta}} \frac{(2r-1)(-1)^{r-k}k_{(r-1)}n_{[r]}}{k!(r-k)!n_{(r)}}, \quad (6)$$

where $n_{[r]}=n(n-1)(n-2),\ldots,(n-r+1)$ and $n_{(r)}=n(n+1),\ldots,(n+r-1)$ (Tavaré 1984).

When x is an "external" branch (adjacent to a leaf) in the species tree, $\mathbf{n}_x^{\mathrm{B}}$ equals the number of samples from the species corresponding to that branch. Let n_x denote this number of samples. Then

$$\Pr[\mathbf{n}_x^{\mathrm{B}}=n] = \begin{cases} 1 & \text{if } n=n_x, \\ 0 & \text{otherwise.} \end{cases} \quad (7)$$

Suppose that x is an internal or external branch in the species tree. Let m_x denote the maximum possible value for $\mathbf{n}_x^{\mathrm{B}}$ or $\mathbf{n}_x^{\mathrm{T}}$, equal to the total number of sampled individuals summed across populations at or below x in the species tree. Suppose that $\Pr[\mathbf{n}_x^{\mathrm{B}}=k]$ has been computed for all k from 1 to m_x. The distribution of $\mathbf{n}_x^{\mathrm{T}}$ is determined by the value of $\mathbf{n}_x^{\mathrm{B}}$ using the conditional probabilities in equation (6):

$$\Pr[\mathbf{n}_x^{\mathrm{T}}=n] = \sum_{k=n}^{m_x} \Pr[\mathbf{n}_x^{\mathrm{B}}=k]\Pr[\mathbf{n}_x^{\mathrm{T}}=n|\mathbf{n}_x^{\mathrm{B}}=k]. \quad (8)$$

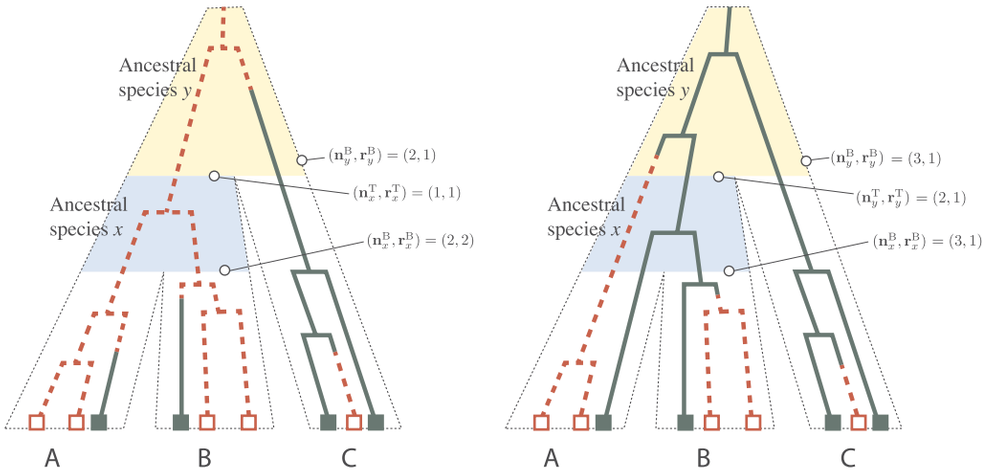

FIG. 1. Gene trees in species trees. Each branch in the species trees corresponds to a species that is either contemporary (A,B,C) or ancestral (x, y). The present-day samples are represented by green (solid) and red (hollow) squares along the lower edge of the tree. The red (dashed) and green (solid) lines trace out two possible gene trees for these individuals, the red–green coloring indicating which allele is carried by a lineage at any particular time. The random variables $\mathbf{n}_x^{\mathrm{B}}$ and $\mathbf{r}_x^{\mathrm{B}}$ equal the number of lineages and the number of red lineages, respectively, at the bottom of the branch for ancestral species x. The corresponding values at the top of this branch are denoted $\mathbf{n}_x^{\mathrm{T}}$ and $\mathbf{r}_x^{\mathrm{T}}$, respectively.

Now suppose that x is an internal branch in the species tree and that branches y and z are attached to the base of branch x. There is no time for coalescent events between the tops of the branches y and z and the bottom of the branch above x. Hence, $\mathbf{n}_x^{\mathrm{B}} = \mathbf{n}_y^{\mathrm{T}} + \mathbf{n}_z^{\mathrm{T}}$ and

$$\Pr[\mathbf{n}_x^{\mathrm{B}} = n] = \sum_{k=0}^{n} \Pr[\mathbf{n}_y^{\mathrm{T}} = k] \Pr[\mathbf{n}_z^{\mathrm{T}} = n - k]. \quad (9)$$

Equations (7)–(9) together provide a method for computing $\Pr[\mathbf{n}_x^{\mathrm{B}} = n]$ for all nodes x in the species tree and all $n \geqslant 1$. In our implementation, the nodes are visited in a "postorder traversal" in order from the leaves up to the root, so that a node is always visited after the required probabilities for the children have already been computed. When considering a branch attached to a leaf in the species tree, we use equation (7) to compute $\Pr[\mathbf{n}_x^{\mathrm{B}} = n]$ for all n and equation (8) to compute $\Pr[\mathbf{n}_x^{\mathrm{T}} = n]$ for all n. At an internal branch, we use equation (9) to compute $\Pr[\mathbf{n}_x^{\mathrm{B}} = n]$ for all n and equation (8) to compute $\Pr[\mathbf{n}_x^{\mathrm{T}} = n]$ for all n.

Computing the Partial Likelihoods

We now introduce mutation. We associate the two colors red and green to the two alleles and color each branch of the gene tree according to the allele state along the branch. Hence, a gene tree node will be red or green, depending on whether the corresponding lineage carries a red or green allele at that point. A mutation along a lineage is represented by a change in color along the branch, and a branch can have multiple color changes.

Recall that $\mathbf{n}_x^{\mathrm{B}}$ denotes the number of gene tree lineages at the bottom of a particular branch x in the species tree. We let $\mathbf{r}_x^{\mathrm{B}}$ denote the number of these lineages that carry the red allele at that point, so that $0 \leqslant \mathbf{r}_x^{\mathrm{B}} \leqslant \mathbf{n}_x^{\mathrm{B}}$. In the same way, we let $\mathbf{r}_x^{\mathrm{T}}$ denote the number of lineages carrying the red allele at the top of the branch x, so that $0 \leqslant \mathbf{r}_x^{\mathrm{T}} \leqslant \mathbf{n}_x^{\mathrm{T}}$.

Let r_z denote the number of red alleles set observed in the species associated with an external branch z. Our objective is to compute the joint probability of $(\mathbf{r}_z^{\mathrm{B}} = r_z)$ over all leaves z in the species tree, conditional on the species tree, sample sizes, and model parameters. To this end, we define a "partial likelihood" equal to the corresponding conditional likelihood for a subtree of the species tree. Let $\mathcal{R}_x$ denote the event that $(\mathbf{r}_z^{\mathrm{B}} = r_z)$ holds for every external branch z that is a descendant of branch x in the species tree. That is, $\mathcal{R}_x$ is shorthand for the event that the allele counts below x correspond to those observed in the data for a single genetic marker. For every node x of the species tree, and every choice of n and r, we define

$$\mathbf{F}_x^{\mathrm{B}}(n, r) = \Pr[\mathcal{R}_x | \mathbf{n}_x^{\mathrm{B}} = n, \mathbf{r}_x^{\mathrm{B}} = r] \Pr[\mathbf{n}_x^{\mathrm{B}} = n] \quad (10)$$

and

$$\mathbf{F}_x^{\mathrm{T}}(n, r) = \Pr[\mathcal{R}_x | \mathbf{n}_x^{\mathrm{T}} = n, \mathbf{r}_x^{\mathrm{T}} = r] \Pr[\mathbf{n}_x^{\mathrm{T}} = n]. \quad (11)$$

We will see that the values $\mathbf{F}_x^{\mathrm{B}}(n, r)$ and $\mathbf{F}_x^{\mathrm{T}}(n, r)$ can be computed by starting at the leaves and working upward toward the root, just as in Felsenstein's pruning algorithm. Furthermore, when x is the root of the tree, the probability for the entire marker can be determined from the values of $\mathbf{F}_x^{\mathrm{B}}(n, r)$. Technically speaking, $\mathbf{F}_x^{\mathrm{B}}(n, r)$ is not a partial likelihood; rather, it is the product of a partial likelihood $(\Pr[\mathcal{R}_x | \mathbf{n}_x^{\mathrm{B}} = n, \mathbf{r}_x^{\mathrm{B}} = r])$ and a probability $(\Pr[\mathbf{n}_x^{\mathrm{B}} = n])$. This latter term simplifies the mathematics further on and makes the computation more numerically stable. In the same way, $\mathbf{F}_x^{\mathrm{T}}(n, r)$ is the product of a partial likelihood $(\Pr[\mathcal{R}_x | \mathbf{n}_x^{\mathrm{T}} = n, \mathbf{r}_x^{\mathrm{T}} = r])$ and a probability $(\Pr[\mathbf{n}_x^{\mathrm{T}} = n])$. We will show that these quantities can be computed using dynamic programming and that they can then be used to compute the probability of the marker.

We note that the computation can be extended to multifurcating species trees by converting a multifurcating tree

into a bifurcating tree. This extension is performed by replacing any multifurcation with a series of bifurcations, all separated by branches of length 0. The probabilities of the marker will be unchanged.

Partial Likelihoods for a Leaf

The simplest case for computing the partial likelihood is when the branch x is attached to a leaf (that is, when x is external). The number of samples from the associated species is n_x, and the number of individuals carrying the red allele is r_x. Hence,

$$\mathbf{F}_x^{\mathrm{B}}(n,r) = \begin{cases} 1 & \text{if } n = n_x \text{ and } r = r_x, \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

Partial Likelihoods along a Branch

Let y be a branch for which $\mathbf{F}_y^{\mathrm{B}}(n_b, r_b)$ has already been computed for all n_b and r_b. We carefully manipulate the conditional probabilities to obtain an expression for $\mathbf{F}_y^{\mathrm{T}}(n_t, r_t)$. As before, we let m_y denote the number of individuals sampled from species at or below y. Starting with the definition of $\mathbf{F}_y^{\mathrm{T}}$, we have

$$\begin{aligned} &\mathbf{F}_y^{\mathrm{T}}(n_t, r_t) \\ &\quad = \Pr[\mathcal{R}_y | \mathbf{n}_y^{\mathrm{T}} = n_t, \mathbf{r}_y^{\mathrm{T}} = r_t] \Pr[\mathbf{n}_y^{\mathrm{T}} = n_t] \\ &\quad = \sum_{n_b=n_t}^{m_y} \sum_{r_b=0}^{n_b} \Pr[\mathbf{n}_y^{\mathrm{T}} = n_t] \\ &\qquad \times \Pr[\mathcal{R}_y | \mathbf{n}_y^{\mathrm{T}} = n_t, \mathbf{r}_y^{\mathrm{T}} = r_t, \mathbf{n}_y^{\mathrm{B}} = n_b, \mathbf{r}_y^{\mathrm{B}} = r_b] \\ &\qquad \times \Pr[\mathbf{n}_y^{\mathrm{B}} = n_b, \mathbf{r}_y^{\mathrm{B}} = r_b | \mathbf{n}_y^{\mathrm{T}} = n_t, \mathbf{r}_y^{\mathrm{T}} = r_t]. \end{aligned} \tag{13}$$

We now use the fact that $\mathcal{R}_y$ is conditionally independent of $\mathbf{n}_y^{\mathrm{T}}$ and $\mathbf{r}_y^{\mathrm{T}}$ given $\mathbf{n}_y^{\mathrm{B}}$ and $\mathbf{r}_y^{\mathrm{B}}$, so that $\Pr[\mathcal{R}_y | \mathbf{n}_y^{\mathrm{T}}, \mathbf{r}_y^{\mathrm{T}}, \mathbf{n}_y^{\mathrm{B}}, \mathbf{r}_y^{\mathrm{B}}] = \Pr[\mathcal{R}_y | \mathbf{n}_y^{\mathrm{B}}, \mathbf{r}_y^{\mathrm{B}}]$, and

$$\begin{aligned} \mathbf{F}_y^{\mathrm{T}}(n_t, r_t) &= \sum_{n_b=n_t}^{m_y} \sum_{r_b=0}^{n_b} \Pr[\mathcal{R}_y | \mathbf{n}_y^{\mathrm{B}} = n_b, \mathbf{r}_y^{\mathrm{B}} = r_b] \\ &\qquad \times \Pr[\mathbf{n}_y^{\mathrm{B}} = n_b, \mathbf{r}_y^{\mathrm{B}} = r_b | \mathbf{n}_y^{\mathrm{T}} = n_t, \mathbf{r}_y^{\mathrm{T}} = r_t] \\ &\qquad \times \Pr[\mathbf{n}_y^{\mathrm{T}} = n_t] \\ &= \sum_{n_b=n_t}^{m_y} \sum_{r_b=0}^{n_b} \frac{\mathbf{F}_y^{\mathrm{B}}(n_b, r_b)}{\Pr[\mathbf{n}_y^{\mathrm{B}} = n_b]} \\ &\qquad \times \Pr[\mathbf{n}_y^{\mathrm{B}} = n_b, \mathbf{r}_y^{\mathrm{B}} = r_b | \mathbf{n}_y^{\mathrm{T}} = n_t, \mathbf{r}_y^{\mathrm{T}} = r_t] \\ &\qquad \times \Pr[\mathbf{n}_y^{\mathrm{T}} = n_t]. \end{aligned}$$

After rearranging the conditional probabilities, we have

$$\Pr[\mathbf{n}_y^{\mathrm{B}}, \mathbf{r}_y^{\mathrm{B}} | \mathbf{n}_y^{\mathrm{T}}, \mathbf{r}_y^{\mathrm{T}}] = \Pr[\mathbf{r}_y^{\mathrm{B}} | \mathbf{n}_y^{\mathrm{B}}, \mathbf{n}_y^{\mathrm{T}}, \mathbf{r}_y^{\mathrm{T}}] \Pr[\mathbf{n}_y^{\mathrm{B}} | \mathbf{n}_y^{\mathrm{T}}, \mathbf{r}_y^{\mathrm{T}}],$$

which simplifies to $\Pr[\mathbf{r}_y^{\mathrm{B}} | \mathbf{n}_y^{\mathrm{B}}, \mathbf{n}_y^{\mathrm{T}}, \mathbf{r}_y^{\mathrm{T}}] \Pr[\mathbf{n}_y^{\mathrm{B}} | \mathbf{n}_y^{\mathrm{T}}]$, as the number of red lineages at the top of the branch ($\mathbf{r}_y^{\mathrm{T}}$) is conditionally independent of the number of red lineages at the bottom ($\mathbf{r}_y^{\mathrm{B}}$), given the total number of lineages at the top ($\mathbf{n}_y^{\mathrm{T}}$). Applying Bayes rule

$$\Pr[\mathbf{n}_y^{\mathrm{B}} | \mathbf{n}_y^{\mathrm{T}}] \frac{\Pr[\mathbf{n}_y^{\mathrm{T}}]}{\Pr[\mathbf{n}_y^{\mathrm{B}}]} = \Pr[\mathbf{n}_y^{\mathrm{T}} | \mathbf{n}_y^{\mathrm{B}}],$$

we obtain

$$\begin{aligned} \mathbf{F}_y^{\mathrm{T}}(n_t, r_t) &= \sum_{n_b=n_t}^{m_y} \sum_{r_b=0}^{n_b} \mathbf{F}_y^{\mathrm{B}}(n_b, r_b) \\ &\qquad \times \Pr[\mathbf{r}_y^{\mathrm{B}} = r_b | \mathbf{n}_y^{\mathrm{B}} = n_b, \mathbf{n}_y^{\mathrm{T}} = n_t, \mathbf{r}_y^{\mathrm{T}} = r_t] \\ &\qquad \times \Pr[\mathbf{n}_y^{\mathrm{T}} = n_t | \mathbf{n}_y^{\mathrm{B}} = n_b]. \end{aligned} \tag{14}$$

The term $\Pr[\mathbf{n}_y^{\mathrm{T}} = n_t | \mathbf{n}_y^{\mathrm{B}} = n_b]$ is evaluated using equation (6) above. Computing $\Pr[\mathbf{r}_y^{\mathrm{B}} | \mathbf{n}_y^{\mathrm{B}}, \mathbf{n}_y^{\mathrm{T}}, \mathbf{r}_y^{\mathrm{T}}]$ is more involved. In the special case that $u = v = 0$ a closed-form expression exists for this probability (Slatkin 1996). To our knowledge, a closed-form expression in the general case has not previously been derived. We express this probability using a matrix exponential.

Define the matrix $\mathbb{Q}$ with rows and columns indexed by pairs (n, r) and entries given by

$$\begin{aligned} \mathbb{Q}_{(n,r);(n,r-1)} &= (n - r + 1)v, \quad 0 < r \leqslant n, \\ \mathbb{Q}_{(n,r);(n,r+1)} &= (r + 1)u, \quad 0 \leqslant r < n, \\ \mathbb{Q}_{(n,r);(n-1,r)} &= \frac{(n - 1 - r)n}{\theta}, \quad 0 \leqslant r < n, \\ \mathbb{Q}_{(n,r);(n-1,r-1)} &= \frac{(r - 1)n}{\theta}, \quad 0 < r \leqslant n, \\ \mathbb{Q}_{(n,r);(n,r)} &= -\frac{n(n-1)}{\theta} \\ &\qquad -(n - r)v - ru, \quad 0 \leqslant r \leqslant n. \end{aligned} \tag{15}$$

All other entries in the matrix are 0. Here, n ranges from 1 to the number of individuals sampled, whereas for all n, we have $0 \leqslant r \leqslant n$. Hence, $\mathbb{Q}$ has

$$\sum_{n=1}^{m} (n + 1) = \frac{1}{2} m(m + 3)$$

rows and columns. We use $\mathbb{Q}_{(n,r);(n',r')}$ to denote the entry of $\mathbb{Q}$ in the row corresponding to pair (n, r) and column corresponding to pair (n', r'). We note that $\mathbb{Q}$ is not the generator of a process, and the connection with the coalescent and mutation processes is somewhat indirect. The most important feature (and justification) of the matrix is its role in computing the conditional allele probabilities required for the partial likelihoods.

Suppose that $\mathbf{n}^{\mathrm{B}}$ individuals are sampled from a Wright–Fisher population. Let $\mathbf{n}^{\mathrm{T}}$ denote the number of ancestral lineages at some time t in the past, and let $\mathbf{r}^{\mathrm{T}}$ be the number of these lineages that carry the red allele at that time. In the appendix, we use a result of Griffiths and Tavare (1997) to show that

$$\Pr[\mathbf{r}^{\mathrm{B}} = r | \mathbf{n}^{\mathrm{B}} = n, \mathbf{n}^{\mathrm{T}} = n_t, \mathbf{r}^{\mathrm{T}} = r_t] = \frac{\exp(\mathbb{Q}t)_{(n,r);(n_t,r_t)}}{\Pr[\mathbf{n}^{\mathrm{T}} = n_t | \mathbf{n}^{\mathrm{B}} = n]}. \tag{16}$$

Substituting equations (16) into (14), we can compute the values for $\mathbf{F}_y^{\mathrm{T}}(n_t, r_t)$ at the top of the branch given the respective values for $\mathbf{F}_y^{\mathrm{B}}(n_b, r_b)$ at the bottom of the branch.

Partial Likelihoods at a Speciation

Suppose that a branch x represents a population that diverges into two populations, corresponding to branches y and z. Each of the n lineages at the bottom of branch x came

up either from the top of branch y or from the top of branch z. If n_y is the number that came up from branch y, then $n-n_y$ is the number from branch z. The conditional joint distribution of $\mathbf{n}_y^{\mathsf{T}}$ and $\mathbf{n}_z^{\mathsf{T}}$ is then

$$\begin{aligned}\Pr[\mathbf{n}_y^{\mathsf{T}} = n_y, \mathbf{n}_z^{\mathsf{T}} = n - n_y | \mathbf{n}_x^{\mathsf{B}} = n] \\ = \frac{\Pr[\mathbf{n}_y^{\mathsf{T}}=n_y]\Pr[\mathbf{n}_z^{\mathsf{T}}=n-n_y]}{\Pr[\mathbf{n}_x^{\mathsf{B}}=n]}\end{aligned} \tag{17}$$

and is computed using equations (7), (8), and (9). Assuming $\mathbf{n}_x^{\mathsf{B}} = n$ and $\mathbf{r}_x^{\mathsf{B}} = r$ we have that the conditional distribution of red allele counts is given by the hypergeometric distribution:

$$\Pr[\mathbf{r}_y^{\mathsf{T}} = r_y, \mathbf{r}_z^{\mathsf{T}} = r - r_y | \mathbf{n}_y^{\mathsf{T}} = n_y, \mathbf{n}_z^{\mathsf{T}} = n - n_y] = \frac{\binom{n_y}{r_y}\binom{n-n_y}{r-r_y}}{\binom{n}{r}}. \tag{18}$$

The value of n_y can range from $n_y = 1$ (one lineage coming from branch y) to $n_y = n-1$ (all but one lineage coming from branch y). Combining equations (17) and (18), summing over n_y and r_y, and applying equations (10) and (11), we obtain

$$\begin{aligned}\mathbf{F}_x^{\mathsf{B}}(n,r) \quad = \quad & \sum_{n_y=1}^{(n-1)} \sum_{r_y=0}^{r} \mathbf{F}_y^{\mathsf{T}}(n_y, r_y)\mathbf{F}_z^{\mathsf{T}}(n-n_y, r-r_y) \\ & \times \frac{\binom{n_y}{r_y}\binom{n-n_y}{r-r_y}}{\binom{n}{r}}.\end{aligned} \tag{19}$$

This equation gives the joint probability of the allele counts in the subtree below branch y and the subtree below branch z, conditional on the sum of the lineage counts equalling n and the sum of the red lineage counts equalling r. Note, however, that the way we have defined $\mathbf{F}_y^{\mathsf{T}}(n,r)$ and $\mathbf{F}_z^{\mathsf{T}}(n,r)$ means these quantities are not partial likelihoods as they also include the lineage count probabilities at the nodes (see equations [10] and [11]).

Root Probabilities

Let ρ denote the root of the species tree. The probability of the observed data at a genetic marker, conditional on the species tree and model parameters, is

$$\begin{aligned}\Pr[\mathcal{R}_\rho] &= \sum_{n=1}^{m_\rho}\sum_{r=0}^{n} \Pr[\mathcal{R}_\rho | \mathbf{n}_\rho^{\mathsf{B}} = n, \mathbf{r}_\rho^{\mathsf{B}} = r]\Pr[\mathbf{n}_\rho^{\mathsf{B}} = n] \\ &\quad \times \Pr[\mathbf{r}_\rho^{\mathsf{B}} = r | \mathbf{n}_\rho^{\mathsf{B}} = n] \\ &= \sum_{n=1}^{m_\rho}\sum_{r=0}^{n} \mathbf{F}_\rho^{\mathsf{B}}(n,r)\Pr[\mathbf{r}_\rho^{\mathsf{B}} = r | \mathbf{n}_\rho^{\mathsf{B}} = n].\end{aligned} \tag{20}$$

Here, m_ρ is the total number of individuals sampled. The value for the probability $\Pr[\mathbf{r}_\rho^{\mathsf{B}} = r | \mathbf{n}_\rho^{\mathsf{B}} = n]$ depends on the choice of assumptions about what happens in the population above the root.

Using diffusion models, it can be shown that the allele frequencies in a single population have approximately a beta distribution (see, e.g., Ewens 2004, p. 174), and this is the distribution used for the root allele probabilities in RoyChoudhury et al. (2008). Here, we use the formulae for these probabilities under the coalescent model derived by Sawyer et al. (1987), Lundstrom (1990), and Griffiths and Tavare (1997), simplified to the case of biallelic markers. Let $\mathbf{N}$ and $\mathbf{R}$ be the (random) numbers of lineages and red lineages sampled from a single population of constant size. Let $\mathbb{Q}$ be the matrix defined in equation (15) and let $\mathbf{x}$ be be the nonzero solution for $\mathbb{Q}\mathbf{x} = \mathbf{0}$ such that $\mathbf{x}_{(1,0)} + \mathbf{x}_{(1,1)} = 1$. The vector $\mathbf{x}$ is indexed by pairs in the same way as $\mathbb{Q}$. Then, $\Pr[\mathbf{R} = r | \mathbf{N} = n] = \mathbf{x}_{(n,r)}$ for all n and r. The matrix $\mathbb{Q}$ is highly structured, and $\mathbf{x}$ can be computed using a simple recurrence in $O(m^2)$ time, where m is the maximum value for n.

Dominant Markers

AFLPs are dominant markers: heterozygotes cannot be distinguished from homozygotes for the dominant band, an issue that creates statistical difficulties (Lynch and Milligan 1994; Krauss 2000). We can include dominance explicitly within the likelihood calculation. Consider a biallelic locus for which the red allele is dominant. Given a sample of n diploid individuals from the population at leaf x, we consider the allele counts within the sample of $2n$ chromosomes. Suppose that there are r chromosomes with the red allele and r_x individuals with at least one copy of the red allele. It follows that exactly $r - r_x$ of the individuals will be homozygotes for the red allele and $2r_x - r$ will be heterozygotes. The remainder will be homozygotes for the green allele. Hence, the number of ways of distributing the r red alleles among the $2n$ chromosomes so that exactly r_x individuals have at least one red copy is

$$\frac{n!}{(r-r_x)!(2r_x-r)!(n-r_x)!}2^{2r_x-r}.$$

The exponential term, 2^{2r_x-r}, results from the two ways that the red allele can be assigned in each of the $2r_x - r$ heterozygotes. The probability that one of the $\binom{2n}{r}$ ways of assigning r alleles to the $2n$ chromosomes gives r_x individuals with at least one copy is then

$$\begin{aligned}\Pr[\mathbf{R}_x^{\text{dom}} &= r_x | \mathbf{R}_x = r] \\ &= \frac{n!}{(r-r_x)!(2r_x-r)!(n-r_x)!}2^{2r_x-r}\binom{2n}{r}^{-1},\end{aligned}$$

where $\mathbf{R}_x^{\text{dom}}$ is the observed number of individuals with at least one red allele in the population corresponding to leaf x.

To analyze dominant markers in SNAPP, we need only make two changes to the likelihood algorithm described above. First, the sample size for each population is doubled to reflect the fact that we are counting each chromosome as an individual. Second, the likelihood calculation for a leaf is modified. Suppose that the number of individuals from the associated species is n_x and the number of individuals carrying at least one copy of the red allele is r_x. Then

$$\mathbf{F}_x^{\mathsf{B}}(n,r) = \begin{cases} \frac{n!}{(r-r_x)!(2r_x-r)!(n-r_x)!}2^{2r_x-r}\binom{2n}{r}^{-1} & \text{if } n = 2n_x \\ & \text{and } r_x \leqslant r \leqslant 2r_x \\ 0 & \text{otherwise.}\end{cases}$$

Algorithm snappLikelihood
Computes the log-likelihood for biallelic data at a genetic marker

for each branch x of the species tree in a *post-order* traversal
  compute $\Pr[\mathbf{n}_x^{\mathrm{B}} = n]$ for all n, using (7) if x is external and (9) otherwise.
  if not at the root, compute $\Pr[\mathbf{n}_x^{\mathrm{T}} = n]$ for all n using (8).
end(for)
compute $\Pr[\mathbf{R}_\rho = r | \mathbf{N}_\rho = n]$ for all n, r
for each marker i
  for each branch x of the species tree in a *post-order* traversal
    compute $\mathbf{F}_x^{\mathrm{B}}(n, r)$ for all n, r, using (12) if x is external and (19) otherwise.
    if not at the root, compute $\mathbf{F}_x^{\mathrm{T}}(n, r)$ for all n, r using (14).
  end(for)
  compute $L_i = \Pr[\mathcal{R}_\rho]$ using (20).
end(for)
return $\sum_i \log(L_i)$.

FIG. 2. High-level outline of the algorithm to compute the log-likelihood of a set of unlinked biallelic markers, given the species tree. A branch *x* in the species tree is external if it is adjacent to a leaf; otherwise, it is internal. In equations (9) and (19), we use *y* and *z* to denote the branches attached to the base of branch *x*.

Below we explore the effect of including, or ignoring, this correction for dominant markers.

Time Complexity

Figure 2 gives a high-level description of the algorithm for computing the likelihood of a species tree given a collection of unlinked biallelic markers. The time complexity of the algorithm is dominated by two calculations. The first is the evaluation of $\mathbf{F}_x^{\mathrm{B}}(n, r)$ for all n, r using equation (19). A direct implementation of the formula would require $O(n^4)$ time per marker, per branch in the species tree, where n is the number of individuals sampled. However, the application of two-dimensional convolution algorithms reduces this complexity to $O(n^2 \log n)$ by using the fast Fourier transform (see Bracewell 2000).

The second time-consuming calculation is the computation of $\exp(\mathbb{Q}t)$, which is required for the application of equation (14). We found that standard diagonalization techniques were both computationally expensive and numerically unstable. Instead, we use the fact that after rearranging equation (14), we only need to be able to evaluate $\exp(\mathbb{Q}t)\mathbf{v}$ for different vectors $\mathbf{v}$. For this computation, we implemented a Carathéodory-Fejér approximation based on Schmelzer and Trefethen (2007) that runs in $O(n^2)$ time per species tree node. To check numerical accuracy, we also implemented the expokit algorithm of Sidje (1998), which is slower than the method of Schmelzer and Trefethen (2007) but has more numerical safety checks.

In summary, the time complexity of our likelihood calculation, per marker, is $O(sn^2 \log n)$, where n is the number of individuals and s is the number of species. We implemented a dynamic cache-based system to store partial likelihood values for different subtrees and multithreading to take advantage of parallel computation on multiple core machines or graphics processing units.

The SNAPP Sampler

We implemented our likelihood algorithm as the core of an MCMC software package SNAPP, which takes biallelic data (e.g., SNPs or AFLP) at multiple loci in a set of species and returns samples from the joint posterior distribution of

1. species phylogenies,
2. species divergence times,
3. effective population sizes at the root and along each branch of the species tree.

Note that our method does not sample gene trees; it only samples the species tree and its parameters.

The software is open source and is available for download from http://snapp.otago.ac.nz.

We have implemented a range of simple priors in the SNAPP package.

1. The stationary allele proportions are fixed at the observed proportions of red and green alleles in the data. In our experience, the posterior distribution of these parameters was tightly peaked at the observed value. These observed proportions also determine mutation rates u and v since we measure time in units of expected mutations. The user can also specify different values for the allele proportions or allow these to be sampled within the MCMC.
2. Following Drummond and Rambaut (2007), we assume a pure birth (Yule) model for the species tree topology and species divergence times, with a hyperparameter λ equal to the birth rate of the species tree. This hyperparameter is either fixed or allowed to vary with an improper uniform hyperprior.
3. Following Rannala and Yang (2003), we use independent gamma prior distributions for the population size parameters θ. We use fixed values for the shape α and inverse scale β parameters for the gamma distribution.

Later, we tabulate the parameters for the prior distributions that we used for simulations and for the analysis of the *Ourisia* data.

SNAPP gives the user a great deal of control over the exact combinations of priors and prior parameters used. A wide range of distributions is available for the model parameters,

and it is relatively straightforward to implement new prior distributions.

The MCMC proposal functions implemented in SNAPP are standard and are a subset of those available in BEAST (Drummond and Rambaut 2007) when sampling from molecular clock trees. Briefly, we implemented moves that raise or lower single nodes in the species tree, a move that swaps subtrees in the species tree, moves that alter θ values for single or multiple populations, and several moves that alter branch lengths and θ values simultaneously. See Drummond et al. (2002) for a detailed discussion of these moves, and Huelsenbeck et al. (2001) for a general overview of the use of MCMC methods to sample phylogenies. The SNAPP manual provides the most up-to-date list of MCMC proposals available.

The execution of the MCMC, and outputs, is controlled via a BEAST-style XML-file, which can be constructed using a graphical user interface. The user has the option of outputting a range of parameters and statistics computed from the Markov chain; many more are available by passing the output tree files through TreeStat (http://tree.bio.ed.ac.uk/software/treestat/). Convergence can be assessed for several statistics (e.g., likelihood values, tree length, tree height, and summary θ values) visually using Tracer (Rambaut and Drummond 2007) on multiple chains and using the Gelman–Rubin diagnostic (Gelman and Rubin 1992). Credibility sets are determined by ranking the topologies in the sample by decreasing sample frequency, and keeping as many trees as were necessary to obtain a total of 95% of the sample (after burn-in).

In principle, the model implemented in SNAPP can identify both divergence times and population sizes, parameters that are confounded in the models of Nielsen et al. (1998) and RoyChoudhury et al. (2008). However, care must be taken when interpreting the estimated θ values. If the amount of mutation is low, the data will satisfy the assumptions underlying the model of Nielsen et al. (1998) and RoyChoudhury et al. (2008), making the absolute θ values nonidentifiable. In these situations, scaling the divergence times and θ values by the same amount makes little or no change to the likelihood, and apparent precision in the estimates could be due to the prior rather than the data. For this reason, we recommend performing an additional analysis using modified priors for the tree height and θ values, so that their prior expectations are either both increased or reduced. We do this for our simulations and the *Ourisia* analysis.

Simulations

We have tested the likelihood algorithm and the SNAPP software extensively. The likelihood algorithm and sampler were originally implemented in C++ and reimplemented in Java. Core calculations, such as those required to apply equation (16), were independently reimplemented and tested in MATLAB. The likelihood values returned by the algorithm for two species and a small number of individuals were identical to those obtained analytically from gene tree probabilities.

We have implemented a simulator called SimSNAPP that generates polymorphic biallelic markers on a species tree according to the multispecies coalescent. Internally, SimSNAPP generates a gene tree within the species tree and then evolves the marker along that gene tree. If the marker is not polymorphic, both gene tree and marker are discarded. We note that MCMC-COAL (Rannala and Yang 2003) and Mesquite (Maddison WP and Maddison DR 2010) could also have been used to generate gene trees, though the fact that most gene trees are discarded made it more efficient for us to combine gene tree simulation and character simulation within a single program.

Simulation is a blunt tool for analyzing and comparing methods of inference, particularly in a context like this one where the parameter space for the model is large and complex. Here, we use simulation fairly conservatively. First, we run a check that when used as a means to infer species tree topologies, SNAPP returns a credibility set containing the tree used to simulate data. A failure to do this would indicate problems with the likelihood algorithm or implementation. In our second experiment, we demonstrate by example that SNAPP, in at least one case, is able to infer absolute θ values and divergence times. This ability represents a qualitative difference between SNAPP and the methods of Nielsen et al. (1998) and RoyChoudhury et al. (2008). A more difficult and complex problem, and one beyond the scope of this paper, would be to properly characterize the situations in which the θ values can be reliably inferred.

In our simulations, we began with the two four-species trees and two eight-species trees used in simulation experiments of Liu and Pearl (2007) (fig. 3). Two of the trees were classified as "hard" due to a short internal branch that would be difficult to resolve: We would expect to need more individuals or more markers to accurately infer these trees. We used the same θ values as Liu and Pearl (2007) for all internal branches and $\theta = 0.006$ for the root and external branches. The average θ values over all populations in these four trees are (A) 0.0060, (B) 0.0057, (C) 0.0061, and (D) 0.0047. The respective tree heights are (A) 0.024, (B) 0.014, (C) 0.018, and (D) 0.018 expected mutations per site.

Ability to Recover the Species Tree

We tested whether, as the number of sites increased, the species tree used to generate data appeared in the credibility set produced by SNAPP. This experiment tests the likelihood algorithm and the sampling algorithm simultaneously. As the number of sites increases, the likelihood function should concentrate around the true value (assuming identifiability) and, consequently, so should the posterior distribution. We also ran SNAPP with four choices of prior. For the θ values, we considered a "correct" prior with expectation close to the values used in simulation and an "incorrect" prior with values averaging 10% of the true values. In the same way, we considered a "correct prior" value for the speciation rate λ in the Yule model giving expected tree heights roughly equal to the tree heights of the simulated trees as well as an incorrect prior value for λ giving tree heights of around 50% of the true value. The parameter values used

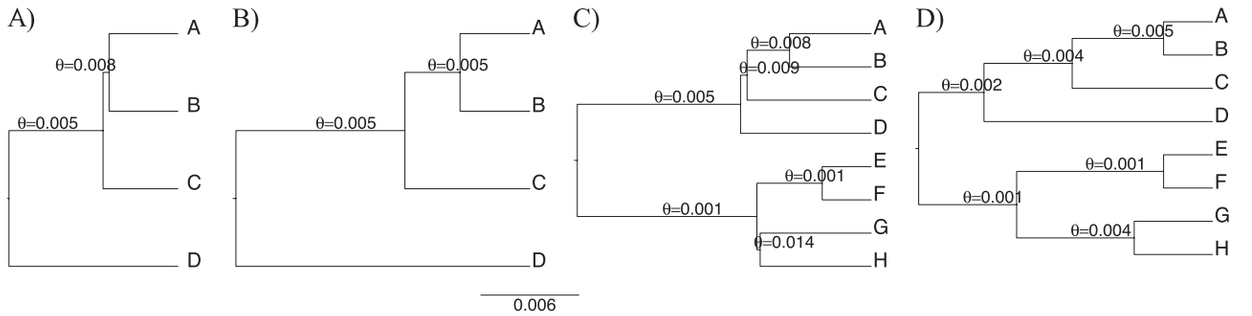

FIG. 3. Species trees used for simulations. These trees are identical to those used by Liu and Pearl (2007) to assess the reconstruction of species trees from known gene trees. θ values are indicated on the tree, and branch lengths are drawn to scale with respect to expected number of mutations (scale bar at base of figure). All external branches have $\theta = 0.006$. Branch lengths are drawn to scale to depict times (not θ values). (A) A "hard" four-taxa tree, the difficulty stemming from the short branch separating taxa A and B from C. (B) An "easy" four-taxon tree. (C) A hard eight-taxon tree, made difficult by the two short branches. (D) An easy eight-taxon tree.

for the priors are summarized in table 1. Chains were started from random trees and initial parameter values drawn from the prior, with chain lengths of 200,000 for the four-taxon case and 400,000 for the eight-taxon case, these lengths being determined by convergence tests on preliminary runs. For this experiment, we sampled only one individual per species. All simulations were run using an Opteron 24-core computer.

To examine the effect of correcting for dominant markers, we repeated the experiment using simulated dominant markers and ran SNAPP both with and without the dominant marker correction as described above. We also repeated the experiment using a modified version of SNAPP that assumed that no mutations occur along the branches, thereby emulating the model of Nielsen et al. (1998) and RoyChoudhury et al. (2008).

Ability to Recover Parameters

For the second experiment, we examined the posterior distribution of divergence time and θ values on a fixed tree, using the "easy" four-taxon tree (fig. 3*B*). We simulated 10,000 polymorphic loci for 40 individuals, with 10 individuals for each of the four species. We used the same priors for θ and λ as before (correct and incorrect) and generated chains of length 200,000.

Analysis of *Ourisia* AFLP Data

Meudt et al. (2009) investigated the utility of AFLP markers for species delimitation and reconstruction of evolutionary relationships between New Zealand populations of *Ourisia* (Plantaginaceae), the native foxglove. Molecular evidence suggests that *Ourisia* species have radiated fairly recently (between 0.4 and 1.3 Ma), adapting rapidly to a range of habitats, from sea level to alpine herbfields (Meudt et al. 2009).

AFLP markers are a readily available source of whole-genome information, well suited to the analysis of closely related species, particularly in the absence of whole-genome sequences (see review in Meudt and Clarke 2007). The analysis in Meudt et al. (2009) used a collection of 2,555 nonconstant AFLP markers for 193 *Ourisia* individuals, sampled from 100 locations in New Zealand and 3 locations in Australia. Several contrasting tree-based and cluster-based methods were applied, identifying 15 distinct meta-populations. Meudt et al. also detected strong evidence for a split between a "large-leaved group" and a "small-leaved group," the molecular signal for the split being consistent with differences in both morphology and habitat. The relationship between the species within each of these groups was not well resolved by any of their methods. Meudt et al. (2009) argued that this lack of resolution was not due to introgression or insufficient diversity. Two plausible explanations provided were the potentially low ratio of phylogenetic signal to noise in AFLP data and the effect of incomplete lineage sorting.

Here, we applied SNAPP to AFLP data from all members of the large-leaved group, producing a data matrix of 69 individuals from 6 populations and 1,997 characters. We used a diffuse gamma prior for the θ values ($\alpha = 10$, $\beta = 100$), with independent θ values on each branch. We used a pure birth (Yule) prior for the species tree, with birth rate λ sampled from an improper uniform hyperprior. We generated one chain of 790,000 iterations (sampling every 500 iterations) with additional smaller chains to check convergence and the effect of the prior. To assess the impact of the computationally intensive correction for dominant markers, we also generated a chain of 1.4 million iterations without this correction.

Table 1. Values Used for Priors in Simulations.

Prior	Version	Distribution	Expectation
θ	Correct (c)	Gamma ($\alpha = 1$, $\beta = 200$)	0.005
	Incorrect (i)	Gamma ($\alpha = 1$, $\beta = 2000$)	0.0005
Tree	Correct (c)	Yule ($\lambda = 40$)	0.014/0.021
	Incorrect (c)	Yule ($\lambda = 80$)	0.027/0.042

NOTE.—Note that the correct prior for θ values gives a prior expectation roughly the same as the average θ value in the input trees, whereas the incorrect prior has expectation 10% of that. The prior expectations for tree heights are given for four-taxon (first value) and eight-taxon (second value) trees. Height is measured in expected number of mutations per site (along a single lineage).

Table 2. The Size of the Credibility Sets in the Simulation Used to Assess Recovery of the Species Tree.

	Four Taxa								Eight Taxa							
Tree	Easy				Hard				Easy				Hard			
θ-Prior	c		i		c		i		c		i		c		i	
Tree Prior	c	i	c	i	c	i	c	i	c	i	c	i	c	i	c	i
100	1	1	1	1	3	3	3	3	1	1	1	1	22	21	12	14
200	1	1	1	1	3	3	3	3	1	1	1	1	9	9	8	8
300	1	1	1	1	3	3	2*	3	1	1	1	1	3*	3*	1*	1*
400	1	1	1	1	3	3	3	3	1	1	1	1	9	9	8	8
500	1	1	1	1	3	3	3	3	1	1	1	1	6	6	4	5
600	1	1	1	1	3	3	3	3	1	1	1	1	8	8	6	6
700	1	1	1	1	3	3	3	3	1	1	1	1	7	6	4	4
800	1	1	1	1	3	3	3	3	1	1	1	1	5	5	3*	3*
900	1	1	1	1	2	2	2	2	1	1	1	1	4	4	3	3
1,000	1	1	1	1	3	3	3	3	1	1	1	1	8	9	8	8
10,000	1	1	1	1	1	1	1	1	1	1	1	1	3	3	5	5
100,000	1	1	1	1	1	1	1	1	1	1	1	1	3	3	3	3
1,000,000	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2

NOTE.—"Tree" indicates which of the trees in figure 3 was used to generate data. Letters "c" and "i" indicate whether correct or incorrect priors were used on the θ values and on the speciation rate. Numbers 100 to 1,000,000 indicate the number of polymorphic sites generated. Values in the table are the numbers of distinct tree topologies in the 95% credibility set. The seven instances where the true tree was not contained within this set are marked by an asterisk (*).

Results

Simulation Experiments: Recovering the Species Tree

For the first experiment, we tested whether SNAPP would recover the species tree used to generate the data. We expected the true tree to be in the 95% credibility set for almost all replicates and that the size of the credibility set would shrink as the number of sites increased. (See table 2 for a summary of the sizes of the credibility sets constructed).

Four Taxa with an "Easy" Tree

In all simulations, the 95% credibility set contained the true tree and no other trees.

Four Taxa with a "Hard" Tree

The true tree was in the 95% credibility set for all except one instance. The credibility set contained three trees (the three resolutions of the short branch) for data sets with 100 to 1,000 sites. The credible set contained only the true tree with 10,000 or more sites. The choice of prior had little effect on outcomes.

Eight Taxa with an Easy Tree

In all simulations, the 95% credibility set contained the true tree and no other trees.

Eight Taxa with a Hard Tree

The true tree was in the 95% credibility set for all except two simulated data sets with 300 and 800 sites. Otherwise, the true tree was contained in the credibility set. There were at least three trees in the credibility set even for data sets with 1 million sites: SNAPP was unable to resolve the short edge in the species tree.

Overall, the credibility sets contained the true species tree topology in nearly all the experiments, and in many cases, only the true species tree. This result provides a good indication that the likelihood computation is working correctly. The processes were run on single threads. Each simulation (chain length 200,000) took around 4050 s for the four-taxon cases and between 2 and 4 min for the eight-taxon cases. Note that there was only one sample per species.

When dominant, rather than codominant, markers were simulated, little difference was observed in the ability of SNAPP to recover the species trees. Credibility sets had similar sizes to those seen in the case of codominant markers (supplementary data, Supplementary Material online).

Restricting SNAPP to a model with no mutation along the branches also did not have a noticeable impact on the ability of the software to reconstruct species trees (data not shown). This result was unexpected since, under this model, mutations that actually occurred along the branch would need to be explained by a polymorphism maintained all the way back to the root. It would be useful to explore the practical implications of excluding mutation in the model, both in theory and in application, although a thorough investigation of this issue falls beyond the scope of this paper.

Simulation Experiments: Recovering Parameters

The marginal posterior distributions for the node height and θ parameters appear in figure 4. We used two sets of priors, a correct prior centered on the true values and an incorrect prior where the prior expectations of the θ values were approximately one-tenth of the true values and the prior expectations of the tree height one-half of the true value. The posterior distributions given by the correct prior are indicated using solid lines, whereas those for the incorrect prior are given by dashed lines. The two posterior distributions almost coincide. This result gives strong evidence that the posterior distributions are influenced far more by the data than by the priors and also that in some situations, SNAPP is able to accurately estimate both θ values and branch lengths unlike earlier methods (Nielsen et al. 1998; RoyChoudhury et al. 2008).

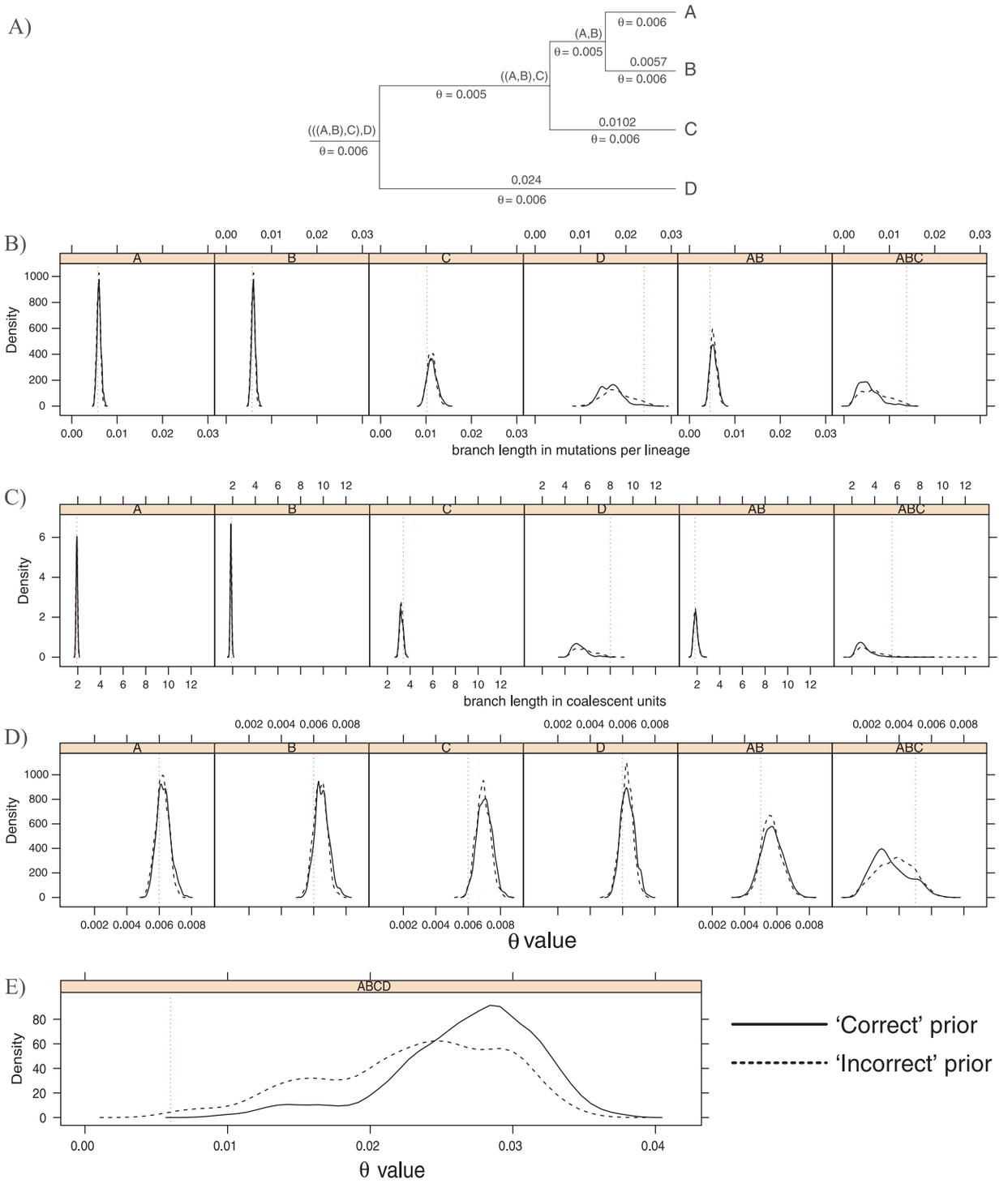

FIG. 4. Posterior distribution for θ and node heights for data simulated on a four-taxon tree. θ values are offset by subtracting off the corresponding true value on the model tree. Solid lines indicate use of the correct prior; dashed lines indicate use of the incorrect prior. True values are indicated by vertical gray lines. (*A*) The model tree used for the second simulation; (*B*) the posterior distribution for branch lengths in units of expected number of mutations; (*C*) posterior distributions for branch lengths in coalescent units; (*D*) posterior distributions for θ on internal branches; and (*E*) posterior distribution for θ at the root.

The posterior variances for the θ and branch length estimates differ considerably in different parts of the tree: The posterior variance for the θ value at the root is an order of magnitude greater than that for the other branches. We suspect that an important determinant for the posterior variance is the number of coalescent events occurring along each branch. When simulating the data, we recorded the numbers of lineages at each node in the

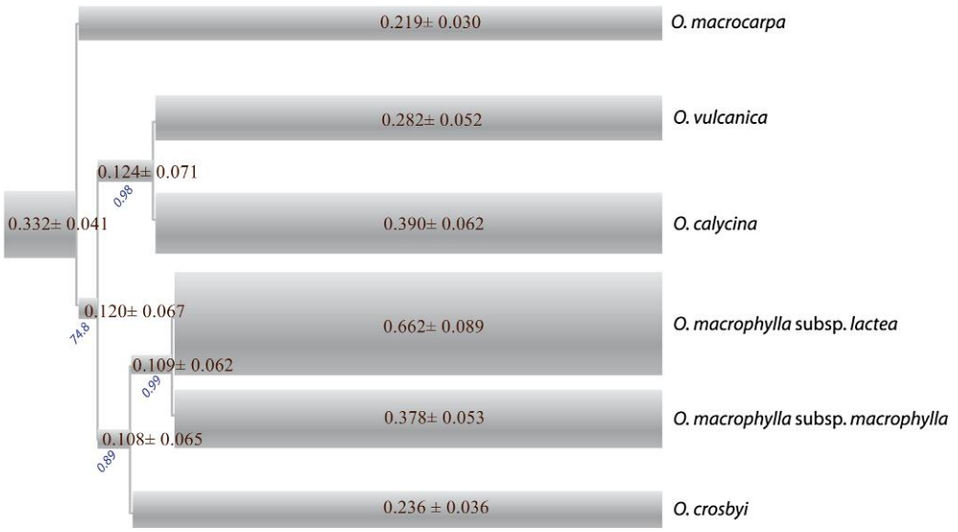

FIG. 5. Species tree with the highest posterior probability (73%) for six "large-leaf" *Ourisia* species. The thicknesses of bars are proportional to θ values for the respective populations. θ values for each population are printed on the pipes. The posterior probabilities for internal nodes are printed on an angle.

species tree. There were 10 lineages sampled from each population (A), (B), (C), and (D). There were, on average, 2.8 lineages at the base of population (A,B), 2.3 lineages at the base of (A,B,C), and 2.0 lineages at the base of the root population. Hence, very few coalescent events occurred along internal branches of the species tree, explaining the more diffuse posterior distributions for the corresponding θ values.

It took a little under 3 h to generate a chain of length 200,000 for the data sets with four species and ten samples per species, using four threads for each run.

Ourisia Data

We ran 790,000 iterations of the SNAPP MCMC algorithm in ~1,200 h of computing time on Opteron desktop computers, with three machines running 10 threads each. We note that although the data included 69 diploid individuals, the correction for dominance meant that, so far as computation was concerned, we were effectively analyzing 138 haploid individuals. Additional shorter chains were run to test convergence and the effect of the prior, and results were compared with an earlier run of 1.4 million iterations that did not correct for dominance. We removed 10% of the chains as burn-in and retained one in every 500 iterations for the sample.

The species tree topology represented in figure 5 has a posterior probability of 73%. The second most probable topology (15%) differed only by the position of the root. These trees had similar posterior probabilities when we ran SNAPP without the dominance correction. Interestingly, the AFLP phylogeny of Meudt et al. (2009), which was obtained using a MrBayes analysis that ignored lineage sorting, had <5% posterior probability.

SNAPP found significant support for several relationships between species in the large-leaved group. The (*O. vulcanica*, *O. calycina*) clade and the (*O. macrophylla*, *O. crosbyi*) clade both had significant posterior probability. The former clade also appears in the MrBayes tree of Meudt et al. (2009) though the latter does not.

One feature of the tree in figure 5 is that the divergence times for all the clades are early relative to the age of the tree. This result is consistent with a rapid species radiation at the base of the large-leaved group where an initial swift expansion was followed by a period of consolidation.

The θ estimates reported in figure 5 are shown with estimates of the posterior standard deviations. The standard errors on the sample means for the θ values were around 0.001, with the standard correction for autocorrelation (Ripley 1987, p. 143). The mean values differed little from those computed without the dominance correction, and from those for which the prior expectations for tree length and θ values were halved.

One anomaly is the θ estimate for *O. macrophylla* subsp. *lactea*, which is at least twice that of other species. A Neighbor-Net (Bryant and Moulton 2004) of the AFLP data reveals considerable substructure, and *O. macrophylla* subsp. *lactea* is not monophyletic in neighbor joining or parsimony analyses (Meudt et al. 2009). Hence, the high θ value could well represent fragmentation within the subspecies or poor delimitation of the subspecies with respect to *O. macrophylla* subsp. *macrophylla* rather than a large population.

In summary, by taking lineage sorting into account, we have been able to extract a well-supported phylogenetic tree for the large-leaved group of *Ourisia* species. Earlier tree-based and cluster-based analyses were unable to extract such a clear signal from these data (Meudt et al. 2009). Our analysis did not support the same tree as a Bayesian phylogenetic analysis that ignored incomplete lineage sorting. Furthermore, our θ estimates indicate potential fragmentation or poor delimitation in *O. macrophylla* subspecies *lactea*.

Discussion

We have presented a method that takes biallelic markers sampled from multiple individuals from multiple species and computes the likelihood of a species tree topology together with population genetic parameters. Our approach implements a full multispecies coalescent model without having to explicitly integrate or sample the gene trees at each locus. With our MCMC sampler, SNAPP, we can concentrate on the parameters of interest: the species tree, population sizes, and divergence times rather than on the problem of traversing through the space of potential gene trees. The likelihood values we compute are exact up to numerical error and do not require a simplification or approximation of the full coalescent model.

The model we implement differs from that of Nielsen et al. (1998) and RoyChoudhury et al. (2008) through its inclusion of mutation in populations other that the root population. Although mutation is rare in SNP data from the most closely related populations, it can play a significant role in the evolution of markers for more distant species, for trees with multiple species, or when analyzing markers such as AFLP that may have higher mutation rates than SNPs. We showed that modeling mutation permits the inference of both θ values and divergence times from biallelic markers, parameters that cannot be identified under the model of Nielsen et al. (1998) and RoyChoudhury et al. (2008).

To incorporate mutation, we have derived new theoretical results on the evolution of biallelic markers under the coalescent. These formulae extend work of Tavaré (1984) on the distribution of ancestral lineage counts and of Slatkin (1996) on the evolution of markers in a population without mutation. They combine the coalescent process, which operates backward in time, with the mutation process, which works forward in time.

The SNAPP sampler differs from methods such as BEST (Liu and Pearl 2007) and STAR-BEAST (Heled and Drummond 2010), which sample gene trees explicitly. Each of the methods is suited for different kinds of unlinked loci. BEST and STAR-BEAST can analyze loci with multiple sites, provided that no recombination within loci has occurred. SNAPP assumes that each locus is a single biallelic site. Unlike the other methods, SNAPP can analyze tens of thousands of unlinked markers, a feature that is not practical for methods that explicitly jointly sample one gene tree for each marker in a Monte Carlo algorithm.

We have reported some of the analyses we have performed to validate the algorithm and our implementation and to assess the ability of the method to infer phylogenetic and demographic parameters. Considerable scope exists for a more extensive investigation into the strengths, weaknesses and characteristics of the methodology with respect to other approaches. A wide variety of factors affect the performance of our method, or indeed any method, when inferring trees and parameters. (i) Sufficient mutation must have occurred but not so much mutation as to cause loss of signal; (ii) θ values can only be reliably inferred for ancestral populations if sufficiently many coalescent events occur within these populations; (iii) if θ values are too high, then there will be no coalescent events along the branches of the species tree and all coalescences will occur within the root population. In this case, little or no information is available to infer the tree or θ values. Alternatively, if the θ values are too low, then all coalescences will occur along pendant branches (as in the examples above), and only patchy information will be available about phylogenetic relationships and population sizes closer to the root of the species tree. These issues are likely to be faced not only by SNAPP but also by any method inferring phylogenies and population sizes.

The computational advances underlying SNAPP are made possible by some fairly stringent assumptions regarding both genetic and demographic processes. First, we assume genealogies for different markers are conditionally independent given the species tree. Violations of this assumption may not necessarily bias the analysis, though they are likely to bias measures of variability (RoyChoudhury 2011). Further, by assuming markers are unlinked, SNAPP fails to take into account patterns of linkage disequilibrium that can provide valuable information about demographics and genetic relationships. It may be possible to use theoretical advances combining the coalescent with recombination (Griffiths et al. 2008) to incorporate a model including linkage in future versions of SNAPP.

Second, we assume that gene dynamics within populations are well described by the (neutral) Wright–Fisher model, approximated by a coalescent process. In some cases, it might be possible to incorporate alternative models or variations simply by modifying the transition formula (15) for the allele frequencies. Other cases will demand a completely new approach.

Third, we assume that there is no gene flow between populations. Incorporating gene flow will be difficult mainly because the dynamic programming algorithm used within SNAPP relies on a lack of gene flow between descendent populations. Here, approaches based on diffusion processes are especially promising (Gutenkunst et al. 2009; Siren et al. 2010). The use of diffusions also makes it far easier to include selection and migration into the model (Gutenkunst et al. 2009). This flexibility incurs a computational cost however. The numerical methods used to evaluate likelihoods by Gutenkunst et al. (2009) require grid sizes (and hence running times) that grow exponentially in the number of populations, although careful approximations, perhaps along the lines of those used by Siren et al. (2010), might address this exponential explosion in runtime.

In many ways, SNAPP is itself an example of a combination of coalescent theory and diffusion theory: The model is based on the coalescent process and yet no explicit sampling of gene trees takes place. What remains to be developed is a combination of the computational advances of SNAPP and the flexibility of diffusion-based methods that provides a tractable method for implementing a full and rich model of demographic and genetic change.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank John Bryant, Alexei Drummond, David Fletcher, Colin Fox, Joseph Heled, Pete Lockhart, William Martin, Heidi Meudt, Melanie Pierson, Zachary Szpiech, and Elisabeth Thompson for help with numerous aspects of this work. D.B. received funding from the Alexander von Humboldt foundation, the NZ Marsden fund, and the Allan Wilson Centre for Molecular Ecology and Evolution. A.R. was supported in part by National Institutes of Health (NIH) program project (GM-45344; R01 GM071639-01A1) (PI: J Felsenstein). J.F. was funded by NIH (R01 GM071639; R01 HG004839) (PI: Mary Kuhner) and by interim "life support" funds from the Department of Genome Sciences, University of Washington. N.A.R. received funding from National Science Foundation (DBI-1146722), NIH (R01 GM 081441), and the Burroughs Wellcome Fund.

Appendix

For $0 \leqslant r \leqslant n$ and $t \geqslant 0$, we define

$$f_t(n,r) = \Pr[\mathbf{R}_t = r | \mathbf{N}_t = n].$$

If we write $\mathbf{x} = (r, n-r)$, then we have $f_t(n,r) = q(\mathbf{x},t)$ in the notation of Griffiths and Tavare (1997). Griffiths and Tavare (1997) show that $\frac{df}{dt} = \mathbb{Q}f$, where $\mathbb{Q}$ is the matrix defined in equation (15). Let $g_t(n,r)$ denote the conditional probability

$$\begin{aligned} g_t(n,r) &= \Pr[\mathbf{R}_t = r | \mathbf{N}_t = n, \mathbf{R}_\tau = r_\tau, \mathbf{N}_\tau = n_\tau] \\ &\quad \times \Pr[\mathbf{N}_\tau = n_\tau | \mathbf{N}_t = n] \\ &= \Pr[\mathbf{R}_t = r | \mathbf{N}_t = n, \mathbf{R}_\tau = r_\tau, \mathbf{N}_\tau = n_\tau] \\ &\quad \times \frac{\Pr[\mathbf{R}_\tau = r_\tau | \mathbf{N}_t = n, \mathbf{N}_\tau = n_\tau]}{\Pr[\mathbf{R}_\tau = r_\tau | \mathbf{N}_t = n, \mathbf{N}_\tau = n_\tau]} \\ &\quad \times \Pr[\mathbf{N}_\tau = n_\tau | \mathbf{N}_t = n] \\ &= \frac{\Pr[\mathbf{R}_t = r | \mathbf{N}_t = n]}{\Pr[\mathbf{R}_\tau = r_\tau | \mathbf{N}_\tau = n_\tau]} \\ &= \frac{f_t(n,r)}{f_\tau(n_\tau, r_\tau)}. \end{aligned}$$

Hence, $\frac{dg}{dt} = \mathbb{Q}g$ and equation (16) follows by solving this ordinary differential equation for $0 \leqslant t \leqslant \tau$ with the boundary conditions

$$g_\tau(n,r) = \begin{cases} 1 & n = n_\tau \text{ and } r = r_\tau, \\ 0 & \text{otherwise.} \end{cases}$$

References

Bracewell R. 2000. The Fourier transform and its applications. New York: McGraw-Hill.

Bryant D, Moulton V. 2004. NeighborNet: an agglomerative algorithm for the construction of planar phylogenetic networks. *Mol Biol Evol.* 21:255–265.

Degnan JH, Rosenberg NA. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol Evol.* 24:332–340.

Degnan JH, Salter LA. 2005. Gene tree distributions under the coalescent process. *Evolution* 59:24–37.

Donnelly PJ, Kurtz TG. 1996. A countable representation of the Fleming–Viot measure-valued diffusion. *Ann Appl Probab.* 24:698–742.

Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W. 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* 161:1307–1320.

Drummond AJ, Rambaut A. 2007. Beast: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol.* 7:214.

Efromovich S, Kubatko LS. 2008. Coalescent time distributions in trees of arbitrary size. *Stat Appl Genet Mol Biol.* 7:2.

Ewens WJ. 2004. Mathematical population genetics. 2nd ed. New York: Springer.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol.* 17:368–376.

Felsenstein J. 1988. Phylogenies and quantitative characters. *Annu Rev Ecol Syst.* 19:445–471.

Felsenstein J. 2004. Inferring phylogenies. Sunderland (MA): Sinauer Associates Inc.

Gelman A, Rubin DB. 1992. Inference from iterative simulation using multiple sequences. *Stat Sci.* 7:457–472.

Griffiths RC, Jenkins PA, Song YS. 2008. Importance sampling and the two-locus model with subdivided population structure. *Adv Appl Probab.* 40:473–500.

Griffiths RC, Tavare S. 1997. Computational methods for the coalescent. In: Progress in population genetics and human evolution. Berlin (Germany): Springer-Verlag. p. 165–182.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* 5:e1000695.

Hein J, Schierup MH, Wiuf C. 2005. Gene genealogies, variation and evolution: a primer in coalescent theory. Oxford: Oxford University Press.

Heled, J, Drummond, AJ. 2010. Bayesian inference of species trees from multilocus data. *Mol Biol Evol.* 27:570–580.

Hey J, Nielsen R. 2007. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proc Natl Acad Sci U S A.* 104:2785–2790.

Hudson RR. 1983. Testing the constant-rate neutral allele model with protein sequence data. *Evolution* 37:203–217.

Huelsenbeck JP, Ronquist F, Nielsen R, Bollback J. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* 294:2310–2314.

Krauss SL. 2000. Accurate gene diversity estimates from amplified fragment length polymorphism (AFLP) markers. *Mol Ecol.* 9: 1241–1245.

Kubatko LS, Carstens BC, Knowles LL. 2009. Stem: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics* 25:971–973.

Liu L, Pearl DK. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Syst Biol.* 56:504–514.

Lundstrom R. 1990. Stochastic models and statistical methods for DNA sequence data [PhD thesis]. [Salt Lake City (UT)]: University of Utah.

Lynch M, Milligan BG. 1994. Analysis of population genetic structure with RAPD markers. *Mol Ecol.* 3:91–99.

Maddison WP, Maddison DR. 2010. Mesquite: a modular system for evolutionary analysis. Version 2.75. [cited 2012 Apr 12]. Available from: http://mesquiteproject.org

Meudt HM, Clarke AC. 2007. Almost forgotten or latest practice? AFLP applications, analyses and advances. *Trends Plant Sci.* 12: 106–117.

Meudt HM, Lockhart PJ, Bryant D. 2009. Species delimitation and phylogeny of a New Zealand plant species radiation. *BMC Evol Biol.* 9:111.

Nei M. 1987. Molecular evolutionary genetics. New York: Columbia University Press.

Nielsen R. 1998. Maximum likelihood estimation of population divergence times and population phylogenies under the infinite sites model. *Theor Popul Biol.* 53:143–151.

Nielsen R, Mountain JL, Huelsenbeck JP, Slatkin M. 1998. Maximum-likelihood estimation of population divergence times and population phylogeny in models without mutation. *Evolution* 52: 669–677.

Pamilo P, Nei M. 1988. Relationships between gene trees and species trees. *Mol Biol Evol.* 5:568–583.

Rambaut A, Drummond AJ. 2007. Tracer v1.4. Available from: Phttp://beast.bio.ed.ac.uk/Tracer

Rannala B, Yang ZH. 2003. Bayesian estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* 164:1645–56.

Ripley B. 1987. Stochastic simulation. New York: Wiley.

Rosenberg NA, Nordborg M. 2002. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat Rev Genet.* 3:380–390.

RoyChoudhury A. 2006. Likelihood inference for population structure, using the coalescent [PhD thesis]. [Seattle (WA)]: University of Washington.

RoyChoudhury A. 2011. Composite likelihood-based inferences on genetic data from dependent loci. *J Math Biol.* 62:65–80.

RoyChoudhury A, Felsenstein J, Thompson EA. 2008. A two-stage pruning algorithm for likelihood computation for a population tree. *Genetics* 180:1095–1105.

Sawyer SA, Dykhuizen DE, Hartl DL. 1987. Confidence interval for the number of selectively neutral amino acid polymorphisms. *Proc Natl Acad Sci U S A.* 84:6225–6228.

Schmelzer T, Trefethen LN. 2007. Evaluating matrix functions for exponential integrators via Carathéodory-Fejér approximation and contour integrals. *Electron Trans Numer Anal.* 29:1–18.

Sidje RB. 1998. Expokit: a software package for computing matrix exponentials. *ACM Trans Math Softw.* 24:130–156.

Siren J, Marttinen P, Corander J. 2010. Reconstructing population histories from single nucleotide polymorphism data. *Mol Biol Evol.* 28:673–683.

Slatkin M. 1996. Gene genealogies within mutant allelic classes. *Genetics* 143:579–587.

Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105:437–460.

Takahata N. 1989. Gene genealogy in three related populations: consistency probability between gene and population trees. *Genetics* 122:957–66.

Takahata N, Nei M. 1985. Gene genealogy and variance of interpopulational nucleotide differences. *Genetics* 110:325–344.

Tavaré, S. 1984. Line-of-descent and genealogical processes, and their applications in population genetics models. *Theor Popul Biol.* 26:119–164.

Wakeley, J. 2009. Coalescent theory: an introduction. Greenwood Village (CO): Roberts and Company.

Wilson IJ, Weale ME, Balding DJ. 2003. Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. *J R Stat Soc A.* 166:155–201.

Wu Y. 2011. Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. *Evolution*. 66:763–775.