bioRxiv preprint doi: https://doi.org/10.1101/010199; this version posted March 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Species delimitation and phylogeny estimation under the multispecies coalescent

Graham Jones

March 22, 2015

Abstract

This article describes a Bayesian method for inferring both species delimitations and species trees under the multispecies coalescent model using DNA sequences from multiple loci. The focus here is on species delimitation with no a priori assignment of individuals to species, and no guide tree. The method uses a new model for the population sizes along the branches of the species tree, and three new operators for sampling from the posterior using the Markov chain Monte Carlo (MCMC) algorithm. The correctness of the moves is demonstrated both by proofs and by tests of the implementation. Current practice, using a pipeline approach to species delimitation under the multispecies coalescent, has been shown to have major problems on simulated data [10]. The same simulated data set is used to demonstrate the accuracy and efficiency of the present method. The method is implemented in a package called STACEY for BEAST2.

1 Introduction

Species delimitation is the problem of assigning a number of individual organisms to one or more species. The word 'delimitation' is also used to refer to a particular assignment or clustering of the individuals into groups or clusters. There are many approaches to this important problem and this article concentrates on the use of genetic data to infer such a clustering. The basic idea can be understood by considering two gene copies sampled at present time. Tracing their history back in time, they can coalesce within a group of interbreeding organisms according to a coalescent model, whereas between different groups, they cannot coalesce until the groups have merged. This idea is made precise by the multispecies coalescent model [15, 13, 2, 3].

There are two main types of 'noise' which interfere with inference of delimitation and phylogeny: mutational variance and incomplete lineage sorting. Mutational variance is a problem because the organisms are often closely related with few mutations separating them. This means there is a large amount of uncertainty about coalescence times. Incomplete lineage sorting refers to gene copies which fail to coalesce until after the (single) species to which they both

bioRxiv preprint doi: https://doi.org/10.1101/010199; this version posted March 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

belong has merged with another species.The degree of incomplete lineage sorting depends the effective population size along a branch, divided by the branch length measured in generations.

1.1 Previous work

Over the past ten years, the multispecies coalescent model has become the standard approach for species tree estimation using sequences from multiple loci. It accounts for incomplete lineage sorting, a ubiquitous source of discord between gene trees. More recently, it has been used for species delimitation, in BPP [16, 14, 17] and DISSECT [9]. More details and discussion of alternative approaches can be found in [9], [10], and [18].

DISSECT uses a prior for the species tree in which the usual birth-death model is replaced by one which incorporates a spike near zero in the density for node heights, a 'birth-death-collapse' model. This is a computational approximation to a model in which the dimensionality of the parameter space changes as the number of species changes.

In [10] sequences were simulated under the multispecies coalescent model, and then analyzed the data using a standard 'pipeline' method using Structurama [12], [8], *BEAST [5], and BPP [14]. The analysis showed there were major problems with the method.

1.2 Overview of the present method

The method presented here replaces this pipeline with a single analysis. It is implemented as a package called STACEY (Species Tree And Classification Estimation, Yarely) for BEAST2 [1]. STACEY is aimed mainly at species delimitation, but can also be used as an alternative to *BEAST [5]. It incorporates a new model for the populations along the branches of the SMC-tree, and three new MCMC moves for exploring the posterior when the multispecies coalescent model is assumed. It uses the same birth-death-collapse model as DISSECT.

The multispecies coalescent model requires a model for the populations along the branches of the SMC-tree. The simplest option is to assume that the population in all branches is identical and constant along each branch. Another option, as used in *BEAST, is to introduce one or more population parameters for each branch and sample these using the Markov chain Monte Carlo (MCMC) algorithm. Here is assumed that each branch in has a population parameter which is constant along the branch, and that these parameters are independent and identically distributed. Instead of sampling these parameters, they are integrated out. The method caters for variation among branches, and is similar to the 'piecewise constant' option in *BEAST but does not allow individual populations to be estimated. The hope is that this simplification makes the posterior easier to sample from.

To achieve this sampling, operators with the right statistical properties (MCMC moves) are needed. Their design is important for the efficiency of the method. The moves described here were designed with species delimitation

bioRxiv preprint doi: https://doi.org/10.1101/010199; this version posted March 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

in mind, although all three moves are also applicable to species tree estimation with a fixed species delimitation. Species delimitation presents a difficult challenge for the MCMC algorithm, since the MCMC moves must be capable of efficiently exploring all possible delimitations, and for each delimitation, all the usual parameters. In the multi-species coalescent model, there is one species tree and one or more gene trees. Each gene tree must 'fit inside' the species tree. A change to the species tree or to a gene tree may result in an incompatibility between the species tree and one or more gene trees. If an MCMC move makes such a change it must be rejected, and if such rejections are common the move will be inefficient. The three moves described here preserve compatibility between the species tree and the gene trees.

The first MCMC move, called NodesNudge, changes the height of a node in a SMC-tree, and changes the height of certain 'nearby' nodes in the gene trees. It does this in a way that leaves the all tree topologies unchanged, and preserves the compatibility of the gene trees with the SMC-tree. It is a subtle move, in that it typically changes the node heights by a small amount, but it appears to have a large beneficial effect on the convergence, at least on some data sets.

The second move, called CoordinatedPruneRegraft, is a subtree-prune-and-regraft move which makes coordinated topological changes to the species tree and gene trees. The CoordinatedPruneRegraft move can be seen as an extension of the nearest neighbor interchange (NNI) move described in [17], which makes a coordinated set of fixed node height NNI moves to the species tree and to the gene trees. When viewed this way, the CoordinatedPruneRegraft extends the NNI move to the more general subtree-prune-and-regraft move. It can also be seen as an extension to the 'Fixed Nodeheight Prune and Regraft' (FNPR) as described in [6]. The FNPR move changes the topology of a single tree, whereas the move described here makes a coordinated set of FNPR moves to the species tree and to the gene trees in order to maintain compatibility between the trees.

The third move, called FocusedScaler, scales node heights whilst preserving topologies. The scaling is 'focused' on a node in the species tree. This node is scaled by the largest amount. The further away a node is from the focus (in a sense to be made precise later), the less it is affected by the move. Once the relative amount by which each node should be scaled by has been chosen, the maximum range of scaling consistent with compatibility is found. The actual scaling is then chosen from this range.

The population model is described first, followed by the MCMC moves. Two sets of tests on simulated data are then described. Firstly, the method is tested for correctness by sampling from prior distributions in cases where some analytic results are available. Finally, the simulated data set of [10] is re-analyzed.

2 Conventions and notation

All trees are rooted and binary. Time is measured backwards from zero at present, and all tree nodes have a time, referred to as a node height. A tree topology should be understood as a labeled topology, that is, it includes the as-

bioRxiv preprint doi: https://doi.org/10.1101/010199; this version posted March 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

signment of labels to tips. In the context of species delimitation using STACEY, the species tree has tips which represent **minimal clusters** of individuals [9]. These minimal clusters may be merged but not split to form potential species. At its most flexible, there is just one individual in each minimal cluster, so the possible number of species ranges from one to the number of individuals. Thus 'species tree' is not a good name for this tree, and instead I will refer to it a the **SMC-tree**, as a shorthand for 'species or minimal clusters tree'.

Lower case letters are used for gene tree nodes, and upper case for SMC-tree nodes and in situations where the type of tree does not matter. For either type of node X, its parent is denoted by $\text{anc}(X)$ and its node height by $t(X)$. The branch that leads from $\text{anc}(X)$ to X is referred to as 'the branch X'. The 'subtree of X' contains X, all its descendants, and the branch X, but not the node $\text{anc}(X)$ which is the origin of the subtree.

For a node X in the SMC-tree, let $I(X)$ denote the set of minimal clusters belonging to X (that is, assigned to a tip node which is a descendant of X). For a node x in a gene tree, let $I(x)$ denote the set of the minimal clusters which yielded a sequence belonging to x. Furthermore, if X is not a tip, let $R(X)$ and $L(X)$ denote the set of minimal clusters belonging to the two children of X. Note that $I(X) = L(X) \cup R(X)$ for both SMC-tree nodes and gene tree nodes, so they can be calculated recursively from the tips, and all these sets are unions of minimal clusters. In the SMC-tree the unions are disjoint, and a node is uniquely identified by its set of minimal clusters. In the gene tree case, neither of these is true in general. However, the set $I(x)$ and height $t(x)$ for a gene tree node x are enough to assign x to a unique branch in the SMC-tree, as follows. If the SMC-tree is cut across at height $t(x)$, this will intersect some branches $X_1, X_2, \ldots, X_n$ say. All the $I(X_i)$ are pairwise disjoint, and $I(x)$ cannot intersect more than one of them non-trivially or the gene tree would be incompatible with the SMC-tree. Thus $I(x) \subset I(X_i)$ for some i thus identifying the branch X_i as the one which contains x.

The notion that part of a gene tree is 'inside' a branch of a SMC-tree is intuitively obvious from diagrams, but a formal definition is required for algorithms and proofs. Suppose X is a node in the SMC-tree and x is a gene tree node and $t \in [t(x), t(\text{anc}(x))]$. Then the point (x, t) is **inside** the branch X if $I(x) \subset I(X)$ and $t \in [t(x), t(\text{anc}(X))]$.

We also formally define the notion of compatibility. Firstly, if A is a set of minimal clusters, and X is a node in the SMC-tree, we say that A **straddles** X if X is not a tip, $A \cap L(X) \neq \emptyset$, and $A \cap R(X) \neq \emptyset$. If X is a node in the SMC-tree and x is a gene tree node, then the pair (X, x) is **compatible** if $t(x) \geq t(X)$ or $I(x)$ does not straddle X. A gene tree is compatible with the SMC-tree if every pair of nodes (X, x) is compatible.

We define a pair of nodes (X, x) with X in the SMC-tree and x in a gene tree to be **hitched** if $I(x)$ straddles X, but neither $L(x)$ nor $R(x)$ straddle X. See Figure 1 for an example. The hitched nodes are the minimal set of nodes that need to be checked for compatibility to ensure the SMC-tree and the gene tree are compatible. See Figure 1.

bioRxiv preprint doi: https://doi.org/10.1101/010199; this version posted March 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

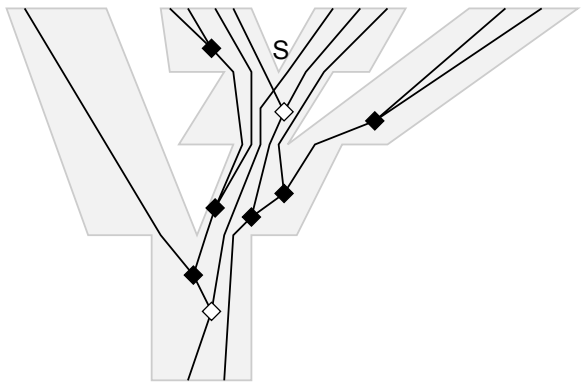

Figure 1: Example of hitched nodes. The SMC-tree is pale gray. A gene tree is shown inside it. Gene tree nodes which are hitched to the SMC-tree node S are shown as white diamonds, and other nodes as black diamonds.

3 The population model

Consider a single gene and a single branch. The coalescent model of Kingman (see Chapters 26-28 of [4]) is assumed. The probability density for the coalescent times takes the following form (simplified from equation (3), p572, of [5]):

$$f_L(L|P) =$$
$$\prod_{i=0}^{k-1} P^{-1} \prod_{i=0}^{k} \exp\left(- \int_{t_i}^{t_{i+1}} \binom{n-i}{2} P^{-1} \mathrm{d}t \right) =$$
$$P^{-k} \exp\left(- \left[\sum_{i=0}^{k} (t_{i+1} - t_i) \binom{n-i}{2} \right] P^{-1} \right) \tag{1}$$

where L is the lineage history of a gene tree within a single branch, and P is the effective number of gene copies in the population for this branch, which is assumed constant along the branch in this paper. Thus P is the expected number of generations for a pair of gene copies to coalesce. The lineage history L consists of the number n of lineages at the tipward end of the branch, the number k of coalescences within the branch, plus the times $(t_0 < t_1, ...t_k < t_{k+1})$ where t_0 is the node height at the tipward end, t_{k+1} is the node height at the rootward end, and $(t_1, ...t_k)$ are the coalescence times within the branch. Between t_i and t_{i+1} there are $n - i$ lineages. The complete multispecies coalescent probability density is the double product, over genes and over branches, of terms like this.

As usual, we convert P into substitution units by multiplying by the mutation rate measured in substitutions per site per generation. Denote the effective population in branch b by N_b and the mutation rate by μ_b. The effective number of gene copies is obtained from N_b by multiplying by a factor p_j (sometimes

bioRxiv preprint doi: https://doi.org/10.1101/010199; this version posted March 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

called the 'ploidy') for gene j. This p_j depends on the type of gene involved, and is 2 for the common case of autosomal nuclear genes in diploid species. Exceptions include genes from sex chromosomes and organelles. For gene j in branch b, we thus need to replace P by $p_j N_b \mu_b$ in equation (1).

To write down the full expression, some more notation is needed. The branches in the SMC-tree are indexed by b. A sum or product over b should be understood as being over all branches. Note that this includes the root, so that all gene lineages eventually coalesce. The number of branches is B. Set $\theta_b = N_b \mu_b$. The vector $(\theta_1, \theta_2, \ldots, \theta_B)$ is denoted by Θ. The genes are indexed by j. A sum or product over j should be understood as being over all genes. The number of coalescences of gene j within branch b is denoted by k_{jb}. The number of lineages in gene tree j at the tipward end of branch b is denoted by n_{jb}. The number of lineages in gene tree j at the rootward end of branch b is thus $n_{jb} - k_{jb}$. The time interval between the tipward and rootward branch b is divided into $k_{jb} + 1$ intervals by the coalescent times of gene j. These $k_{jb} + 1$ intervals are denoted by c_{jbi} $(0 \leq i \leq k_{jb})$. There are $n_{jb} - i$ lineages in gene tree j, branch b during the time interval c_{jbi}. Let G denote all the lineage histories of all the genes in all the branches. The complete multispecies coalescent probability density is

$$\begin{aligned} f_G(G|\Theta) &= \prod_j \prod_b (p_j \theta_b)^{-k_{jb}} \exp\left(- \left[\sum_{i=0}^{k_{jb}} c_{jbi} \binom{n_{jb} - i}{2} \right] (p_j \theta_b)^{-1} \right) \\ &= \prod_b r_b \theta_b^{-q_b} \exp\left(- \gamma_b \theta_b^{-1} \right) \end{aligned}$$

where

$$q_b = \sum_j k_{jb}, \qquad r_b = \prod_j p_j^{-k_{jb}}, \qquad \text{and} \quad \gamma_b = \sum_j p_j^{-1} \sum_{i=0}^{k_{jb}} c_{jbi} \binom{n_{jb} - i}{2}. \tag{2}$$

For each b this has the form of an unnormalised inverse gamma density for θ_b. The normalised inverse gamma density is

$$\mathcal{IG}(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} \exp(-\beta x^{-1}) \mathbf{1}_{[0,\infty)}$$

where α and β are parameters in $(0, \infty)$. If, a priori, the θ_b are assumed independent and are assumed to have an inverse gamma density it is possible to integrate out the θ_b analytically. In fact the prior can be more general than a single inverse gamma density: an overall scaling parameter σ can be introduced, together with hyperprior $\pi_\sigma(\sigma)$ for it; and a mixture of inverse gamma densities can be used. This mixture takes the form

$$h(x|\sigma) = \sum_{c=1}^{C} \lambda_c \mathcal{IG}(x;\, \alpha_c,\, \sigma\beta_c)$$

bioRxiv preprint doi: https://doi.org/10.1101/010199; this version posted March 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Here C, the λ_c, the α_c, and the β_c ($1 \leq c \leq C$) are user-chosen values, which are constant for the analysis. The λ_c are positive and sum to one, and the α_c and β_c are arbitrary positive numbers. The density π_σ is also user-chosen and can be any density with support contained in $[0, \infty)$. Each θ_b is then an independent draw from the density h. So the joint prior density for Θ and σ is

$$\begin{aligned}
&\pi_\Theta(\Theta|\sigma)\pi_\sigma(\sigma) = \\
&\pi_\sigma(\sigma)\prod_b h(\theta_b|\sigma) = \\
&\pi_\sigma(\sigma)\prod_b \sum_{c=1}^{C} \lambda_c(\sigma\beta_c)^{\alpha_c}\Gamma(\alpha_c)^{-1}\theta_b^{-\alpha_c-1}\exp(-\sigma\beta_c\theta_b^{-1})\mathbf{1}_X
\end{aligned} \tag{3}$$

where X is the positive orthant in $\mathbf{R}^B$.

Then combining (2) and (3), the posterior density for the multispecies coalescent is

$$\begin{aligned}
&f_G(G|\Theta)\pi_\Theta(\Theta|\sigma)\pi_\sigma(\sigma) = \\
&\pi_\sigma(\sigma)\prod_b \sum_{c=1}^{C} f(\sigma, \lambda_c, \alpha_c, \beta_c, \theta_b, q_b, \gamma_b, r_b)\mathbf{1}_X
\end{aligned}$$

where $f(\sigma, \lambda_c, \alpha_c, \beta_c, \theta_b, q_b, \gamma_b, r_b) =$

$$\begin{aligned}
&\lambda_c\frac{(\sigma\beta_c)^{\alpha_c}}{\Gamma(\alpha_c)}\theta_b^{-\alpha_c-1}\exp(-\sigma\beta_c\theta_b^{-1})r_b\theta_b^{-q_b}\exp\big(-\gamma_b\theta_b^{-1}\big) = \\
&\frac{\lambda_c r_b(\sigma\beta_c)^{\alpha_c}}{\Gamma(\alpha_c)}\theta_b^{-\alpha_c-1-q_b}\exp\big(-(\sigma\beta_c+\gamma_b)\theta_b^{-1}\big) = \\
&\frac{\lambda_c r_b(\sigma\beta_c)^{\alpha_c}}{(\sigma\beta_c+\gamma_b)^{\alpha_c+q_b}}\frac{\Gamma(\alpha_c+q_b)}{\Gamma(\alpha_c)}\frac{(\sigma\beta_c+\gamma_b)^{(\alpha_c+q_b)}}{\Gamma(\alpha_c+q_b)}\times \\
&\qquad\qquad \theta_b^{-(\alpha_c+q_b)-1}\exp\big((\sigma\beta_c+\gamma_b)\theta_b^{-1}\big) = \\
&\frac{\lambda_c r_b(\sigma\beta_c)^{\alpha_c}}{(\sigma\beta_c+\gamma_b)^{\alpha_c+q_b}}\frac{\Gamma(\alpha_c+q_b)}{\Gamma(\alpha_c)}\mathcal{IG}(\theta_b; \alpha_c+q_b, \sigma\beta_c+\gamma_b).
\end{aligned}$$

Now Θ can be integrated out from the posterior, using the fact that $\mathcal{IG}$ integrates to 1 to obtain

$$\begin{aligned}
&\int_X f_G(G|\Theta)\pi_\Theta(\Theta|\sigma)\pi_\sigma(\sigma)\mathrm{d}\Theta = \\
&\pi_\sigma(\sigma)\prod_b r_b\sum_{c=1}^{C}\lambda_c\frac{(\sigma\beta_c)^{\alpha_c}}{(\sigma\beta_c+\gamma_b)^{\alpha_c+q_b}}\frac{\Gamma(\alpha_c+q_b)}{\Gamma(\alpha_c)}.
\end{aligned} \tag{4}$$

Equations (4) and (2) provide the information needed to implement the method.

bioRxiv preprint doi: https://doi.org/10.1101/010199; this version posted March 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

4 The NodesNudge move

4.1 The algorithm

We describe a more general algorithm than the move which is currently implemented, since it may be useful to use variants of the move. It uses the concept of a connected component from graph theory. Given a subset Δ of the nodes in a gene tree, we first remove the nodes not in Δ, then divide what is left into the connected components. Figure 2 illustrates the idea. On the left is a gene tree, in which nodes are shown by solid diamonds if they are in Δ and open diamonds otherwise. On the right, the three connected components in Δ are shown as diamonds and solid lines. For any gene tree node $x \in \Delta$, let $C(x)$ denote the connected component in the gene tree to which x belongs. Furthermore, define the set $C^*(x)$ to be the set of nodes c in the gene tree such that $\text{anc}(c) \in C(x)$ and $c \notin C(x)$. Their positions are at the tops of the dotted lines in the right of the figure, and can be thought of as the 'children' of $C(x)$. Finally let $r(x)$ be the oldest node in $C(x)$, the root of $C(x)$. Here is the algorithm:

1. Choose uniformly and at random any node S in the SMC-tree which is not a tip and not the root.

2. Let $\Delta(S) = \{s_1, \ldots, s_n\}$ be a set of internal gene tree nodes defined by a criterion which only depends on S, and the topologies of the SMC-tree and gene trees.

3. Let $d_0 = \max_X\{t(X) : \text{anc}(X) = S\}$, which is the time of the most ancient of the two child nodes of S, and let $u_0 = t(\text{anc}(S))$.

4. For $1 \leq i \leq n$, let

$$d_i = t(s_i) + \max_c\{t(c) - t(\text{anc}(c)) : c \in C^*(s_i)\}$$

   and let $u_i = \infty$ if s_i is the root, otherwise let $u_i = t(s_i) + t(\text{anc}(r(s_i))) - t(r(s_i))$.

5. Let

$$D = \max_{0 \leq i \leq n}(d_i - t(s_i) + t(S))$$

   and

$$U = \min_{0 \leq i \leq n}(u_i - t(s_i) + t(S)).$$

6. Choose a new node height $t'(S)$ for S uniformly in $[D, U]$.

7. Change the height of all gene tree nodes in $\Delta(S)$ by the same amount, that is, by $t'(S) - t(S)$.

We formally prove some properties of this move below; here is an informal description of the ideas behind the proofs. Firstly, note that the value d_0 (step

bioRxiv preprint doi: https://doi.org/10.1101/010199; this version posted March 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

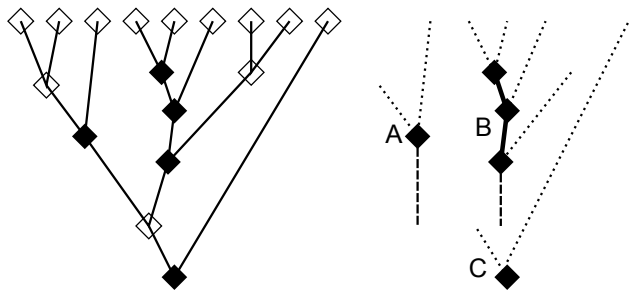

Figure 2: Example of connected components.

3) can be written as $t(S) + \max_X\{t(X) - t(S) : \text{anc}(X) = S\}$ which emphasizes the similarity with the other d_i (step 4). Next, note that since S is not he root, u_0 is finite, so that $[D, U]$ is a finite interval, and step 6 makes sense.

Now we explain the role of the connected components. Returning to Figure 2, note that the minimum length of the dotted edges leaving each connected component determines the maximum amount by which the nodes in the connected component can move forward in time. The oldest node in each connected component usually provides a limit (the length of the dashed line) on how far back in time the connected component can move; the exception is if it is the root node of the gene tree, as in the case of connected component C. The key property of connected components is that the limit of movement back and forwards in time of one connected component is determined by the times of nodes which cannot belong to another connected component. This ensures that the definitions of D and U are unaffected by the move. Also note that all nodes are moved by the same amount, so the internal structure of $C(s_i)$ does not change.

4.2 Properties

Proposition 1 *The NodesNudge move preserves all the tree topologies and keeps all branch lengths nonegative.*

Proof. From step 6, $D \leq t'(S) \leq U$, and from step 5, $d_0 \leq t'(S) \leq u_0$, and so it follows from step 3 that

$$\max_X\{t(X) : \text{anc}(X) = S\} \leq t'(S) \leq t(\text{anc}(S))$$

hence the new height for S lies between that of S's oldest child and its parent.

Now assume $1 \leq i \leq n$. From step 7, the new height $t'(s_i)$ is $t(s_i) + t'(S) - t(S)$. From steps 5 and 6,

$$d_i - t(s_i) + t(S) \leq t'(S) \leq u_i - t(s_i) + t(S)$$

so

$$d_i \leq t'(s_i) \leq u_i.$$

bioRxiv preprint doi: https://doi.org/10.1101/010199; this version posted March 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

The next step is to show that the definition of d_i and u_i in step 4 preserves the topologies and keeps all branch lengths non-negative in the gene trees. The condition $c \in C^*(s_i)$ identifies all pairs of nodes $(c, \mathrm{anc}(c))$ such that $\mathrm{anc}(c)$ is in the connected component and c outside, so the maximum of $t(c) - t(\mathrm{anc}(c))$ over such c is the biggest negative value by which this connected component can move. Likewise $t(\mathrm{anc}(r(s_i))) - t(r(s_i))$ is the maximum positive value by which this connected component can move. Thus any new times for the nodes $s_j \in C(s_i)$ that are in $[d_j, u_j]$ for all j such that $s_j \in C(s_i)$ will preserve this connected component.

Put $\delta = t'(S) - t(S)$, the amount by which the node times are changed. Then from step 6,

$$D \leq t'(S) \leq U$$

so

$$D - t(S) \leq \delta \leq U - t(S)$$

and for all i,

$$d_i - t(s_i) + t(S) - t(S) \leq \delta \leq u_i - t(s_i) + t(S) - t(S)$$

hence

$$d_i \leq t(s_i) + \delta \leq u_i$$

as required.

Proposition 2 *The NodesNudge move is symmetric.*

Proof. First note that the choice of S in step 1 has the same probability for the reverse move. Then, the key property of connected components described earlier together with Proposition 1, ensures that $\Delta(S)$ is unaffected by the move. It only remains to show that the interval $[D, U]$ is unaffected by the move. Primes ($'$) are used to denote the various quantities after the move. From step 3, $d'_0 = d_0$ and $u'_0 = u_0$, and for $i > 0$, from step 4 we have

$$d_i = t(s_i) + \min_c \{t(c) - t(\mathrm{anc}(c)) : c \in C^*(s_i)\}$$

and since all the $t(c)$ values are unaffected by the move, and all $t(\mathrm{anc}(c))$ are changed by δ, as is $t(s_i)$, it follows that

$$\begin{aligned} d'_i &= t'(s_i) + \\ &\quad \min_c \{t'(c) - t'(\mathrm{anc}(c)) : c \in C^*(s_i)\} \\ &= t(s_i) + \delta + \\ &\quad \min_c \{t(c) - \delta - t(\mathrm{anc}(c)) : c \in C^*(s_i)\} \\ &= d_i. \end{aligned}$$

Similarly, $u'_i = u_i$, and it follows that $[D', U'] = [D, U]$. Since the choice of $t'(S)$ in $[D, U]$ is uniform, the move is symmetric.

bioRxiv preprint doi: https://doi.org/10.1101/010199; this version posted March 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

4.3 The definition of $\Delta(S)$

In the NodesNudge move as currently implemented, $\Delta(S)$ is defined to be all the internal gene tree nodes s such that the pair (S, s) is hitched. For this definition of $\Delta(S)$, it is not possible for a node and its parent to both belong to $\Delta(S)$, so all the connected components $C(x)$ in the algorithm consist of single nodes. This simplifies the implementation, and can be used to simplify the proofs of the two Propositions. However it is likely that future versions of STACEY will exploit the more general case.

5 The coordinated subtree and regraft move

5.1 The subtree prune and regraft move for one tree

First we describe the fixed height subtree prune and regraft algorithm [6] as it applies to a single tree. This is to make precise the algorithm used here, since there are variants of the main idea. Figure 3 illustrates the process. The algorithm prunes a subtree S and regrafts it into a branch D in both SMC-tree and gene trees. The requirements are that neither S nor $\text{anc}(S)$ is the root of the tree, none of S, $\text{anc}(S)$, or the sibling of S can be D, and $t(D) \leq t(\text{anc}(S)) \leq t(\text{anc}(D))$. It is possible for the D and $\text{anc}(S)$ to be siblings (so in the figure, Y can be equal to Z). Note that there is no change in the set of node heights; the existing heights are re-used. The subtree prune and regraft algorithm for one tree follows.

1. Let B be the sibling of S, Y the parent of $\text{anc}(S)$, and Z the parent of D.
2. Remove child D from Z. Remove child B from $\text{anc}(S)$. Remove child $\text{anc}(S)$ from Y.
3. Add $\text{anc}(S)$ as child of Z. Add D as child of $\text{anc}(S)$. Add B as child of Y.

5.2 The algorithm

The idea is to make a subtree prune and regraft move on the SMC-tree and a co-ordinated set of subtree prune and regraft moves on each of the gene trees in order to make them compatible with the new SMC-tree. Figure 4 shows an example. The node S is the subtree to be pruned and the branch D is the destination branch into which S is regrafted. Here is the algorithm.

1. Choose at random any node S such that neither S nor $\text{anc}(S)$ is the root. Let B be the sibling of S.
2. Choose at random any node D which is none of S, $\text{anc}(S)$ or B, and such that $t(D) \leq t(\text{anc}(S)) \leq t(\text{anc}(D))$.

bioRxiv preprint doi: https://doi.org/10.1101/010199; this version posted March 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

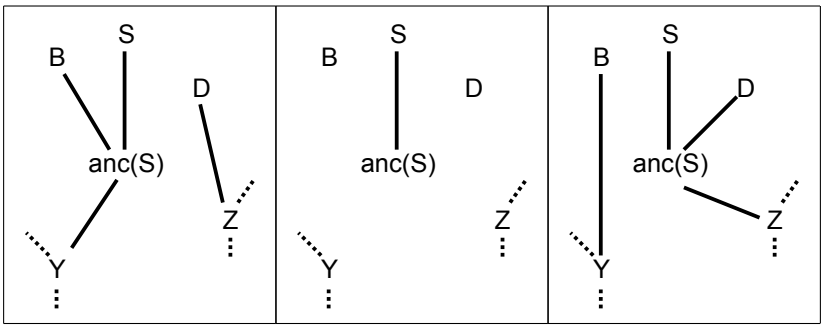

Figure 3: Fixed height subtree prune and regraft subroutine for one tree, shown in 3 stages from left to right.

3. Find the most recent common ancestor node M of S and D.

4. For each gene tree G, find all the nodes s of G such that $\text{anc}(s)$ is inside one the SMC-tree branches between the node $\text{anc}(S)$ and the node M such that $I(s) \subset I(S)$ and the sibling node x of s satisfies $I(x) \not\subset I(S)$. Denote by $\text{Src}(G, S)$ the set of such nodes s.

5. For each gene tree G, for each $s \in \text{Src}(G, S)$, calculate the set of branches $\text{Dest}(G, s)$ using the subroutine below.

6. Prune subtree S and regraft into branch D.

7. For each gene tree G, for each $s \in \text{Src}(G, S)$, choose a member d of $\text{Dest}(G, s)$ at random then carry out the prune and regraft operations for each pair.

8. For each transformed gene tree G', let $\text{Src}(G', S)$ be defined as in step 4. For each $s' \in \text{Src}(G', S)$, calculate the set of branches $\text{Dest}(G', s')$ using the subroutine below. Return

$$\sum_{G} \sum_{s \in \text{Src}(G,S)} \log(|\text{Dest}(G, s)|) \; - \sum_{G'} \sum_{s' \in \text{Src}(G',S)} \log(|\text{Dest}(G', s')|)$$

as the logarithm of the Hastings ratio.

In step 4, all the gene tree nodes which need to be pruned and regrafted are identified. These are nodes s all of whose sequences belong to S, and whose parent nodes $\text{anc}(s)$ are the first nodes going back in time at which sequences from S are joined to those not belonging to S. In step 5, the set of possible destination branches is identified for each s. In steps 6 and 7, the prune and regraft operations are carried out, first for the SMC-tree, then all gene trees. In step 8, the number of choices for destination branches is found for the reverse move. The subroutine for the calculation of $\text{Dest}(G, s)$ in steps 5 and 8 follows.

bioRxiv preprint doi: https://doi.org/10.1101/010199; this version posted March 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

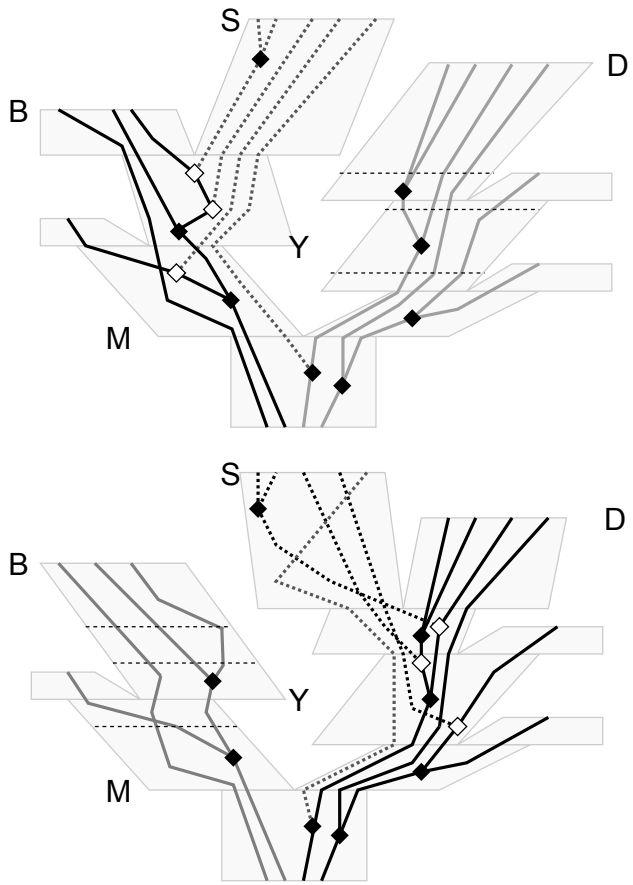

Figure 4: Example of the CoordinatedPruneRegraft move. The state before the move is at the top, and after the move below. The SMC-tree is pale gray. Gene tree branches whose sequences all belong to $I(S)$ are dotted. Before the move, gene tree branches whose sequences are descendants of the same (left) child of M as S but do not belong entirely to $I(S)$ are black. The white nodes are the origins (anc(s) in the text) of subtrees that need to be pruned and regrafted, and the thin horizontal dotted lines cut across the available destination branches. After the move, the colors and styles are reversed to illustrate the reverse move.

bioRxiv preprint doi: https://doi.org/10.1101/010199; this version posted March 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

1. Set $\mathrm{Dest}(G, s) = \emptyset$.
2. Suppose the chain of nodes leading from D back to M is $D_0 = D, D_1, \ldots, D_m = M$. Let x be a node in G. Find d such that $t(D_d) \leq t(\mathrm{anc}(S)) \leq t(D_{d+1})$
3. For all nodes x in G, add x to $\mathrm{Dest}(G, s)$ if and only if $t(x) \leq t(\mathrm{anc}(s)) \leq t(\mathrm{anc}(x))$ and $I(x) \subset I(D_d)$.

Note that the number of choices in step 7 can differ for the reverse move. In the example of Figure 4, the numbers are 3,4,4 for the forward move, and 3,3,3 for the reverse move.

5.3 Properties

Proposition 3 *The coordinated subtree and regraft move preserves compatibility*

Proof. Pruning a subtree cannot produce an incompatibility, so we just need to show that the new nodes created by regrafting do not result in incompatibilities. The proof for each gene tree is identical, so we just consider one gene tree G. Suppose that the subtree S has been pruned and regrafted, and that all the nodes in $\mathrm{Src}(G, S)$ have been pruned but not yet regrafted. (The algorithm is not carried out in this order, but it is convenient for the proof.) The definition of $\mathrm{Src}(G, S)$ ensures that no incompatibility exists in this state. This is because the remaining nodes x in G either satisfy $I(x) \cap I(S) = \emptyset$ or $t(x) \geq t(M)$.

It remains to consider the new nodes which are created when the gene subtrees are regrafted. Suppose node x is regrafted into branch y. The definition of destination branches $\mathrm{Dest}(G, s)$ ensures that x is created inside the branches between $\mathrm{anc}(S)$ and M and that $I(x)$ only contains $I(y)$ plus members of $I(S)$. This shows that the new gene nodes are compatible, and completes the proof.

Proposition 4 *The coordinated subtree and regraft move is reversible with Hastings ratio as given in step 8.*

Proof. Given the source subtree and destination branch, each individual FNPR move is symmetric, so the only asymmetry arises in the choice of the set of moves. Since the move does not change heights, the number of branches whose duration includes a particular height t is unaffected by the move. It follows that the number of choices for S and D (steps 1 and 2) are the same for the reverse move: that is, the probability of choosing D for the forward move is the same as the probability of choosing B for the reverse move. Furthermore, it follows from the definition of $\mathrm{Src}(G, s)$ in step 4 that $|\mathrm{Src}(G, S)| = |\mathrm{Src}(G', S)|$. Finally, the sizes of $|\mathrm{Dest}(G, s)|$ are accounted for in step 8.

bioRxiv preprint doi: https://doi.org/10.1101/010199; this version posted March 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

6 The focused scaler move

6.1 The algorithm

We assume that the SMC-tree has at least 4 tips, so that the first step below is possible.

1. Choose at random any node S in the SMC-tree which is not the root or a tip, and has at least one child which is not a tip.

2. For any node X in the SMC-tree, let $\mathrm{dist}(X)$ be the number of branches from S to X.

3. For each gene tree G, find all the nodes s of G such that (S, s) are hitched.

4. For each node s hitched to S, define $\mathrm{dist}(s)$ to be 1 if $I(s) \subset I(S)$ and 2 otherwise. For other nodes x in gene trees, $\mathrm{dist}(x) = \infty$ initially. Then $\mathrm{dist}(x)$ is defined recursively using the following rule. If y is adjacent to x (that is, if $y = \mathrm{anc}(x)$ or y is a child of x), then $\mathrm{dist}(x) = \min(\mathrm{dist}(x), 1 + \mathrm{dist}(y))$.

5. For each tree (SMC-tree or gene tree) T let $f_T : \mathbb{N} \to [0, 1]$ be a function from the nonnegative integers such that $f_T(0) = 1$ and $f_T(d) < 1$ for $d > 0$. Let $w(X) = f_T(\mathrm{dist}(X))$ for all nodes X of T.

6. Let Λ be the set of pairs of nodes defined as follows. Firstly, Λ contains all hitched pairs of nodes (Y, y) for any SMC-tree node Y and any node y in any gene tree. Secondly Λ contains all pairs $(X, \mathrm{anc}(X))$ where X is in any tree. Thus Λ contains all hitched pairs and all branches.

7. Let $\Lambda^+ = \{(A, B) \in \Lambda : w(A) > w(B)\}$ and $\Lambda^- = \{(A, B) \in \Lambda : w(A) < w(B)\}$. Set
$$D = \max\left\{-\frac{\log(t(B)/t(A))}{w(B) - w(A)} : (A, B) \in \Lambda^-\right\}$$
and
$$U = \min\left\{\frac{\log(t(B)/t(A))}{w(A) - w(B)} : (A, B) \in \Lambda^+\right\}.$$

8. Choose η uniformly from the interval $[D, U]$ and scale the height of every internal node X which has a nonzero weight by $\exp(w(X)\eta)$. Return the sum of all the $w(X)\eta$ values as the logarithm of the Hastings ratio.

The conditions on S in step 1 ensure that there are two nodes adjacent to S, one with a bigger height (its parent) and one with a smaller but nonzero height (one of its children). Both these have distance 1 from S (step 2) so get a smaller weight than S due to the conditions $f_T(0) = 1$, $f(1) < 1$ in step 5. Thus the maximum and minimum in step 7 are taken over a non-empty sets, so D and U and hence η are all finite.

bioRxiv preprint doi: https://doi.org/10.1101/010199; this version posted March 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

The nodes given a distance of 1 in step 4 are 'topologically closer' to S than those given distance 2. Usually, they are closer in height as well. There is freedom to choose a wide variety of functions for f_T in step 5. In the current implementation, a decreasing function is used, which becomes zero at the root of each tree. The weights $w(X)$ are thus zero whenever $\text{dist}(X) \geq \text{dist}(R)$ where R is the root of the tree T containing X.

6.2 Properties

Proposition 5 *The focused scaler move keeps all branch lengths nonnegative and all gene trees compatible with the SMC-tree.*

Proof. Consider two nodes X and Y with $(X, Y) \in \Lambda$. Note that $t(X) \leq t(Y)$. Let $g(X) = \log(t(X))$ and $g(Y) = \log(t(Y))$. After the move the heights are $t(X)\exp(w(X)\eta)$ and $t(Y)\exp(w(Y)\eta)$ (step 8) so the logarithms of the heights after the move are $g'(X) = g(X) + w(X)\eta$ and $g'(Y) = g(Y) + w(Y)\eta$. If $w(X) = w(Y)$, we obviously have $g'(Y) \geq g'(X)$. Suppose that $w(X) > w(Y)$ so that $(X, Y) \in \Lambda^+$. We have from step 7 that

$$\eta \leq U \leq \frac{\log(t(Y)/t(X))}{w(X) - w(Y)} = \frac{g(Y) - g(X)}{w(X) - w(Y)}$$

so

$$\eta(w(X) - w(Y)) \leq g(Y) - g(X)$$

so the difference between the logarithms of heights after the move is

$$g'(Y) - g'(X) = g(Y) - g(X) - \eta(w(X) - w(Y)) \geq 0$$

so it remains nonnegative. The case $w(X) < w(Y)$ is similar.

Since Λ includes all branches in all trees (step 6), the move keeps all branch lengths nonnegative. It remains to show that compatibility is preserved. For pairs $(X, x) \in \Lambda$ where X is a SMC-tree node and x a gene tree node, compatibility follows in the same way as for branch lengths. Suppose (X, x) is not in Λ. Then either $I(x)$ does not straddle X, or at least one of $R(x)$ or $L(x)$ does straddle X. In the first case, X and x are compatible since there is no conflict between the minimal clusters belonging to them. In the second case, some descendant y of x must be in Λ, and the compatibility of (X, y) implies that of (X, x).

Proposition 6 *The focused scaler move is reversible with Hastings ratio as stated in step 8.*

Proof. The choice of S in step 1 is based on topological criteria only, and the move does not change the topology so the probabilities of choosing S for the move and reverse move are identical. The definition of dist() only depends on topologies and assignments at tips, so given the choice of S, the values $\text{dist}(X)$

bioRxiv preprint doi: https://doi.org/10.1101/010199; this version posted March 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

and hence $w(X)$ for all nodes X are the same for the reverse move. Denoting quantities for the reverse move with primes ($'$) we have

$$\frac{\log(t'(A)) - \log(t'(B))}{w'(B) - w'(A)}$$

$$= \frac{\log(t(A)) + w(A)\eta - \log(t(B)) - w(B)\eta}{w(B) - w(A)}$$

$$= \frac{\log(t(A)) - \log(t(B))}{w(B) - w(A)} - \eta$$

whenever $w(B) \neq w(A)$, from which it follows that

$$D' = D - \eta \text{ and } U' = U - \eta.$$

Now $D \leq 0$ and $U \geq 0$, so $0 \in [D, U]$ and so $-\eta \in [D', U']$, so a choice of $-\eta$ is available for the reverse move, which will restore the original state. The two intervals $[D, U]$ and $[D', U']$ have the same size, and the choice of η is made uniformly, so there is no contribution to Hastings ratio here. This leaves only the scaling of the node heights for which step 8 provides the Hastings ratio.

7 Tests of correctness

In order to check the theory and the implementation of these moves, some tests were carried out by sampling from prior distributions. Full details of these tests and the results are in the supplementary information. The following is a brief summary. There were two sets of tests. One has an unknown number of species (between 1 and 8). The other uses a fixed number (8) of species and samples from the prior on the species tree. Although there is no sequence data, the assumptions about the number of species constitute some 'meta-data'. In both sets of tests, there was one gene tree with no data, that is with a sequence "?" at each tip. Since the operators change the gene trees simultaneously with the SMC-tree, it is important to include at least one gene tree.

BEAST2 XML files were generated for the two sets of tests, with various combinations of operators. These were then run in BEAST2. The sampled SMC-trees were examined for agreement with theoretical distributions for the node heights; the species tree topology in the case of fixed species assignments; and for clusterings in the case of delimitation.

Some results in which the CoordinatedPruneRegraft move is used together with the usual BEAST operators are shown. These are included to illustrate the type of tests used, but for full details see the supplementary information.

Figure 5a is for the estimated delimitation case. It shows estimated and theoretical values for the 22 partitions of the number 8. Each partition of 8 represents one or more clusterings of 8 objects. There are a total of 4140 clusterings of 8 objects, and these can be grouped into 22 sets corresponding to the partitions of 8. For example suppose the 8 objects are a, b, c, d, e, f, g, h.

bioRxiv preprint doi: https://doi.org/10.1101/010199; this version posted March 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

One clustering is $\{\{a,b,c\},\{d\},\{e,f,g,h\}\}$, which corresponds to the partition 4+3+1 of 8. There are 280 clusterings with the shape 4+3+1, and whenever one of these are visited during the MCMC, it counts towards the posterior probability of this partition of 8.

The other three plots in Figure 5 are for the fixed species delimitation case. Figure 5b shows how this combination of operators sample from the 23 possible topologies with 8 tips. The x-axis annotations show the topologies in the following format. A tip is shown by *****. A cherry **(*,*)** is denoted as **C**, a 3-tip tree **(*,(*,*))** as **T**, an unbalanced 4-tip tree as **U** and a balanced 4-tip tree as **B**. Otherwise, the Newick format is used. For example **(T,(*,U))** is short for **((*,(*,*)),(*,(*,(*,(*,*)))))**. Finally the sampled node heights are compared to theoretical densities in (5c,d).

The scenarios are sufficiently simple that various marginal aspects of the prior distribution can be calculated analytically, but sufficiently complicated to provide a meaningful test of the operators. No problems were found (after two bugs were found and fixed during preliminary testing). However, it is not possible to give a 100% guarantee of correctness. There may be problems which are not revealed by the marginal aspects of the distribution that were analyzed here. There may be problems which only produce an undetectable bias in these tests but which become more serious in other scenarios.

8 Results on the simulated data of Olave et al.

As a proof-of-concept demonstration, I re-analyzed the simulated data provided in the supplementary material of [10]. This data set contains 50 replicates for each of twelve configurations. In all cases there are 40 individuals, 2 sequences of length 1000bpp per individual, and 8 true species each consisting of 5 individuals. There are two tree shapes, symmetric and asymmetric, two amounts of incomplete lineage sorting, and three values 4, 8, or 14 for the number of loci.

The data was incorporated into XML files for BEAST2. Version 2.2.0 of BEAST2 and 1.0.1 of STACEY were used. For each replicate, the program was run for 5, 7, or 10 million generations for the 4, 8, and 14 loci cases respectively. The first 1 million discarded as burnin. Samples were taken every 1000 generations, so there were 4000, 6000, or 9000 SMC-trees on which to base the species delimitations using SpeciesDelimitationAnalyser [9].

8.1 Priors and other settings

For the population variability among branches, a single inverse gamma component with mean and standard deviation 1 was used. (In equation (3), $C = 1, \alpha_1 = 3.0, \beta_1 = 2.0$.) A lognormal(-7.0,2.0) was used for the hyperprior π_σ for the overall population scaling factor. (Parameters to the lognormal are given in log space.) The value of p_j was set to 2 for all genes. The HKY model was assumed for the substitution model. It was assumed that there was no site rate heterogeneity (although the data set does contain such heterogeneity). The

bioRxiv preprint doi: https://doi.org/10.1101/010199; this version posted March 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

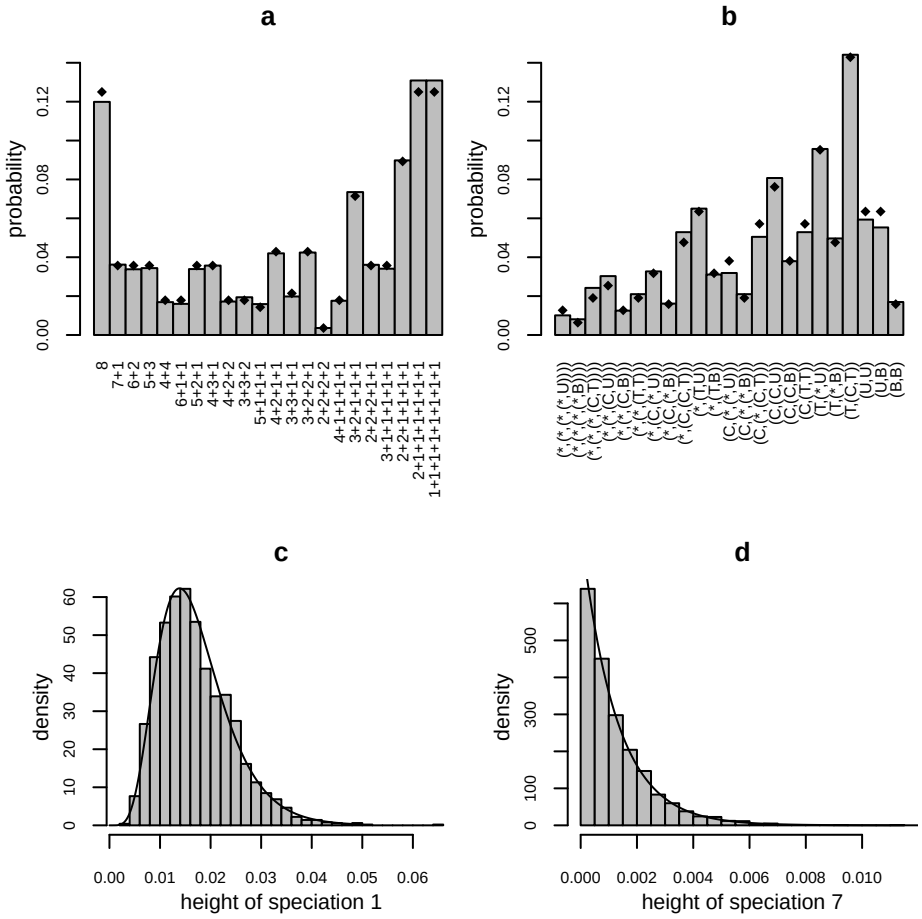

Figure 5: Plot (a) shows some results for the case of sampling from the prior distribution with an unknown number of species between 1 and 8. Estimated posterior probabilities are shown in the gray bars for the 22 partitions of the integer 8 (see text for further explanation). The theoretical values are shown as diamonds. Plots (b),(c), and (d) are some results for the case of sampling from the prior distribution with the number of species fixed at 8. In (b), the gray bars show estimated posterior probabilities for each of the 23 unlabeled rooted topologies. The theoretical values are shown as diamonds. The x-axis annotations enumerate the topologies (see text for details). In plots (c) and (d), the gray bars show a histogram of samples from the posterior for the first and 7th speciation height, that is, the root height and most recent speciation. The curves show the theoretical densities for these speciation heights.

bioRxiv preprint doi: https://doi.org/10.1101/010199; this version posted March 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

relative clock rates of the genes other than the first were estimated; a lognormal(0.0,1.0) prior was assumed for these. A birth-death model was assumed for the species tree, with a lognormal(4.6,2) hyperprior for the growth rate, and a Beta(3,1) hyperprior for the relative death rate. The prior on the collapse weight was uniform on $[0, 1]$ so that there was a flat prior on the number of species, and the collapse height ϵ was set to 0.0001. The 40 individuals were used as minimal clusters (containing two sequences each) in STACEY. (See [9] for definitions of 'minimal cluster', 'collapse weight' and 'collapse height'.)

8.2 Results

The results are shown in Figure 6. The clustering with the largest posterior probability (that is, a MAP estimator) was used to estimate the species delimitation. All errors in this estimate were false splits. Usually just one of the true species was split; in five replicates, two true species were split; and in replicate 47 from YH4 and replicate 34 from ZH4, three true species were split. In all 600 replicates, the true clustering was in the 0.95 credible set. The highest posterior probability assigned to a erroneous clustering was 0.83 (replicate 16 from YE4).

The estimated sample sizes (ESSs) for the posterior, as reported by Coda [11], had means of 250, 215, and 215 for the the 4, 8, and 14 loci cases. Some individual replicates had ESSs below 100, with a minimum of 72 over all 600 replicates.

9 Discussion

Based on tests so far, including some results not reported here, STACEY converges much faster than DISSECT. The difference was particularly apparent in the time to 'burn-in' in cases where there was little ILS and many loci. Although these are the cases where signal is strongest, DISSECT can take a very long time to converge, as reported in [9] for the case of 27 loci. It is not yet clear how much of this improvement is due to the new model and how much to the new moves. It seems likely that the new moves will improve convergence in *BEAST in some cases at least, but this has not been tried.

When analyzing the data [10], the number of generations in the MCMC chains were chosen so that all 600 replicates could be run in a reasonable amount of time with limited computational resources (2 weeks on a desktop computer with 4 cores). This resulted in lower ESS values than desirable on some replicates. Given the main purpose of this analysis, this does not seem important: if anything longer runs would be expected to improve accuracy. When used 'for real', several longer runs are strongly recommended.

In the context of phylogeny estimation, the relative importance of the two kinds of noise, namely mutational variance and incomplete lineage sorting, was studied in [7]. In their scenarios, up to 75% of the errors in maximum likelihood estimates of species trees were attributable to mutational variance. It seems very likely that similar conclusions apply to Bayesian species delimitation. The

bioRxiv preprint doi: https://doi.org/10.1101/010199; this version posted March 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

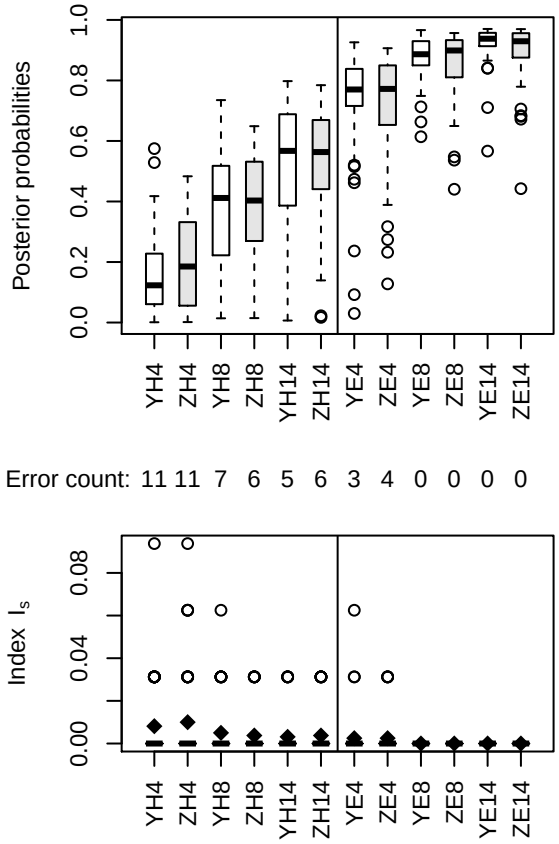

Figure 6: The upper boxplots show the posterior probabilities of the true clustering over 50 replicates for twelve configurations YH4,... ZE14. In the labels for the configurations, the first letter Y or Z denotes the tree shape, with Y for symmetric and Z for asymmetric; the second letter denotes the degree of incomplete lineage sorting, with H for N=0.4 ('hard') and E for N=4 ('easy'); this is followed by the number of loci: 4, 8, or 14. The numbers between the boxplots are the number of times out of 50 that the clustering with the largest posterior probability was not the true clustering. The lower boxplots show the measure of over-splitting of true species lineages using the index I_s of Olave et al (2014). The black diamonds show the mean values. Note that the vertical scale is much smaller than that of Figure 3 in Olave et al.

bioRxiv preprint doi: https://doi.org/10.1101/010199; this version posted March 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

simulated data sets of [10] have low mutational variance. The species tree branch lengths, measured in substitutions, range from 0.004 to 0.028 in the N=0.4 case and from 0.04 to 0.28 in the N=4 case. Since there are two sequences of length 1000bp per individual, the expected number of substitutions per individual per locus along a branch is always at least $0.004 * 2000 = 8$. However, in many empirical data sets the difficulties due to incomplete lineage sorting will be compounded with large amounts of mutational variance. The simulations used in [9] were much harder in terms of the mutational variance: the sequences were 500bp, there was only one sequence per individual, and the shortest branch lengths were 0.001, so that the expected number of substitutions along the shortest branches is only 0.5 instead of 8. The results of that paper may be a better guide to to the accuracy of the approach on many empirical data sets.

The results here should dispel some of the pessimism expressed in [10] about DNA-based species delimitation. It is usually the case that geographical and morphological information is available as well [18], but it is rare that this provides certainty about the assignment of individuals to clusters or populations. I think that a more promising way ahead is to include the extra information in a Bayesian analysis. The location data and morphological characters could be included alongside the genetic data. Alternatively, taxonomists could formalize their knowledge in the form of a prior on the space of all possible clusterings. A program like STACEY can then explore the full space, taking into account the extra information. The space of all clusterings is huge, and it is not easy to construct sensible probability distributions for it which reflect expert knowledge about the organisms. Research is needed to find good ways of doing this.

Acknowledgments

I thank the developers of BEAST for making this work feasible, and Remco Bouckaert in particular for helpful advice on writing the STACEY package. I thank the authors of [10] for making their simulated data readily available, and Melisa Olave for supplying extra details about the simulations.

References

[1] Bouckaert, R., Heled, J., Khnert, D., Vaughan, T., Wu, C.H., Xie, D., Suchard, M.A., Rambaut, A., Drummond, A.J.: BEAST 2: A software platform for bayesian evolutionary analysis. PLoS Comput Biol **10**(4), e1003,537 (2014). DOI 10.1371/journal.pcbi.1003537

[2] Degnan, J.H., Rosenberg, N.A.: Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. **24**, 332–340 (2009)

[3] Edwards, S.V.: Is a new and general theory of molecular systematics emerging? Evolution **63**, 1–19 (2009)

bioRxiv preprint doi: https://doi.org/10.1101/010199; this version posted March 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

[4] Felsenstein, J.: Inferring Phylogenies. Sinauer Associates (2003). DOI http://dx.doi.org/10.1016/S0022-0000(02)00003-X

[5] Heled, J., Drummond, A.: Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. **27**, 570–580 (2010)

[6] Höhna, S., Defoin-Platel, M., Drummond, A.J.: Clock-constrained tree proposal operators in bayesian phylogenetic inference. 8th IEEE International Conference on BioInformatics and BioEngineering, 2008 pp. 1–7 (2008)

[7] Huang, H., He, Q., Kubatko, L.S., Knowles, L.L.: Sources of error for species-tree estimation: impact of mutational and coalescent effects on accuracy and implications for choosing among different methods. Syst. Biol. **59**, 573–583 (2010)

[8] Huelsenbeck, J.P., Andolfatto, P.: Inference of population structure under a Dirichlet process model. Genetics **175**, 1787–1802 (2007)

[9] Jones, G., Aydin, Z., Oxelman, B.: DISSECT: an assignment-free Bayesian discovery method for species delimitation under the multispecies coalescent. Bioinformatics (2014). DOI 10.1093/bioinformatics/btu770

[10] Olave, M., Solà, E., Knowles, L.L.: Upstream analyses create problems with DNA-based species delimitation. Syst Biol **63**, 263–271 (2014). DOI 10.1093/sysbio/syt106

[11] Plummer, M., Best, N., Cowles, K., Vines, K.: Coda: Convergence diagnosis and output analysis for mcmc. R News **6**(1), 7–11 (2006). URL http://CRAN.R-project.org/doc/Rnews/

[12] Pritchard, J.K., Stephens, M., Donnelly, P.J.: Inference of population structure using multilocus genotype data. Genetics **155**, 945–959 (2000)

[13] Rannala, B., Yang, Z.: Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics **164**, 1645–1656 (2003)

[14] Rannala, B., Yang, Z.: Improved reversible jump algorithms for Bayesian species delimitation. Genetics **194**, 245–253 (2013)

[15] Yang, Z.: Likelihood and bayes estimation of ancestral population sizes in hominoids using data from multiple loci. Genetics **162**, 1811–1823 (2002)

[16] Yang, Z., Rannala, B.: Bayesian species delimitation using multilocus sequence data. Proceedings of the National Academy of Sciences of U.S.A **107**, 9264–9269 (2010)

[17] Yang, Z., Rannala, B.: Unguided species delimitation using dna sequence data from multiple loci. Mol. Biol. Evol. **31**(12), 3125–3135 (2014). DOI 10.1093/molbev/msu279

bioRxiv preprint doi: https://doi.org/10.1101/010199; this version posted March 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

[18] Zhang, C., Rannala, B., Yang, Z.: Bayesian species delimitation can be robust to guide-tree inference errors. Syst. Biol. **0**, 1–12 (2014). DOI 10.1093/sysbio/syu052