J. Math. Biol.
DOI 10.1007/s00285-016-1034-0

Mathematical Biology

Algorithmic improvements to species delimitation and phylogeny estimation under the multispecies coalescent

Graham Jones[1]

Received: 26 March 2015 / Revised: 31 May 2016

Abstract The focus of this article is a Bayesian method for inferring both species delimitations and species trees under the multispecies coalescent model using molecular sequences from multiple loci. The species delimitation requires no a priori assignment of individuals to species, and no guide tree. The method is implemented in a package called STACEY for BEAST2, and is a extension of the author's DISSECT package. Here we demonstrate considerable efficiency improvements by using three new operators for sampling from the posterior using the Markov chain Monte Carlo algorithm, and by using a model for the population size parameters along the branches of the species tree which allows these parameters to be integrated out. The correctness of the moves is demonstrated by tests of the implementation. The practice of using a pipeline approach to species delimitation under the multispecies coalescent, has been shown to have major problems on simulated data (Olave et al. in Syst Biol 63:263–271. doi:10.1093/sysbio/syt106, 2014). The same simulated data set is used to demonstrate the accuracy and improved convergence of the present method. We also compare performance with *BEAST for a fixed delimitation analysis on a large data set, and again show improved convergence.

Keywords Species delimitation · Multispecies coalescent · Bayesian analysis · Markov chain Monte Carlo

Electronic supplementary material The online version of this article (doi:10.1007/s00285-016-1034-0) contains supplementary material, which is available to authorized users.

✉ Graham Jones
art@gjones.name

[1] Department of Biological and Environmental Sciences, University of Gothenburg, Box 461, 405 30 Göteborg, Sweden

Published online: 10 June 2016

Mathematics Subject Classification 92B10 · 62P10

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 the area has seen much development recently (Flot 2015; Rannala 2015). 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 the 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 time of divergence of the two groups. This idea is made precise by the multispecies coalescent model (Yang 2002; Rannala and Yang 2003; Degnan and Rosenberg 2009; Edwards 2009).

There are two main types of 'noise' which interfere with inference of delimitation and phylogeny: mutational variation and coalescent variation. Mutational variation is due to the stochastic nature of mutations: for example there may be one substitution in a very short branch, but no substitutions in a somewhat longer branch. For species delimitation in particular, organisms are often closely related with few mutations separating them, so this can be a major problem, and means there is a large amount of uncertainty about the node heights and topologies of the gene trees. Coalescent variation is due to the stochastic nature of mating. This can result in incomplete lineage sorting, where two gene copies fail to coalesce (going back in time) until the group to which they both belong has merged with another group. The expected amount of incomplete lineage sorting depends on the effective population size along a branch, divided by the branch length measured in generations.

1.1 Previous work

Over the past 10 years, the multispecies coalescent model has become a 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 (Yang and Rannala 2010; Rannala and Yang 2013; Yang and Rannala 2014) and DISSECT (Jones et al. 2014). More details and discussion of alternative approaches can be found in Jones et al. (2014), Olave et al. (2014), and Zhang et al. (2014).

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 Olave et al. (2014) sequences were simulated under the multispecies coalescent model, and then analyzed using a standard 'pipeline' method using Structurama (Pritchard et al. 2000; Huelsenbeck and Andolfatto 2007), *BEAST (Heled and Drum-

mond 2010), and BPP (Rannala and Yang 2013). The analysis showed there were major problems with the method. The accuracy was in general low, and in some cases, the support for the wrong number of species became stronger with more loci.

1.2 Overview of the present method

The method presented here replaces the pipeline used in Olave et al. (2014) with a single analysis. It is implemented as a package called STACEY (Species Tree And Classification Estimation, Yarely) for BEAST2 (Bouckaert et al. 2014). STACEY is aimed mainly at species delimitation, but can also be used as an alternative to *BEAST (Heled and Drummond 2010).

The probability of the sequence data for a single alignment can be found given the gene tree topology and node times, plus other parameters such as those for the substitution model. This is the usual 'Felsenstein likelihood' for the alignment. The product of these likelihoods over all the gene trees is the likelihood function for the analysis. A prior probability density is needed for all the parameters, and the density for the gene tree topologies and node times is provided by the multispecies coalescent model. This in turn requires a prior for the species tree, and a model for the population sizes along the branches of the species tree.

In the context of species delimitation using STACEY, the species tree has tips which represent **minimal clusters** of individuals (Jones et al. 2014). 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 we will refer to it a the **SMC-tree**, as a shorthand for 'species or minimal clusters tree'.

In *BEAST and DISSECT, one or more population size parameters for each branch in the SMC-tree are introduced, and these are sampled using the Markov chain Monte Carlo (MCMC) algorithm. The difference between *BEAST and DISSECT is the prior for the species tree. In DISSECT, the birth-death model for the species tree is replaced by a 'birth-death-collapse' model. This effectively extends the model to allow all possible species delimitations, as well as all species trees for each delimitation. A parameter (the 'collapse weight') in the birth-death-collapse model can be set to zero which removes the spike in the prior, and makes DISSECT almost identical to *BEAST. This setting is used when the species delimitation is fixed in a DISSECT or STACEY analysis.

The method presented here incorporates a different model for the population sizes along the branches of the SMC-tree to that used in *BEAST. It is assumed that each branch has a population size 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 allows for variation among branches, and is similar to the 'piecewise constant' option in *BEAST but does not allow individual population sizes 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 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 multispecies 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 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 FocusedScaler, scales node heights whilst preserving topologies. The scaling is 'focused' on a node in the SMC-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 third move, called CoordinatedPruneRegraft, is a subtree-prune-and-regraft move which makes coordinated topological changes to the SMC-tree and gene trees. The CoordinatedPruneRegraft move can be seen as an extension of the nearest neighbor interchange (NNI) move described in Yang and Rannala (2014), 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 Höhna et al. (2008). 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 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 sets of Olave et al. (2014) and Giarla and Esselstyn (2015) are re-analyzed.

2 Conventions and notation

All trees are rooted, binary, and ultrametric. 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 assignment of labels to tips. The tips of the SMC-tree are labeled with the names of minimal clusters, and the tips of the gene trees are labeled with sequence names.

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. 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 $t \in [t(x), t(\text{anc}(X))]$. If X is a node in the SMC-tree and x is a gene tree node, then the pair (X, x) is **compatible** if $(x, t(x))$ is inside a branch of the SMC-tree. A gene tree is compatible with the SMC-tree if every pair of nodes (X, x) is compatible.

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. 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 Fig. 1 for an example. The node x is not hitched to S because it does not straddle S. The node y is not hitched to S because one of its children does straddle S. 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.

3 The population model

The model for the population size parameters is similar to that of Liu et al. (2008), but has a few extensions. For each locus, a 'ploidy' factor is explicitly included, so

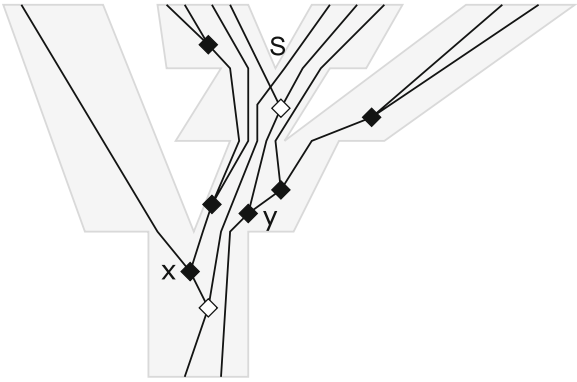

Fig. 1 Example of hitched nodes. The SMC-tree is *pale gray*. A gene tree is shown inside it. Gene tree nodes which are hitched to the SMC-tree node S are shown as *white diamonds*, and other nodes as *black diamonds*

that different types of sequences (e.g., autosomal nuclear genes, genes from sex chromosomes or organelles) can be simultaneously analyzed. Furthermore, a parameter representing an overall scaling factor is introduced, and the single inverse gamma prior is replaced by an arbitrary mixture of inverse gammas.

Consider the coalescent process for a single locus within a single branch on the species tree. The coalescent model of Kingman (see Chapters 26–28 of Felsenstein 2003) is assumed. The probability density for the coalescent times takes the following form (simplified from Eq. 3, p 572, of Heled and Drummond 2010):

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

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 $(t_0 < t_1, \ldots t_k < t_{k+1})$ where t_0 is the node height at the tipward end, t_{k+1} is the node height at the rootward end, and $(t_1, \ldots t_k)$ 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 size 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 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. For gene j in branch b, we thus need to replace P by $p_j N_b \mu_b$ in Eq. (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} $(0 \le i \le k_{jb})$. 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

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

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

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

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 (Hey and Nielsen 2007; Liu et al. 2008). 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

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

Here C, the λ_c, the α_c, and the β_c $(1 \le c \le C)$ 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

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

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

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

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

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

$$= \lambda_c \frac{(\sigma\beta_c)^{\alpha_c}}{\Gamma(\alpha_c)}\theta_b^{-\alpha_c-1}\exp\left(-\sigma\beta_c\theta_b^{-1}\right) r_b\theta_b^{-q_b}\exp\left(-\gamma_b\theta_b^{-1}\right)$$
$$= \frac{\lambda_c r_b(\sigma\beta_c)^{\alpha_c}}{\Gamma(\alpha_c)}\theta_b^{-\alpha_c-1-q_b}\exp\left(-(\sigma\beta_c+\gamma_b)\theta_b^{-1}\right)$$
$$= \frac{\lambda_c r_b(\sigma\beta_c)^{\alpha_c}}{(\sigma\beta_c+\gamma_b)^{\alpha_c+q_b}}\frac{\Gamma(\alpha_c+q_b)}{\Gamma(\alpha_c)}$$
$$\times \frac{(\sigma\beta_c+\gamma_b)^{(\alpha_c+q_b)}}{\Gamma(\alpha_c+q_b)}\theta_b^{-(\alpha_c+q_b)-1}\exp\left((\sigma\beta_c+\gamma_b)\theta_b^{-1}\right)$$
$$= \frac{\lambda_c r_b(\sigma\beta_c)^{\alpha_c}}{(\sigma\beta_c+\gamma_b)^{\alpha_c+q_b}}\frac{\Gamma(\alpha_c+q_b)}{\Gamma(\alpha_c)}\mathcal{IG}(\theta_b; \alpha_c+q_b, \sigma\beta_c+\gamma_b).$$

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

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

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

4 The NodesNudge move

We begin by making two general comments which apply to the NodesNudge move and the FocusedScaler move, as well as others that could be implemented. Assume the MCMC move consists of applying continuous functions to the heights of nodes. Denote by $m_X(t;\eta)$ the function applied to node X, where t is the height and η is a parameter shared by all the m_X, such that $m_X(t;0) = t$ and $m_X(t;\eta)$ is monotonically

increasing as a function of t for all X and η. By choosing η small enough, the move will be 'valid', that is, it will keep all branch lengths non-negative, and all the gene trees compatible with the SMC-tree. It is not necessary for all the nodes to change height: for some nodes X, we can have $m_X(t, \eta) = t$ for all η. Both the NodesNudge and FocusedScaler moves are of this general pattern. We also ensure that for large enough negative and large enough positive values of η, the move becomes invalid. This means that there a finite range $[\eta_{min}, \eta_{max}]$, such that the move is valid if and only if $\eta \in [\eta_{min}, \eta_{max}]$. This condition makes it easier to ensure that the move is reversible.

NodesNudge and FocusedScaler share another characteristic. The functions m_X are chosen for nodes using 'topological information' only. This means the topologies of the individual trees, as well as whether a pair of nodes is hitched. Since the moves do not change these criteria, the m_X are not affected by the moves, so the same ones are used for the reverse 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. A child of $C(x)$ is a gene tree node c which is not in $C(x)$, but whose parent $\text{anc}(c)$ is in $C(x)$. The root of $C(x)$ is the oldest node in $C(x)$.

1. Choose uniformly and at random any node S in the SMC-tree which is not a tip and not the root.
2. Let $\Delta(S)$ be all the internal gene tree nodes s such that the pair (S, s) is hitched.
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 $u_0 = t(\text{anc}(S))$.

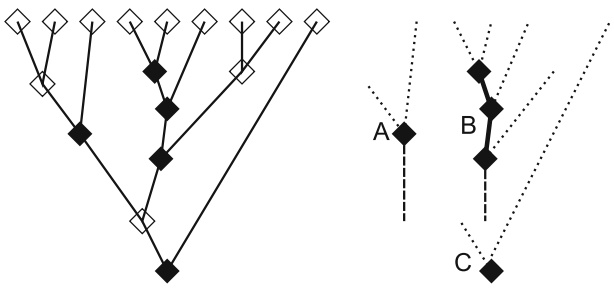

Fig. 2 Example of connected components

4. For $1 \leq i \leq n$, let
$$d_i = t(s_i) + \max_c \{t(c) - t(\text{anc}(c)) : c \text{ is a child of } C(s_i)\}.$$
Let r be the root of $C(s_i)$. Let $u_i = \infty$ if r is the root of the gene tree, otherwise let $u_i = t(s_i) + t(\text{anc}(r)) - t(r)$.
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 $[D, U]$, and let $\eta = t'(S) - t(S)$.
7. Change the height of all gene tree nodes in $\Delta(S)$ by η.

For this move, we have $m_X(t, \eta) = t + \eta$ for all the nodes which change height. These consist of S and the gene tree nodes which are hitched to S. Note that the value d_0 (step 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). Also, note that since S is not the root, u_0 is finite, so that $[D, U]$ is a finite interval, and step 6 makes sense.

4.2 Properties

Returning to Fig. 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.

The choice of S in step 1 has the same probability for the reverse move. From the general comments above regarding m_X, and the key property of connected components it follows $\Delta(S)$ is unaffected by the move. furthermore, the interval $[D, U]$ is unaffected by the move. It follows that the NodesNudge move is symmetric, so no Hastings ratio is required.

In the NodesNudge move as currently implemented, the 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. However it is likely that future versions of STACEY will exploit the more general case.

5 The focused scaler move

This is a 'larger' move than NodesNudge, in that many SMC-tree nodes, as well as gene tree nodes, can change height. The method based on connected components does not work here, so we examine all branches and hitched pairs explicitly to find the allowed range of movement.

5.1 The algorithm

We assume that the SMC-tree has at least 5 tips, which ensures that the first step below is always 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.

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.

5.2 Properties

Suppose a pair of nodes (X, Y) is in Λ. 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.

Given this, it follows that the move keeps all branch lengths nonnegative, and hitched pairs (X, x) in Λ stay compatible. If (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), and hence, that the move as a whole preserves compatibility.

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.

6 The coordinated subtree and regraft move

6.1 The subtree prune and regraft move for one tree

First we describe the fixed height subtree prune and regraft algorithm (Höhna et al. 2008) 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

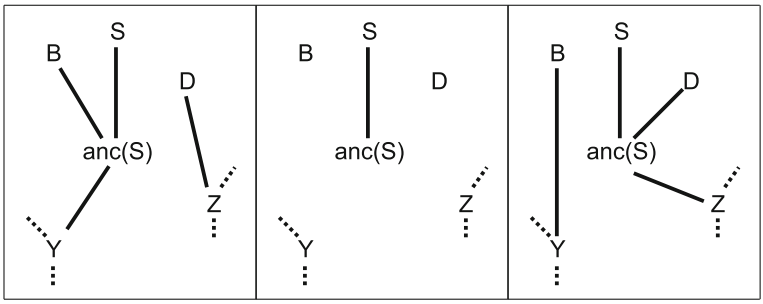

Fig. 3 Fixed height subtree prune and regraft subroutine for one tree, shown in 3 stages from *left* to *right*

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.

6.2 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 $t(D) \leq t(\text{anc}(S)) \leq t(\text{anc}(D))$.
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 $I(s) \subseteq I(S)$ and the sibling node x of s satisfies $I(x) \nsubseteq I(S)$. 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 algorithm 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.

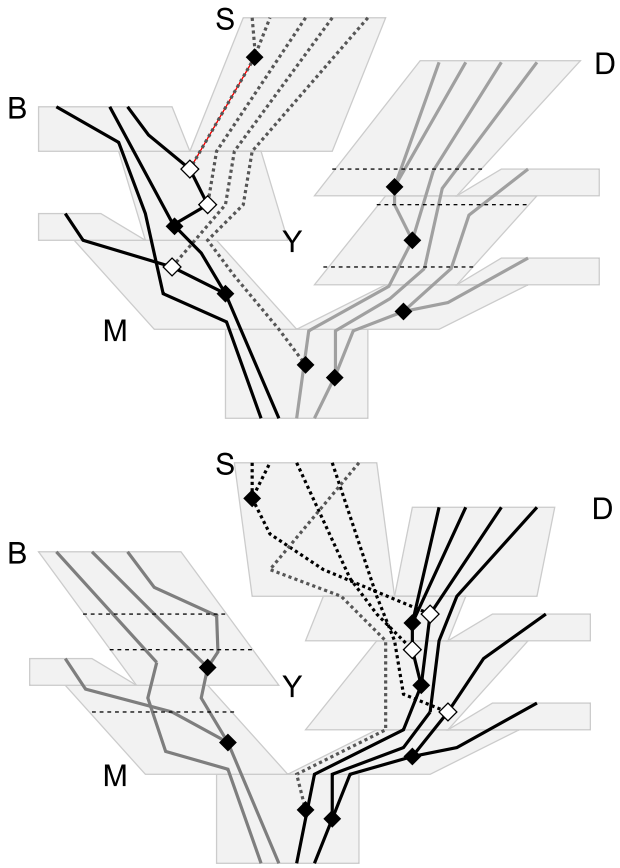

Fig. 4 Example of the CoordinatedPruneRegraft move. The state before the move is at the *top*, and after the move *below*. The SMC-tree is *pale gray*. Gene tree branches whose sequences all belong to $I(S)$ are *dotted*. Before the move, gene tree branches whose sequences are descendants of the same (*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

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

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

as the logarithm of the Hastings ratio.

The algorithm for calculating of $\mathrm{Dest}(G, s)$ in steps 5 and 8 follows.

1. Set $\mathrm{Dest}(G, s) = \emptyset$.
2. Suppose the chain of nodes leading from D back to M is $D_0 = D, D_1, \ldots, D_m = M$. 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 Fig. 4, the numbers are 3, 4, 4 for the forward move, and 3, 3, 3 for the reverse move.

6.3 Properties

In step 4, all the gene tree nodes which need to be pruned and regrafted are collected in $\mathrm{Src}(G, S)$. Note that $\mathrm{Src}(G, S)$ may be empty. We now explain how the choice of $\mathrm{Src}(G, S)$ and $\mathrm{Dest}(G, s)$ keep the gene trees compatible with the SMC-tree. 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 we can consider it.) 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 $t(x) \geq t(M)$. Now 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 tree nodes are compatible.

Finally, we explain the Hastings ratio. 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.

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

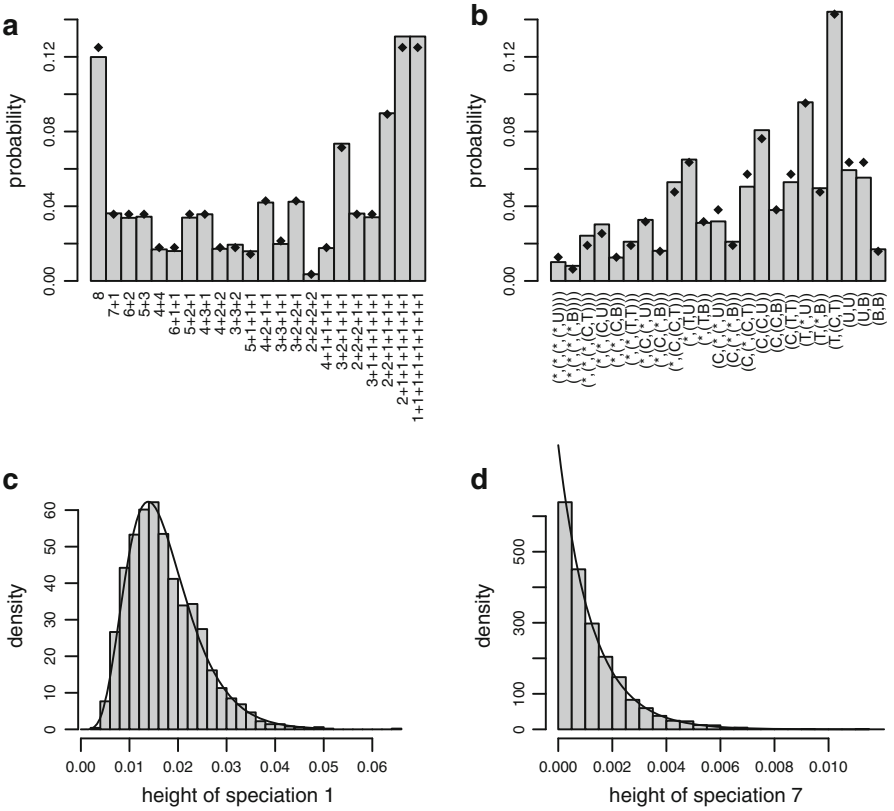

Fig. 5 Plot (**a**) shows some results for the case of sampling from the prior distribution with an unknown number of species between 1 and 8. Estimated posterior probabilities are shown in the gray bars for the 22 partitions of the integer 8 (see text for further explanation). The theoretical values 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

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. 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 Fig. 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 Fig. 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 in the tests. 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 Tests on simulated data

As a proof-of-concept demonstration, we re-analyzed the simulated data provided with (Olave et al. 2014). This re-analysis will be referred to as 'Test-I'. It contains 50 replicates for each of twelve configurations. In all cases there are 40 individuals, 2 sequences of length 1000 bp per individual, and 8 true species each consisting of 5 individuals. There are two tree shapes, symmetric and asymmetric, two amounts of coalescent variation, 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 (Jones et al. 2014).

We also re-analyzed the simulated data of Giarla and Esselstyn (2015), and refer to this as 'Test-II'. This data has 19 individuals in 9 species. There are 500 loci, each 700 bp long, with no site rate heterogeneity, and the Jukes-Cantor substitution model. The root height is about 0.0017, and two branches are very short ($\simeq 0.00005$). The version (1.10 beta) of STACEY used for this data has improvements which were implemented since the original submission of this paper. The implementation is faster and two operators have been added. One is a modification of the *BEAST operator NodeReheight which samples new heights non-uniformly; the other implements the 'rubber-band' move of (Rannala and Yang 2003). The main test used all 500 loci, for which we used four runs with different seeds, each of length 500 M, and with the first 10 % of each discarded as burnin. Other runs with subsets of 100 and 50 loci were also undertaken in order to compare convergence times with those reported for *BEAST in (Giarla and Esselstyn 2015).

8.1 Priors and other settings

The following settings were common to Test-I and Test-II. A single inverse gamma component with mean and standard deviation 1 was used for the prior for the per-branch population size parameters (In Eq. (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 scaling factor for population sizes. (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 Test-I data does contain such heterogeneity). The 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.

For the Test-I data, the HKY kappa parameter was estimated, as were the base frequencies. A Beta(3,1) hyperprior was used for the relative death rate, which is the extinction rate divided by the growth 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 Jones et al. 2014 for definitions of 'minimal cluster', 'collapse weight' and 'collapse height'.)

For the Test-II data, a HKY substitution model with kappa fixed at 1, and empirical base frequencies were used. The relative death rate had a beta(1,8) prior, since few extinctions are expected in a rapid radiation, and the collapse weight was fixed at 0.

8.2 Results for test-I

The results are shown in Fig. 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 (2006), had means of 250, 215, and 215 for the 4, 8, and 14 loci cases. Some individual replicates had ESSs below 100, with a minimum of 72 over all 600 replicates. The run time was about 17 days on a desktop computer with 4 cores.

8.3 Results for test-II

In the test with 500 loci, the four runs appeared to converge to similar distributions, and they were combined to produce the following results. Figure 7 shows the estimated (maximum clade credibility) species tree. The topology is correct except for the poorly supported (*sp, beatus*) clade. The ESS for the posterior, as reported by Tracer, was 1744; for the coalescent probability it was 1826; and for the likelihood it was 206.

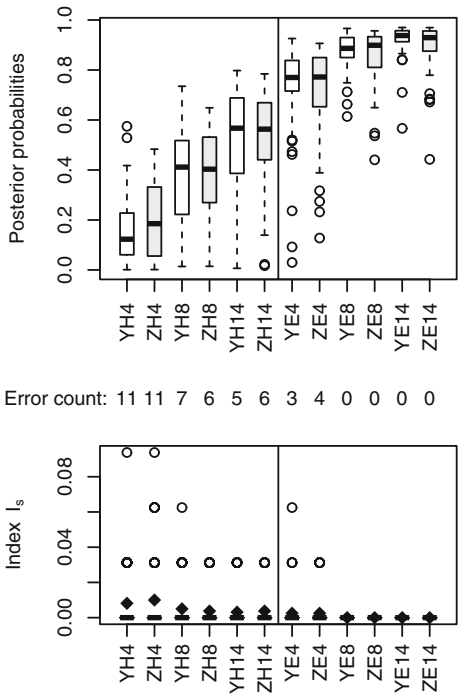

Fig. 6 The *upper boxplots* show the posterior probabilities of the true clustering over 50 replicates for twelve configurations YH4, . . . ZE14. In the *labels* for the configurations, the *first letter* Y or Z denotes the tree shape, with Y for symmetric and Z for asymmetric; the *second letter* denotes the degree of coalescent variation, 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 Fig. 3 in Olave et al

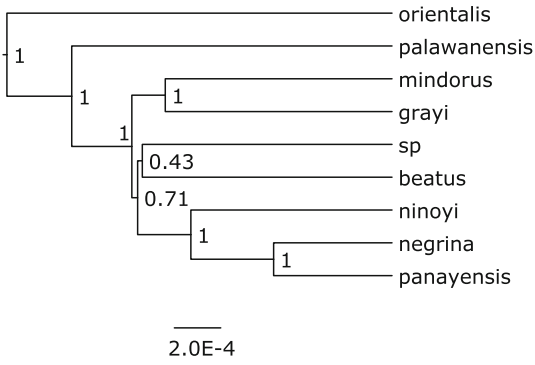

Fig. 7 The estimated SMC-tree for test-II. Posterior clade probabilities are shown at nodes

Most of the other ESSs were high, and all were above 100 except for 18 of the 500 gene tree likelihoods which had ESSs in the range 50 to 100. The run time was about 7 days. For subsets of 100 loci, 100M generations (about 2.5 h) were sufficient to achieve satisfactory EESs in most cases. For subsets of 50 loci, 50M generations (about 0.5 h) were enough.

9 Discussion

Based on tests so far, including some results not reported here, STACEY converges much faster than DISSECT for species delimitation. The difference was particularly apparent in the time to 'burn-in' in cases where there was little incomplete lineage

sorting and many loci in Test-I. Although these are the cases where signal is strongest, DISSECT can take a very long time to converge, as reported in Jones et al. (2014) for the case of 27 loci. With regard to the Test-II data, Giarla and Esselstyn Giarla and Esselstyn (2015) report that 'we observed little sign of convergence in analyses of more than 50 loci in *BEAST, even after billions of MCMC generations.' The Test-II performance is thus also very promising, though tests using different data sets are needed before strong conclusions can be drawn. It is not yet clear how much of the improvement is due to the population model and how much to the new moves.

When analyzing the Test-I data, the number of generations in the MCMC chains were chosen so that all 600 replicates could be run in a reasonable amount of time. 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, as we used for Test-II.

In the context of phylogeny estimation, the relative importance of the two kinds of noise, namely mutational and coalescent variation, was studied in Huang et al. (2010), where they are referred to as 'mutational variance' and 'coalescent variance'. In their scenarios, up to 75 % of the errors in maximum likelihood estimates of species trees were attributable to mutational variation. It seems very likely that similar conclusions apply to Bayesian species delimitation. The simulated data sets of Olave et al. (2014) have low mutational variation. 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 1000 bp 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 coalescent variation will be compounded with large amounts of mutational variation. The simulations used in Jones et al. (2014) were much harder in terms of the mutational variation: the sequences were 500 bp, 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 Olave et al. (2014) about DNA-based species delimitation. It is usually the case that geographical and morphological information is available as well (Zhang et al. 2014), 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. Solís-Lemus et al. present one such approach in Solís-Lemus et al. (2015). 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 Olave et al. (2014) and Giarla and Esselstyn (2015) for making their simulated data readily available, and for supplying extra details about their simulations. I thank two anonymous reviewers for valuable comments on an earlier version of this paper.

References

Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D, Suchard MA, Rambaut A, Drummond AJ (2014) BEAST 2: A software platform for Bayesian evolutionary analysis. PLoS Comput Biol 10(4):e1003,537. doi:10.1371/journal.pcbi.1003537

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

Edwards SV (2009) Is a new and general theory of molecular systematics emerging? Evolution 63:1–19

Felsenstein J (2003) Inferring phylogenies. Sinauer Associates, Sunderland. doi:10.1016/S0022-0000(02)00003-X

Flot JF (2015) Species delimitation's coming of age. Syst Biol 64(6):897–899

Giarla T, Esselstyn J (2015) The challenges of resolving a rapid, recent radiation: empirical and simulated phylogenomics of Philippine shrews. Syst Biol 64(5):727–740. doi:10.1093/sysbio/syv029

Heled J, Drummond A (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 104:2785–2790

Höhna S, Defoin-Platel M, Drummond AJ (2008) Clock-constrained tree proposal operators in Bayesian phylogenetic inference. In: 8th IEEE international conference on bioinformatics and bioengineering, Athens, Greece, pp 1–7, 8–10 Oct 2008

Huang H, He Q, Kubatko LS, Knowles LL (2010) 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

Huelsenbeck JP, Andolfatto P (2007) Inference of population structure under a Dirichlet process model. Genetics 175:1787–1802

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

Liu L, Pearl DK, Brumfield RT, Edwards SV (2008) Estimating species trees using multiple allele DNA sequence data. Evolution 62(8):2080–2091

Olave M, Solà E, Knowles LL (2014) Upstream analyses create problems with DNA-based species delimitation. Syst Biol 63:263–271. doi:10.1093/sysbio/syt106

Plummer M, Best N, Cowles K, Vines K (2006) CODA: convergence diagnosis and output analysis for MCMC. R News 6(1), 7–11. http://CRAN.R-project.org/doc/Rnews/

Pritchard JK, Stephens M, Donnelly PJ (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959

Rannala B (2015) The art and science of species delimitation. Curr Zool 61:846–853

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

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

Solís-Lemus C, Knowles LL, Ane C (2015) Bayesian species delimitation combining multiple genes and traits in a unified framework. Evolution 69:492–507

Yang Z (2002) Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. Genetics 162:1811–1823

Yang Z, Rannala B (2010) Bayesian species delimitation using multilocus sequence data. Proc Natl Acad Sci USA 107:9264–9269

Yang Z, Rannala B (2014) Unguided species delimitation using DNA sequence data from multiple loci. Mol Biol Evol 31(12):3125–3135. doi:10.1093/molbev/msu279

Zhang C, Rannala B, Yang Z (2014) Bayesian species delimitation can be robust to guide-tree inference errors. Syst Biol 63:993–1004. doi:10.1093/sysbio/syu052