STACEY: species delimitation and phylogeny estimation under the multispecies coalescent

Graham Jones, www.indriid.com

October 9, 2014

Abstract

This article describes a new package called STACEY for BEAST2 which is capable of both species delimitation and species tree estimation using DNA sequences from multiple loci. The focus in this article is on species delimitation. STACEY is based on the multispecies coalescent model, and builds on earlier software (DISSECT), which uses a 'birth-death-collapse' prior to deal with delimitations without the need for reversible-jump Markov chain Monte Carlo moves. Like DISSECT, it requires no a priori assignment of individuals to species or populations, and no guide tree. This paper introduces two innovations. The first is a new model for the populations along the branches of the species tree, and the second is a new MCMC move for exploring the posterior when the multispecies coalescent model is assumed. The main benefit of STACEY over DISSECT is much better convergence. Current practice, using a pipeline approach to species delimitation under the multispecies coalescent, has been shown to have major problems on simulated data. The same simulated data set is used to demonstrate the accuracy and efficiency of STACEY.

1 Introduction

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 major 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) and DISSECT (Jones and Oxelman, 2014). DISSECT provided the first program which required no a priori assignment of individuals to species or populations, and no guide tree. More details and discussion of alternative approaches can be found in Jones and Oxelman (2014), Olave et al. (2014), and Zhang et al. (2014).

The birth-death-collapse model in 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. This is a computational approximation to a model in which the dimensionality of the parameter space changes as the number of species changes. As explained in Jones and Oxelman (2014), the birth-death-collapse model effectively converts the species tree into a tree in which the tip nodes are **minimal clusters** of individuals. Other nodes represent unions of these clusters, including the inferred species. I will refer to this tree as the **SMC-tree** meaning the 'species or minimal clusters tree'.

Olave et al. (2014) simulated sequences under the multispecies coalescent model, and then analyzed the data using a standard 'pipeline' method using Structurama (Pritchard et al., 2000; Huelsenbeck and Andolfatto, 2007), *BEAST (Heled and Drummond, 2010), and BPP (Rannala and Yang, 2013). 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 (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). A beta version of STACEY can be found at `www.indriid.com`. It incorporates a new model for the populations along the branches of the SMC-tree, and a new MCMC move for exploring the posterior when the multispecies coalescent model is assumed.

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 is to introduce one or more population parameters for each branch. The method described here is between these. It 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, but does not allow individual populations to be estimated. The method is analogous to the common one for modeling site rate heterogeneity where it is assumed that each site independently 'chooses' a rate from a gamma (or other) distribution. Unlike the site heterogeneity case, there is no need to approximate the integral.

The 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 details of the population model and the MCMC move are in sections 2 and 3, followed by the results obtained on the data set of Olave et al. (2014). All trees are rooted and binary. Times are measured backwards from zero at present.

2 The population model

Consider a single gene and a single branch. 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 equation (3), p572, of Heled and Drummond (2010)):

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

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. 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 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 most 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}. Thus the number of lineages in gene tree j at the rootward end of branch b is $n_{jb} - k_{jb}$. 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

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

where

$$q_b = \sum_j k_{jb}, \quad r_b = \prod_j p_j^{-k_{jb}}, \quad \text{and}$$

$$\gamma_b = \sum_j p_j^{-1} \sum_{i=0}^{k_{jb}} c_{jbi} \binom{n_{jb} - i}{2}. \qquad (3)$$

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

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

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

$$\begin{aligned} &\pi_\Theta(\Theta|\sigma)\pi_\sigma(\sigma) = \\ &\pi_\sigma(\sigma) \prod_b h(\theta_b|\sigma) = \\ &\pi_\sigma(\sigma) \prod_b \sum_{c=1}^{C} \lambda_c (\sigma\beta_c)^{\alpha_c} \Gamma(\alpha_c)^{-1} \theta_b^{-\alpha_c - 1} \times \\ &\qquad\qquad \exp(-\sigma\beta_c\theta_b^{-1}) \mathbf{1}_{\mathbf{R}_+^B} \end{aligned} \qquad (4)$$

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

Then combining (2) and (4), 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 $f(\sigma, \lambda_c, \alpha_c, \beta_c, \theta_b, q_b, \gamma_b, r_b) =$

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

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

$$\begin{aligned} &\int_X f_G(G|\Theta)\pi_\Theta(\Theta|\sigma)\pi_\sigma(\sigma) \mathrm{d}\Theta = \\ &\pi_\sigma(\sigma) \prod_b r_b \sum_{c=1}^{C} \lambda_c \frac{(\sigma\beta_c)^{\alpha_c}}{(\sigma\beta_c + \gamma_b)^{\alpha_c + q_b}} \frac{\Gamma(\alpha_c + q_b)}{\Gamma(\alpha_c)}. \end{aligned} \qquad (5)$$

Equations (5) and (3) provide the information needed to implement the method.

3 The NodesNudge move

3.1 Conventions and notation

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 $\mathrm{anc}(X)$ and its node time by $t(X)$. The branch that leads from $\mathrm{anc}(X)$ to X is referred to as 'the branch X'. A tree topology should be understood as a labeled topology, that is, it includes the assignment of labels to tips.

For a node X in the SMC-tree, let $I(X)$ denote the set of the individuals 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 individuals which provided a sequence belonging to x. Furthermore, if X is not a tip, let $R(X)$ and $L(X)$ denote the set of individuals 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 of individuals are unions of minimal clusters. In the SMC-tree the unions are disjoint, and a node is uniquely identified by its set of individuals. In the gene tree case, neither of these is true in general. However, the set $I(x)$ and time $t(x)$ 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 time $t(x)$, 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.

3.2 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 1 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 $\mathrm{anc}(c) \in C(x)$ 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:

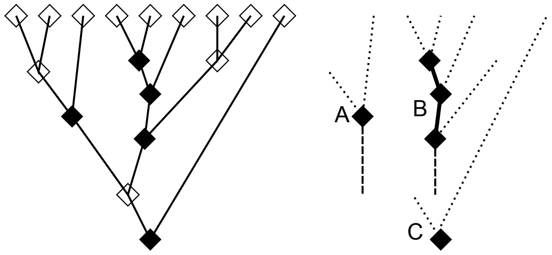

Figure 1: Example of connected components

1. Choose uniformly and at random any internal node S in the SMC-tree. S may be the root.

2. Let $\Delta(S) = \{s_1, \dots, s_n\}$ 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 $d_0 = \max_X\{t(X) : \mathrm{anc}(X) = S\}$, which is the time of the most ancient of the two child nodes of S, and let $u_0 = \infty$ if S is the root, otherwise let $u_0 = t(\mathrm{anc}(S))$.

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

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

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

5. Let

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

   and

$$U = \min\{U_{\max}, \min_{0\le i\le n}(u_i - t(s_i) + t(S))\}$$

   where $U_{\max}$ is an arbitrary positive constant.

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. The value d_0 (step 3) can be written as $t(S) + \max_X\{t(X) - t(S) : \mathrm{anc}(X) = S\}$ which emphasizes the similarity with the other d_i (step 4). The constant $U_{\max}$ (step 5) is introduced to deal with the case that S is the root of the SMC-tree, and $\Delta(S)$ is either empty or consists only of root nodes of gene trees. In this case there are no constraints on how large $t(S)$ could become, so an 'artificial' constraint is introduced to avoid dealing with an infinite interval.

Returning to Figure 1, 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.

3.3 Properties

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

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

$$\max_X\{t(X) : \mathrm{anc}(X) = S\} \le t'(S) \le t(\mathrm{anc}(S))$$

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

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

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

so

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

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 nonnegative 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,

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

so

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

and for all i,

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

hence

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

as required.

Proposition 2. *The move is reversible.*

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 4, before the move,

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

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 reversible.

3.4 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:

$$\Big(\big(L(s) \subset L(S)\big) \wedge \big(R(s) \subset R(S)\big)\Big) \quad \vee$$
$$\Big(\big(L(s) \subset R(S)\big) \wedge \big(R(s) \subset L(S)\big)\Big)$$

and

$$\big(I(s) \not\subset L(S)\big) \wedge \big(I(s) \not\subset R(S)\big).$$

The first condition says that one of $L(s)$ and $R(s)$ is contained in one of $L(S)$ and $R(S)$, and the other one of $L(s)$ and $R(s)$ is contained in the other one of $L(S)$ and $R(S)$. The second condition says that $I(s)$ is not contained in either $L(S)$ or $R(S)$. The nodes in $\Delta(S)$ are the 'first meetings' between sequences from $L(S)$ and $R(S)$. 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.

4 Results on simulated data

As a proof-of-concept demonstration, I analyzed the simulated data provided in the supplementary material of Olave et al. (2014). The data was incorporated into XML files for BEAST2. Version 2.1.3 of BEAST2 and 0.1.0 of STACEY were used. The program was run for 4 million generations on each replicate with the first 1 million discarded as burnin. Samples were taken every 1000 generations, so there were 3000 SMC-trees on which to base the species delimitations using SpeciesDelimitationAnalyser (Jones and Oxelman, 2014).

4.1 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 (2), $C = 1, \alpha_1 = 3.0, \beta_1 = 2.0$.) 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 'coalescent factor' 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 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 Jones and Oxelman (2014) for definitions of 'minimal cluster', 'collapse weight' and 'collapse height'.)

4.2 Results

The results are shown in Figure 2. The clustering with the largest posterior probability was used as a point estimate of the species delimitation. All errors in this point estimate were false splits. Usually just one of the true species was split, and occasionally, two true species were split. In all 600 replicates, the true clustering was in the 0.95 credible set. The highest posterior probability assigned to an erroneous clustering was 0.82 (replicate 16 from YE4). The estimated sample sizes (ESSs) as reported by Tracer (Rambaut et al., 2014) were low, generally around 70-140 for the 4-gene scenarios and 25-70 for the 14-gene scenarios.

5 Discussion

Based on tests so far, including some results not reported here, STACEY converges much faster than DISSECT, especially when there is a strong signal in the data. I estimate it would have taken around ten times as much computation to obtain similar results as those presented here if DISSECT had been used. It is not yet clear how much of this improvement is due to the new model and how much to the new move. It seems likely that the new move will improve convergence in *BEAST in some cases at least, but this has not been tried.

The number of generations (4 million) in the MCMC chain was 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. Given the purpose of this paper, 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.

There are two main types of 'noise' which interfere with inference of delimitation and phylogeny: mutational variance and incomplete lineage sorting. In the context of phylogeny estimation, the relative importance of these was studied in Huang et al. (2010). 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 simulated data sets of Olave et al. (2014) 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 Jones and Oxelman (2014) 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 Olave et al. (2014) about DNA-based species

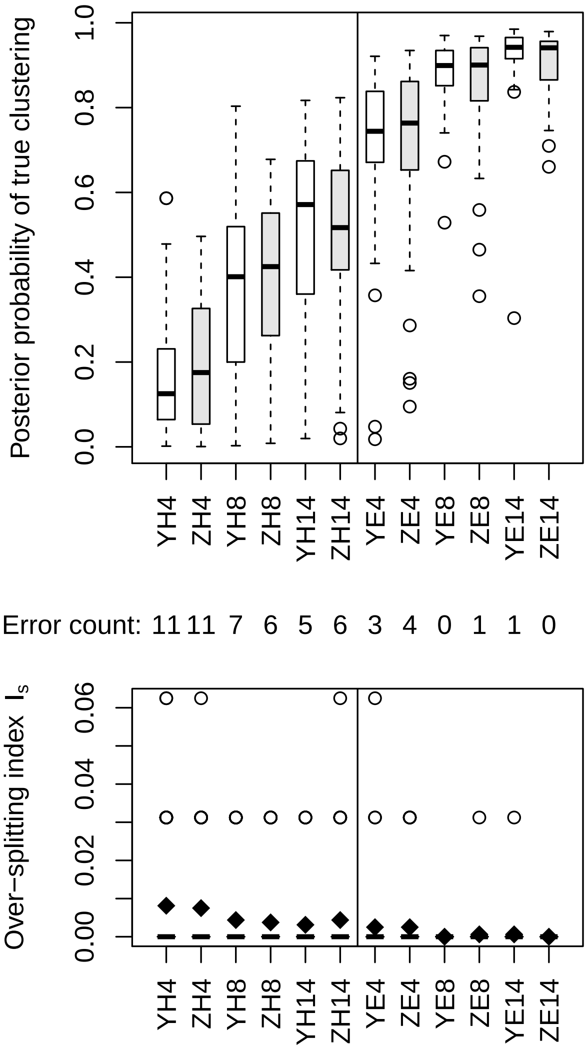

Figure 2: The upper boxplots show the posterior probability of the true clustering over 50 replicates for eight configurations YH4,... ZE14. In the labels for the configurations, the first letter Y or Z denotes the tree shape, with Y for symmetric and Z for asymmetric; the second letter denotes the degree 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 about one tenth of Fig 3 in Olave et al.

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 express the geographical and morphological information in a Bayesian 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) for making their simulated data readily available, and Melisa Olave for supplying extra details about the simulations.

References

Remco Bouckaert, Joseph Heled, Denise Khnert, Tim Vaughan, Chieh-Hsi Wu, Dong Xie, Marc A Suchard, Andrew Rambaut, and Alexei J. Drummond. BEAST 2: A software platform for bayesian evolutionary analysis. *PLoS Comput Biol*, 10(4):e1003537, 2014. doi: 10.1371/journal.pcbi.1003537.

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

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

H Huang, Q He, L S Kubatko, and L L Knowles. 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.

J P Huelsenbeck and P Andolfatto. Inference of population structure under a Dirichlet process model. *Genetics*, 175: 1787–1802, 2007.

G Jones and B Oxelman. DISSECT: an assignment-free Bayesian discovery method for species delimitation under the multispecies coalescent. *bioRχiv*, 2014. doi: 10.1101/003178.

Melisa Olave, Eduard Solà, and L Lacey Knowles. Upstream analyses create problems with DNA-based species delimitation. *Syst Biol*, 63:263–271, 2014. doi: 10.1093/sysbio/syt106.

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

A Rambaut, M A Suchard, D Xie, and A J Drummond. Tracer v1.6, 2014. URL `http://beast.bio.ed.ac.uk/Tracer`.

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

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

Chi Zhang, B Rannala, and Z Yang. Bayesian species delimitation can be robust to guide-tree inference errors. *Syst. Biol.*, 0:1–12, 2014. doi: 10.1093/sysbio/syu052.