Using STACEY on larger data sets, part 2

Graham Jones

2015-11-01, February 4, 2016

This is a second (but still preliminary) report on testing STACEY on data sets with a large number of individuals for species delimitation, or a large number of loci for phylogeny estimation.

The version 1.1.0 (beta) of STACEY used here has improvements compared to that in Jones (2015). The `StaceyNodeReheight` operator is a variant of BEAST's `NodeReheight` which samples new heights non-uniformly. A new operator `ThreeBranchAdjuster` has been added. There are some further tuning of the operator choices and weights. The implementation has also been made more efficient (using classes `UnionArrays` and `FitsHeights`) both for the calculation of the coalescent, and for some of the operators.

Brief results:

- Estimated delimitation. 200 individuals, 5 loci, high mutational variance: STACEY fails to separate very close species.
- Estimated delimitation. 200 individuals, 30 loci, high mutational variance: this is enough data to get good results, but STACEY was very slow at exploring different delimitations.
- Fixed delimitation. 19 individual, 9 species. Convergence problems with 1112 loci. OK with 278 loci.

I am using a desktop computer with a i7-4790 CPU @ 3.6GHz, 16Gb RAM.

1 Estimated delimitation

1.1 Simulated data with 200 individuals, 5 loci

This data is similar to that used in Jones et al. (2014) but with more species and individuals. There are 10 true species, each containing 20 individuals. Fig 1 shows the true species tree. There is one sequence of length 500bpp per individual, and 5 loci. These numbers were chosen to roughly match some analyses that others have attempted with STACEY. The sequences were generated using a HKY substitution model, with no site rate heterogeneity, and the same clock rate for all loci. Fig 3 shows histograms of the genetic distances between pairs of sequences taken from individuals from f and g, for the 5 loci. (I used `dist.dna` from package `ape`, default arguments.) The third and fourth columns show the within-species and between-species values.

The estimation used the HKY substitution model, with no site rate heterogeneity. Independent substitution parameters (kappa and base frequencies) were estimated for each locus. The relative clock rates were also estimated.

I used two runs with different seeds, each of length 1000M. The first 10% of each was discarded as burnin, and the remaining samples combined. Samples were taken every 100,000, meaning there were 18000 samples in total. Time: about 4m20s/Msamples, when running 2 simultaneously. Total about 72h.

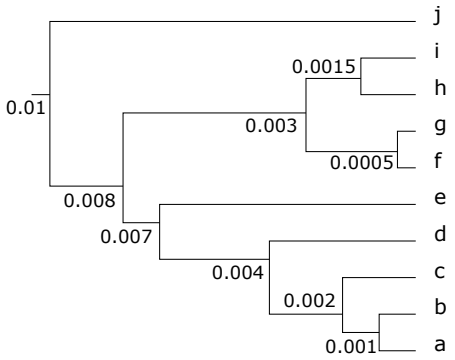

Figure 1: Species tree for a scenario for estimated delimitation, with individuals in each of the ten species a-j. Node heights are in substitutions.

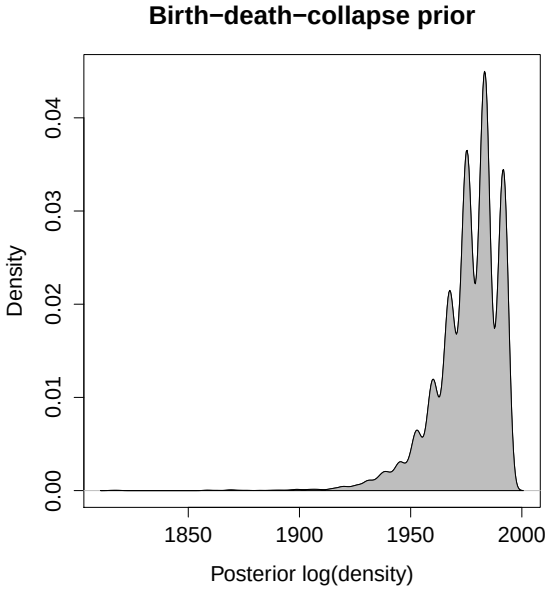

Figure 2: Posterior distribution of log-density of birth-death-collapse prior.

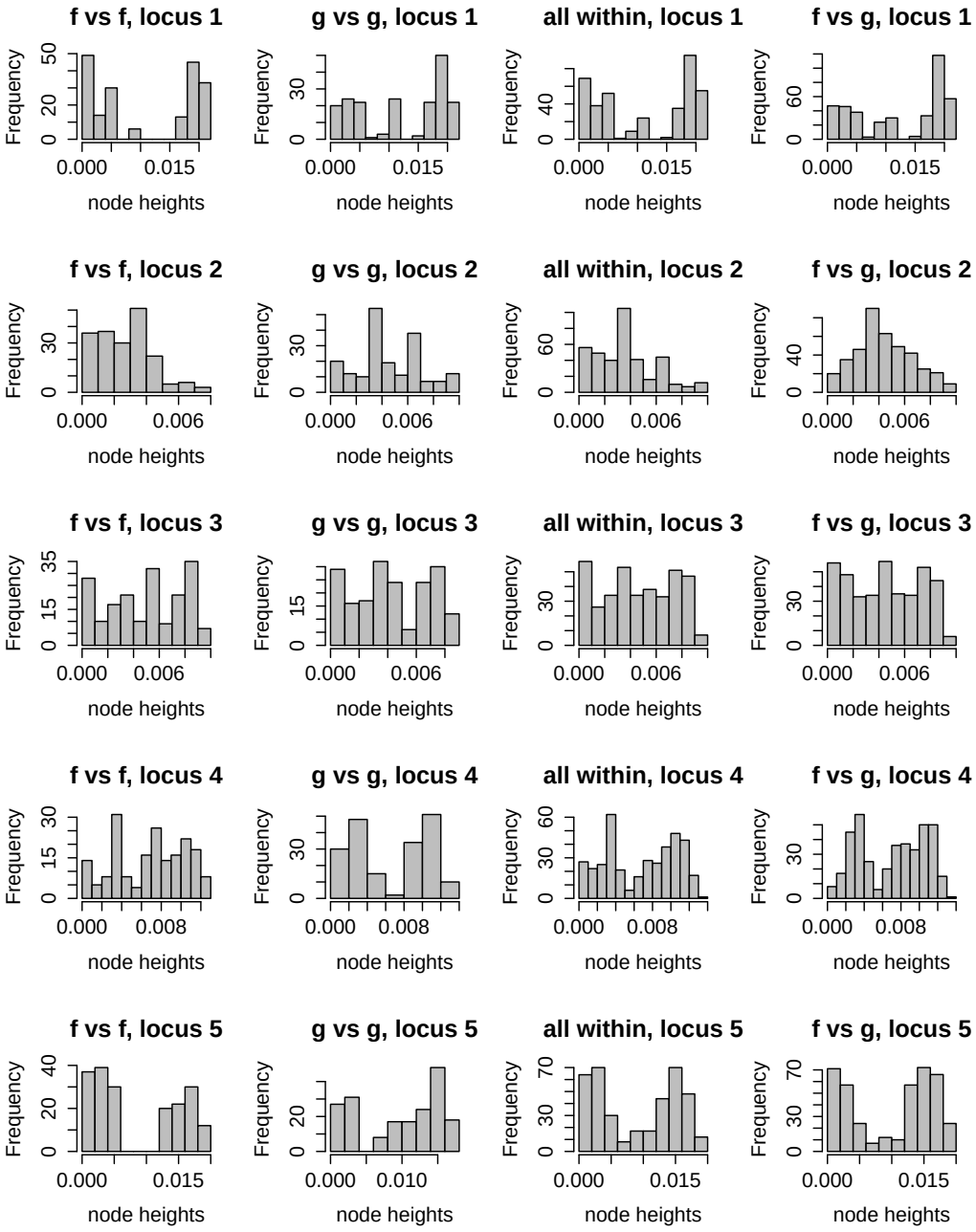

Figure 3: Histograms of genetic distances between pairs from individuals from species f (first column), species g (second column). All within-species pairs, that is, the first two columns combined, are in the third column and between-species pairs in the fourth column. Each row is for a different locus.

Results. The ESS values for the posterior, as reported by Tracer (Rambaut et al., 2014), were 710 and 848, combined 1557. The ESS values for the coalescent probability were 242 and 660, combined 524. The ESS values for most other values were large, mostly above 1000. The exceptions were the prior (83 and 88, combined 130), the birth-death-collapse prior (84 and 88, combined 129), and the number of clusters (82, 92, combined 133). These three are closely related: the birth-death-collapse prior is a major part of the overall prior, and its value depends on the number of clusters. The distribution of the birth-death-collapse prior (Fig 2) was multimodal. (See Figure 2 in the DISSECT paper, and imagine sampling density values, to see that this is expected with little signal in the data.)

The correct species delimitation did not appear in the posterior samples. The main problem was that the two very close species f and g were never correctly separated. The delimitation with the highest PP (0.16) had a false join of all individuals from species f and g, and a mis-assignment of one individual a02 with the b's. The second (PP 0.064) also had the false join of f's and g's, but no other errors. The posterior mean of the number of clusters was 10.9. About 9800 delimitations appeared in the posterior samples in total, mostly just once. About 1160 appeared twice or more.

1.2 Simulated data with 200 individuals, 30 loci

This data set is the same as the previous one, except there are 30 loci, not 5. The settings for teh analysis were also the same.

I used two runs with different seeds, each of length 1000M. The first 10% of each was discarded as burnin, and the remaining samples combined. Samples were taken every 100,000, meaning there were 18000 samples in total. Time: about 9m30s/Msamples, when running 2 simultaneously. Total about 7 days.

Results. The ESS values for the posterior, as reported by Tracer (Rambaut et al., 2014), were 122 and 111, combined 246. The ESS values for the coalescent probability were 120 and 133, combined 250. The ESS values for most other values were large, mostly above 400. The exceptions were the prior the birth-death-collapse prior and the number of clusters: all these were around (100, 28, combined 14).

There were no false joins and very few mis-assignments sampled, so the true delimitation and false splits heavily dominated the posterior samples. The first run visited the true delimitation about 0.6 of the time, delimitations with 11 species (≈ 0.3), rest (≈ 0.1). The second run visited the true delimitation about 0.1 of the time, delimitations with 11 species (≈ 0.4), 12 (≈ 0.3), rest (≈ 0.2). The second run spent a long time with species 'e' split into two, and this accounts for most of the difference between the runs.

The posterior probability of the true delimitation was 0.34. The second highest posterior probability was 0.036. About 2800 delimitations were visited, about 740 more than once. The runs were not long enough to explore the delimitations well, so these numbers might change with longer runs.

2 Fixed delimitation

2.1 UCE data from Giarla and Esselstyn (2015)

The UCE data from Giarla and Esselstyn (2015) has 19 individuals in 9 species (2+2+1+4+1+5+1+2+1). There are 1112 loci. Unlike Giarla et al, I did not remove any alignments such as those which are phylogenetically uninformative. As I understand it, 'phylogenetically uninformative' means 'uninformative about topology in the context of unrooted trees'. They may still be informative about the topology of a rooted tree, and the continuous

parameters also matter. Even an alignment which consists of identical sequences influences the estimates of some of these.

I used a HKY substitution model with kappa fixed at 3.0, and empirical frequencies. The value of 3.0 was roughly estimated from runs using less data. There was no site rate heterogeneity. Relative clock rates for each gene tree except the first were estimated with `lnorm(meanlog=0, sdlog=1)` priors. These were the only per-gene parameters apart from the gene trees. For the birth-death-collapse model, the growth rate had a `lnorm(meanlog=4.6, sdlog=2)` prior, relative death rate a `beta(alpha=1, beta=8)` prior, and the collapse weight was fixed at 0. The population scaling factor in the STACEY coalescent had a `lnorm(meanlog=-7, sdlog=2)` prior.

I tried using all 1112 loci, and also dividing into 4 sets each containing 278 loci. For the all loci case, I used four runs with different seeds, each of length 900M. The runs were halted at this point due to convergence problems. For the subsets of loci, I used two runs with different seeds for each subset, and each run was 1000M with 200M discarded as burnin.

Topologies in the results. All the sampled species trees when using all loci, and almost all when using the subsets, had the following clades in common:

- all but orientalis
- all but orientalis+palawanensis
- mindorus+grayi
- ninoyi+negrina+panayensis
- negrina+panayensis

Given this, almost all the topologies can be described as one of the 15 rooted trees on these four clades:

- mindorus+grayi (mg)
- ninoyi+negrina+panayensis (nnp)
- sp (s)
- beatus (b)

Results for all 1112 loci. Time: about 22m/Msamples, when running 4 simultaneously. Took about 14 days. There were mixing problems. Posterior ESSs were not bad (combined 133) but the likelihood ESSs were very low. Further, it was clear that the runs were exploring different topologies very slowly, and that different runs were sampling different distributions of topologies.

Results for subsets of 278 loci. Time: about 5m/Msamples, when running 4 simultaneously. The eight runs took about 7 days. Figure 4 and table 2 shows how the topologies were explored. For each of the four subsets, there is good agreement between the runs with different seeds. The posterior ESSs were much better than the 1112-loci case (combined 1680, 868, 767, 1671) though there were still problems with low ESSs for likelihoods (45, 70, 157, 282). The results are summarised further in tables 3 and 4, where:

- q1 means first quarter, loci 1-278
- q2 means loci 279-556
- q3 means foci 557-834
- q4 means loci 835-1112
- s42, s43 refer to two different seeds.

Index	Topology
0	X
1	(((mg,s),b),nnp)
2	(((mg,b),s),nnp)
3	(((s,b),mg),nnp)
4	(((mg,s),nnp),b)
5	(((mg,nnp),s),b)
6	(((s,nnp),mg),b)
7	(((mg,b),nnp),s)
8	(((mg,nnp),b),s)
9	(((b,nnp),mg),s)
10	(((s,b),nnp),mg)
11	(((s,nnp),b),mg)
12	(((b,nnp),s),mg)
13	((mg,s),(b,nnp))
14	((mg,b),(s,nnp))
15	((mg,nnp),(s,b))

Table 1: Numbers assigned to topologies in Figure 4. Nearby topologies do not always get nearby numbers. X is a topology not containing all of the clades listed above.

Topology	q1s42	q1s43	q2s42	q2s43	q3s42	q3s43	q4s42	q4s43
X	0	0	10	121	0	0	0	0
(((mg,s),b),nnp)	0	0	369	518	0	1	0	0
(((mg,b),s),nnp)	0	0	0	1	0	1	0	0
(((s,b),mg),nnp)	42	0	0	0	70	20	486	126
(((mg,s),nnp),b)	0	5	5126	5463	0	0	0	0
(((mg,nnp),s),b)	0	0	121	277	0	0	0	0
(((s,nnp),mg),b)	0	0	459	732	2	94	0	0
(((mg,b),nnp),s)	0	0	0	5	0	3	0	0
(((mg,nnp),b),s)	0	0	0	20	0	6	0	0
(((b,nnp),mg),s)	209	671	91	45	67	125	0	0
(((s,b),nnp),mg)	621	832	0	0	1698	1320	6592	7184
(((s,nnp),b),mg)	539	257	0	4	4670	3424	574	469
(((b,nnp),s),mg)	2227	2272	33	4	1408	2685	50	92
((mg,s),(b,nnp))	4309	3964	1792	766	6	233	0	0
((mg,b),(s,nnp))	0	0	0	8	69	35	0	0
((mg,nnp),(s,b))	54	0	0	37	11	54	299	130

Table 2: Raw counts of samples of topologies, over 4 subsets of loci (q1-q4) and two seeds (42 and 43).

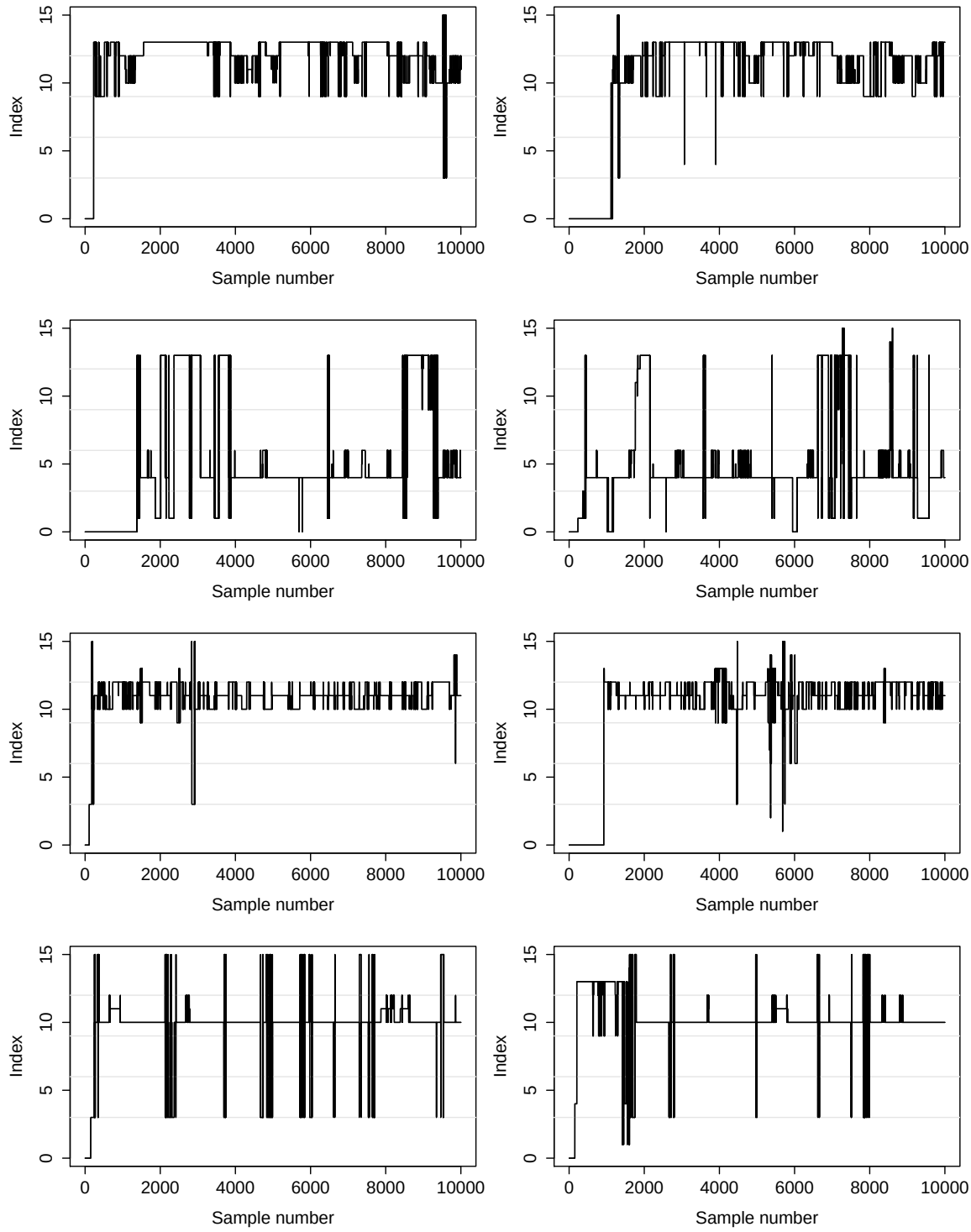

Figure 4: Traces of topologies for 2 runs with different seeds, for each of four subsets of loci. Each row is for a different subset of 278 loci. See table 1 for the meaning of the index numbers.

Topology	q1	q2	q3	q4
X	0	1	0	0
(((mg,s),b),nnp)	0	6	0	0
(((mg,b),s),nnp)	0	0	0	0
(((s,b),mg),nnp)	0	0	1	4
(((mg,s),nnp),b)	0	66	0	0
(((mg,nnp),s),b)	0	2	0	0
(((s,nnp),mg),b)	0	7	1	0
(((mg,b),nnp),s)	0	0	0	0
(((mg,nnp),b),s)	0	0	0	0
(((b,nnp),mg),s)	5	1	1	0
(((s,b),nnp),mg)	9	0	19	86
(((s,nnp),b),mg)	5	0	51	7
(((b,nnp),s),mg)	28	0	26	1
((mg,s),(b,nnp))	52	16	1	0
((mg,b),(s,nnp))	0	0	1	0
((mg,nnp),(s,b))	0	0	0	3

Table 3: PPs×100 of topologies, combined over seed, for 4 subsets of loci (q1-q4).

Topology	PPs×100
(((mg,s),nnp),b)	17
(((s,b),nnp),mg)	29
(((s,nnp),b),mg)	16
(((b,nnp),s),mg)	14
((mg,s),(b,nnp))	17

Table 4: PPs×100 of topologies, combined over seeds, and loci. Small PPs (< 0.03) are ignored.

References

TC Giarla and JA Esselstyn. The challenges of resolving a rapid, recent radiation: empirical and simulated phylogenomics of Philippine shrews. *Systematic Biology*, 64(5):727–40, 2015. doi: 10.1093/sysbio/syv029.

G Jones. Species delimitation and phylogeny estimation under the multispecies coalescent. *bioRχiv*, 2015. doi: 10.1101/010199. URL `http://biorxiv.org/content/early/2015/03/22/010199`.

Graham Jones, Zeynep Aydin, and Bengt Oxelman. DISSECT: an assignment-free Bayesian discovery method for species delimitation under the multispecies coalescent. *Bioinformatics*, 2014. doi: 10.1093/bioinformatics/btu770.

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. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, 164:1645–1656, 2003.