Using STACEY on larger data sets

Graham Jones

2015-07-21, August 31, 2015

This is a 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. I am using a development version of STACEY, and attempting to improve it at the same time. This increases the risk of bugs or other mistakes, so the results here are very preliminary.

I am using a desktop computer with a i7-4790 CPU @ 3.6GHz.

1 Estimated delimitation

1.1 Simulated data with 100 individuals

The version of STACEY used here is similar to that in Jones (2015). The implementation is faster, using `UnionArrays` and `FitsHeights` objects, and a replacement version of `NodeReheight` operator (`StaceyNodeReheight`), but the algorithm is essentially identical. There are also changes in the XML (operator choices and weights) compared to Jones (2015).

This data is similar to that used in Jones et al. (2014) but larger. There are ten true species, each containing ten individuals. Fig 1 shows the true species tree. There is one sequence of length 500bpp per individual, and 30 loci. The sequences were generated using a HKY substitution model, with no site rate heterogeneity, and the same clock rate for all loci.

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 four runs with different seeds, each of length 550M. The first of each 10% was discarded as burnin, and the remaining samples combined. Samples were taken every 200,000, meaning there were 9900 samples in total. Time: about 5m30s/Msamples, when running 4 simultaneously. Took about 50 hours.

Results. The ESS values for the posterior, as reported by Tracer (Rambaut et al., 2014), were 118, 90, 123, 43, combined 250. The ESS values for the coalescent probability were very similar. All other ESSs were at least as big, mostly above 1000.

The correct species delimitation had a posterior probability (PP) of 0.173, the largest for any delimitation. The delimitation with second highest PP (0.086) had a false split in species e with individuals e01 and e07 separated from the others. The third highest PP was 0.010 and all other delimitations had PP less than 0.01. All erroneous delimitations were false splits. About 5200 delimitations appeared in the posterior samples in total, mostly just once. About 600 appeared twice or more.

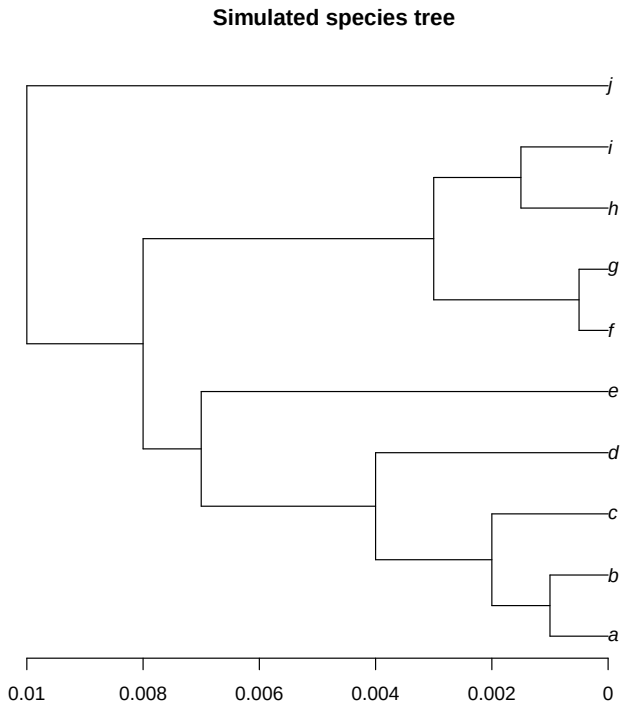

Figure 1: Species tree for a scenario for estimated delimitation, with ten individuals in each of the ten species a-j.

2 Fixed delimitation

2.1 Simulated data from Giarla and Esselstyn (2015)

The version of STACEY used here has improvements compared to that in Jones (2015). `StaceyNodeReheight` operator samples new heights non-uniformly, and a new operator `ThreeBranchAdjuster` has been added. I committed it to local Git repository on 31 August. There are some further tuning of the operator choices and weights.

The simulated data (Giarla and Esselstyn, 2015b) from (Giarla and Esselstyn, 2015a) has 19 individuals in 9 species (2+2+1+4+1+5+1+2+1). There are 500 loci, each 700bp, no site rate heterogeneity, JC substitution model. The root height is about 0.0017, and two branches are very short, I think around 0.00005.

I used a HKY substitution model with kappa fixed at 1, and empirical frequencies. 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 used four runs with different seeds, one of length 400M, the other three 200M. The first 50M of each was discarded as burnin, and the remaining samples combined. Samples were taken every 100,000, meaning there were 3500+3*1500=8000 samples in total. Time: about 10m30s/Msamples for one run, about 19m/Msamples when running 4 simultaneously. Took about 35h for one run of 200M, then 63h to resume this together with three others.

Results. The ESS values for the posterior, as reported by Tracer, were 113, 235, 242, 196, combined 414. The ESS values for the coalescent probability were very similar. The ESSs for the likelihood were worse: 46, 18, 114, 76, combined 80. About 54 of the 500 gene tree likelihoods had combined ESSs below 100. About 7 of the relative clock rates had combined ESSs below 100. ('About' because manual counts.) See Figure 2 for traces of some values. CODA (Plummer et al., 2006) reports higher ESSs.

Figure 3 shows the estimated (maximum clade credibility) species tree for the combined run, which has the correct topology. Figures 4, 5, 6, 7 show the estimated species tree for the individual runs. Only the last one has correct topology. Note that the scale bar shows 0.0003 where it should be 0.00025, etc., in all these figures.

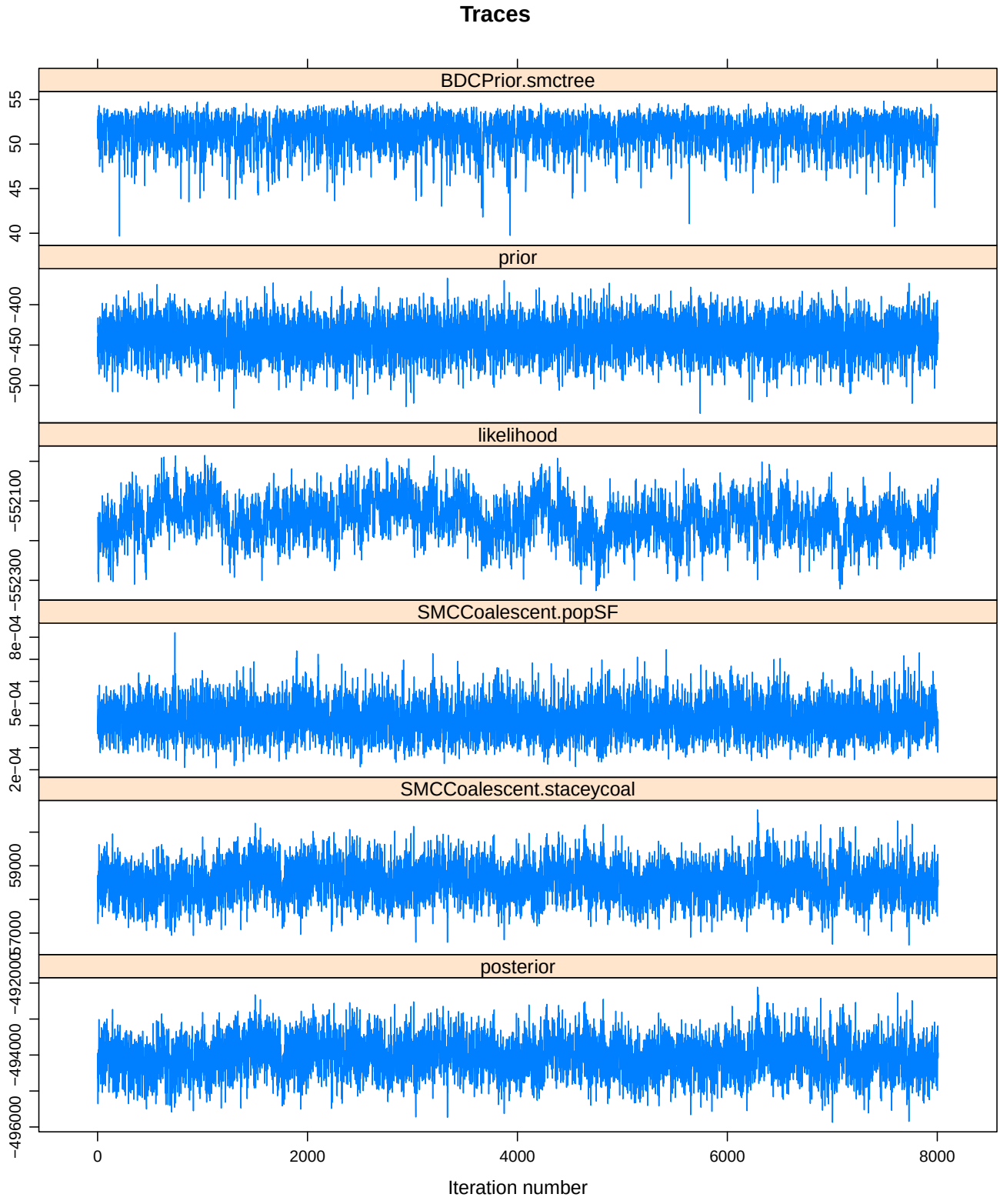

Figure 2: Traces for 4 runs with different seeds combined. One run is 400M, the other three are 200M. Burnin is 50M in all cases.

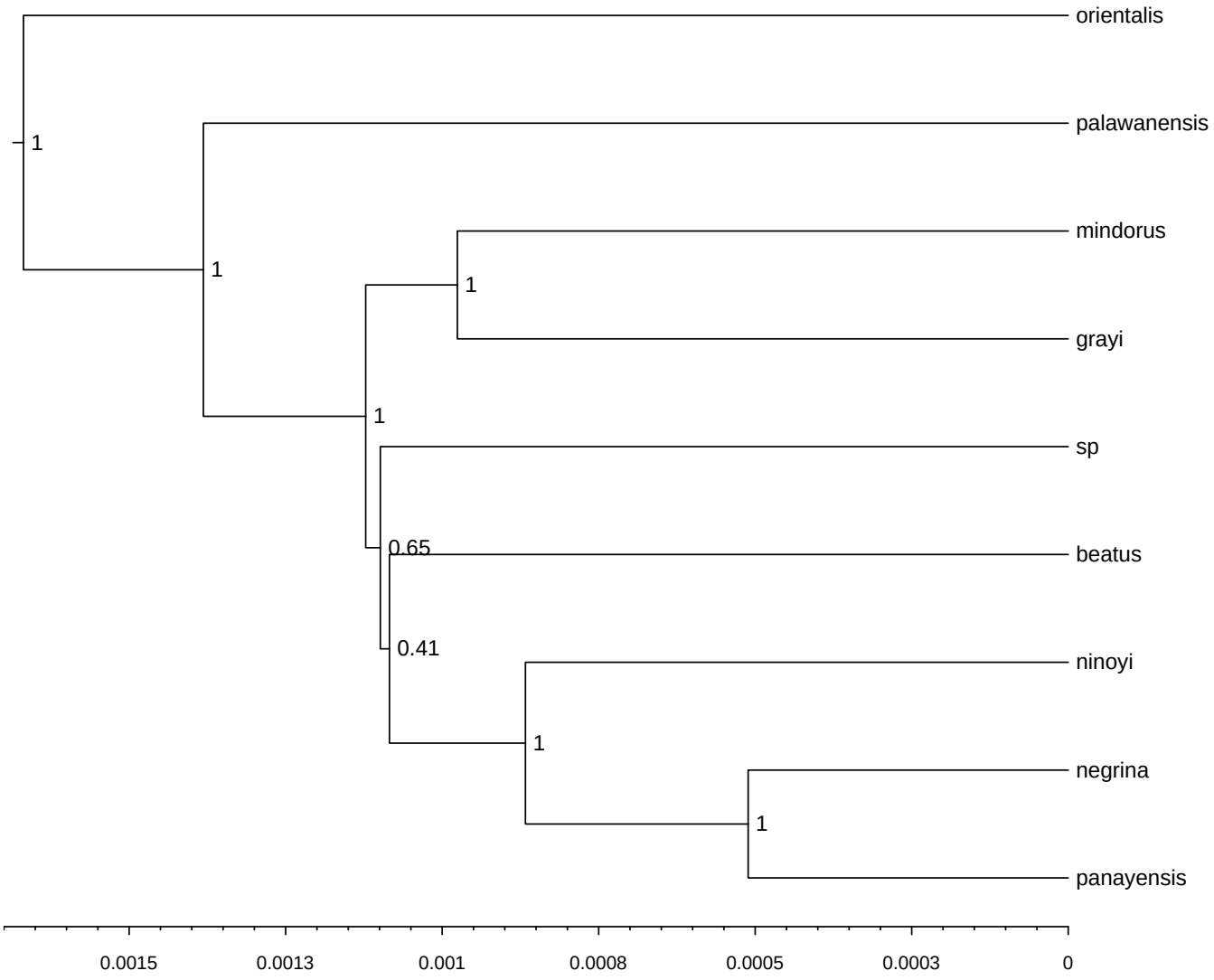

Figure 3: Species tree from the four runs combined with first 50M discarded as burnin. (13% of 400M, 25% of each 200M discarded using LogCombiner.)

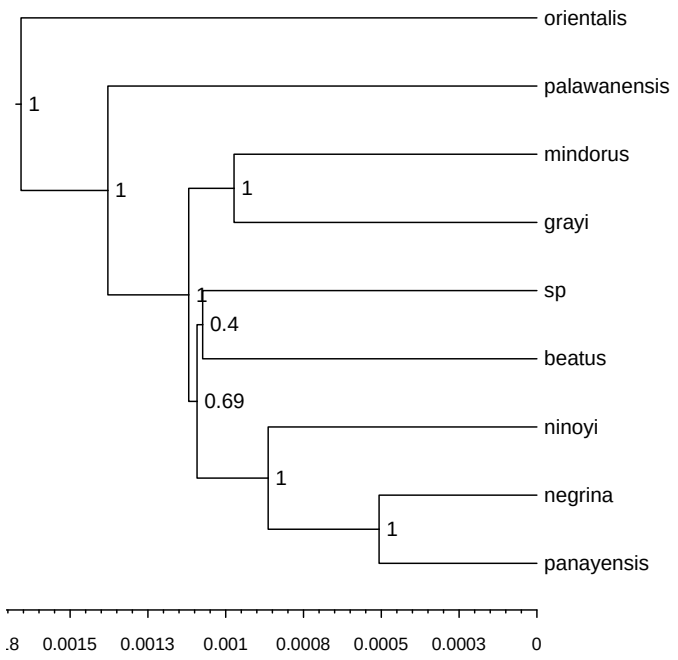

Figure 4: Species tree from the 400M run with 13% burnin.

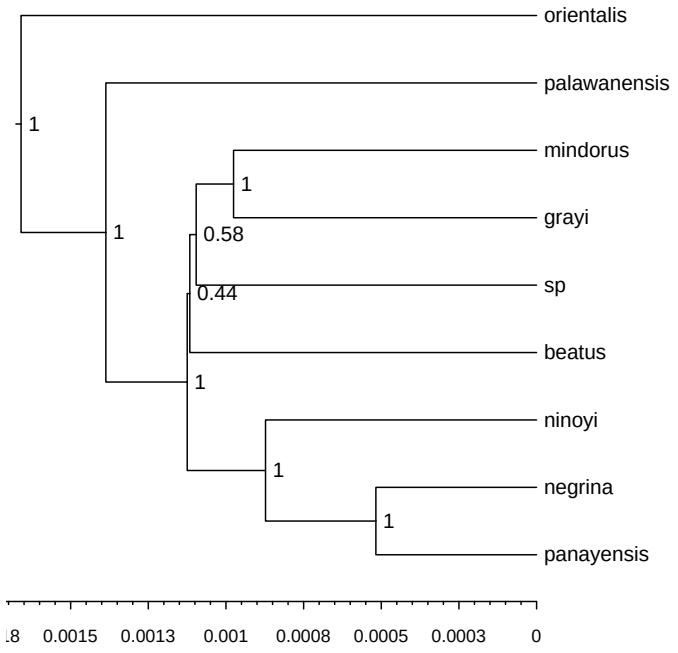

Figure 5: Species tree from the first 200M run with 25% burnin.

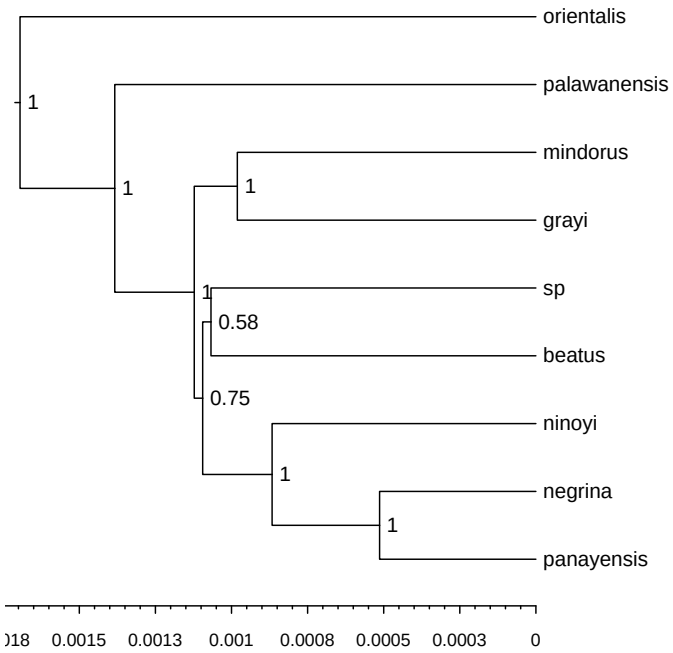

Figure 6: Species tree from the second 200M run with 25% burnin.

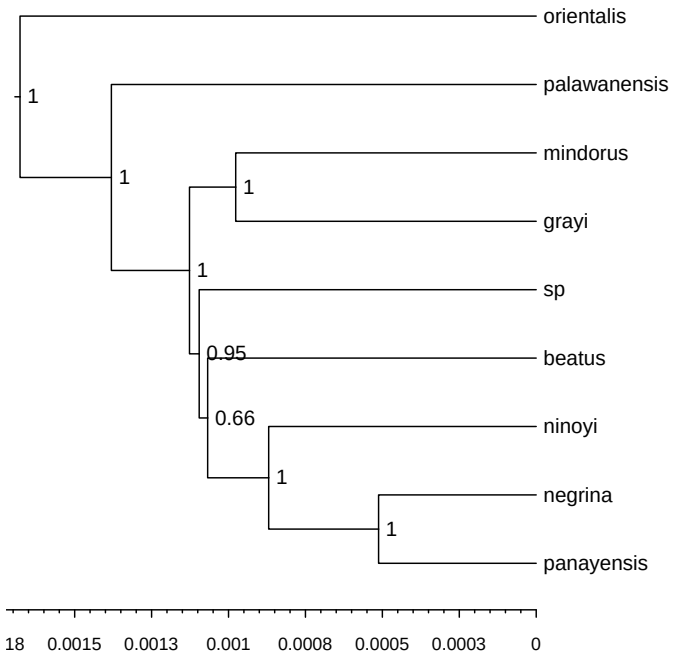

Figure 7: Species tree from the third 200M run with 25% burnin.

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, 2015a. doi: 10.1093/sysbio/syv029.

TC Giarla and JA Esselstyn. Data from: The challenges of resolving a rapid, recent radiation: empirical and simulated phylogenomics of Philippine shrews, 2015b. URL `http://dx.doi.org/10.5061/dryad.b7156`.

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.

Martyn Plummer, Nicky Best, Kate Cowles, and Karen Vines. CODA: Convergence diagnosis and output analysis for MCMC. *R News*, 6(1):7–11, 2006. URL `http://CRAN.R-project.org/doc/Rnews/`.

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