Species Trees and Species Delimitation with SNAPPER

Analyzing SNP or AFLP data

Adam Leaché, Remco Bouckaert

1 Background

This tutorial will help you become familiar with conducting species tree inference and species delimitation in a Bayesian framework using biallelic markers (AFLP or SNP data) using the package `SNAPPER` in BEAST2.

Coalescent methods for estimating species trees typically rely on estimating a gene tree for each locus. Combining hundreds or thousands of gene trees into a coalescent-based species tree inference framework presents some serious computational challenges. SNAPP is a method for estimating species trees that does not require gene trees Bryant et al. 2012, and SNAPPER Stoltz et al. 2021 is a fast and good approximation to SNAPP. This approach has been applied to the problem of species delimitation (Leaché et al. 2014). The species tree estimation method SNAPPER (Bryant et al. 2012; Stoltz et al. 2021) estimates species trees directly from biallelic markers (e.g., SNP or AFLP data), which bypasses the necessity of having to explicitly integrate or sample the gene trees at each locus. The method works by estimating the probability of allele frequency change across ancestor/descendent nodes. The result is a posterior distribution for the species tree, species divergence times, and effective population sizes, all obtained without the estimation of gene trees. The method works well for relatively small numbers of species (minimum = 2; maximum is probably near 20 due to computational constraints). Multiple alleles should be sampled for each species.

Species delimitation using genetic data and coalescent methods are increasing in popularity for good reasons (Fujita et al. 2012). Comparing candidate species delimitation models that contain different numbers of species, or different allocations of populations to species, is relatively easy in a Bayesian framework. The general approach is to estimate the marginal likelihood (Baele et al. 2012) of each competing species delimitation model, rank models by marginal likelihood, and use Bayes factors (Kass and Raftery 1995) to assess support for model rankings. This approach, called Bayes factor delimitation (BFD), was first implemented by Grummer et al. (2013) with DNA sequences in the program *BEAST. The approach was modified to work with SNP data (BFD*) (Leaché et al. 2014) using the SNAPPER model.

BFD* estimates the species tree and evaluates the species delimitation model at the same time, while allowing the user to compare models that contain different numbers of species and different assignments of samples to species. This is useful when the goal is to compare predefined species delimitation models or competing taxonomies. However, one drawback is that the user needs to predefine the number of species and sample assignments. This prevents the method from searching among all possible species assignments, an obvious disadvantage for studies aiming to discover cryptic diversity. Another major limitation is that the method does not explicitly consider gene flow, isolation by distance, selection, or several other important biological processes; however, these limitations are shared by many current methods. For example, failing to sample admixed populations often favours models containing more species, whereas including admixed populations will support more models containing fewer species. Distinguishing among these problematic scenarios requires paying close attention to both sample selection and prior settings. Finally, when evaluating results, remember to consider other aspects of the biology, ecology, and geography of "species" before jumping to conclusions.

2 Programs used in this Exercise

2.0.1 BEAST2 - Bayesian Evolutionary Analysis Sampling Trees 2

BEAST2 is a free software package for Bayesian evolutionary analysis of molecular sequences using MCMC and strictly oriented toward inference using rooted, time-measured phylogenetic trees (Bouckaert et al. 2014). This tutorial uses BEAST2 version 2.7.4.

2.0.2 BEAUti – Bayesian Evolutionary Analysis Utility

BEAUti is a utility program with a graphical user interface for creating BEAST2 input files, which are written in XML. The eXtensible Markup Language (XML) is a general-purpose markup language, which allows for the combination of text and additional information. The use of the XML makes analysis specification very flexible and readable by both the program and people. The XML file specifies all the components of the analysis, including sequences, node calibrations, models, priors, output file names.

2.0.3 TreeAnnotator

TreeAnnotator is used to summarize the posterior sample of trees to produce a maximum clade credibility tree and summarize the posterior estimates of other parameters that can be easily visualized on the tree (e.g. node height). This program is also useful for comparing a specific tree topology and branching times to the set of trees sampled in the MCMC analysis.

2.0.4 Tracer

Tracer is used for assessing and summarizing the posterior estimates of the various parameters sampled by the Markov Chain. This program can be used for visual inspection and assessment of convergence and it also calculates 95% credible intervals (which approximate the 95% highest posterior density intervals) and effective sample sizes (ESS) of parameters. Contrary to the other software in this section, Tracer is not distributed with BEAST2 and needs to be downloaded separately here.

3 Practical: species delimitation

3.1 Dataset

The dataset used in this tutorial contains SNP data for geckos in the *Hemidactylus fasciatus* species complex. Details on how the data were collected are provided in (Leaché et al. 2014). For this tutorial, we will use a data matrix containing 129 SNPs that is also available for download on Dryad. Allopatric divergence seems to be the primary mechanism causing speciation in this group. These geckos are restricted to rainforest habitats, and their distributions match those of the major blocks of rainforest in West and Central Africa (Figure 1).

For this species delimitation example, we will test models based on historical connections between adjacent rainforest blocks. These models differ in the number of species, and how samples are assigned to species. The base model has four species (Figure 1a). The alternative models are grouped into three classes: (1) lumping: populations are collapsed into the same species, (2) splitting: populations are partitioned into separate species, (3) reassigning: population(s) are allocated into a different species.

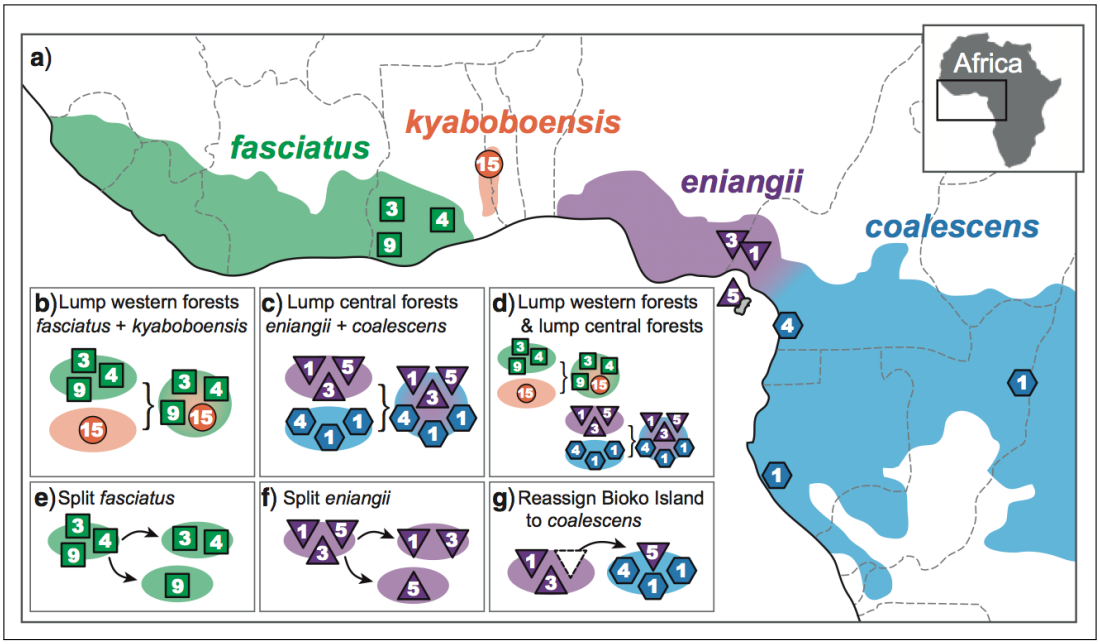

Figure 1: Geographic sampling of geckos (numbers in symbols indicate sample sizes). Starting taxonomy is shown in (a). BFD* is used to test the alternative species delimitation models outlined in (b) – (g) and (s) which combines (b) and (g).

The gecko SNP data is in binary format (necessary for SNAPPER). If you are unsure of how to convert your own SNP data from nucleotide to binary format, please read the documentation A rough guide to SNAPP (Section 4. Preparing Input File). You can find scripts for converting SNP data into SNAPPER input format at the phrynomics project site at GitHub. You can also find help at the BEAST2 google users group.

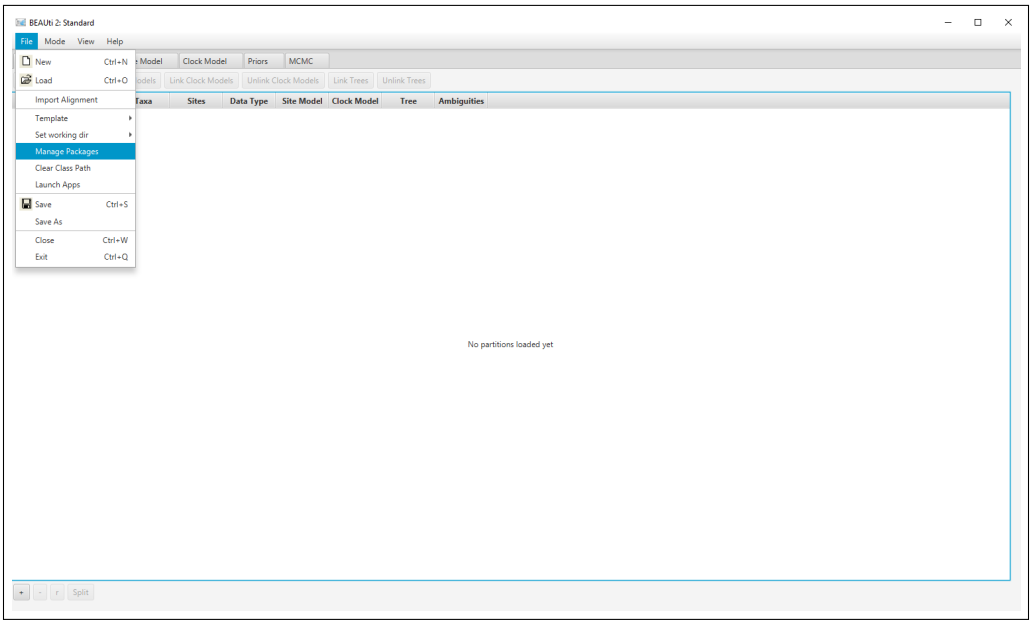

Figure 2: Finding the BEAST2 package manager.

3.2 Setting up the XML file

This section will demonstrate how to create an XML configuration file using BEAUti, which will then be used to run the analysis in BEAST2.

3.2.1 Package installation

We need to install additional packages to run this analysis.

Open the **BEAST2 Package Manager** by navigating to **File > Manage Packages**. (Figure 2)

Install the **SNAPPER** and **Model_Selection** packages by selecting them and clicking the **Install/Upgrade** button. (Figure 3)

Since **SNAPPER** depends on the **SNAPP** package and **Model_Selection** on the **BEASTLabs** package, these packages will be automatically installed as well. BEAUti needs to be restarted for the newly installed packages to be loaded properly.

Close the **BEAST2 Package Manager** and ***restart*** BEAUti to fully load the new packages.

3.2.2 Setting the template

We need to tell BEAUti that we are setting up a SNAPPER analysis, which will change the menu options and allow us to import SNP data.

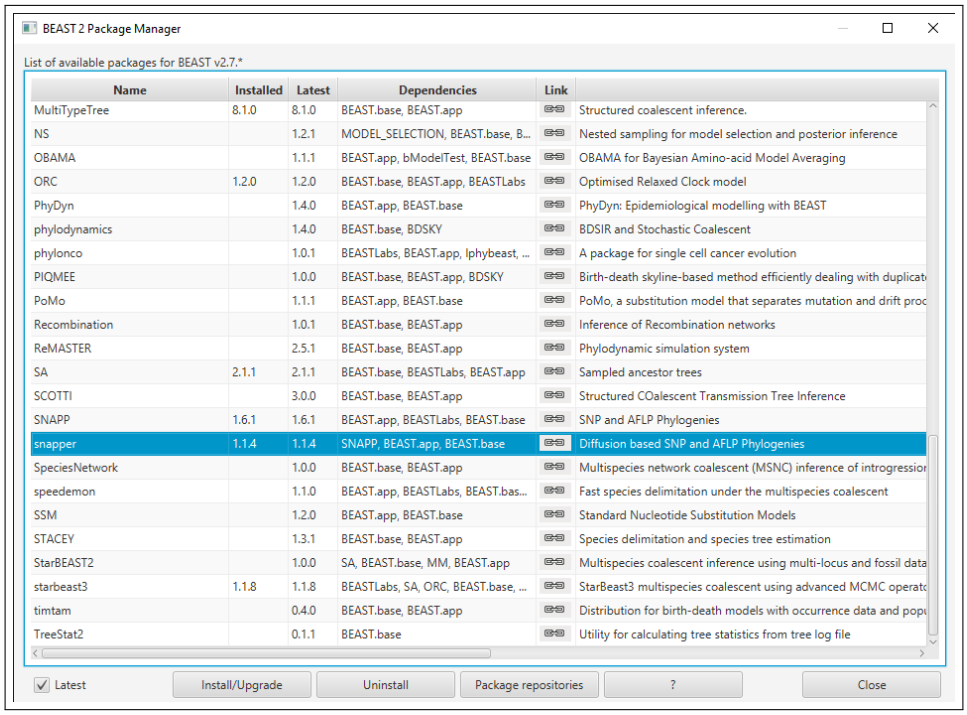

Figure 3: BEAUti package manager for BEAST2.

Select the template by using the drop-down menu **File > Template > SNAPPER**. This should change the appearance of the BEAUti window to look something like Figure 4.

3.2.3 Importing the SNP data

Import the SNP data (the **hemi129.nex** file) using the drop-down menu **File > Import Alignment**.[1]

Once the data are successfully loaded into BEAUti you should see a list of the samples included in the data file (Figure 5.)

3.2.4 Defining species

There are several ways to designate species assignments. You can automatically designate species names using the names already present in the data files. The species names can be pre-defined this way by including a "delimiter" that allows the species name to be parsed from the rest of the sequence name. The gecko data file uses an underscore "_" to separate the species name (on the left) from the rest of the sequence name (on the right) as follows:

```
eng_NG_1
coal_CA1_2
coal_CA1_3
coal_CA1_4
```

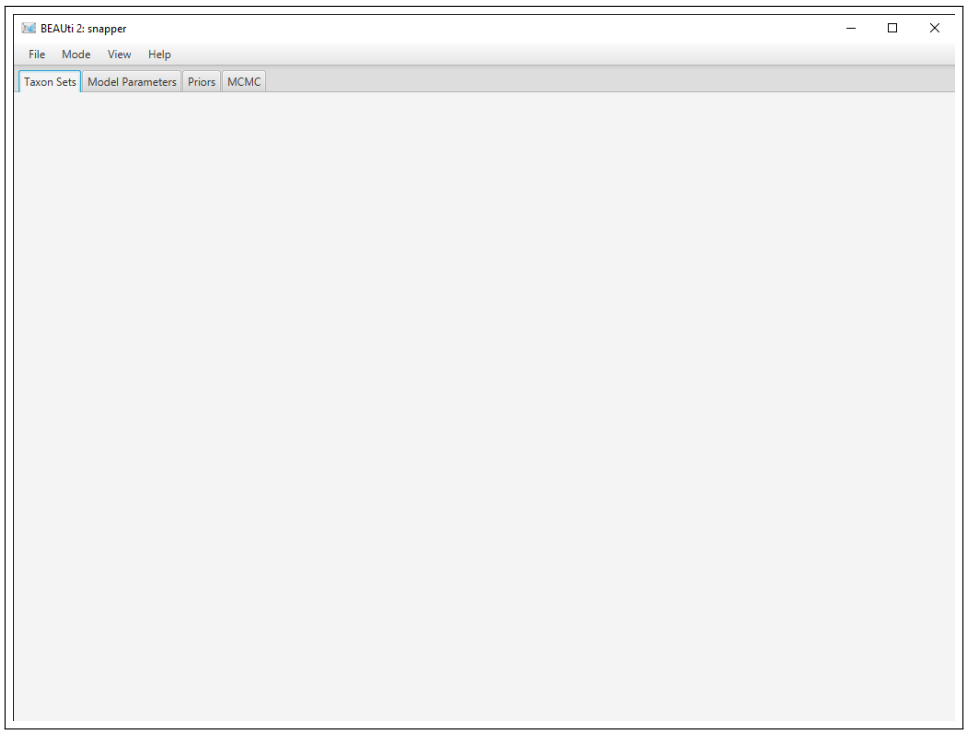

Figure 4: BEAUti window after importing the SNAPPER template. Notice that the menu tabs have changed.

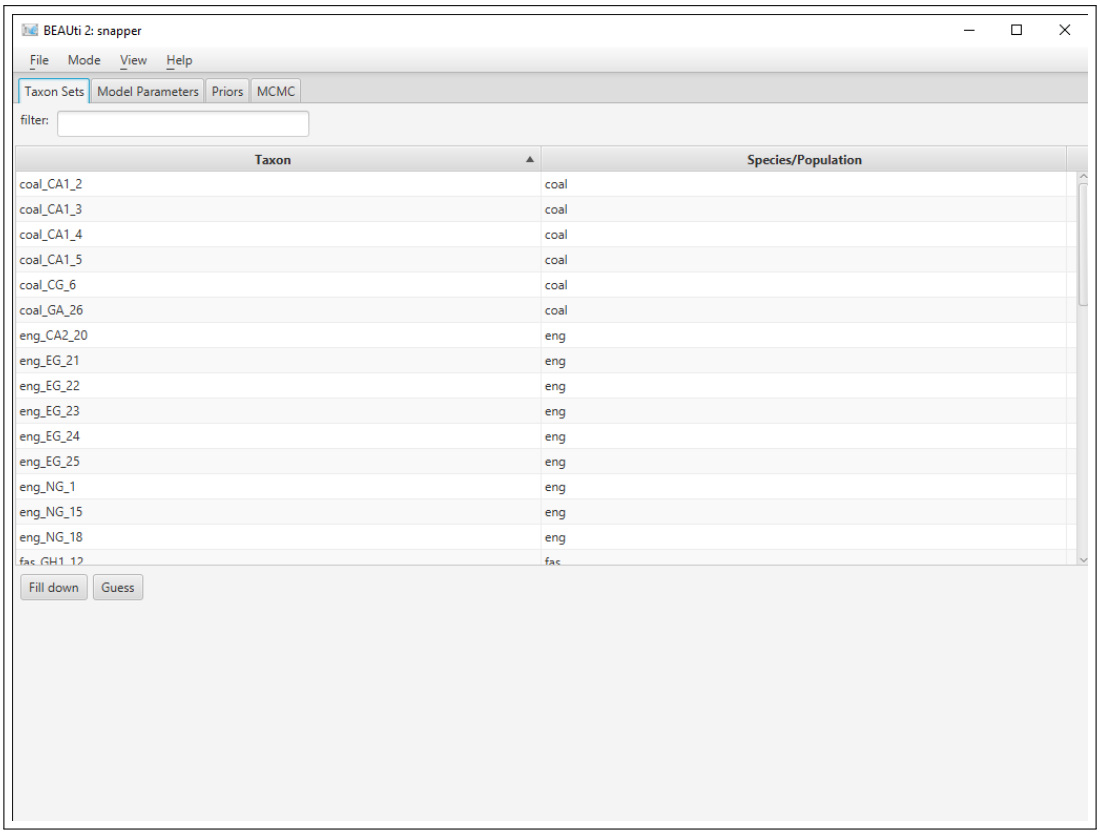

Figure 5: The data successfully loaded by BEAUti.

```
coal_CA1_5
coal_CG_6
kya_GH3_7
kya_GH3_8
...
```

Other options for assigning species names are available using the "Guess" button.

Open the guessing menu by clicking on **Guess** button. The screen should look similar to Figure 6.

How many samples should you include? More samples give more accurate estimates and has not computation cost for SNAPPER (unlike SNAPP, which is very sensitive to the number of samples), so the more samples the better. Keep in mind that the number of species slows down SNAPPER much more than the number of SNPs. If you have too many species and the analysis rusn unbearably slow, you might sub-sample the number of species.

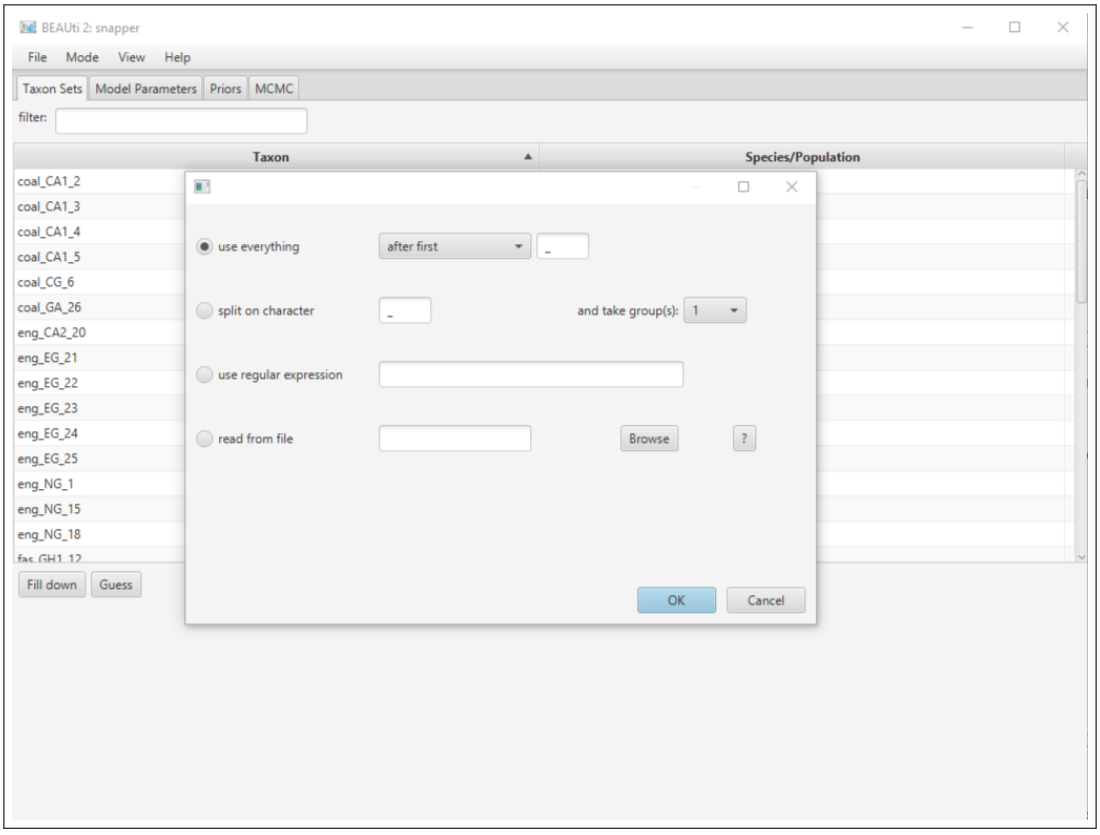

Figure 6: The species assignment options that appears after you select the "Guess' button.

You can import a custom mapping file that links each sample to a species using the "read from file" option.

Click the "Ok" button to return to the **Taxon Sets** window. Make sure that each Taxon has a Species/Population name.

3.2.5 Setting the model parameters

Next, we set up our model under the **Model Parameters** tab.

Be sure to read the documentation A rough guide to SNAPP to learn more about the model options[2]. Briefly, the main parameter of interest is the `Coalescent Rate` which determines the population size parameter with one value for each node in the tree.

Recommendations:

"Coalescent Rate": check the "estimate" box. If you do not sample, then you assume that all population sizes are the same, which is unrealistic. The coalescent rate is 2/theta, and the number is simply the starting value used to initialise the analysis.

"N" is the dimension of Chebyshef functions used for approximating the SNAPP tree likelihood. N should be power of 2 plus 1 (e.g. 33, 17, 65). Higher values are more accurate but slower.

The "Non-polymorphic" checkbox is used in cases where invariant sites have been included in the data. The likelihood calculations are different if SNAPPER assumes that all constant sites have been removed. If you are using a typical SNP dataset that only includes variable sites, then make sure that the box is not checked.

The "Mutation Only At Root" checkbox indicates a conditioning on zero mutations, except at root (default false). As a result, all gene trees will coalesce in the root only, and never in any of the branches. This option allows you to emulate the model used by Nielsen (1998) and RoyChoudhury et al. (2008).

The "Use Log Likelihood Correction" checkbox is for calculating corrected likelihood values for Bayes factor test of different species assignments (the calculation is almost instantaneous, and it will not slow down your analysis).

When the "Use beta root prior" checkbox is checked, instead of using a uniform prior for allele frequencies at the root, a beta root prior is used.

Following these recommendations, the only change we need to make here is to tell BEAUti that our alignment is made of polymorphic sites only.

> Uncheck the **Non-polymorphic** checkbox.

The final setup will appear as in Figure 7.

3.2.6 Defining the priors

Next, we move to the **Prior** tab and specify the priors. Again, read the documentation A rough guide to SNAPP to learn more about these priors. It is important to be aware of the biological meaning of these priors. One problem with SNAPPER (and BEAST2 in general) is that it is deceptively easy to set up an analysis using default options.

[2]For SNAPP, the mutation rates U and V representing instantaneous rate of mutating from the 0 allele to the 1 allele and from 1 to 0 respectively are included in the model. However, keeping U and V equal to 1 seems more appropriate, so these parameters are hidden in BEAUti for the SNAPPER model.

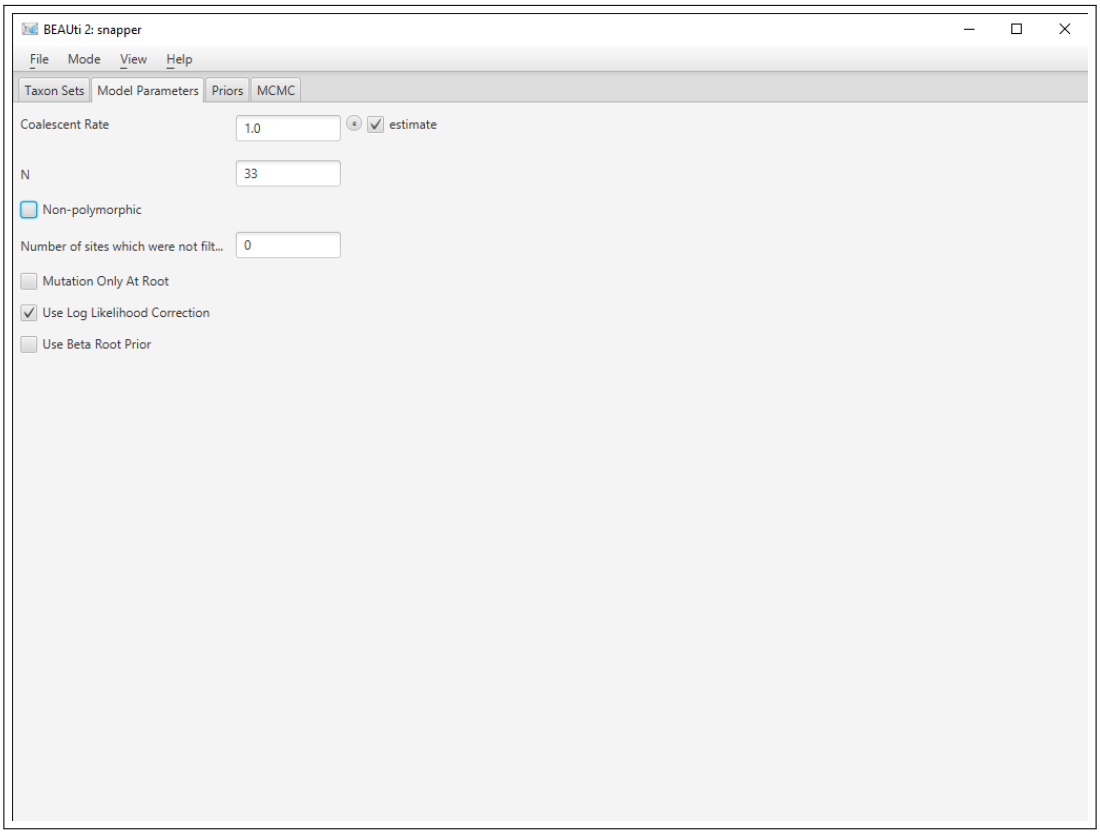

Figure 7: The SNAPPER model options.

The default priors are

- for the tree, a Yule prior representing a pure birth process with birth rate parameter λ.
- a **terrible** uniform prior with range 0 to $+\infty$ for the birth rate λ.
- a gamma prior with shape=2, scale=2 for the coalescent rates.

SNAPPER by default uses a Yule prior for the species tree and branch lengths on the species tree. This prior has a single parameter, λ (Lambda), representing the speciation rate. This rate, in turn, determines the (prior) expected height of the species tree. The higher the speciation rate, the higher the probability of supporting more species during species delimitation.

The default prior on λ can seem non-informative, allowing the data to drive the analysis instead of prior information. However, it is an improper prior. A proper prior is a prior that when integrating over its range sums to 1. Since the uniform prior has infinite bounds, the integral over the range does not do that. The prior can be made proper by defining finite upper and lower bounds.[3] For estimating the marginal likelihood using path sampling, "proper" priors are required. In addition this prior puts too much weight on very high and unrealistic values of the birth rate.

So do we set a prior on the birth rate? The trees that come out of SNAPPER are not time calibrated, thus the branch lengths should be interpreted in units of expected substitutions per site. Translating this into time requires an assumption about the substitution rate. See the Rough Guide to SNAPP for further

[3]The 1/X prior is another prior that is not proper and therefore not recommended.

details and an equation for calculating Lambda. Using a useful script written by Jamie Oaks, called pyule, you can calculate an appropriate lambda value based on prior information about the number of species, and the height of the species tree in expected substitutions per site. If sequence data are available, then you can approximate the height of the species tree obtained by calculating the maximum observed divergence between any pair of taxa divided by two (height = max divergence/2).

Table 1 provides some suggestions for good Lambda values across a range of expected substitutions per site from root to tip (up to 20% sequence divergence from root to any tip), and for relatively small species trees (up to 20 species). In general, different Lambda settings do not influence the results too much, since in most cases the data will dominate.

		Species tree height (expected substitutions per site)					
		0.001 (0.1%)	0.005 (0.5%)	0.01 (1%)	0.05 (5%)	0.10 (10%)	0.2 (20%)
Number of species	2	500	100	50	10	5	0.2
	5	1283	257	128	26	13	6
	10	1928	386	193	39	19	10
	15	2318	464	232	46	23	12
	20	2598	520	260	52	26	13

Table 1: Lambda prior settings across a range of species tree heights (in expected substitutions per site) and sizes (number of species).

In this case, we are going to set a gamma prior with a broad distribution, for example, alpha = 2 and beta = 200. Note that the mean of the gamma distribution for Lambda is calculated as alpha*beta when mode is 'ShapeScale' (rather than alpha/beta when mode='ShapeRate'). If there is prior information about the number of expected substitutions, then this can be used to set up the prior. In general, you should try setting the mean of the gamma distribution to be somewhere in the middle of the extreme lambda values (shown in Table 1) that apply to your study system.

In the **Priors** tab, open the dropdown menu next to the **birthRate** parameter, and change the distribution type from **Uniform** to **Gamma**. Click on the arrow next to **birthRate** to open the prior details and check that **alpha** and **beta** are set to **2.0**. The prior set up should look like Figure 8.

Figure 9 shows a comparison of two different options for setting lambda: using a fixed value (lambda = 5) versus assigning a broad gamma prior as we did. In this case, tree height estimates are similar using either a fixed lambda or a gamma distribution.

Setting the expected divergence (theta) prior. Recall that for a diploid population, theta = 4Nu, where N is the effective population size and u is the per-generation mutation rate. If theta=0.004, you expect to observe 0.4% variation between two randomly sampled alleles in a population. Another way to think about this is in the expected number of substitutions; "theta=0.004" means that for two randomly sampled individuals within a population you expect to observe 4 SNPs in 1,000 bases. In SNAPPER, the gamma prior on theta is parameterised such that the mean is alpha*beta. For example, if alpha=2 and beta = 0.002, the prior mean on theta is 0.004. One way of estimating a reasonable mean for the theta prior is to calculate pairwise sequence divergence among all individuals known to belong to a single species and take the average value as the mean for theta. Higher prior means on theta will favour models with populations grouped together, because they represent an expectation that average divergence within populations is

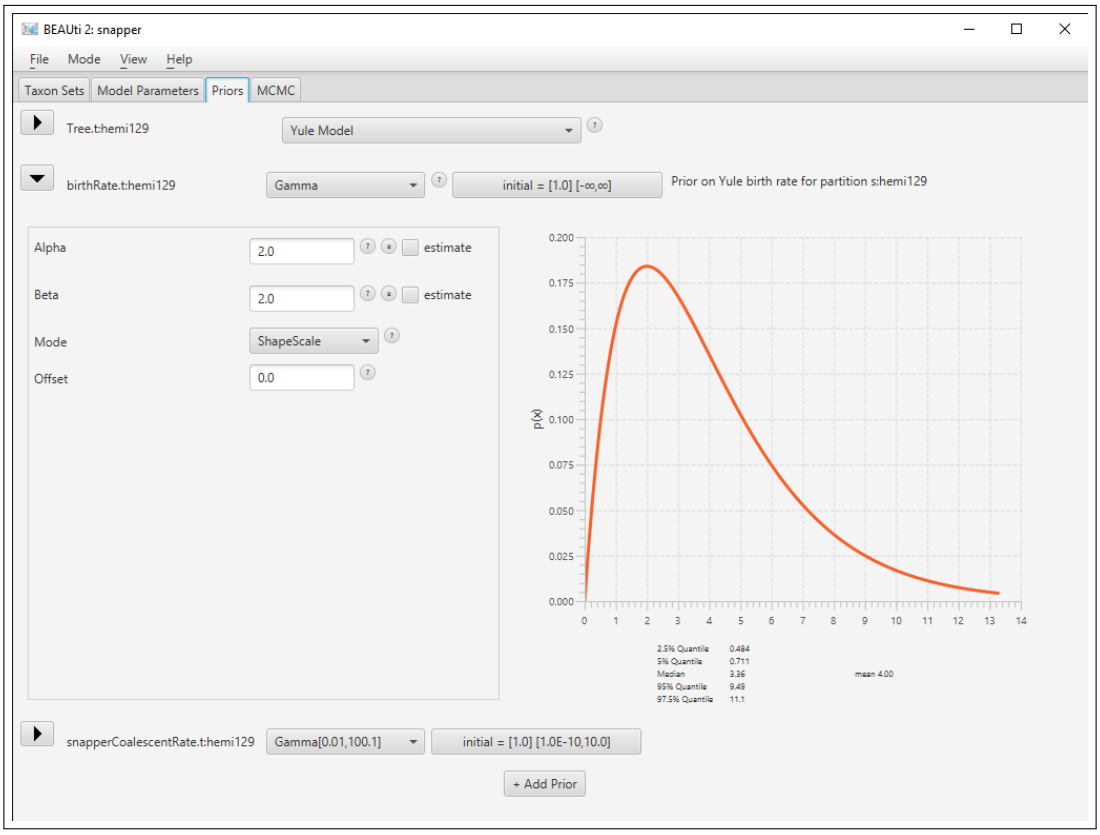

Figure 8: The prior settings: using a gamma prior for the birth rate.

relatively high. You can select different distributions for the theta prior in the **Priors** panel. If the prior is set to "Gamma" (recommended), the mean value of theta is calculated as alpha*beta when using the ShapeScale mode. Note that in some implementations, the mean of a gamma distribution is calculated as alpha/beta rather than alpha*beta. Here we will leave the default prior.

3.2.7 Specify MCMC settings and generate the XML file.

The last step is to configure some run and logging options.

Next, move to the **MCMC** tab. Change the following settings:

- `Chain Length:` **1000**
- `Store Every:` **10**
- `tracelog:Log Every:` **10**
- `treelog:Log Every:` **10**

We leave all the remaining options at their default values (Figure 10).

Note that these MCMC values are way too low, and a thorough analysis requires much more computational time. The MCMC run times are intentionally kept short in this tutorial. These short analyses should run in approximately 2 – 4 minutes depending on the number of processors available on your computer. Thorough

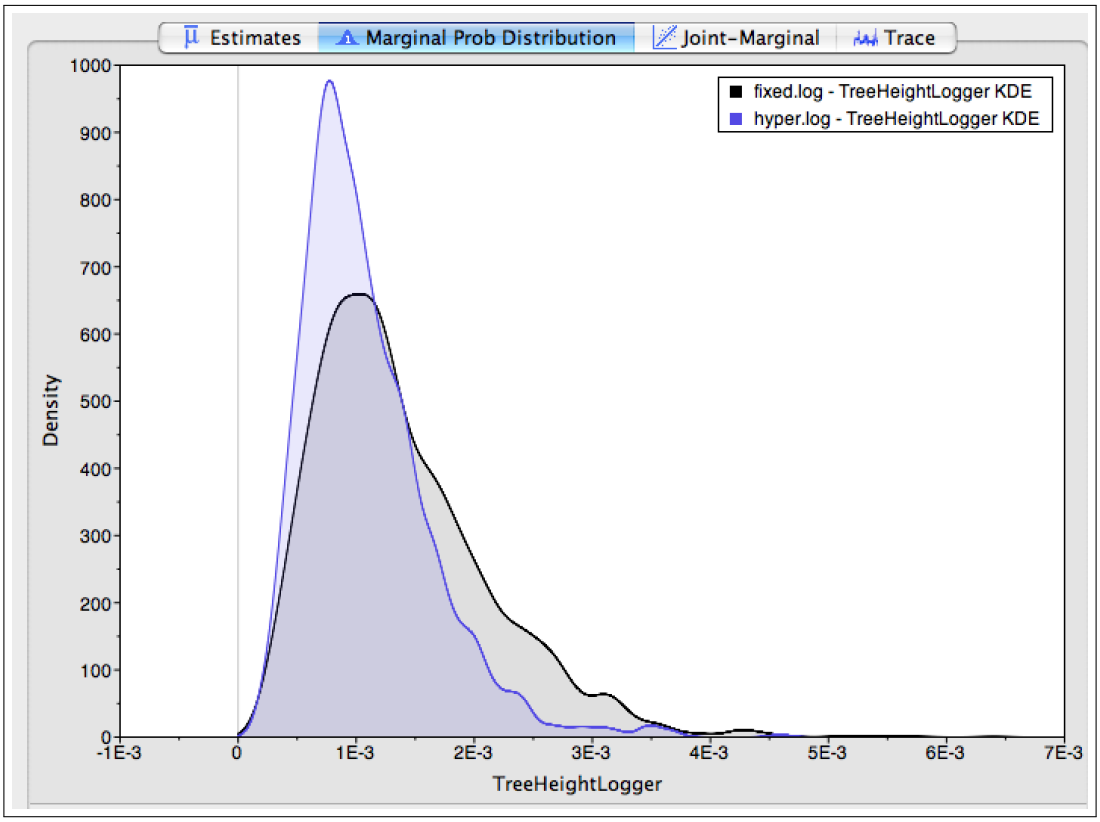

Figure 9: Tree height comparison of analyses using a fixed lambda (lambda = 5) versus a hyperprior on lambda (gamma distribution with mean = 4). The plot shows the marginal probability distributions superimposed for the two estimates, which overlap broadly suggesting that the estimates are quite similar.

analyses of the full data take 2 – 6 days, depending on the number of species in the model, and generally require at least 100,000 generations. Running multiple independent analyses using different starting seeds and comparing results is a good way to ensure that the analyses are converging.

Next, save the file using **File > Save....** Another subwindow will appear for specifying the name and location for saving the XML file. . Name the file `runA.xml` and save it to the `SNAPPER-delimitation-tutorial` folder.

3.3 Running the analysis in BEAST2

You can execute the XML file in BEAST2 using the GUI or the command line. If you are using Mac OSX or Windows, you should be able to launch the BEAST2 GUI by double clicking on the application icon.

In the BEAST2 window, click the **Choose File...** button, and select the XML file you just created (Figure 11). Increase the **Thread pool size** to speed up your analysis. Click the **Run** button.

Running SNAPPER with multiple threads can increase the speed, but experimenting with the number of

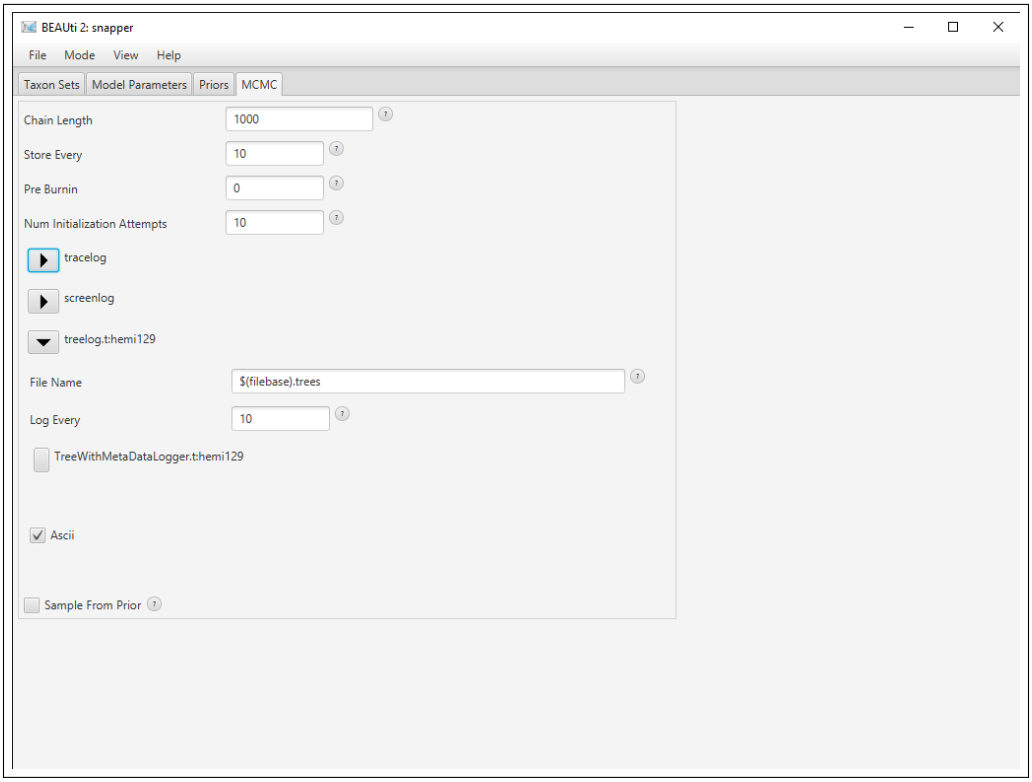

Figure 10: The MCMC settings.

threads is required to get the best performance. The analysis should take about 10 minutes.

You can also run BEAST2 from the command line. Open the **Terminal** Application and navigate to the folder containing your `runA.xml` file. To execute the file, type the following at the command line:

```
/path/to/BEAST2/bin/beast -threads 8 runA.xml
```

or

```
beast -threads 8 runA.xml
```

if you have already moved the BEAST2 executable to your path. Caution: setting the number of `threads` beyond the maximum number available on your computer can have serious drawbacks, and you will probably not have enough memory to support all of those separate analyses (i.e. your computer might slow down badly or crash if you use too many threads).

3.4 Analyzing the output

3.4.1 Output files

Our run has generated 2 different files:

- `RunA.log` which is the general trace log and stores all parameter values.

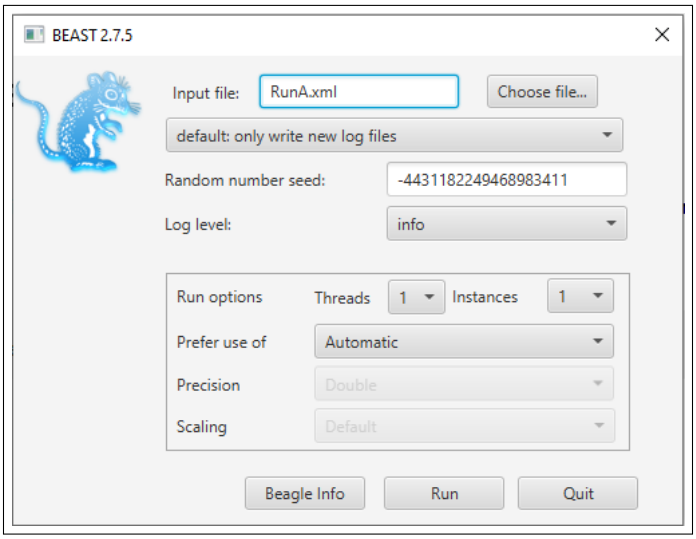

Figure 11: The BEAST2 GUI window.

- `RunA.trees` which recorded the sampled trees in Nexus format.

By loading the log file into Tracer, we can check that as we expected, we are very far from the convergence of the chain. You can also find two sets of prepared log files in the tutorial, one run with the MCMC parameters set as above (`short`) and one with a longer chain length `long`.

3.4.2 Summarising the species tree using TreeAnnotator.

TreeAnnotator will summarize the posterior distribution of species trees and identify the topology with the best posterior support, and summarise the divergence times for each node in the tree.

Launch the **TreeAnnotator** program. Set the **burnin** value to **10%**. For the **Target tree type** field, choose **Maximum clade credibility tree**. For the **Node heights** field, choose **Median heights**. Select the **Input Tree File** button and select the file `runA.trees`. Select the **Output File** button and specify the `output` directory and a file name, `runA-MCC.tre`. Click **Run**

3.4.3 Visualising the species tree in FigTree.

We can look at our summary species tree in FigTree.

Launch the FigTree program, and load the `runA-MCC.tre` file you just created with TreeAnnotator. Check the **Branch Labels** option and select **posterior** for the **Branch labels**Display fields. Check the **Node Bars** option and select **height_95%_HPD** for the **Node bars**Display field.

You can also get a summary of some tree statistics using the TreeSetAnalyser application, which you can launch from the BEAST app launcher.

First, start the BEAST app-store by selecting the **File > Launch apps** menu in BEAUti (alternatively, double click the **AppLauncher** program in the BEAST folder). A window similar to Figure 12 should pop up. Select the **SNAPP tree set analyzer** app, and hit the **Launch** button.

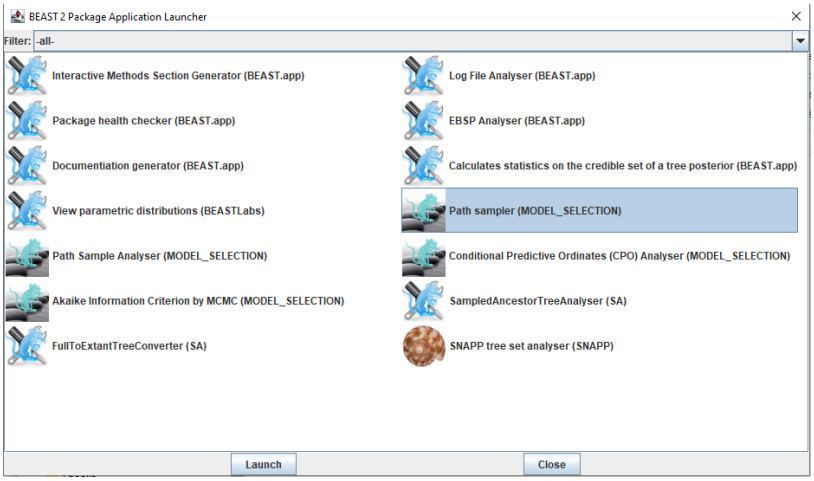

Figure 12: BEAST app launcher.

A window similar to Figure 13 will open.

Figure 13: Tree set analyser utility.

Select the **Browse** button next to the **Trees** input and select the file `runA.trees`. Click on **OK** to process the file.

TreeSetAnalyser will compute a summary of the posterior species tree distribution. It will print how many distinct topologies are present in the full distribution and in the 95% HPD interval, as well as a table of all topologies with their frequency in the posterior distribution, ranked by highest frequency.

3.5 Running the path sampling analysis with BEAST2

Species delimitation using SNPs requires marginal likelihood estimation. There are two ways to set up this analysis; through a GUI, and by editing the XML. The GUI is more convenient but makes it a bit harder to transfer the analysis to a cluster, and since path sampling analyses are typically very computational intensive, it often makes sense to run them on a cluster.

3.5.1 Setting up for marginal likelihood estimation (GUI)

In this subsubsection, we explain how to set up the analysis using the GUI.

As before, start the BEAST app-store by selecting the **File > Launch apps** menu in BEAUti (alternatively, double click the **AppLauncher** program in the BEAST folder). A window similar to Figure 12 should pop up. Select the **Path sampler** app, and hit the **Launch** button.

A new window opens with the GUI for path sampling/stepping stone analysis, similar to Figure 14.

If you prefer to start the app from the command line, you can use the following in a terminal:

```
/path/to/BEAST2/bin/applauncher PathSampler
```

Select the file `RunA.xml` you just set up in BEAUti, and change the following settings:

- `alpha:` **0.3**
- `chainLength:` **1000**
- `burnInPercentage:` **0**
- `preBurnin:` **0**

The **rootdir** option indicates where the run and output files will be placed. Change it to the tutorial folder. Note that checking the **Delete Old Logs** will delete any previous log files in the root directory, so check it only if you are overwriting an old analysis.

The result should look as indicated in Figure 14.

Click on **OK** to start the analysis.

The analysis will take a few minutes (it may look like nothing is happening at the beginning). The output will be printed to the application window.

3.5.2 Editing the XML file for marginal likelihood estimation.

You can also edit the XML file directly to prepare it for path sampling analysis. Detailed instructions for setting up marginal likelihood estimation using path sampling are provided at the BEAST2 website. The procedure involves (1) typing in some short codes in a few places, (2) replacing some words, and (3) copying and pasting some sections around. Specific instructions are below:

Open your XML file in a text editor. Search and replace the opening run statement (located about half way through the file) with an mcmc statement by changing **“<run ...>”** into **“<mcmc ...>”.** Next, type a new closing mcmc statement, **“</mcmc>”**, just before the closing run statement, **“</run>”**, located at the end of the file.

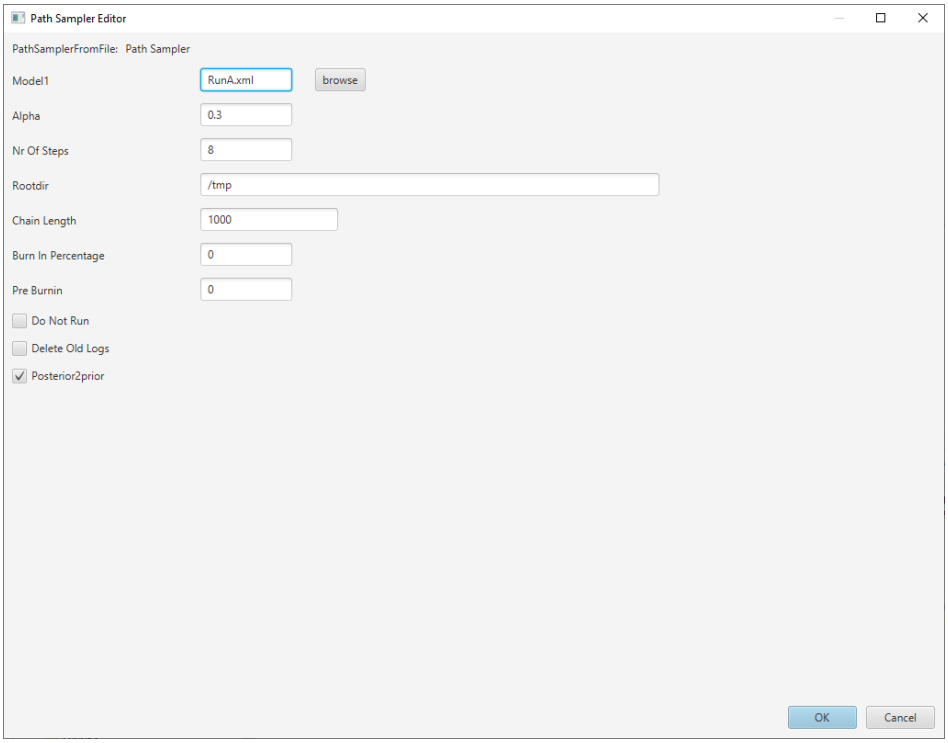

Figure 14: GUI for path sampling/stepping stone analysis.

Now you are ready to insert the path sampling commands. You will need to insert the following block of text into your XML file immediately above the opening "<mcmc ...>"' element:

```
<run spec='beast.inference.PathSampler' chainLength="1000" alpha='0.3'
rootdir='/your/path/goes/here/' burnInPercentage='0' preBurnin="0"
deleteOldLogs='true' nrOfsubsubsections='24'>
```

You can then save the XML, then run the analysis by using the following command in the folder where you saved the file.

```
/path/to/BEAST2/bin/beast path_sampling_runA.xml
```

Important: If you copy and paste this section into your XML file, be sure to check that the symbols paste correctly. The quote symbols (" ' ' ") don't copy as they should, and these will cause problems. Also, make sure that the root directory path (rootdir) exists on your computer.

3.5.3 Parameters of the analysis

The path sampling parameters that you just entered (either through the GUI or in the XML) are as follows:

- `chainLength:` MCMC sample length for each path sampling subsubsection.
- `alpha:` parameter used to space out path sampling subsubsections.
- `rootdir:` directory for storing output. Be sure that the folder exists before starting the run.
- `burnInPercentage:` burn-In percentage used for analyzing the log files.

- `preBurnin:` number of samples that are discarded for the first subsubsection, but not the others.
- `deleteOldLogs:` delete existing log files from rootdir
- `nrOfsubsubsections:` the number of path sampling subsubsections to use

Note that these path sampling parameters are way to low, and a thorough analysis requires much more computational time. Stable marginal likelihood estimates usually require at least 48 subsubsections (sometimes 100), chainLength = 100,000 (sometimes 1,000,000), and preBurnin=10,000 (sometimes 100,000). The MCMC run times are intentionally kept short in this tutorial to obtain quick (but meaningless) results. The run time on a MacBook Pro 2.3GHz i7 processor with 16GB of memory is approximately 2.5 minutes, and this is running 8 concurrent subsubsections (= 8 threads). Increasing the number of threads will speed up the analysis by running more concurrent path sampling subsubsections, but this requires more memory (this analysis uses about 12GB of memory). For large-scale analyses, many users find that they run out of memory before processors.

3.6 Inspecting path sampling results

3.6.1 Inspecting path sampling results.

At the end of your analysis, the path sampling results will be displayed on the screen. An example (from the GUI) is shown in Figure 15. Each row shows the results from one path sampling subsubsection. The example in Figure 15 shows the results from a path sampling analysis with 24 subsubsections. You will use the value after "marginal L estimate" to compare models.

Step	theta	likelihood	contribution	ESS
0	0	-594.4583	-0.0172	8.8139
1	0	-598.2799	-0.157	26.9
2	0.0003	-599.3535	-0.4999	15.2249
3	0.0011	-596.7185	-1.0802	17.8366
4	0.0029	-603.1037	-1.9539	5.8638
5	0.0062	-599.9736	-3.0981	26.1718
6	0.0113	-596.9401	-4.5462	21.1964
7	0.019	-595.7229	-6.3254	13.9231
8	0.0296	-591.576	-8.395	17.6153
9	0.0438	-594.7522	-10.9441	10.7102
10	0.0623	-589.7796	-13.701	20.4862
11	0.0855	-590.4063	-16.9617	22.0646
12	0.1143	-585.6204	-20.423	12.0898
13	0.1493	-581.9506	-24.2965	17.2323
14	0.1911	-578.3612	-28.5113	24.2682
15	0.2406	-573.7403	-33.0335	17.1482
16	0.2983	-573.5769	-38.2293	9.1175
17	0.3651	-537.003	-31.5634	4.2922
18	0.4417	-365.6533	-31.7828	17.6132
19	0.529	-364.111	-35.754	13.381
20	0.6276	-360.7822	-39.867	19.5922
21	0.7384	-359.2399	-44.4018	20.8387
22	0.8623	-356.4251	-49.0131	31.1829
23	1	-355.1386	0	31.2236

marginal L estimate =-444.5553441185161

Figure 15: The path sampling output at the end of the analysis.

Note that when conducting path sampling, SNAPPER generates a posterior distribution of trees for each path sampling step. This means we could have skipped the step earlier where we ran `runA.xml` like a regular analysis, and used the output of the path sampling analysis directly instead. You will typically only want to summarize the tree file in the folder corresponding to theta=1, since the other folders contain trees that were estimated with modified likelihoods. Check the SNAPPER screen output to verify which path sampling subsubsection corresponds to theta=1 (this changes with different versions of SNAPPER).

3.7 Comparing with other species delimitation models

3.7.1 Setting up new XML files for species delimitation.

Now that you have one XML file up and running it is easy to make new XML files for each species delimitation model.

To prepare a new file for species delimitation, make a few slight modifications to the existing `runA.xml` file:

1. save a copy of the xml file as `runB.xml`,
2. change the file logging names in the xml file to **runB.log** and **runB.trees** so that you don't accidentally overwrite any of your previous results,
3. change the species assignments listed in the **"stateDistribution"** element. This last part requires changing the number and/or composition of taxonset features. Each taxonset begins with **"<taxonset ...>"** and ends with **"</taxonset>"** (Figure 16). To lump species, simply combine the taxon names into a single taxonset feature. To split a species, simple create a new taxonset containing the appropriate taxon names. To reassign a taxon to another species you can cut and paste the taxon to a different taxonset.
4. (if you are running directly from the XML instead of using the GUI) edit the path sampling root directory so the output is not overwritten.

XML files containing the species assignments shown in Figure 1 are provided with this tutorial (in the `xml` folder). You don't need to run them, but you can open them in a text editor to see what changes have been made to the species assignments.

```
<taxa dataType="integerdata" id="snap.hemi129" spec="snap.Data">
    <data idref="hemi129" name="rawdata"/>
    <taxonset id="kya" spec="TaxonSet">
        <taxon id="kya_GH3_7" spec="Taxon"/>
        <taxon id="kya_GH3_8" spec="Taxon"/>
        <taxon id="kya_GH3_9" spec="Taxon"/>
    </taxonset>
    <taxonset id="fas" spec="TaxonSet">
        <taxon id="fas_GH2_10" spec="Taxon"/>
        <taxon id="fas_GH2_11" spec="Taxon"/>
        <taxon id="fas_GH1_12" spec="Taxon"/>
        <taxon id="fas_GH4_38" spec="Taxon"/>
        <taxon id="fas_GH4_39" spec="Taxon"/>
        <taxon id="fas_GH4_40" spec="Taxon"/>
    </taxonset>
    <taxonset id="coal" spec="TaxonSet">
        <taxon id="coal_CA1_5" spec="Taxon"/>
        <taxon id="coal_CG_6" spec="Taxon"/>
        <taxon id="coal_GA_26" spec="Taxon"/>
    </taxonset>
    <taxonset id="eng" spec="TaxonSet">
        <taxon id="eng_NG_18" spec="Taxon"/>
        <taxon id="eng_CA2_20" spec="Taxon"/>
        <taxon id="eng_EG_21" spec="Taxon"/>
        <taxon id="eng_EG_22" spec="Taxon"/>
    </taxonset>
</taxa>
```

Figure 16: Example of the taxonset features in the XML file (reduced number of samples to fit on screen, the ready to run XML files for this tutorial contain more samples).

3.7.2 Comparing species delimitation models with Bayes factors.

After you run each of the alternative species delimitation models you can rank them by their marginal likelihood estimate (MLE). You can also calculate Bayes factors to compare the models. The Bayes

factor (BF) is a model selection tool that is simple and well suited for the purposes of comparing species delimitation models. Calculating the BF between models is simple. To do so, simply subtract the MLE values for two models, and then multiply the difference by two (BF = 2 x (model1 - model2). A positive BF value indicates support in favour of model 1. A negative BF value indicates support in favour of model 2.

The strength of support from BF comparisons of competing models can be evaluated using the framework of Kass and Raftery 1995. The BF scale is as follows: $0 < BF < 2$ is not worth more than a bare mention, $2 < BF < 6$ is positive evidence, $6 < BF < 10$ is strong support, and $BF > 10$ is decisive.

The results for the seven gecko models are provided in Table 2. The model that lumps the western forests into one species and splits *eniangii* into two (runS) is the top-ranked model. It has the largest MLE value, and it is supported in favor of the current taxonomy model (runA). The BF in support for model S is decisive compared to model A. It is important to emphasise that these results are *tragically deficient* in terms of the MCMC analysis. Much, much longer runs are required to obtain correct results.

Model	Species	MLE	BF	Rank
runA current taxonomy	4	-2517.7	-	2
runB lump western forests	3	-2527.1	19.4	5
runC lump central forests	3	-2535.1	35.4	6
runD lump western & central forest	2	-2544.2	53.96	7
runE split *fasciatus*	5	-2525.4	16.2	4
runF split *eniangii*	5	-2518.3	2	3
runG reassign Bioko Island	4	-3348.5	1662.2	8
runS lump western and split *eniangii*	4	-2170.4	-693.76	1

Table 2: Path sampling results for the seven species delimitation models shown in Figure 1. MLE = Marginal likelihood estimate, BF = Bayes factor. All BF calculations are made against the current taxonomy model (runA). Therefore, positive BF values indicate support for the current taxonomy model, and negative BF values indicate support for the alternative model. .

Note that as we have seen in this tutorial, Bayesian factor delimitation is computationally expensive, and relies on the user to specify species assignments to test. A recent package **speedemon** proposes a faster method for species delimitation: find more information and a tutorial on the package website.

4 Useful Links

- Bayesian Evolutionary Analysis with BEAST 2 (Drummond and Bouckaert 2014)
- BEAST 2 website and documentation: `http://www.beast2.org/`
- BEAST 1 website and documentation: `http://beast.bio.ed.ac.uk`
- Join the BEAST user discussion: `http://groups.google.com/group/beast-users`

Version dated: June 9, 2024

Relevant References

Baele, G, P Lemey, T Bedford, A Rambaut, MA Suchard, and AV Alekseyenko. 2012. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Molecular Biology and Evolution* 29: 2157–2167.

Bouckaert, R, J Heled, D Kühnert, T Vaughan, CH Wu, D Xie, MA Suchard, A Rambaut, and AJ Drummond. 2014. Beast 2: a software platform for bayesian evolutionary analysis. *PLoS computational biology* 10: e1003537.

Bryant, D, R Bouckaert, J Felsenstein, NA Rosenberg, and A RoyChoudhury. 2012. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Molecular Biology and Evolution* 29: 1917–1932.

Drummond, AJ and RR Bouckaert. 2014. *Bayesian evolutionary analysis with BEAST 2*. Cambridge University Press,

Fujita, MK, AD Leache, FT Burbrink, JA McGuire, and C Moritz. 2012. Coalescent-based species delimitation in an integrative taxonomy. *Trends in Ecology & Evolution* 27: 480–488.

Grummer, JA, RW Bryson, and TW Reeder. 2013. Species delimitation using Bayes factors: simulations and application to the *Sceloporus scalaris* species group (Squamata: Phrynosomatidae). *Systematic Biology* 63: 119–133.

Kass, RE and AE Raftery. 1995. Bayes factors. *Journal of the American Statistical Association* 90: 773–795.

Leaché, AD, MK Fujita, VN Minin, and R Bouckaert. 2014. Species delimitation using genome-wide SNP data. *Systematic Biology* 63: 534–542.

Nielsen, R. 1998. Maximum likelihood estimation of population divergence times and population phylogenies under the infinite sites model. *Theoretical Population Biology* 53: 143–151.

RoyChoudhury, A, J Felsenstein, and EA Thompson. 2008. A two-stage pruning algorithm for likelihood computation for a population tree. *Genetics* 180: 1095–1105.

Stoltz, M, B Baeumer, R Bouckaert, C Fox, G Hiscott, and D Bryant. 2021. Bayesian inference of species trees using diffusion models. *Systematic Biology* 70: 145–161.