Bayes Factor Delimitation of Species (*with genomic data; BFD*): A Tutorial and Worked Example

Adam Leaché
Department of Biology
University of Washington
April 2015

Huw A. Ogilvie
Research School of Biology
Australian National University
January 2016

Contents

1 Objective

This tutorial will help you become familiar with conducting species delimitation in a Bayesian framework using biallelic markers (AFLP or SNP data) using the programs SNAPP and BEAST. We will use example SNP data for geckos (genus *Hemidactylus*) and the software package BEAST version 2 (Bouckaert et al., 2014). We will work through the steps necessary for setting up the required packages on your computer, setting up the XML file, and testing species delimitation models using marginal likelihood estimation and Bayes factors.

2 Version, Author information, and Acknowledgements

This tutorial was written by Adam Leaché for BEAST version 2.1.2, then updated by Huw Ogilvie for BEAST version 2.3.3. Remco Bouckaert helped troubleshoot the tutorial. The layout of the tutorial is a modified version of a divergence time tutorial written by Jamie Oaks (`https://github.com/joaks1`), which he borrowed from Tracy Heath. A similar tutorial is provided at the BEAST website.

3 Background Information

New coalescent-based species delimitation methods are increasing the statistical rigor and objectivity of taxonomy (Fujita et al., 2012). However, expanding these methods to a genome-scale is limited by their reliance on gene trees. Combining hundreds or thousands of gene trees into a single species delimitation framework presents some serious computational challenges. A new method for estimating species trees without gene trees is available (Bryant et al., 2012), and we have leveraged this approach for species delimitation (Leaché et al., 2014). The species tree estimation method `SNAPP` (Bryant et al., 2012) 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.

Comparisons among 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 requires marginal likelihood estimation (MLE) for each competing species delimitation model. Several different MLE approaches are available in BEAST, including path sampling (PS) or stepping-stone (SS) methods (Baele et al., 2012). Once the MLE values are obtained, the models can be ranked from highest to lowest, and Bayes factors (Kass and Raftery, 1995) can be used to compare models. This approach, called Bayes factor delimitation (BFD), was first implemented by Grummer et al. (2014) with DNA sequences in the program *BEAST. The approach was modified to work with genome-wide SNP data (BFD*) using the program SNAPP (Leaché et al., 2014).

One advantage of BFD/BFD* over other species delimitation approaches is the ability to integrate over species trees during the species delimitation procedure, which removes the constraint of specifying a guide tree that represents the true species relationships. In other words, with BFD/BFD* you can estimate the species tree and evaluate the species delimitation model at the same time. Another advantage is the ability to compare models that contain different numbers of species, or different assignments of samples to species. However, the user needs to predefine the number of species and sample assignments, and this prevents the method from searching among all possible species assignments.

4 Programs Used in This Lab

We will be using the free, open-source software package, BEAST (Bayesian Evolutionary Analysis Sampling Trees; http://beast2.org), for estimating species trees. This tutorial is intended to be used with BEAST version 2.3.3. The distribution comes with the BEAUTi, which you will use to manage package plugins (also called add-ons), including SNAPP.

BEAST comes with several other utility programs that we will use to prepare input files (BEAUTi; Bayesian Evolutionary Analysis Utility) and summarize output files TreeAnnotator, and LogCombiner). We will also be using the programs Tracer (http://tree.bio.ed.ac.uk/software/tracer) and FigTree (http://tree.bio.ed.ac.uk/software/figtree) for evaluating, summarizing, and viewing results.

5 The Data

We will be analyzing 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 the species delimitation example, we will test species delimitation 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.

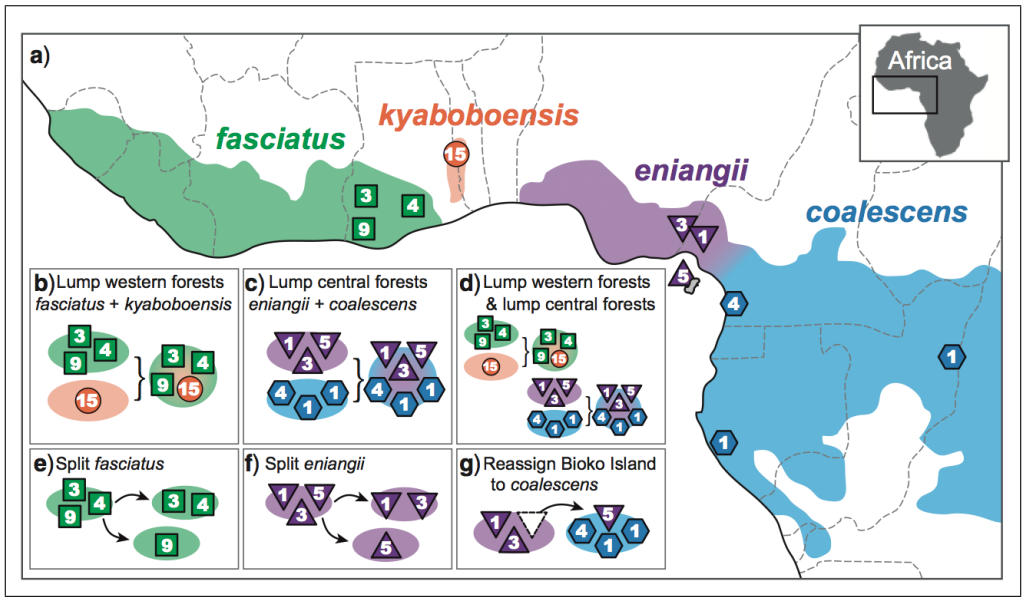

Figure 1: Geographic sampling of geckos (numbers in symbols indicate sample sizes). Starting taxonomy is shown in (a). BFD* is used to test the alternative species delimitation models outlined in (b) – (g).

6 Tutorial

6.1 Downloads and Data

Step 1: Download BEAST from http://beast2.org and install it on your computer. This tutorial is written for the Linux version of BEAST v2.3.3.

You will be using BEAST to run SNAPP, although it is possible to run SNAPP on its own. However, we have to use BEAST in order to combine SNAPP and marginal likelihood estimation into the same analytical framework. Thus, without BEAST we would not be able to conduct Bayes factor delimitation of species with SNAPP.

Step 2: After downloading and unzipping this archive you should have a BFDstar-tutorial folder on your computer. This tutorial contains the files and folders shown in Box 1. The ***data*** folder contains the gecko SNP data in binary format (necessary for SNAPP). 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 SNAPP input format at our phrynomics project site at GitHub. You can also find help at the BEAST google users group. The ***xml*** folder contains seven xml files (named according to the species delimitation models in Figure 1) that are ready to run in BEAST.

- BFDstar-tutorial/
  - BFDstar-tutorial.pdf
  - data/
    - hemi129.nex
  - xml/
    - runA.xml
    - runB.xml
    - runC.xml
    - runD.xml
    - runE.xml
    - runF.xml
    - runG.xml

Box 1: The files included in this tutorial. The data folder contains the SNP data in binary format. Ready-to-run XML files are included in the xml folder.

6.2 Setting up the XML file with BEAUTi

Step 3: Begin by launching the BEAUTi program that comes with BEAST. If you are using Mac OS X or Windows, you should be able to do this by double clicking on the application. On Linux, open a terminal and `cd` into the extracted BEAST folder, then launch BEAUTi using the command `bin/beauti`. If everything is working correctly, a window should appear that looks something like Figure 2.

Step 4: We need to add functionality to BEAST in order to estimate species trees with SNP data and to perform model selection. Begin by using the drop-down menu ***File*** →***Manage Packages.*** A window should appear that looks something like Figure 3. Select and install the packages **SNAPP** and **Model_Selection**. You can then exit the window by clicking the **"Close"** button.

Step 5: We need to tell BEAUTi that we are setting up a SNAPP analysis, which will change the menu options and allow us to import SNP data. Begin by using the drop-down menu ***File*** →***Template, SNAPP*** . This should change the appearance of the BEAUTi window to look something like Figure 4.

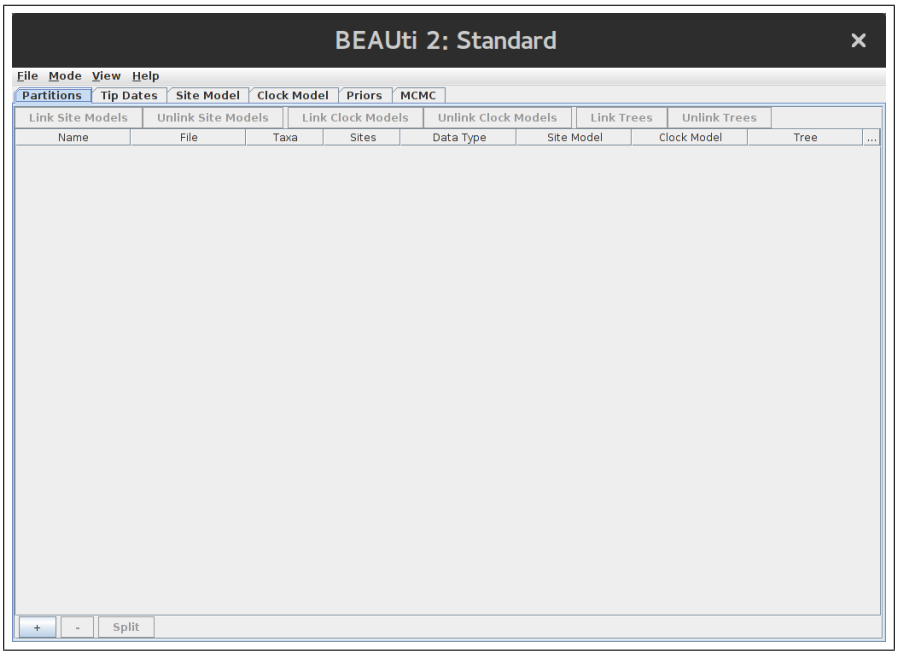

Figure 2: BEAUTi window before any data is loaded.

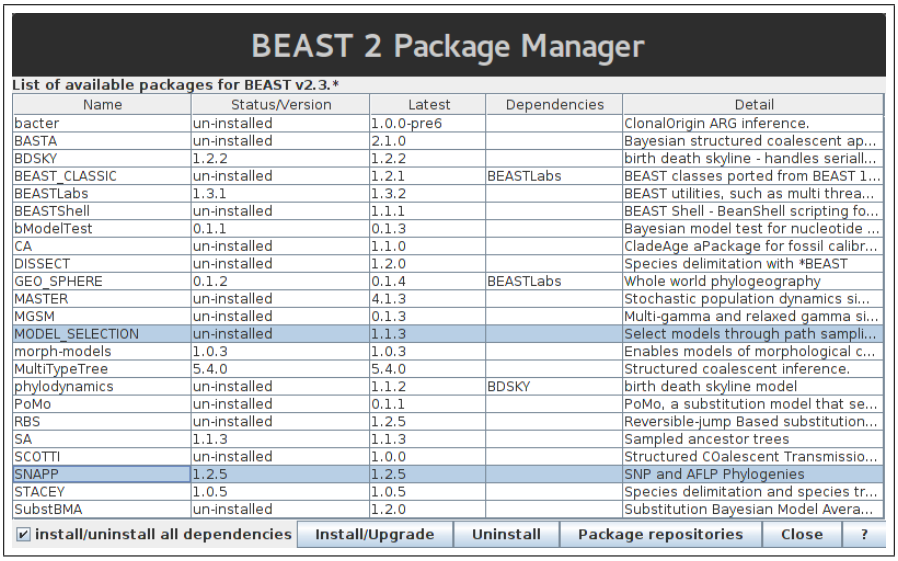

Figure 3: BEAUTi package manager for BEAST.

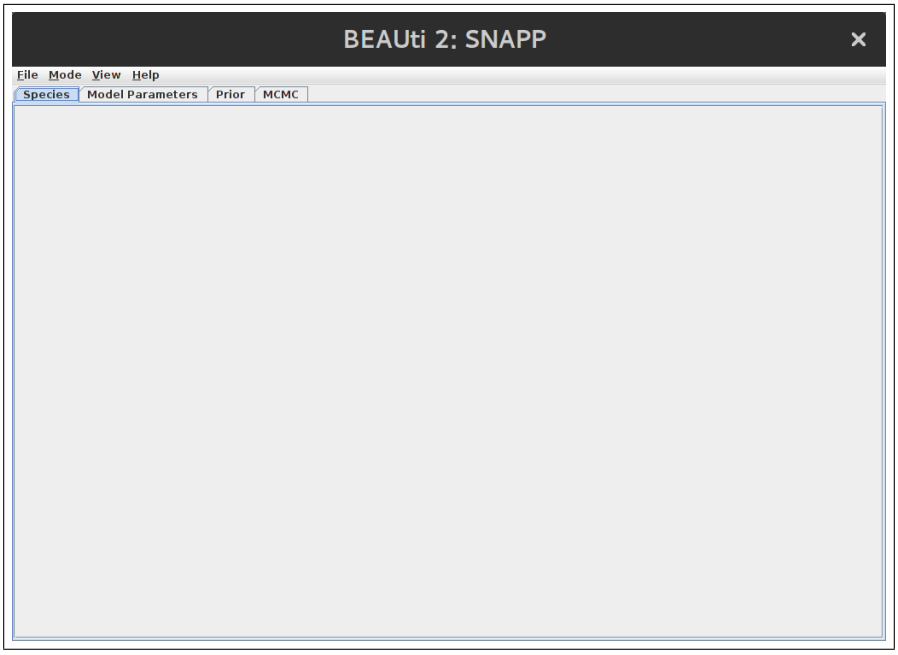

Figure 4: BEAUTi window after importing the SNAPP template. Notice that the menu tabs have changed.

Step 6: Import the SNP data (the **hemi129.nex** file) using the drop-down menu ***File*** →***Add Alignment.***

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

Step 7: There are several ways to designate species assignments. 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
coal_CA1_5
coal_CG_6
kya_GH3_7
kya_GH3_8
...
```

To use this information to designate the correct species assignments, click the "Guess" button. The screen should look similar to Figure 6. The original data include 46 samples, but the XML files included in this tutorial contain a reduced number of samples to speed up the analyses.

Change "use everything" to "before first", and leave the underscore in the text box. This means BEAUTi will use the part of each sample name before the first underscore as the species name. Click OK to make the species assignments.

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 ***Species*** window. Be sure that each Taxon has a Species/Population name.

Step 8: Next, we need to set up our model under the ***Mutation Model*** tab Figure 7. We will use the default options for this tutorial, but you should read the documentation A rough guide to SNAPP to learn

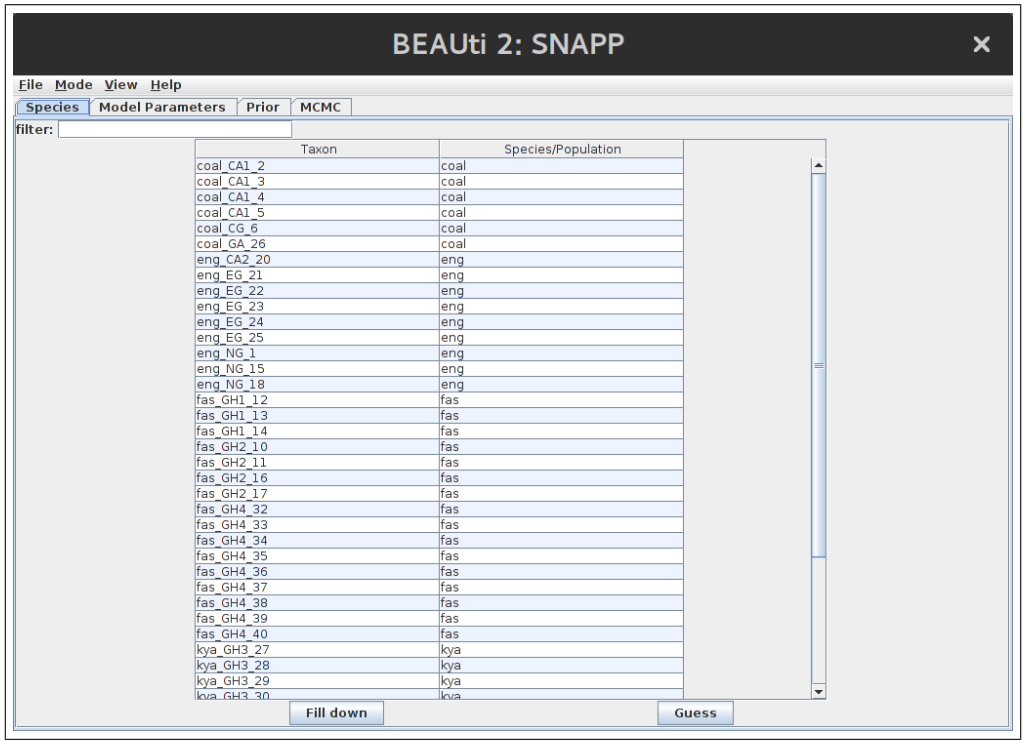

Figure 5: The data successfully loaded by BEAUTi.

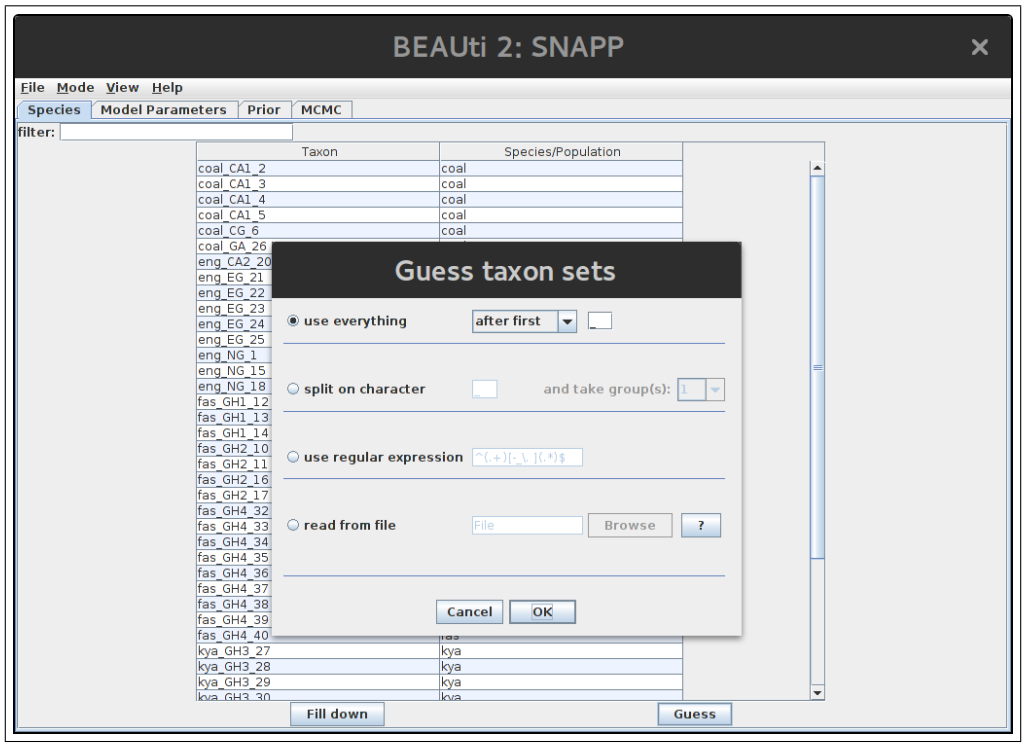

Figure 6: The species assignment options that appears after you select the "Guess" button.

more about the model options. Briefly, the parameters are as follows:

```
Mutation Rate U: instantaneous rate of mutating from the 0 allele to the 1 allele.
Mutation Rate V: instantaneous rate of mutating from the 1 allele to the 0 allele.
Coalescence Rate:  population size parameter with one value for each node in the tree.
```

By default the forward and reverse mutation rates U and V are both set to equal 1.0. These can be sampled during the MCMC, or estimated directly from the alignment. For this tutorial, we will estimate the rates from the alignment. First, click the "Calc mutation rates" button to estimate both rates from the alignment. Then, untick the "Sample" checkbox next to Mutation Rate U.

Untick the "Include non-polymorphic sites" checkbox. This option is used in cases where invariant sites have been included in the data. The likelihood calculations are different if SNAPP assumes that all constant sites have been removed.

Leave the "Mutation Only At Root" checkbox unticked. This option indicates 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.

Leave the "Show Pattern Likelihoods And Quit" checkbox unticked. This option is handy if you just want to print out the likelihoods for all patterns in the starting state and then quit.

Leave the "Use Log Likelihood Correction" checkbox ticked. This option calculates corrected likelihood values for Bayes factor test of different species assignments.

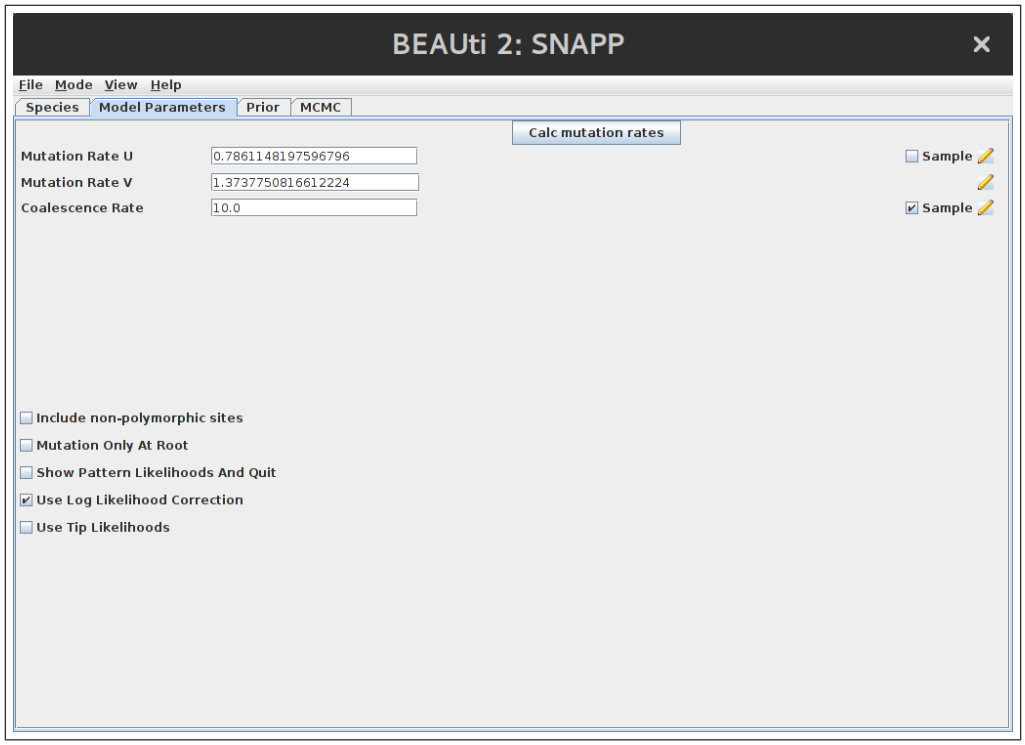

Figure 7: The Mutation Model options.

Step 9: Next, we need to move to the ***Prior*** tab and specify the priors (Figure 8.) We will use the default options for this tutorial, but you should read the documentation A rough guide to SNAPP to learn more about these options. A short description of the priors are provided below: Briefly, SNAPP uses a Yule prior for the species tree and branch lengths on the species tree. This prior has a single parameter, λ (Lambda), which governs the rate that species diverge. This rate, in turn, determines the (prior) expected height of the species tree.

```
Alpha:  shape parameter for the gamma prior on population sizes.
Beta:  scale parameter for the gamma prior on population sizes.
Kappa:  parameter used when selecting the CIR rate prior (below).
```

```
Lambda:  Birth rate for the Yule model prior on the species tree.
Rateprior:  prior on rates can be Gamma, InverseGamma, CIR, or Uniform.
```

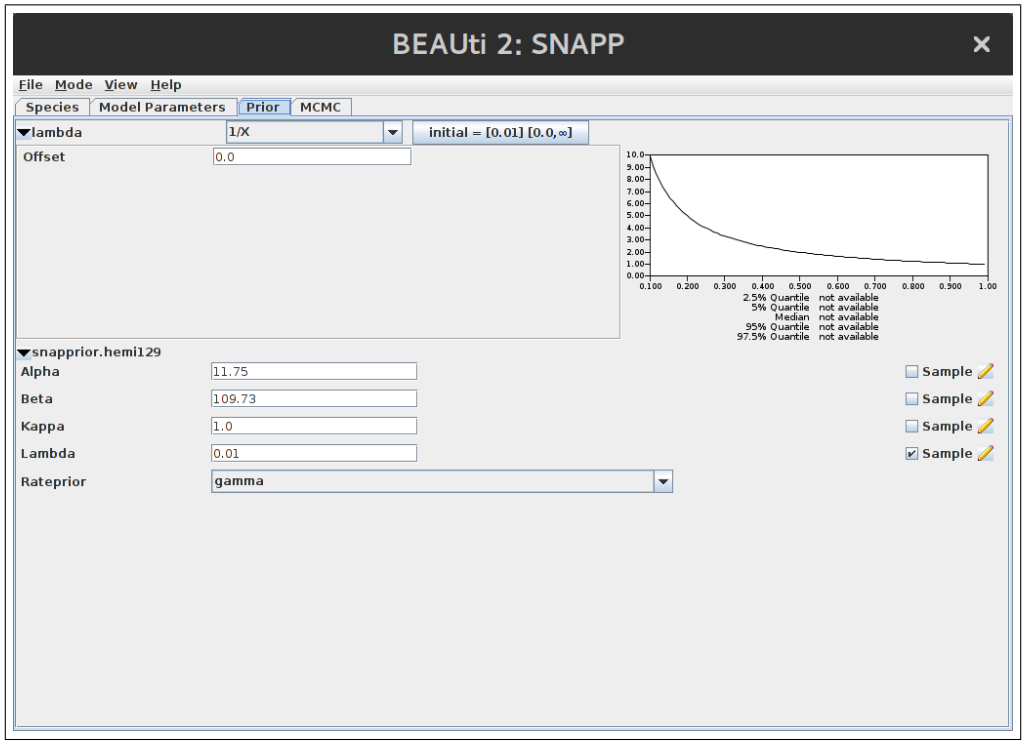

Figure 8: The prior settings.

Step 10: Next, move to the ***MCMC*** tab. Change the following settings:

```
Chain Length:  1000
Store Every:  10
tracelog:File Name:  runA.log
tracelog:Log Every:  10
screenlog:Log Every:  10
treelog:File Name:  runA.trees
treelog:Log Every:  10
```

Leave the remaining options at their default values (Figure 9). These MCMC values are way to low, and a thorough analysis requires much more computational time. The original SNP data include 46 samples, but the files included in this tutorial contain a reduced number of samples to speed up the analyses. The MCMC run times are intentionally kept short (and the data files reduced) in this tutorial. These short analyses should run in approximately 2 – 4 minutes depending on the number of processors available on your computer. Thorough analyses of the full data take 2 – 6 days, depending on the number of species in the model. A SNP matrix with 1,000 loci requires 5 – 20 days.

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 place it in a folder with the name "runA". Save the file to the BFDstar-tutorial folder.

6.3 Editing the XML file for marginal likelihood estimation

Step 11: Species delimitation using SNPs requires marginal likelihood estimation. You will need to edit the XML file to prepare it for analysis in BEAST. Instructions for setting up marginal likelihood estimation using path sampling are provided at the BEAST 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.

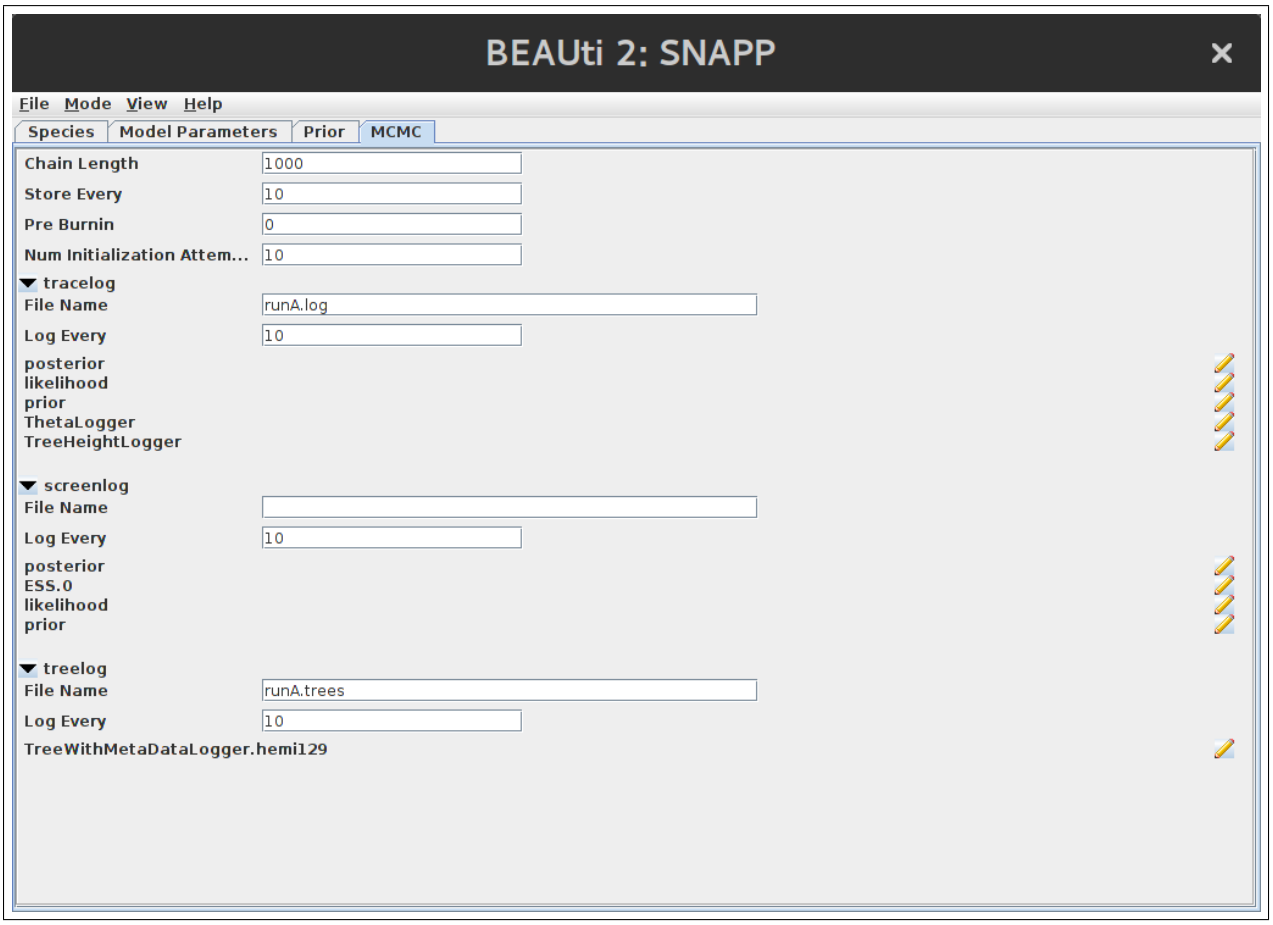

Figure 9: The MCMC settings.

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.

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='/home/desktop/BFDstar-tutorial/runA/'
burnInPercentage='0'
preBurnin="0"
deleteOldLogs='true'
nrOfSteps='12'>
cd $(dir)
java -cp $(java.class.path) beast.app.beastapp.BeastMain $(resume/overwrite) -java -seed $(seed) beast.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 (", ", etc.) don't copy as they should, and these will cause problems. Also, make sure that the root directory path (rootdir) exists on your computer.

These path sampling parameters are way to low, and a thorough analysis requires much more computational time. The MCMC run times are intentionally kept short in this tutorial so that we have time to conduct analyses and discuss the results.

The path sampling parameters that you just entered into your XML file are as follows:

```
chainLength:  MCMC sample length for each path sampling step.
alpha:  parameter used to space out path sampling steps.
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 step, but not the others.
deleteOldLogs:  delete existing log files from rootdir
nrOfSteps:  the number of path sampling steps to use
```

6.4 Running the XML file with BEAST

Step 12: You can execute the XML file in BEAST using the GUI or the command line. If you are using Mac OS X or Windows, you should be able to launch the BEAST GUI by double clicking on the application icon. On Linux, launch the GUI by executing `bin/beast -threads 4` inside the extracted BEAST folder. On Linux, BEAST will immediately ask you to choose the XML file to execute. Select the runA.xml file you just created. By specifying 4 threads, four steps can execute simultaneously, and the entire process should complete within minutes. You can also run BEAST from the command line. Open your computer's Terminal and navigate to the folder containing your **runA.xml** file. To execute the file, type the following at the command line:

```
/path/to/beast/bin/beast -threads 4 runA.xml
```

Set the number of `threads` to equal the number of CPU cores on your computer.

6.5 Inspecting path sampling results

Step 13: At the end of your analysis, the path sampling results will be displayed on the screen. An example is shown in Figure 10. Each row shows the results from one path sampling step. The example in Figure 10 shows the results from a path sampling analysis with 12 steps. You will use the value after "marginal L estimate" to compare models.

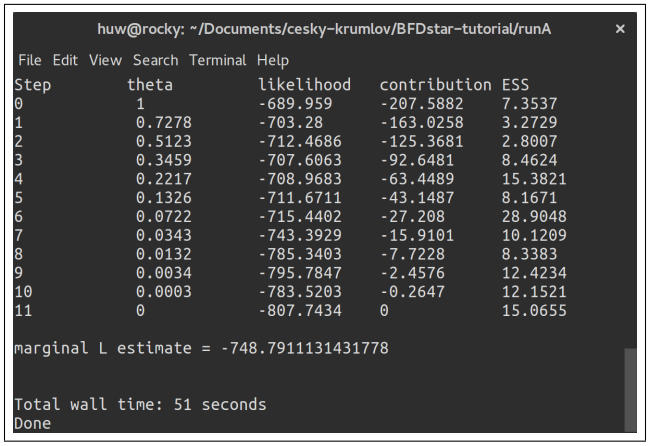

Figure 10: The path sampling output at the end of the analysis.

6.6 Setting up new species delimitation models

Step 14: 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, we have to make a few slight modifications to the existing runA.xml file: (1) save a copy of the xml file as runB.xml and save it in a new folder with the name "runB", (2) to avoid overwriting your runA files use find-and-replace to replace all occurances of "runA" with "runB", (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 11). 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 a different species you can cut and paste the taxon to the new taxonset. XML files containing the species assignments shown in Figure 1 are provided with this tutorial. The XML files included with this tutorial contain a reduced number of samples in the taxonset blocks to help speed up the analyses.

The taxon names in the data and XML files are a little cryptic. Lumping species should be no problem, but for splitting and reassignment you should transfer the following individuals:

- *H. eniangii* population 5 (Bioko Island, Equatorial Guinea). Labeled as `eng_EG_22` in alignment.
- *H. fasciatus* population 9 (western Ghana). Labeled as `fas_GH4_37` in alignment.

Step 15: 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 negative BF value indicates support in favor of model 1. A positive BF value indicates support in favor 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: $0 < \text{BF} < 2$ is not worth more than a bare mention, $2 < \text{BF} < 6$ is positive evidence, $6 < \text{BF} < 10$ is strong support, and $\text{BF} > 10$ is decisive.

The results for the seven gecko models are provided in Table 1. The model that splits *Hemidactylus eniangii* into two species (runF) 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 F is decisive compared to model A.

6.7 Summarizing the trees using TreeAnnotator.

Step 16: TreeAnnotator will summarize the posterior distribution of species trees and identify the topology with

```
<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 11: Example of the taxonset features in the XML file.

Table 1: Path sampling results for the seven species delimitation models.

Model	*Species*	*MLE*	*Rank*	*BF*
runA, current taxonomy	4	−748.79	5	–
runB, lump western forests	3	−704.07	2	−89.45
runC, lump central forests	3	−765.83	7	34.07
runD, lump western & central forests	2	−750.85	6	4.12
runE, split *fasciatus*	5	−726.36	4	−44.85
runF, split *eniangii*	5	−692.83	1	−111.92
runG, reassign Bioko Island	4	−706.73	3	−84.13

MLE = Marginal likelihood estimate, BF = Bayes factor. Model names reflect alternative hypotheses shown in Figure 1.

the best posterior support, and summarize the divergence times for each node in the tree.

Each path sampling step uses a different mixture of the prior and likelihood. Step 0 is the full likelihood, so trees sampled in Step 0 are from the posterior distribution. Step 11 has no likelihood, so trees sampled in Step 11 are from the prior distribution. To summarize the trees in the posterior distribution, follow these steps:

1. Launch the TreeAnnotator program.
2. For the `Target tree type` field, keep **Maximum clade credibility tree**.
3. For the `Node heights` field, keep **Common Ancestor heights**.
4. Select the `Input Tree File` button and select the file runX/step0/runX.trees.
5. Select the `Output File` button and specify the output directory and a file name, runX-MCC.tre.
6. Click `Run`

6.8 Visualizing the tree in FigTree

Step 17: 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.

7 Quick Version of the Tutorial

Step 1: Download and data.

Step 2: Data included with the tutorial.

Step 3: Launch BEAUTi

Step 4: Install SNAPP and model selection packages

Step 5: Converting BEAUTi to SNAPP mode

Step 6: Import the SNP data.

Step 7: Define species.

Step 8: Set the mutation model.

Step 9: Define the priors.

Step 10: Specify MCMC settings and generate the XML file.

Step 11: Editing the XML file for marginal likelihood estimation.

Step 12: Run the XML file in BEAST.

Step 13: Inspecting path sampling results.

Step 14: Setting up new XML files for species delimitation.

Step 15: Comparing species delimitation models with Bayes factors.

Step 16: Summarize the species tree using TreeAnnotator.

Step 17: Visualize the species tree in FigTree.

References

Baele, G., P. Lemey, T. Bedford, A. Rambaut, M. A. Suchard, and A. V. 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, C.-H. Wu, D. Xie, M. A. Suchard, A. Rambaut, and A. J. Drummond. 2014. BEAST 2: A software platform for bayesian evolutionary analysis. PLoS Comput Biol 10:e1003537.

Bryant, D., R. Bouckaert, J. Felsenstein, N. A. 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.

Fujita, M. K., A. D. Leaché, F. T. Burbrink, J. A. McGuire, and C. Moritz. 2012. Coalescent-based species delimitation in an integrative taxonomy. Trends in Ecology & Evolution 27:480 – 488.

Grummer, J. A., R. W. Bryson, and T. W. Reeder. 2014. Species delimitation using bayes factors: Simulations and application to the sceloporus scalaris species group (squamata: Phrynosomatidae). Systematic Biology 63:119–133.

Kass, R. E. and A. E. Raftery. 1995. Bayes factors. Journal of the American Statistical Association 90:773–795.

Leaché, A. D., M. K. Fujita, V. N. Minin, and R. R. Bouckaert. 2014. Species delimitation using genome-wide snp data. Systematic Biology 63:534–542.