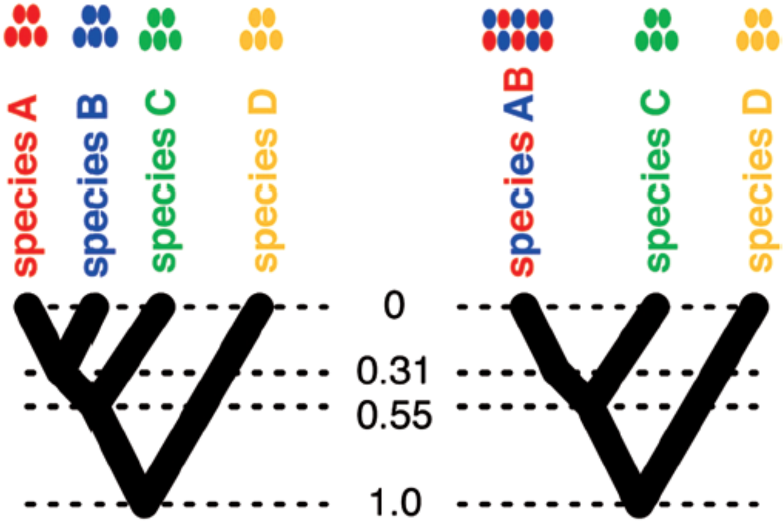
February 6th: Week 5 Tutorial

Objective: To learn how to perform Bayes factor delimitation of species (BFD), one of the many available methods to delimit species. Also, to understand how to calculate and interpret Bayes factors from results of a Bayesian phylogenetic analysis.

Lab Assignment

New methods are becoming available each year to identify/discover and validate "species", primarily with molecular data. Many of these programs utilize the **coalescent model** (discussed last week) to identify the most likely species tree, given a collection of independent gene trees. AND, as you could imagine, both ML (maximum likelihood) and Bayesian approaches exist for species delimitation. Today, we are going to explore a new method of species delimitation (peer-reviewed and produced by your very own TA!) that utilizes the Bayesian package **BEAST** that we used last week. We are going to be working with the same great ape dataset as last week. This species delimitation method takes advantage of the fact that individuals must be assigned to species before performing the species tree analysis.

The method we will use today compares many **speciation models** and determines which model is the best fit to our data. We will consider three speciation models where each speciation model has individuals differently partitioned into species. Consider the image below:

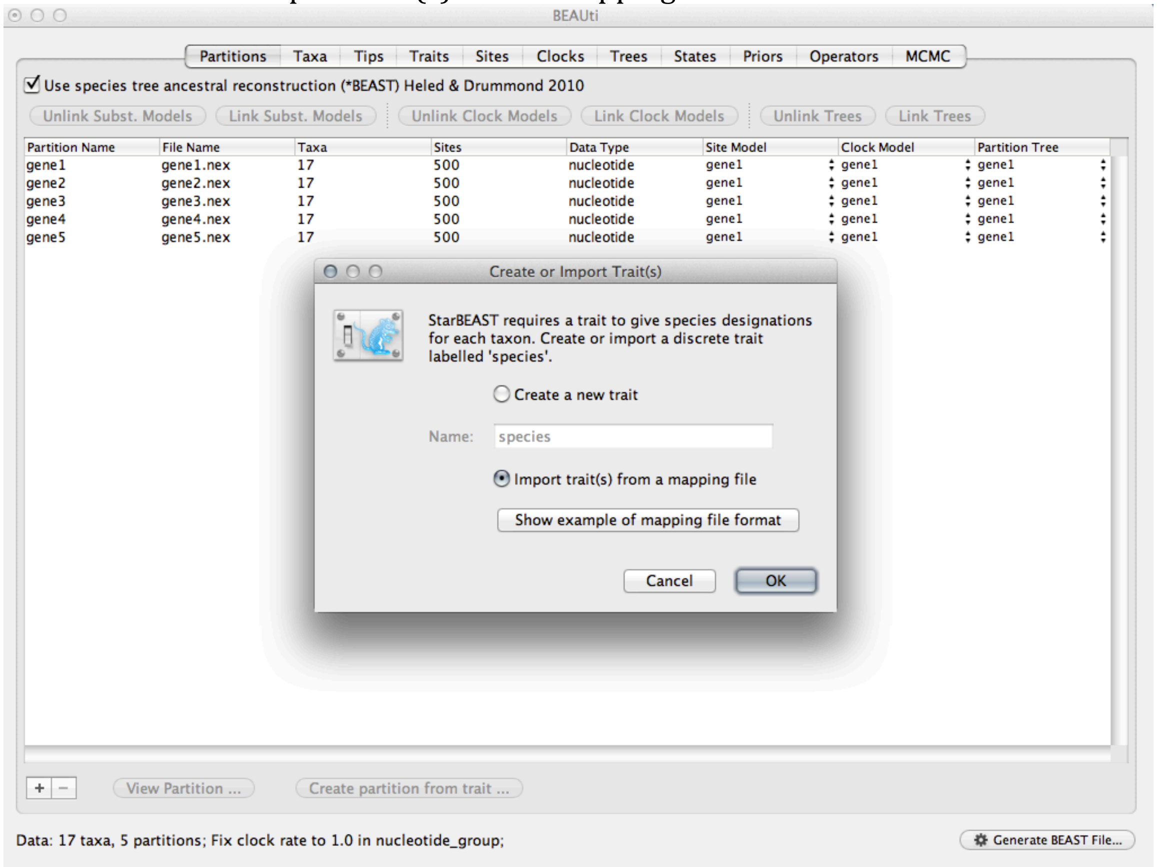


Don't worry about the numbers. What I want you to notice is that there are two speciation models here: one has four species (on the left), and one has three species (right). The model with three species has species A and B "lumped" into a single lineage, whereas the four species model has individuals from species A and B as distinct lineages (= species). This is fundamentally what we are going to be doing today: generating different speciation models in BEAUti (with different **species**

traits files), running the species tree analyses in *BEAST where a **marginal likelihood** score is generated, and then comparing the marginal likelihood scores via **Bayes factors**.

Go to the Google Drive folder and download the five gene alignments that we used last week (in Week 4 folder), as well as the species_traits file in the Week 5 folder that you have been assigned (either lump, split, reassign, or true). Place these six files into a folder on the desktop, named with your choice.

Go to the **BEASTv1.8.0** folder (*different than v1.7.5 from last week!*) within the Phylogenetic Analysis folder, then double-click on BEAUti. As in last week, load in your nexus files by either dragging and dropping or hitting the "+" symbol in the bottom left portion of BEAUti. Next, click on the "Use species tree ancestral reconstruction…" button in the top-left of BEAUti. Differently than last week, you will now click on "Import trait(s) from a mapping file".



☑ Use species tree ancestral reconstruction (*BEAST) Heled & Drummond 2010

Partition Name	File Name	Taxa	Sites	Data Type	Site Model	Clock Model	Partition Tree
gene1	gene1.nex	17	500	nucleotide	gene1	gene1	gene1
gene2	gene2.nex	17	500	nucleotide	gene1	gene1	gene1
gene3	gene3.nex	17	500	nucleotide	gene1	gene1	gene1
gene4	gene4.nex	17	500	nucleotide	gene1	gene1	gene1
gene5	gene5.nex	17	500	nucleotide	gene1	gene1	gene1

Create or Import Trait(s)

StarBEAST requires a trait to give species designations for each taxon. Create or import a discrete trait labelled 'species'.

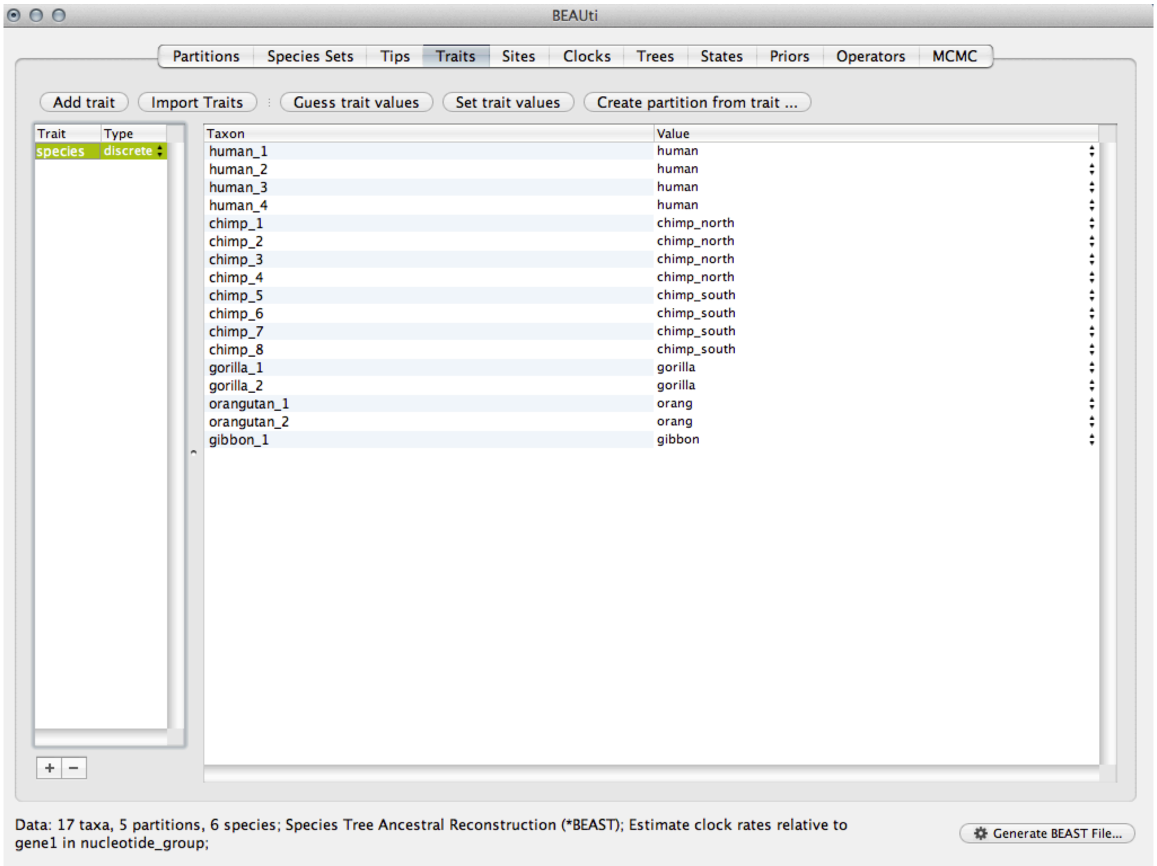
○ Create a new trait

Name: species

● Import trait(s) from a mapping file

Data: 17 taxa, 5 partitions; Fix clock rate to 1.0 in nucleotide_group;

Based on which speciation model you were assigned to analyze, upload the appropriate species_traits file. In the "Traits" tab, you can click on the "species" field under "Trait" to see how your individuals are assigned to species. This is what it looks like for the "species_traits_split" speciation model:



Trait	Type
species	discrete

Taxon	Value
human_1	human
human_2	human
human_3	human
human_4	human
chimp_1	chimp_north
chimp_2	chimp_north
chimp_3	chimp_north
chimp_4	chimp_north
chimp_5	chimp_south
chimp_6	chimp_south
chimp_7	chimp_south
chimp_8	chimp_south
gorilla_1	gorilla
gorilla_2	gorilla
orangutan_1	orang
orangutan_2	orang
gibbon_1	gibbon

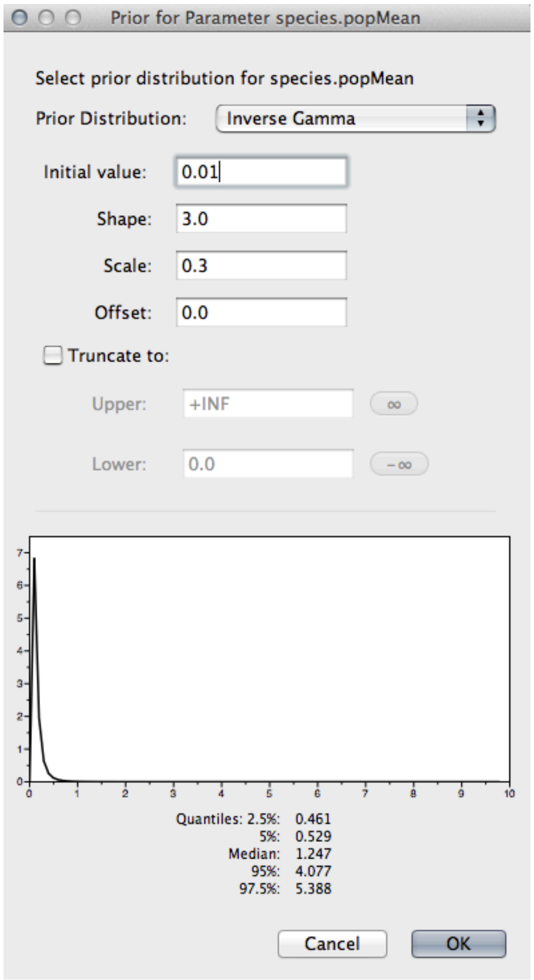
Data: 17 taxa, 5 partitions, 6 species; Species Tree Ancestral Reconstruction (*BEAST); Estimate clock rates relative to gene1 in nucleotide_group;

<u>Important Note!</u>

BEAUti v1.8.0 has a bug that sometimes doesn't allow you to see all parameters in the MCMC tab. Immediately after uploading data alignments, click on the MCMC tab; if you cannot read all values (they are off-screen to the left), quit BEAUti and re-open it. Try again. If you can't read the MCMC values after a couple tries, notify one of us!

Now, if BEAUti is cooperating, we will set up everything identical to how they were last week:

Sites: change base frequencies to "all equal" for each gene
Clocks: make sure all clock rates are being estimated (all boxes under "Estimate" checked), and click "Fix Mean" box in bottom left
Trees: change population size model to piecewise constant; make sure all genes have ploidy type correctly set
Priors: make sure all gene.clock.rate priors are set to an Exponential distribution. Note that the "species.popMean" and "species.yule.birthRate" have improper priors. Change the species.popMean to an Inverse Gamma with an initial value of 0.01, shape value of 3.0, and scale value of 0.3:



Prior for Parameter species.popMean

Select prior distribution for species.popMean

Prior Distribution: Inverse Gamma

Initial value: 0.01

Shape: 3.0

Scale: 0.3

Offset: 0.0

□ Truncate to:

Upper: +INF

Lower: 0.0

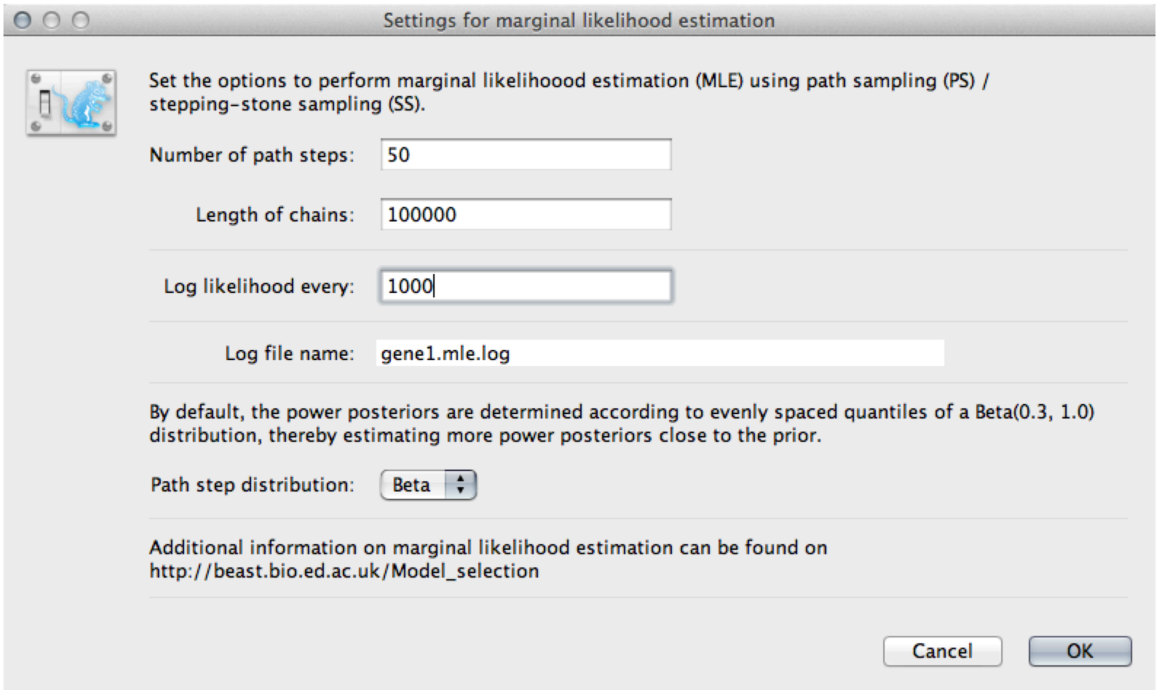
```
Quantiles: 2.5%:  0.461
             5%:  0.529
         Median:  1.247
            95%:  4.077
          97.5%:  5.388
```

Next, change the species.yule.birthRate to an Exponential distribution.

Now, in the MCMC tab, change:

- Echo state to screen every 5000
- Log parameters every 2000
- File name stem to your last name

- Next, click on the last box, "Perform marginal likelihood estimation (MLE)...", then click on the Settings button right below. Change:
    - Number of path steps to 50
    - Length of chains to 100000



Settings for marginal likelihood estimation

Set the options to perform marginal likelihoood estimation (MLE) using path sampling (PS) / stepping-stone sampling (SS).

Number of path steps: 50

Length of chains: 100000

Log likelihood every: 1000

Log file name: gene1.mle.log

By default, the power posteriors are determined according to evenly spaced quantiles of a Beta(0.3, 1.0) distribution, thereby estimating more power posteriors close to the prior.

Path step distribution: Beta

Additional information on marginal likelihood estimation can be found on http://beast.bio.ed.ac.uk/Model_selection

- ○ Now, hit OK.

Click the "Generate BEAST File…" button. If all priors and everything else are properly set, press Continue, then save the file in your desktop folder!

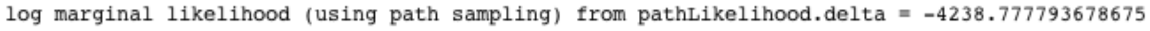
Now, back in the BEASTv1.8.0 folder, double click on BEAST, load in your .xml file, and hit the Run button. While your analysis is running, we will talk with you individually about your research projects.

This analysis should take ~30 mins, which is actually composed of two parts. The first part is a "normal" MCMC analysis, just like what you did in *BEAST last week. After this first part is done, it will begin the second part, the marginal likelihood estimation. The transition between the two parts should look like this:

```
Creating the Marginal Likelihood Estimator chain:
  chainLength=100000
  pathSteps=50
  pathScheme=betaQuantile(0.3)
  If you use these results, please cite:
    Guy Baele, Philippe Lemey, Trevor Bedford, Andrew Rambaut, Marc A. Suchard, and
    2012. Improving the accuracy of demographic and molecular clock model compariso
          phylogenetic uncertainty. Mol. Biol. Evol. 29(9):2157-2167.
    and
    Guy Baele, Wai Lok Sibon Li, Alexei J. Drummond, Marc A. Suchard, and Philippe
    Accurate model selection of relaxed molecular clocks in Bayesian phylogenetics.

Attempting theta (1/51) = 1.0 for 100000 iterations + 10000 burnin.
Attempting theta (2/51) = 0.9348750848354606 for 100000 iterations + 10000 burnin.
Attempting theta (3/51) = 0.872778642317937 for 100000 iterations + 10000 burnin.
Attempting theta (4/51) = 0.8136285485214058 for 100000 iterations + 10000 burnin.
Attempting theta (5/51) = 0.7573432479107086 for 100000 iterations + 10000 burnin.
Attempting theta (6/51) = 0.7038417613775041 for 100000 iterations + 10000 burnin.
Attempting theta (7/51) = 0.6530436945664579 for 100000 iterations + 10000 burnin.
Attempting theta (8/51) = 0.6048692465088115 for 100000 iterations + 10000 burnin.
```

This part will go through 50 steps (actually 51), just as we told it to in the MCMC tab. Once this portion is done, it will generate two marginal likelihood values: one from path sampling and the other from stepping stone sampling. *Write these two values down (to the second decimal place) in a word document and enter them into our graphs on the board (for the speciation model you analyzed)*.



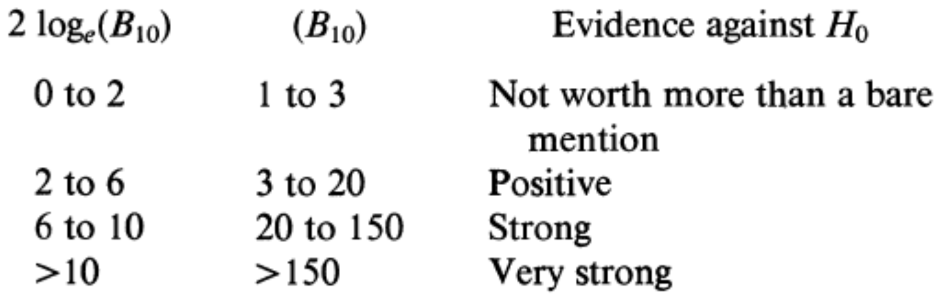
```
log marginal likelihood (using path sampling) from pathLikelihood.delta = -4238.777793678675
```

For completeness, you will also record the harmonic mean estimate of the likelihood value. To find this value, open your .log file in **Tracer** (make sure it's the .log file and not the .mle.log file!). Click on the "likelihood" parameter on the left and record the mean value. *Also write this value on the board*. And remember, now that you have your file in Tracer, check the ESS values and traces to see if your run has converged or not! We tend to run analyses for very short chain lengths in this class because we are limited by time.

And now it's time for the last step of this species delimitation method…model selection! You may remember from our experience with jModelTest that there are a variety of model selection criteria available to use to discern models (we used the Akaike Information Criterion [AIC] and Bayesian Information Criterion [BIC]). Today, we are going to use **Bayes factors**. The Bayes factor is a model selection tool that is (superficially) simple and well suited for the purposes of ranking species delimitation models. Calculating the Bayes factor between models is simple. To do so, arrange your models by likelihood score (your best model is that with the largest (negative) number, closest to zero). Once you have your models arranged, simply subtract the likelihood score of the worst of the two models from the best model, multiply this difference by two, and voila, there's your Bayes factor! Here's an example:

Model	Likelihood score	Bayes factor
Split	-4000	--
True	-4040	80 (Split vs. True)
Lump	-4100	200 (Split vs. Lump)

You assess support through Bayes factors based on the left column in this table (from a highly-cited publication from a UW professor!):



$2\log_e(B_{10})$	(B_{10})	Evidence against H_0
0 to 2	1 to 3	Not worth more than a bare mention
2 to 6	3 to 20	Positive
6 to 10	20 to 150	Strong
>10	>150	Very strong

Construct a table with the following format (lnL=log-likelihood score, PS=path sampling, SS=stepping stone sampling, HME=harmonic mean estimation, Bf=Bayes factor), and include it with your answers. Use either PS or SS to rank your models (not HME!), an example ordering is below:

Model	Rank	PS –lnL	PS Bf	SS –lnL	SS Bf	HME – lnL	HME Bf
True	1						
Split	2						
Lump	3						
Reassign	4						

Enter in values from your run for one row, then choose –lnL values from the board to insert into the other rows. Do this all in Excel, and make a formula that automatically calculates the Bayes factor value for each model (this is easy to do, and ask if you don't know how to!).

Questions to Answer

1.) Was there one speciation model that was the highest ranked across all types of marginal likelihood estimators? If so, which?
2.) Based on the Bayes factor values from your table above (using PS or SS estimators), how strongly favored was your top model?
3.) Do two of your marginal likelihood estimators (PS, SS, and HME) seem to report similar values to one another, for a given speciation model? If so, which two?
4.) What do your results say about the evolutionary history of the great apes, and in particular, humans and two populations of chimpanzees? In essence, are they "good species", given our data and your results?