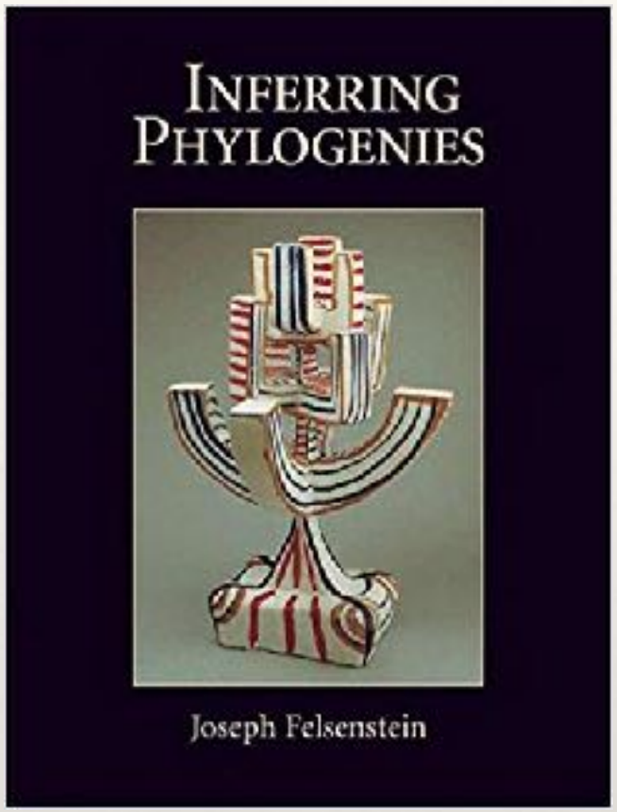
Zach Hancock — PhD student, Ecology & Evolutionary Biology

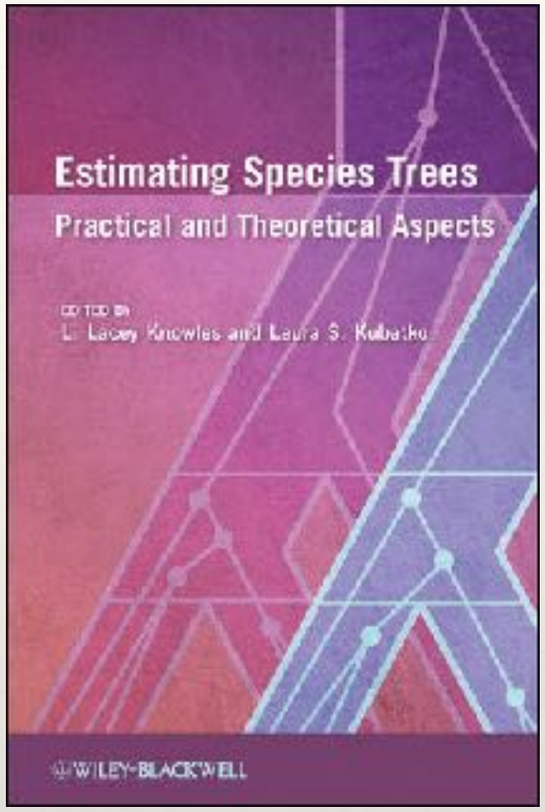
Rousing the BEAST

A molecular phylogenetic pipeline tutorial

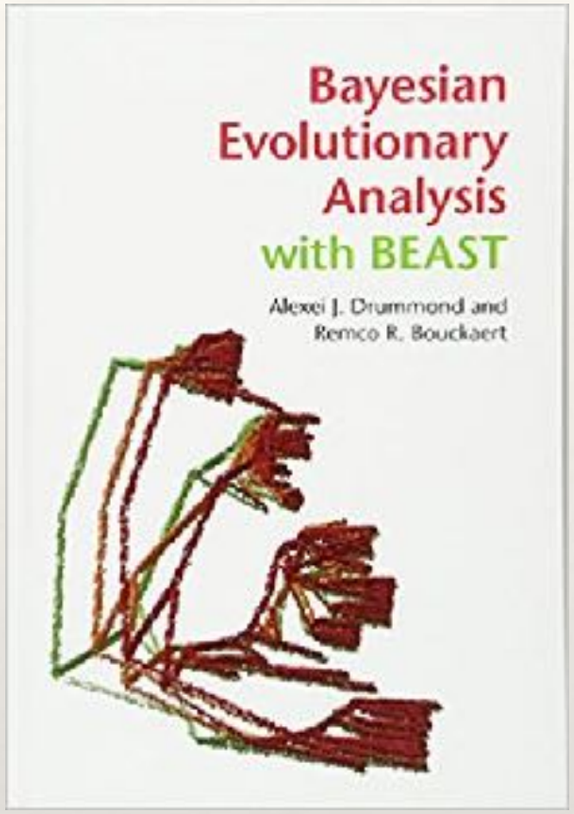
Recommendations



Felsenstein (2003)



Ed. Knowles & Kubatko (2011)



Drummond & Bouckaert (2015)

Recommended Course:
WFSC 646: Quantitative Phylogenetics
Dr. Mariana Mateos

Taming the BEAST Tutorials:

https://taming-the-beast.org/tutorials/

Phyloseminar lectures by Paul O. Lewis: https://www.youtube.com/watch?v=4PWlnNsfz90&t=2408s

Required Software

A good text editor:

- **Windows**: Notepad++ (https://notepad-plus-plus.org)
- **Mac**: BBEdit (http://www.barebones.com/products/bbedit/)

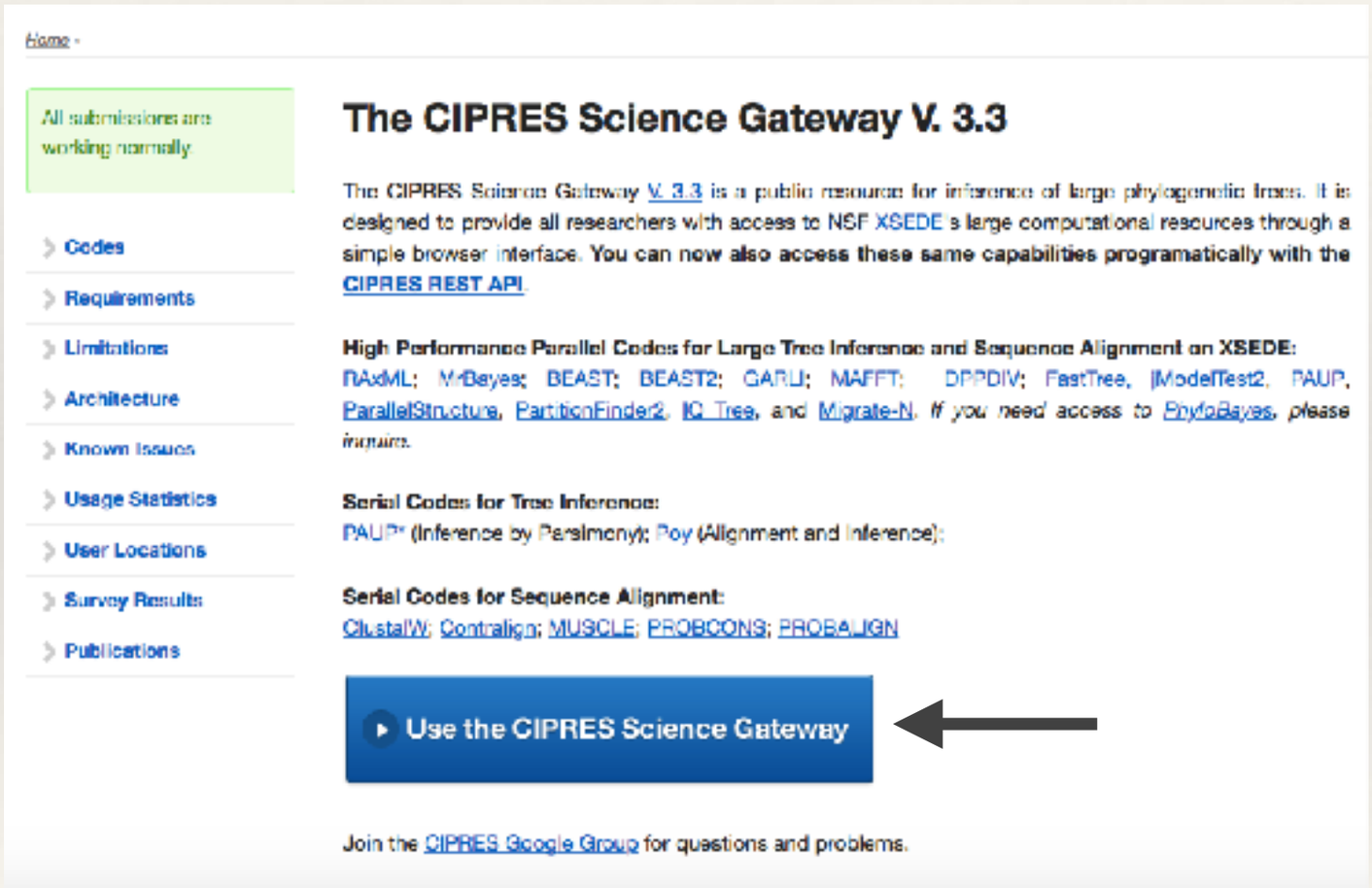
(https://www.mesquiteproject.org)

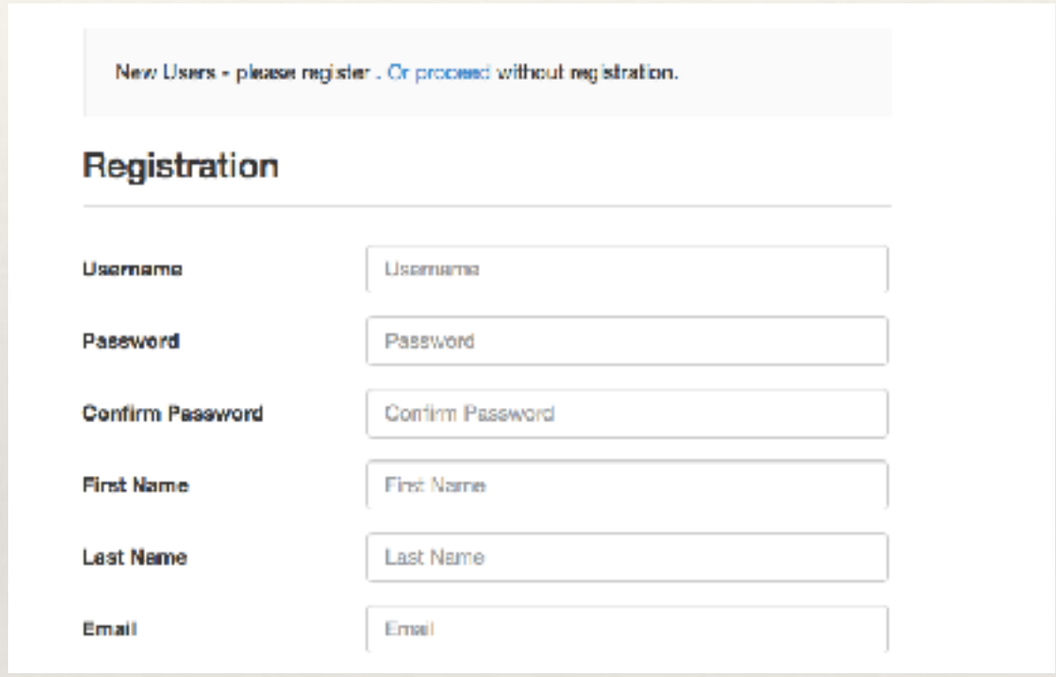
PAUP*
(https://paup.phylosolutions.com)

BEAST v2.5.0
(http://www.beast2.org)

TRACER v1.6
(http://tree.bio.ed.ac.uk/software/tracer/)

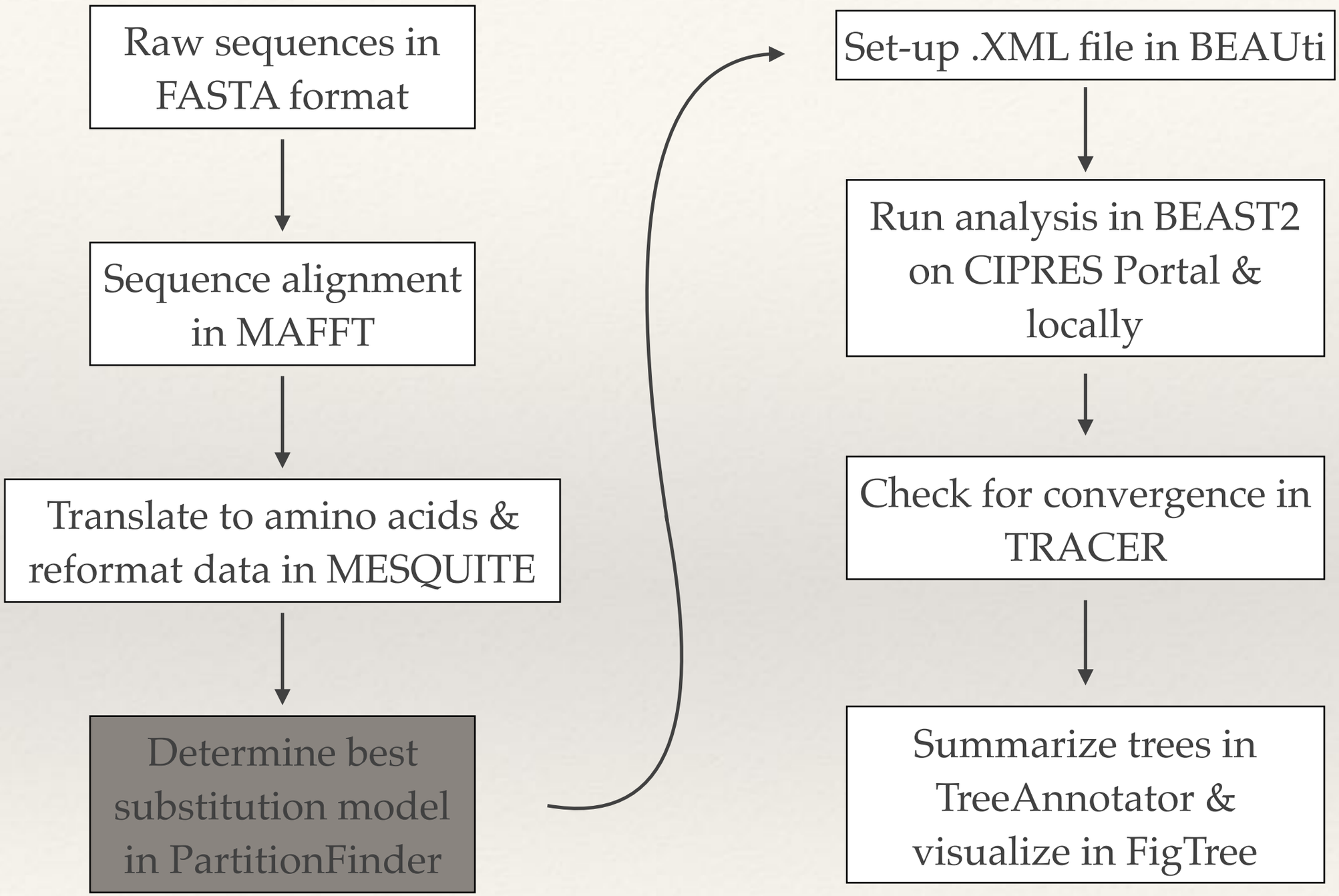
Make an account in CIPRES!
(https://www.phylo.org)





The email you use will be where all your notification on running projects will be sent

Molecular Phylogenetic Pipeline

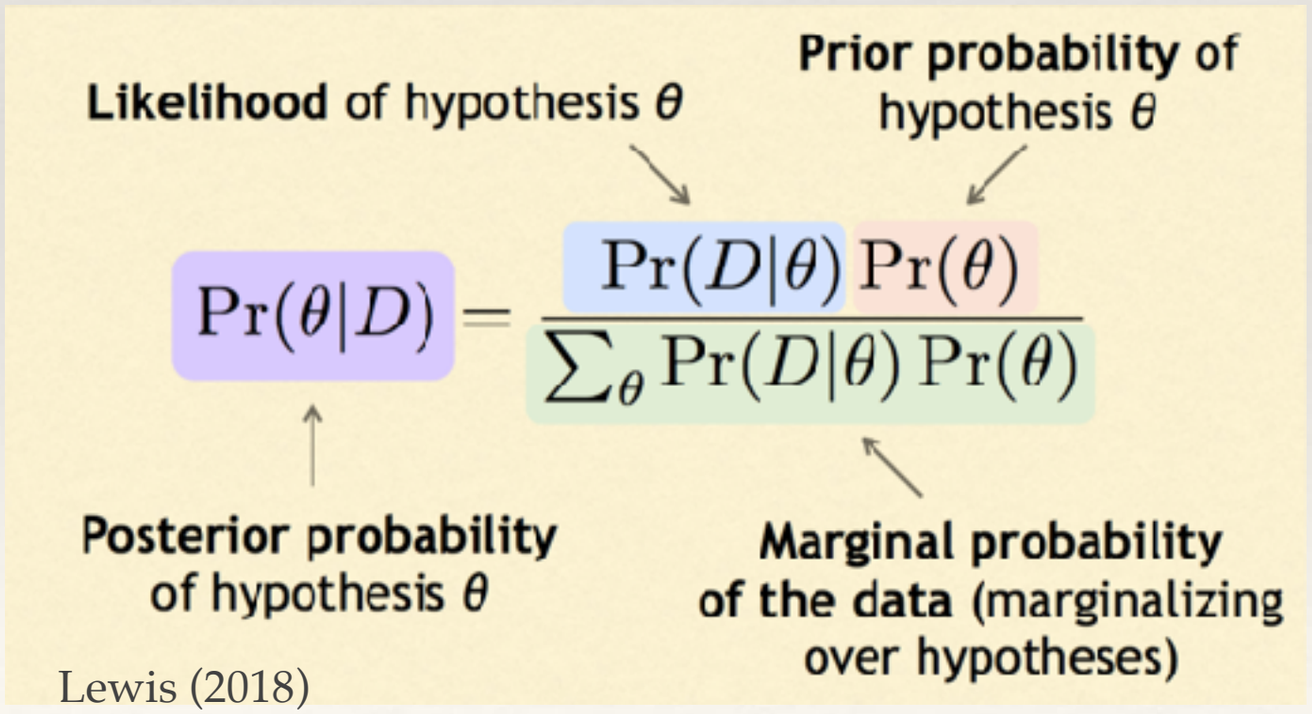


(Very) Brief Intro to Bayesian Inference

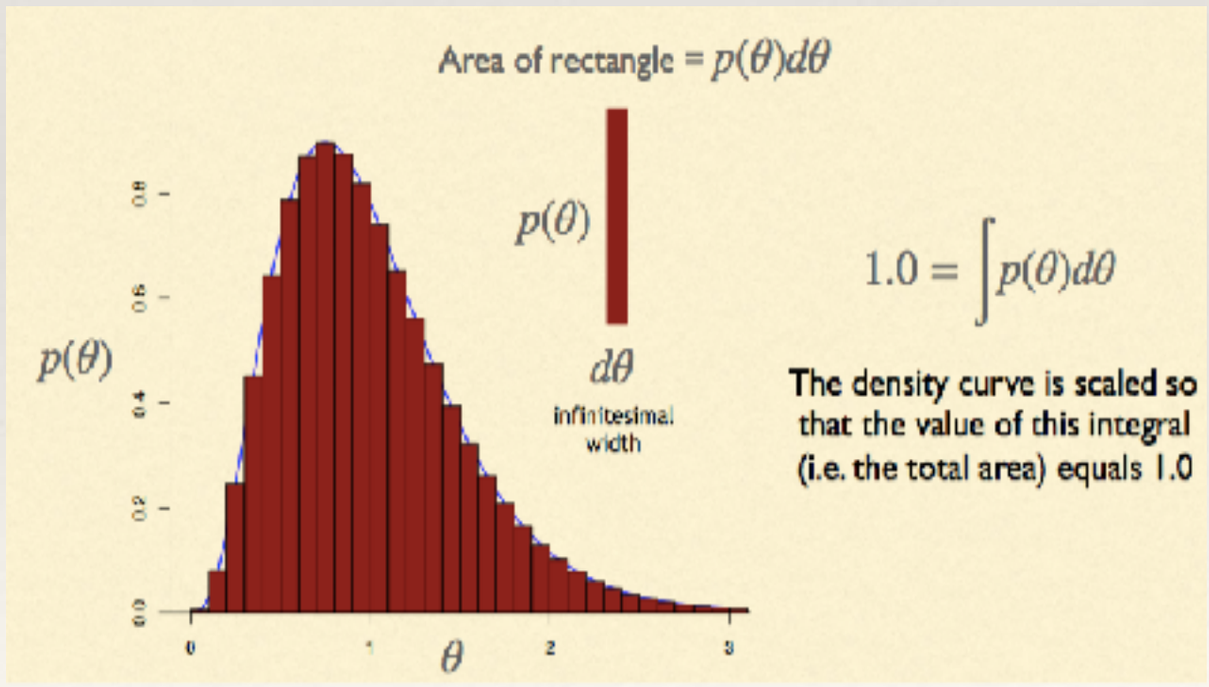
Bayes' theorem is useful in phylogenetics for two reasons:

1. Allows us to take into account prior knowledge about the data
2. Allows us to specifically test the probability of the *hypothesis* given the *data*

Thomas Bayes (1701–1761)

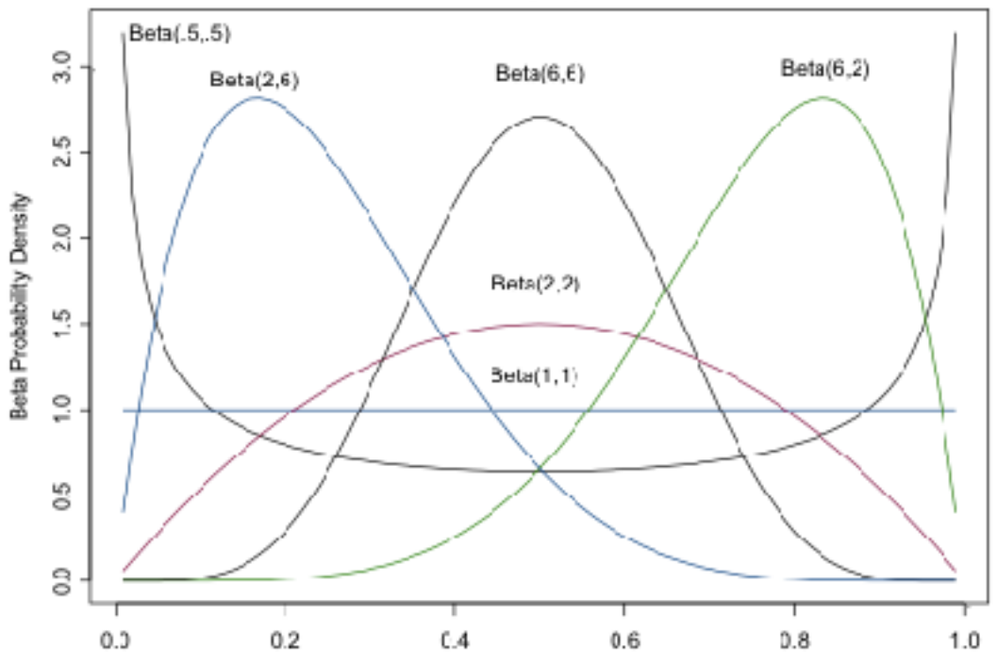


Lewis (2018)

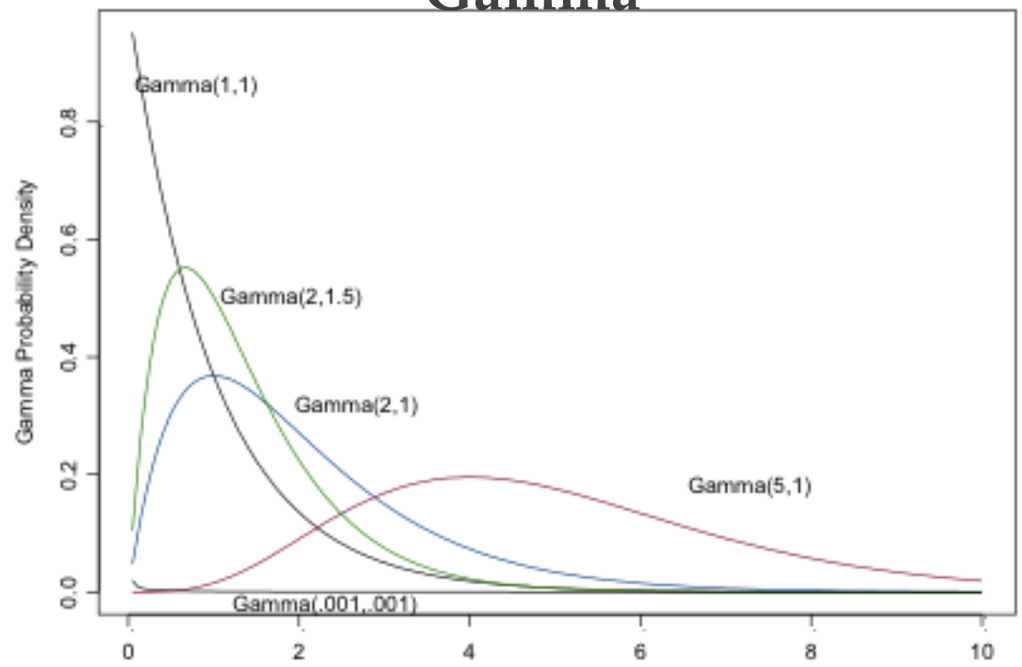


Prior Distributions

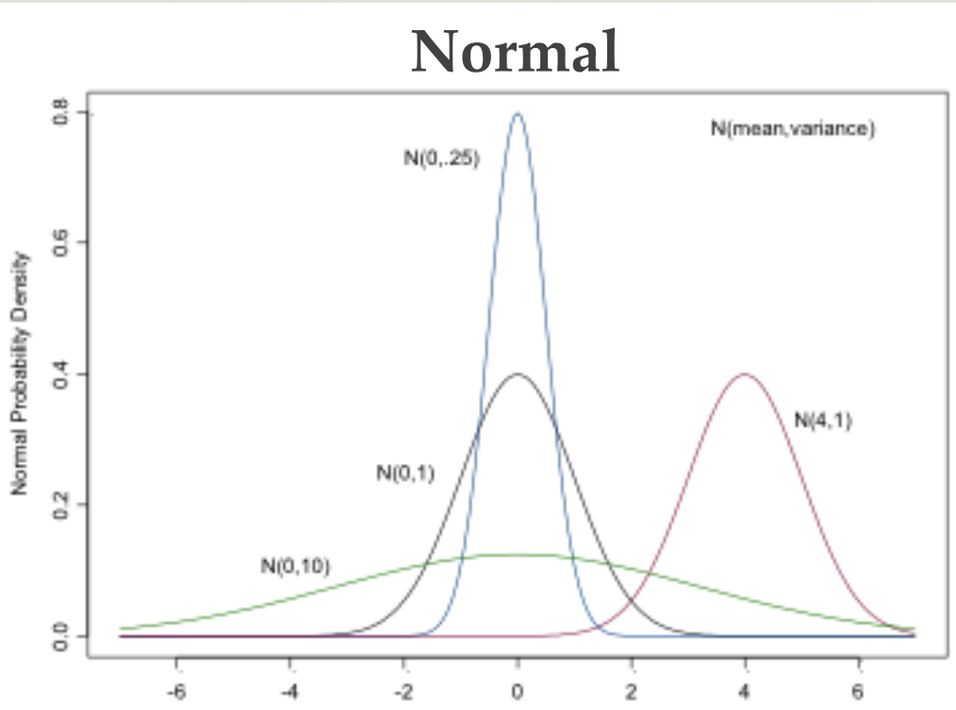
Beta



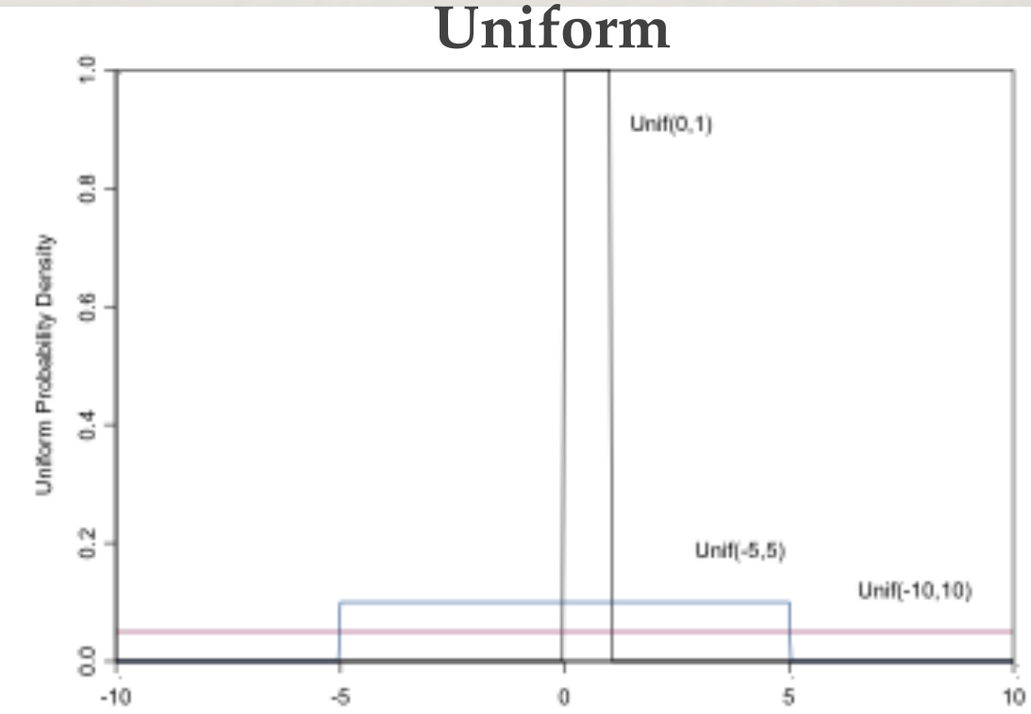
Gamma



Normal

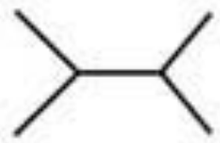


Uniform



http://www.columbia.edu/~cjd11/charles_dimaggio/DIRE/styled-4/styled-11/code-4/

How many trees are there, anyway?

number of taxa	Number of rooted fully bifurcating trees 	Number of unrooted fully bifurcating networks 
1	1	-
2	1	1
3	3	1
4	15	3
5	105	15
20	8,200,794,532,673,891,559,375	221,643,095,476,699,771,875

Forey (https://www.palass.org/publications/newsletter/cladistics-palaeontologists/cladistics-palaeontologists-part-3-tree-building)

If there are more than 50 sequences (not at all uncommon), the number of possible rooted trees exceeds the number of atoms in the observable universe (~10^{78})

For this reason, it's impossible to search *all* trees—instead, we explore **tree-space**

Sampling the Posterior (MCMC)

Estimating 2-parameters:

$$p(\theta, \phi \mid D) = \frac{p(D \mid \theta,\phi)\, p(\phi)p(\theta)}{\int_\theta \int_\phi p(D \mid \theta,\phi)\, p(\phi)p(\theta)\, d\phi\, d\theta}$$

The marginal probability can quickly get out of hand

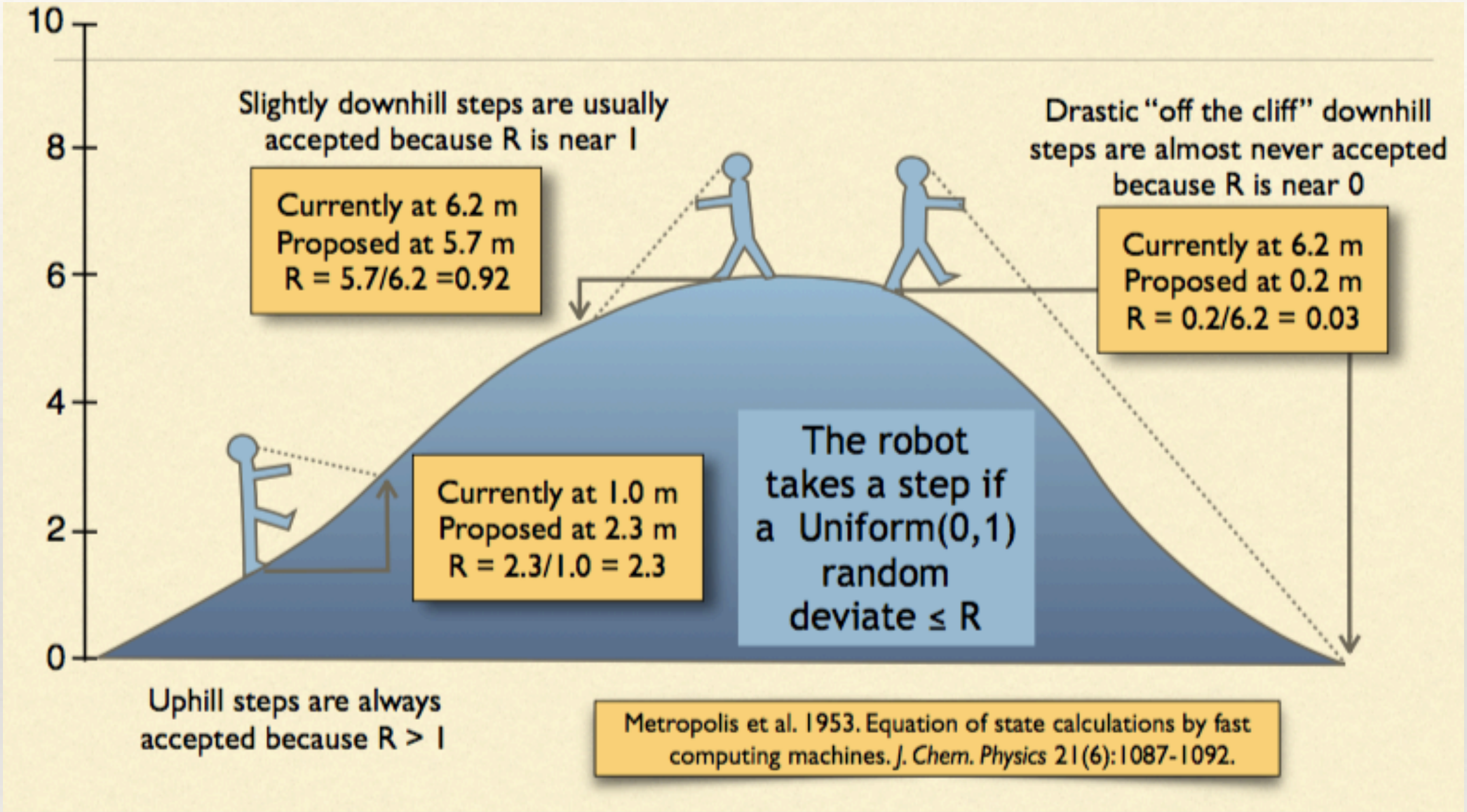
Using the Metropolis et al. (1953) algorithm of the Markov chain Monte Carlo, we can sample the posterior without calculating the marginal.

1. Start with a random parameter estimation and tree.
2. Propose a new estimate—is it better or worse?
3. If better, take the step and repeat. If worse, stay where you are and try again.
4. Continue until you have generated a good sample of the posterior.

$$\frac{\dfrac{p(\theta^* \mid D)p(\theta^*)}{\cancel{p(D)}}}{\dfrac{p(\theta \mid D)p(\theta)}{\cancel{p(D)}}} = \frac{p(\theta^* \mid D)p(\theta^*)}{p(\theta \mid D)p(\theta)}$$

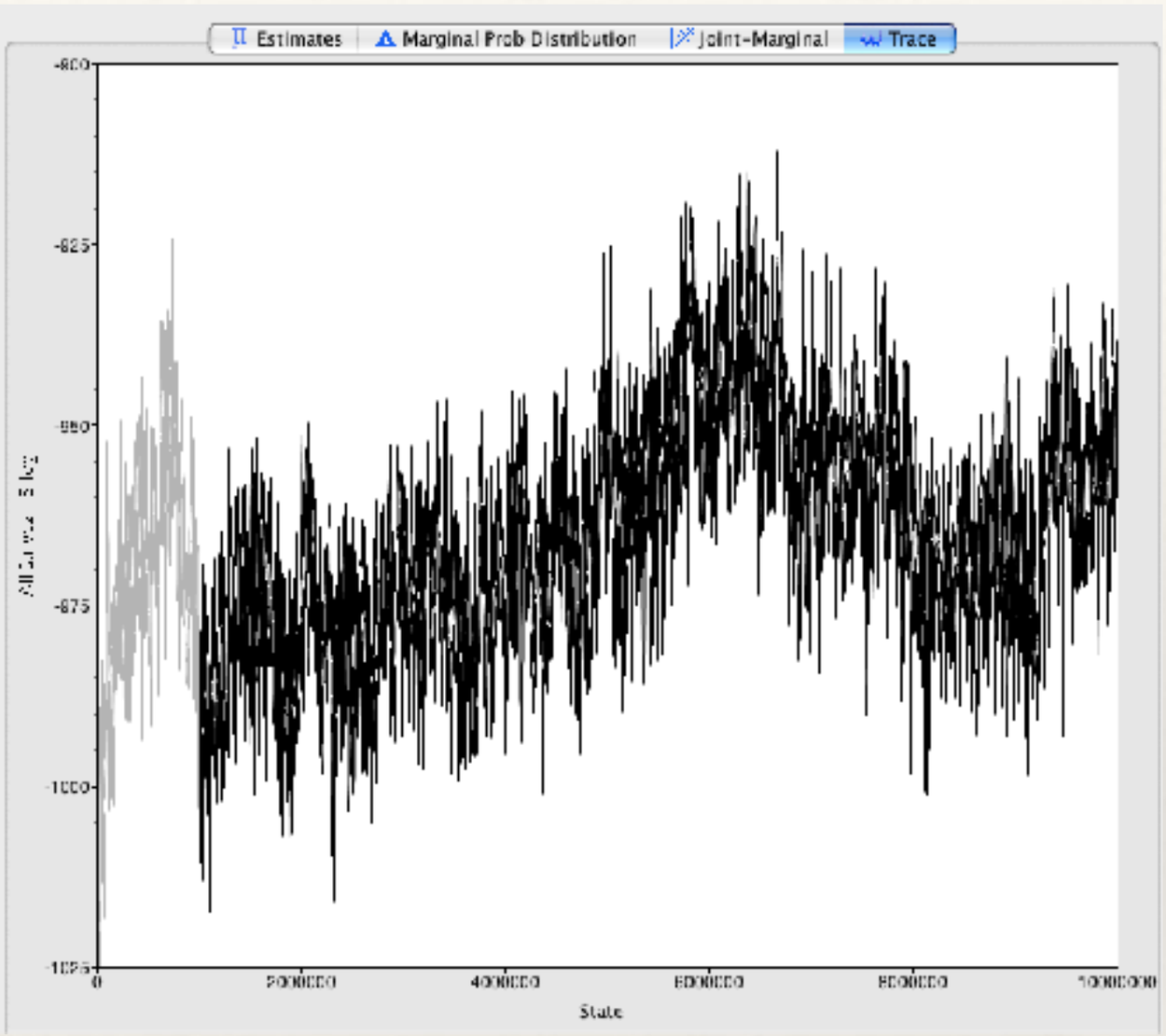
$p(\theta^* \mid D)p(\theta^*)$ ← New proposal

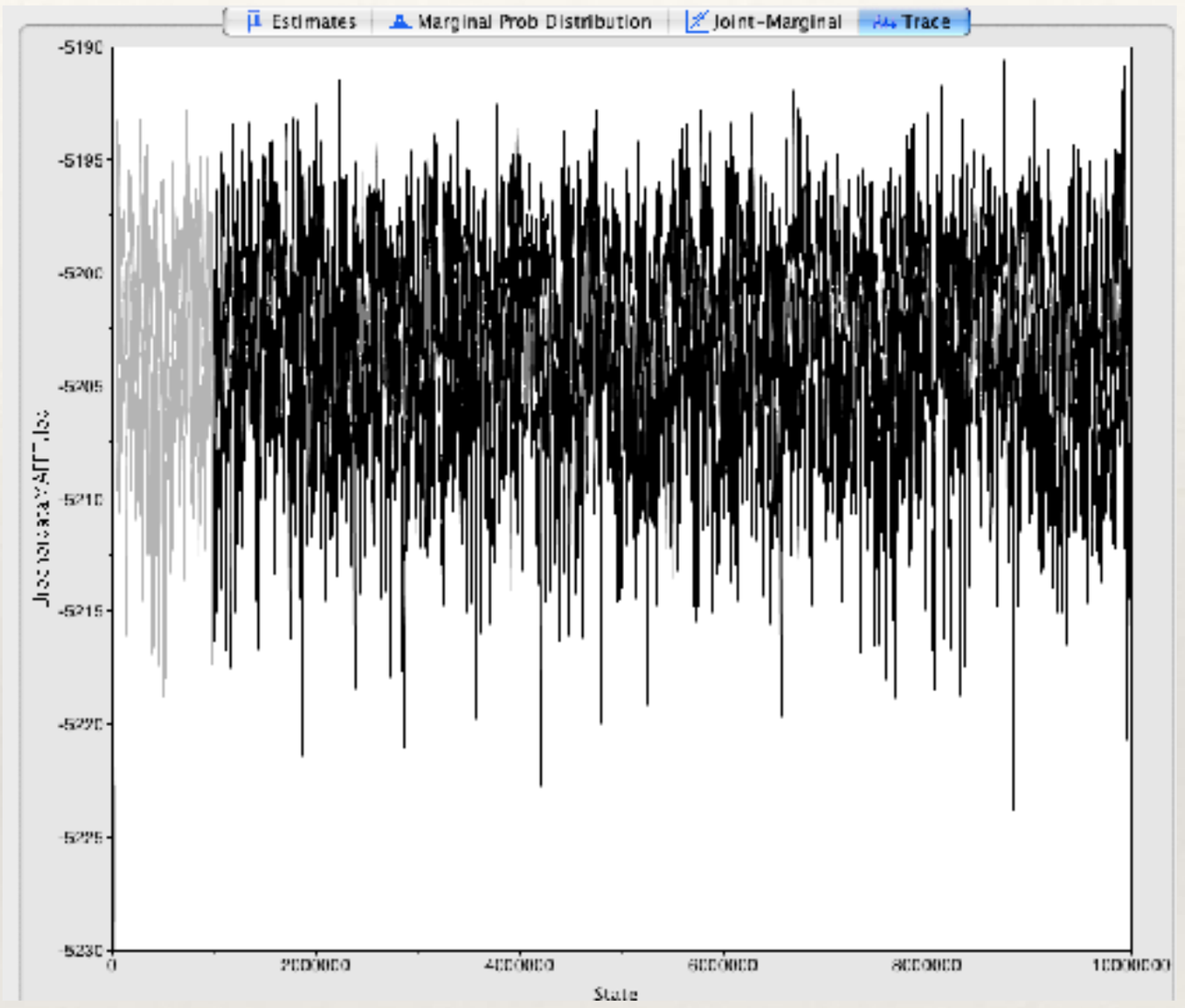
$p(\theta \mid D)p(\theta)$ ← Original proposal


10
8
6
4
2
0
Slightly downhill steps are usually accepted because R is near 1
Currently at 6.2 m
Proposed at 5.7 m
R = 5.7/6.2 =0.92
Drastic "off the cliff" downhill steps are almost never accepted because R is near 0
Currently at 6.2 m
Proposed at 0.2 m
R = 0.2/6.2 = 0.03
Currently at 1.0 m
Proposed at 2.3 m
R = 2.3/1.0 = 2.3
The robot takes a step if a Uniform(0,1) random deviate ≤ R
Uphill steps are always accepted because R > 1
Metropolis et al. 1953. Equation of state calculations by fast computing machines. J. Chem. Physics 21(6):1087-1092.

Lewis (2018)

Are we there yet?





Not reached convergence (ESS = 7)

Likely has reached convergence (ESS = 2084)

We can also use some convergence summary stats such as estimated sample size (ESS). A “good” sample of the posterior is an ESS > 200

Quick word about the dataset

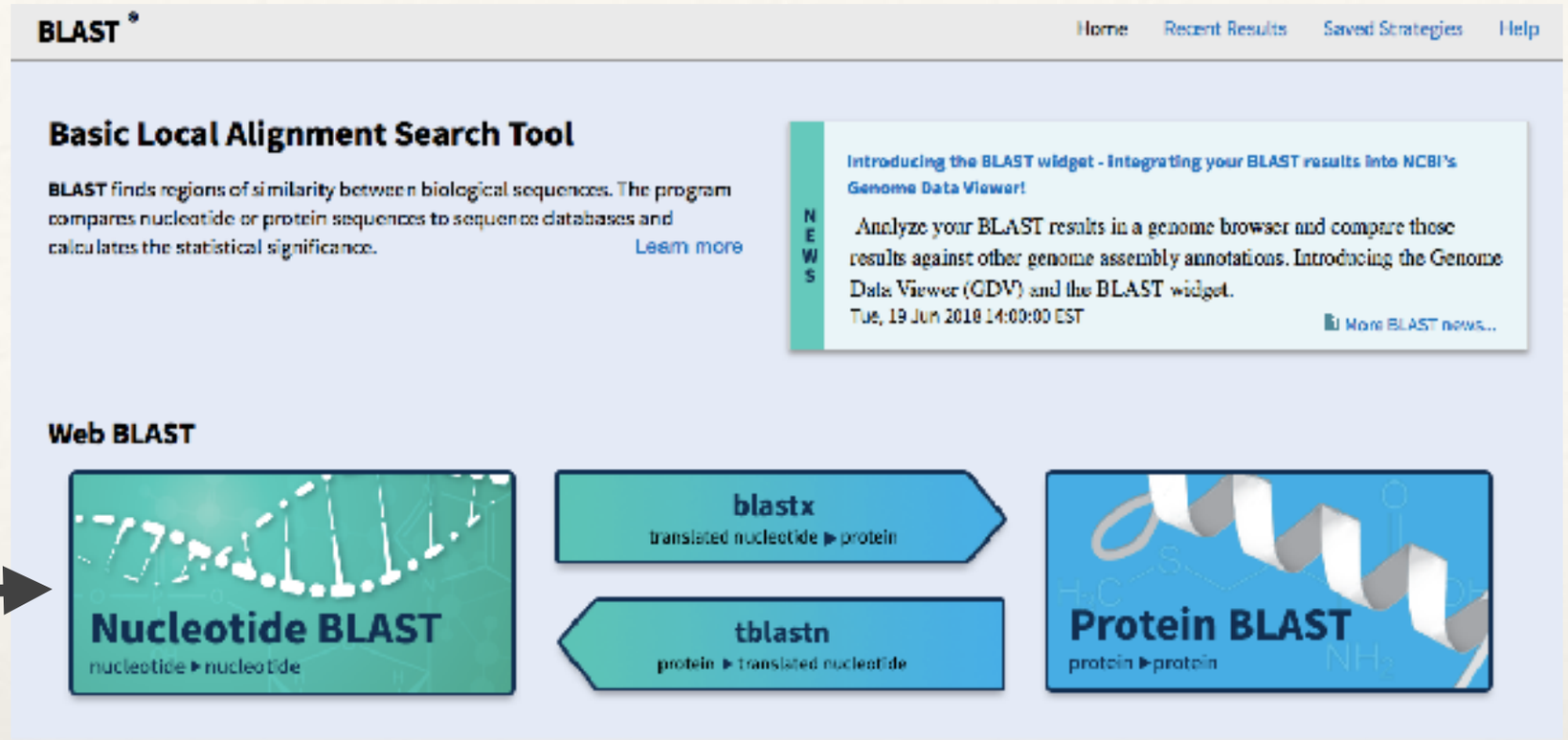
Hancock (2017)

Walkthrough

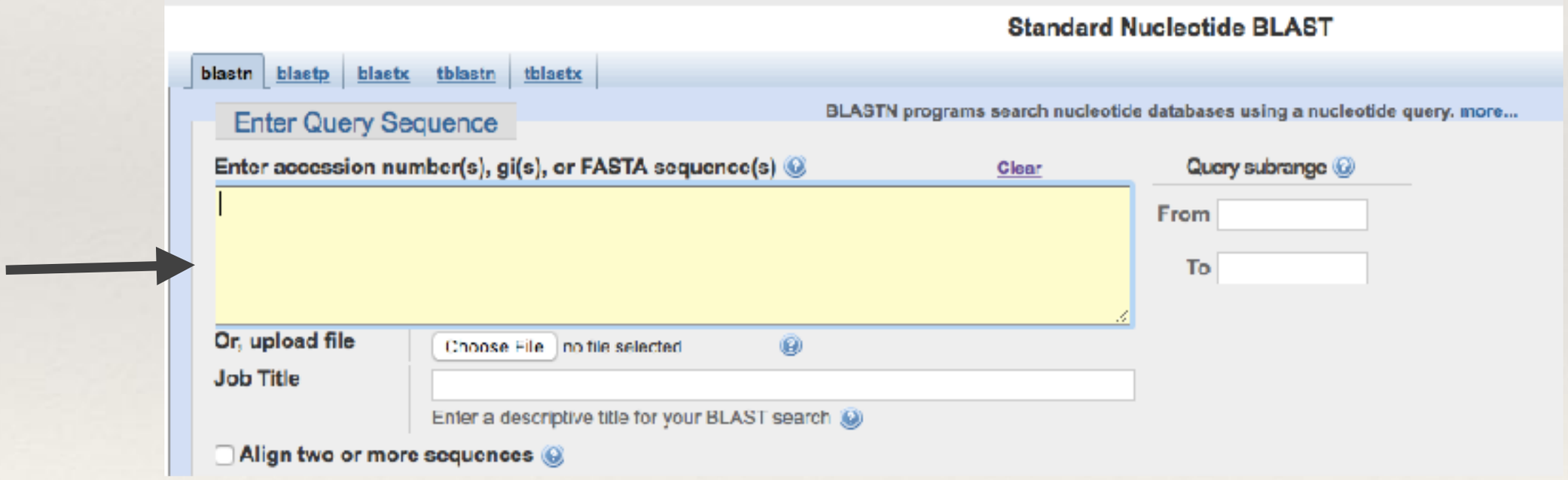
Open the file “RawCO1seqs.txt” in your text editor. This file is in **FASTA** format:

“>Name [ENTER]
Sequence [ENTER]”

Highlight the first sequence in the file (A1_CO1) and copy it to the clipboard.

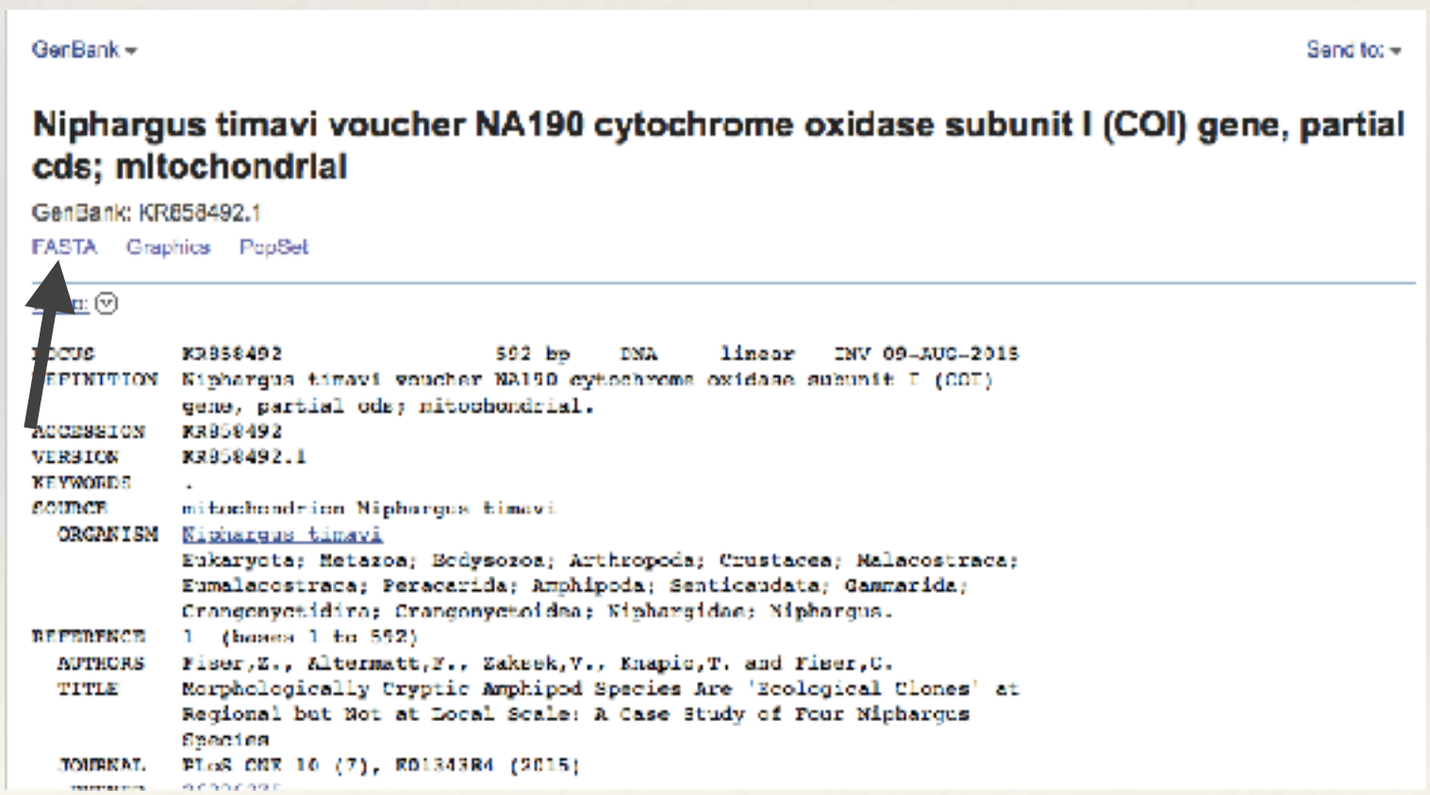


Go to https://blast.ncbi.nlm.nih.gov/Blast.cgi and select “Nucleotide BLAST”



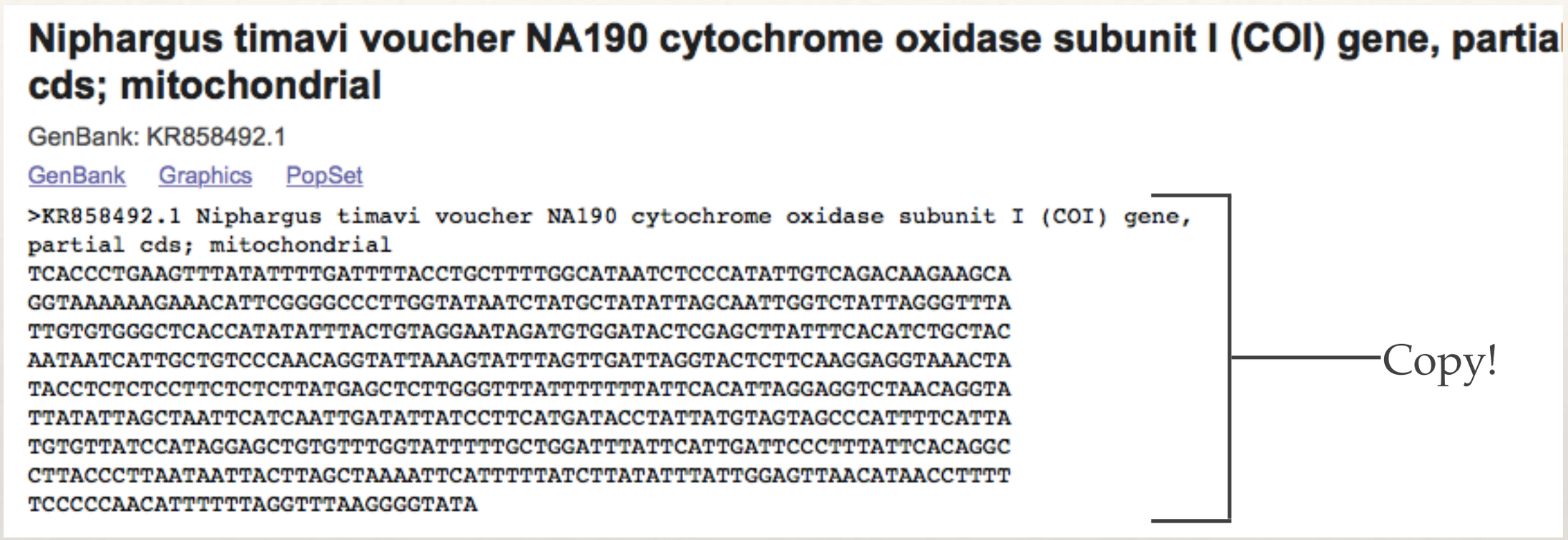
Paste the sequence you copied into the box above, then scroll down and hit BLAST

Scroll down until you see the sequences that best matched the query. The first sequence should be *Niphargus timavi*. Click on the blue link on the far right under the tab "Accession".



Below the identifier is another link that reads "FASTA". Click on this link to view the sequence in a FASTA format.

Highlight the sequence and the identifier and copy it; return to the RawCO1seqs.txt file and paste it just beneath the last sequence. **Be sure there are no extra spaces!** Save the file as RawCO1seqsout.txt to indicate it now has an outgroup.

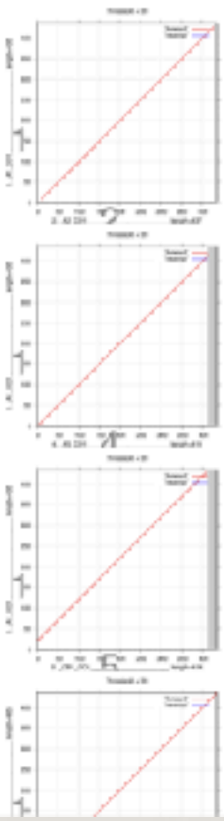


Once you've added the outgroup to the file, go to https://mafft.cbrc.jp/alignment/server/, which is the online version of MAFFT for sequence alignment.

Beneath the box where it says "Choose file", select the RawCO1seqsout.txt file. Scroll down and hit "Submit".

LAST hits (score>39) between the top sequence and the others.

Open all plots



Clustal format | Fasta format | MAFFT result | View | Tree | Refine dataset | Return to home

View

Reformat to GCG, PHYLIP, MSF, NEXUS, uppercase/lowercase, etc. with Readseq

GUIDANCE2 computes the residue-wise confidence scores and extracts well-aligned residues.

Refine dataset

Phylogenetic tree

MAFFT-L-INS-i Result

```
CLUSTAL format alignment by MAFFT (v7.402)


A1_CO1          ------------------------------------------------------------
A4_CO1          ------------------------------------------------------------
T7_CO1          ------------------------------------------------------------
A5_CO7          ------------------------------------------------------------
A-10_CO1        ------------------------------------------------------------
A-9_CO1         ------------------------------------------------------------
ST-4_CO1        ------------------------------------------------------------
T2_CO7          ------------------------------------------------------------
T4_CO1          ------------------------------------------------------------
ST-8_CO1        ------------------------------------------------------------
T3_CO1          ------------------------------------------------------------
T8_CO1          ------------------------------------------------------------
T10_CO1         ------------------------------------------------------------
T5_CO1          ------------------------------------------------------------
```

Your readout should look like the one above. Note that MAFFT produces the alignment in CLUSTAL format. Copy this readout, and then open a new text document in your text editor. Paste the readout and save it as “CO1MAFFT.txt”.

Now, open Mesquite. Go to “File” and select “Open file,” and choose “CO1MAFFT.txt”. A new window will open prompting you to choose what format the file is in. Select “Clustal (DNA/RNA)”.

A second window will open asking you if you want to save the file as “CO1MAFFT.txt.nex”, select “OK” and the alignment should be displayed.

Slide the bar over to the 105th position—the should be the last position of gaps for all in-group taxa. Slide to the end of the alignment, position 499. We need to exclude positions 1–105 and 499–703.

To do this, open the file CO1MAFFT.txt.nex in your text editor.

Scroll down to the end of the alignment—you should see a line that reads “END;” Click next to the semicolon and hit ENTER twice. Now, add:

exclude 1-102 505-703;

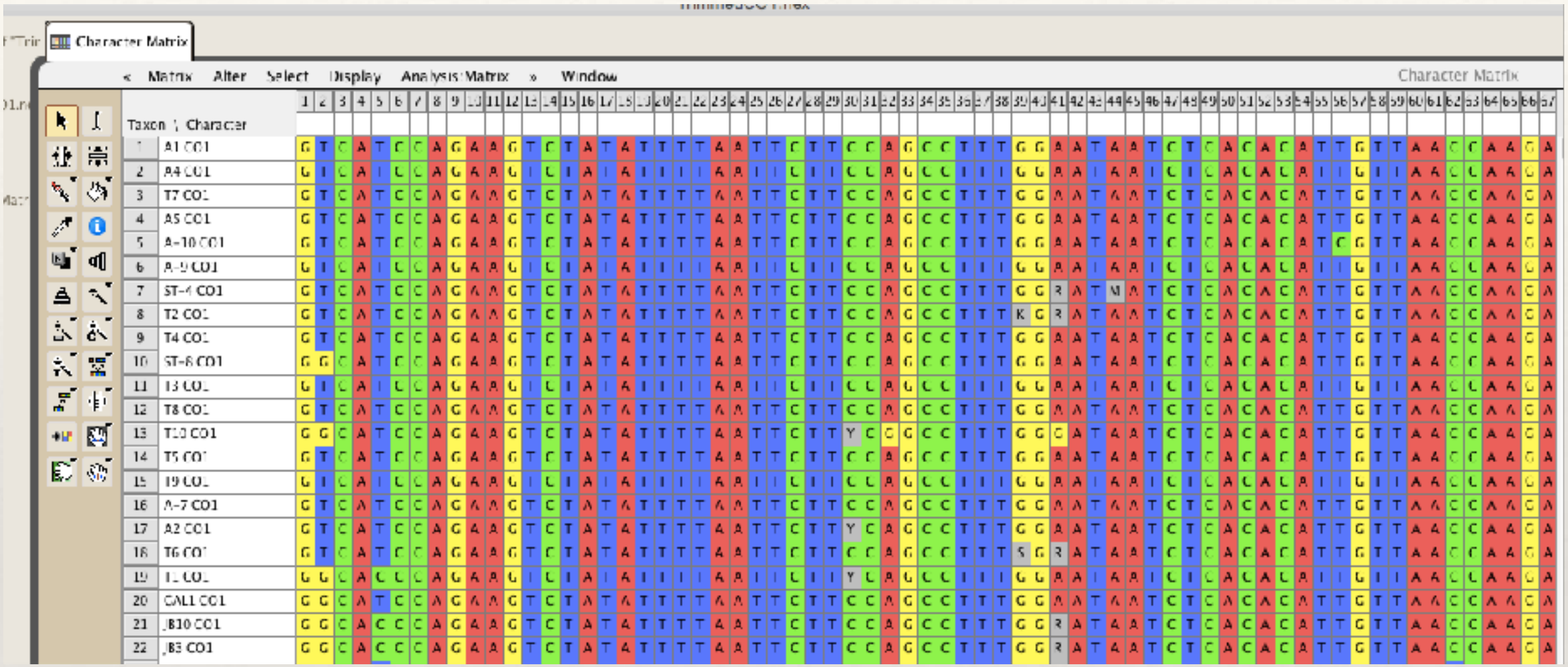
Be sure to add this *exactly* as typed above, with a single space between each and a semicolon at the end. Save the file. Now, open PAUP*. Click File —> Open —> and choose “CO1MAFFT.txt.nex”

```
Running on IA-32 architecture (64-bit word length)
SSE vectorization enabled
SSSE3 instructions supported
Multithreading enabled for likelihood using Pthreads
Compiled using Intel compiler (icc) 11.1.0 (build 20091012)

Processing of file "~/Desktop/0505Workshop/CO1MAFFT.txt.nex" begins...
Data read in DNA format

Data matrix has 81 taxa, 703 characters
Valid character-state symbols: ACGT
Missing data identified by '?'
Gaps identified by '-'
"Equate" macros in effect:
   R,r ==> {AG}
   Y,y ==> {CT}
   M,m ==> {AC}
   K,k ==> {GT}
   S,s ==> {CG}
   W,w ==> {AT}
   H,h ==> {ACT}
   B,b ==> {CGT}
   V,v ==> {ACG}
   D,d ==> {AGT}
   N,n ==> {ACGT}

Character-exclusion status changed:
  310 characters excluded
  Total number of characters now excluded = 310
  Number of included characters = 393

Character types changed:
  310 characters are excluded
  Of the remaining 393 included characters:
    All characters are of type 'unord'
    All characters have equal weight

NOTE: PAUP* does not support the "CodeSet" command.  It has been skipped.

*** Skipping "MESQUITECHARMODELS" block

*** Skipping "MESQUITE" block
```

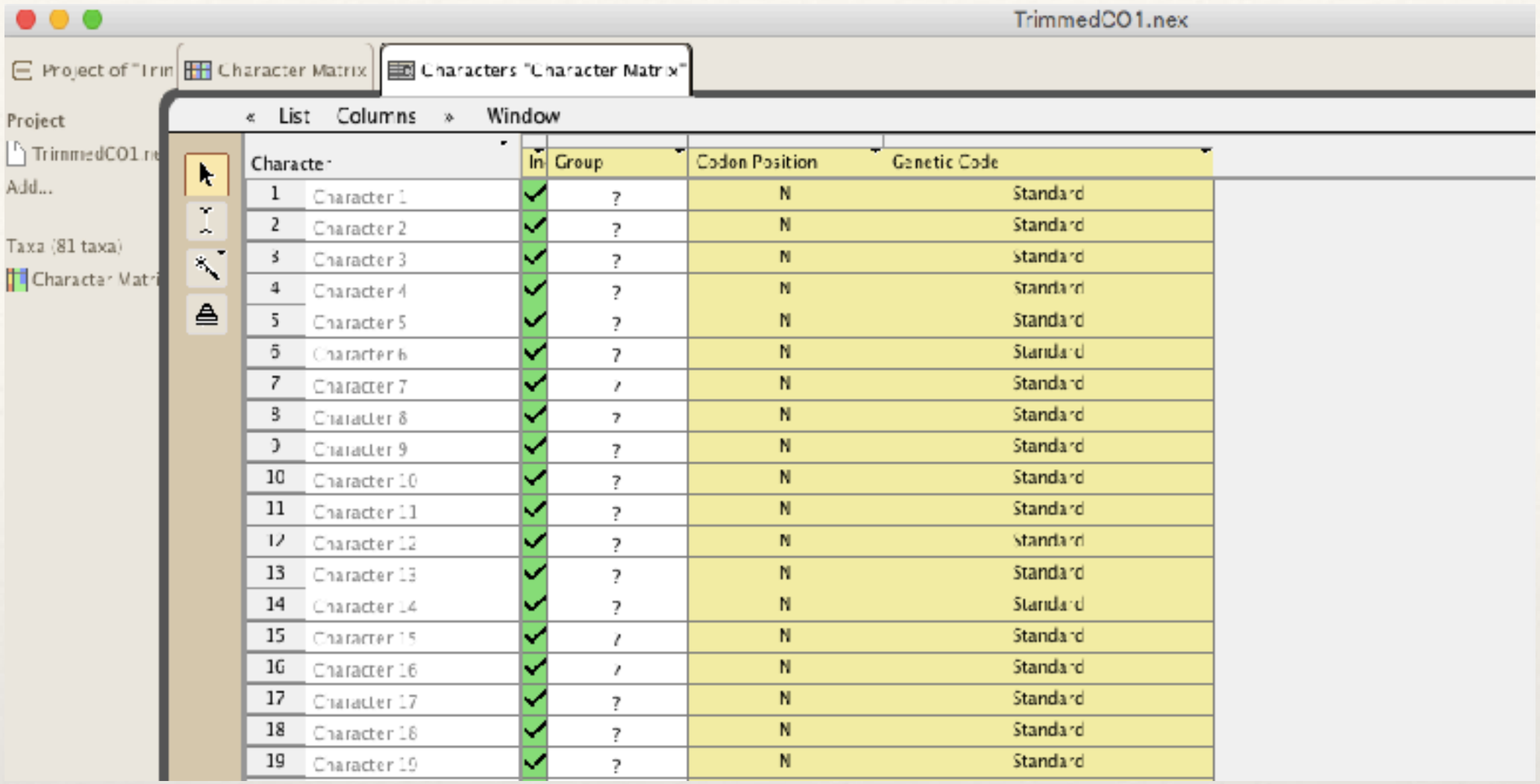
You should get a readout like the one here, showing the number of characters excluded (301) and the new number of characters (402).

Go to File —> Export Data —> Choose to export the file as a “NEXUS” and save it as “TrimmedCO1.nex”

Return to Mesquite. Open the file TrimmedCO1.nex, it should look like:



Now, at the top of the screen, go to the tab that reads “Matrix” and scroll down to select “Genetic Code”

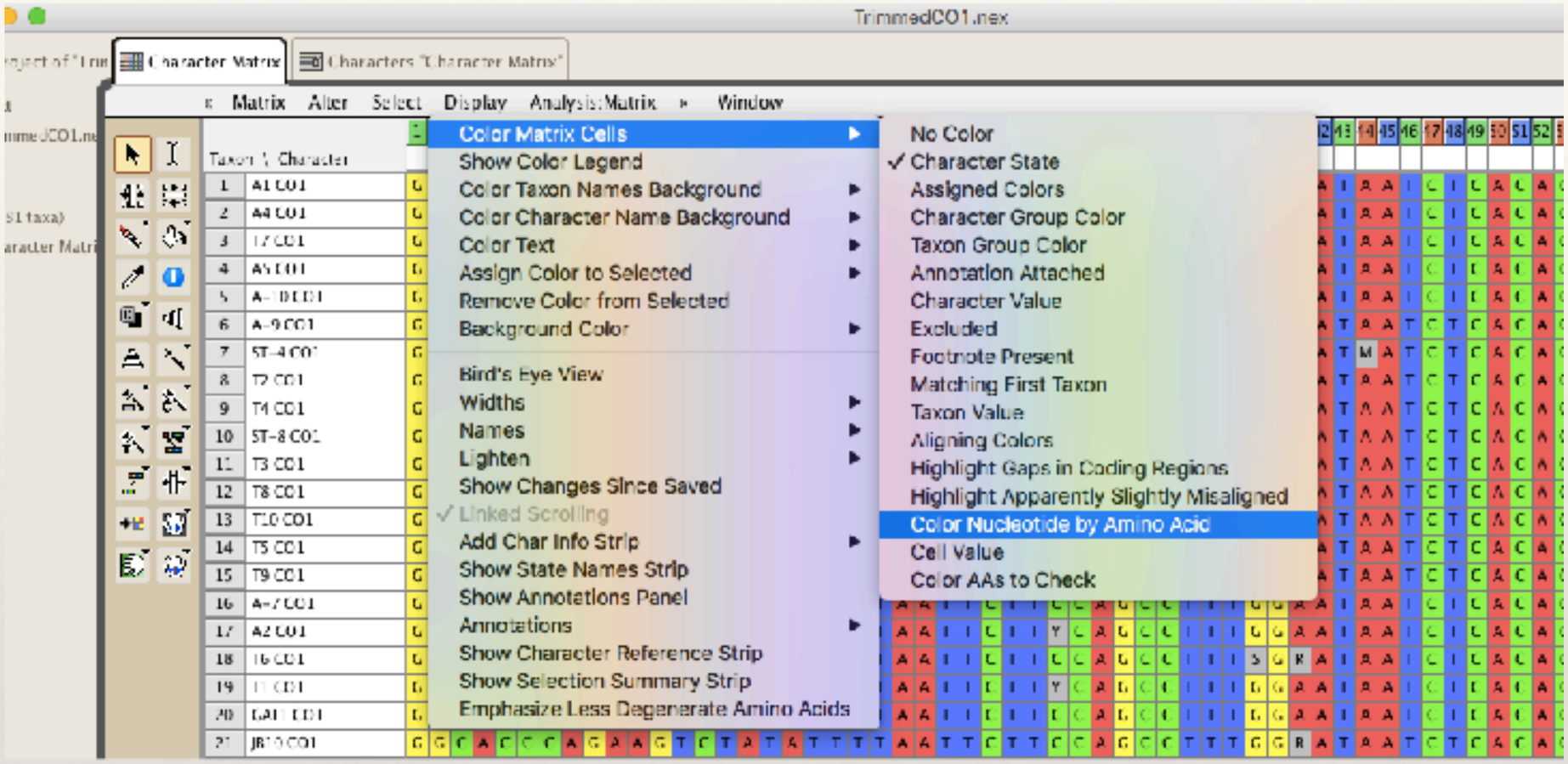


Highlight the entire block using COMMAND+A (or CTRL+A). Then select "Genetic Code", choose "Invertebrate Mitochondria".

Next, select the tab Codon Position —> Set Codon Position —> Minimize Stop Codons

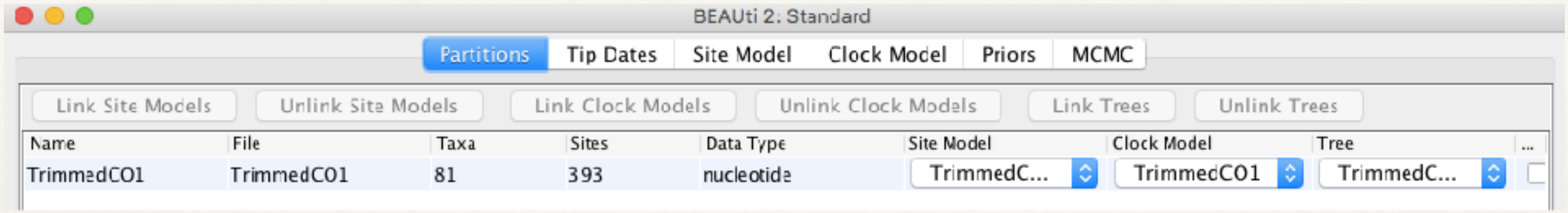
Return to the sequence matrix.

Directly above the matrix is a tab that says “Display”. Select it, choose Color Matrix Cells —> Color Nucleotide by Amino Acid

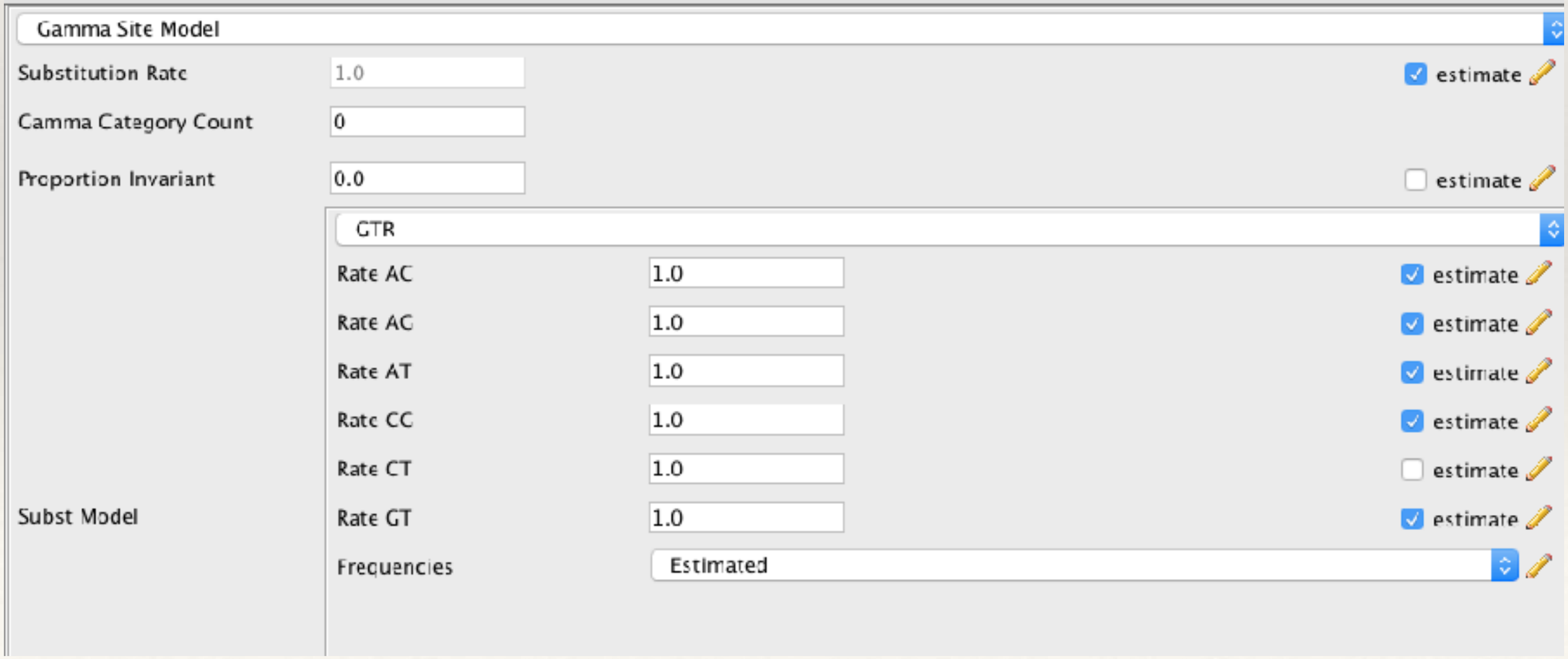


This will change the color of the matrix to match the amino acid coded for. Slide through the matrix to check for black codons—this indicates a premature stop codon (there should not be any!).

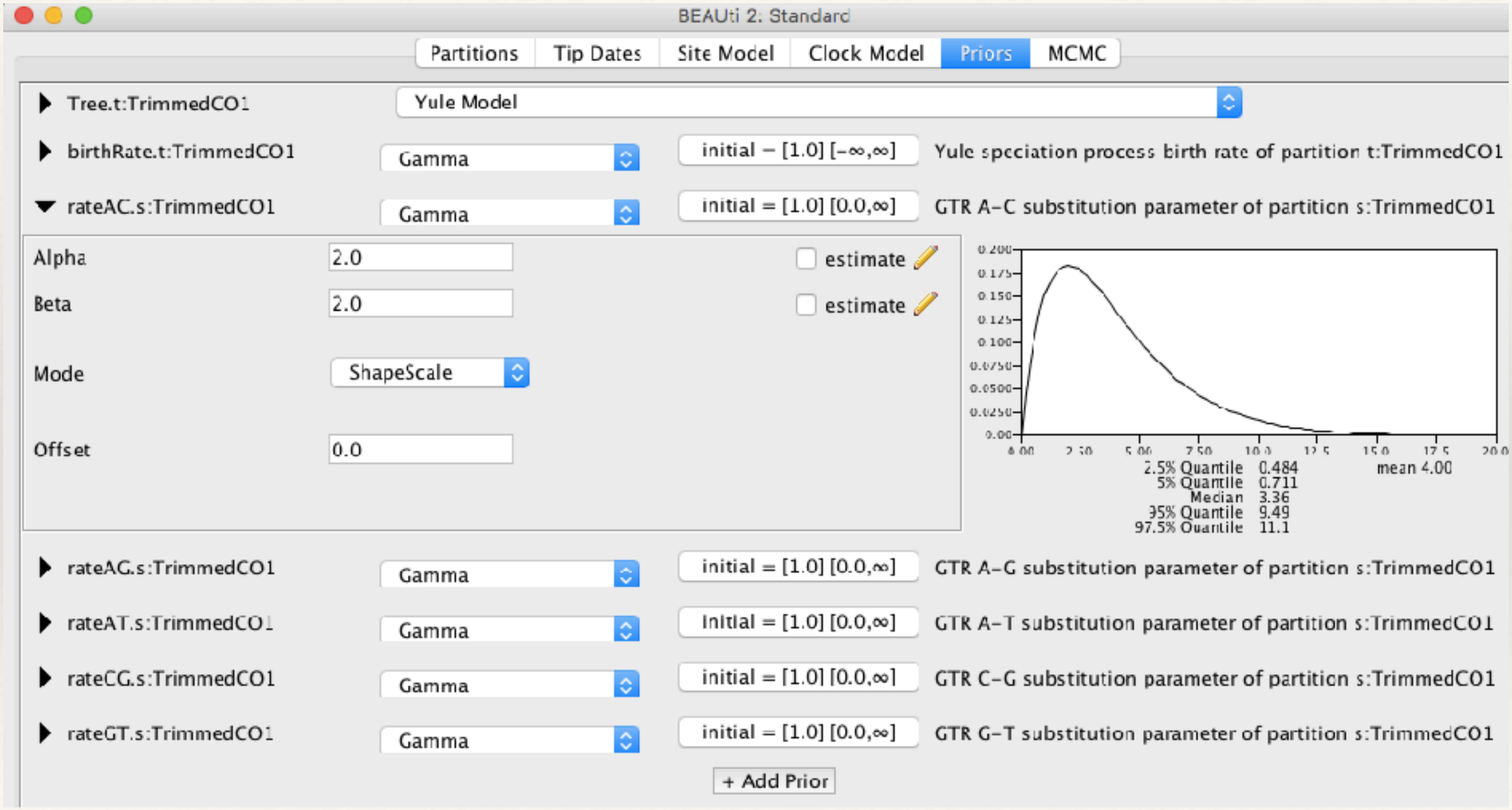
Next, open BEAUti. Select “Import Alignment” and choose TrimmedCO1.nex.



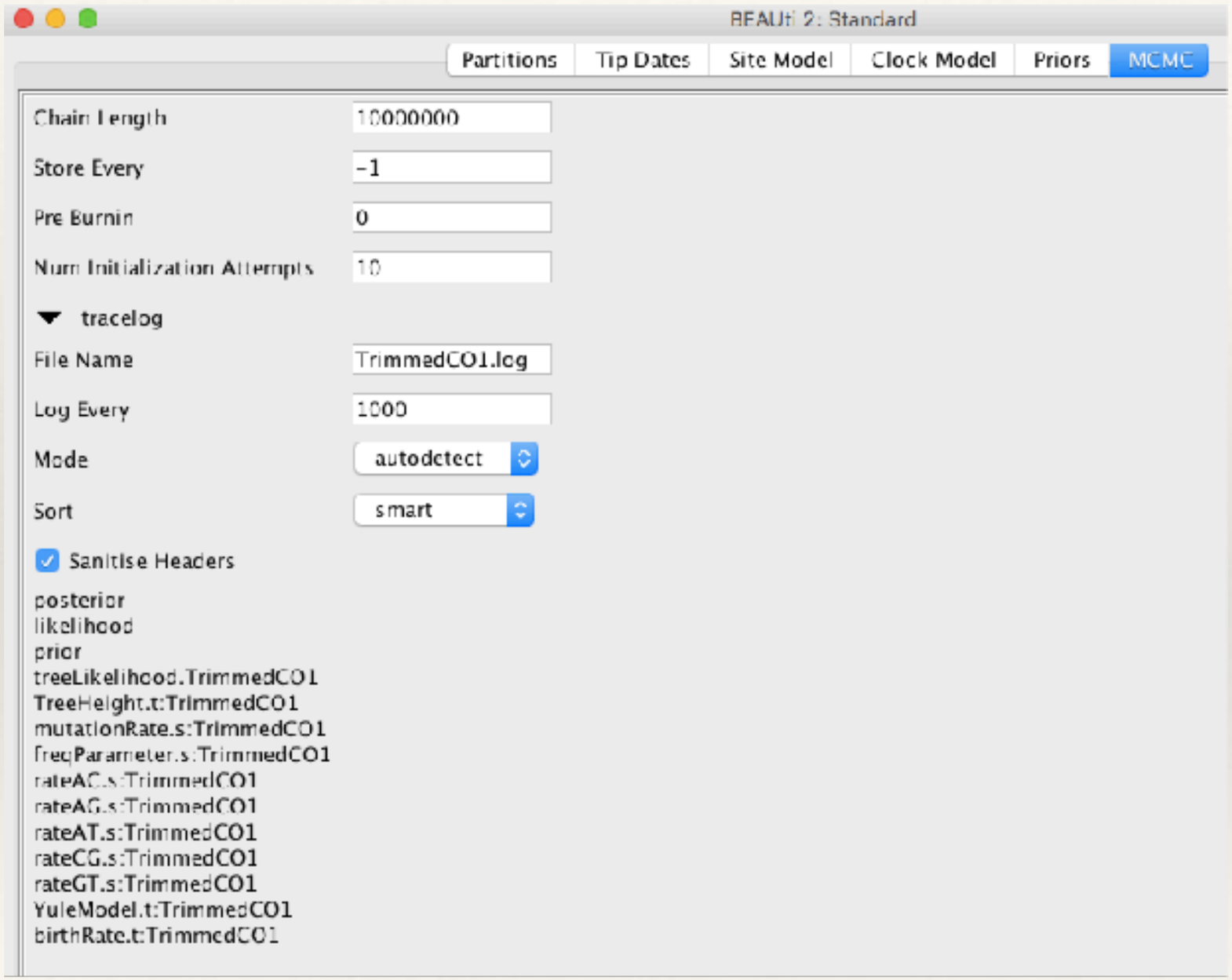
Go over to the tab that says “Site Model”. Check the box that says “estimate” beside “Substitution Rate” and then scroll down and select “GTR” as the substitution model.



Scroll over to the “Priors” tab. We will keep the model the same here, but click on the arrows beside each rate prior to see how different α and β values affect the sampling distribution of the prior. You can also change the shape from gamma.



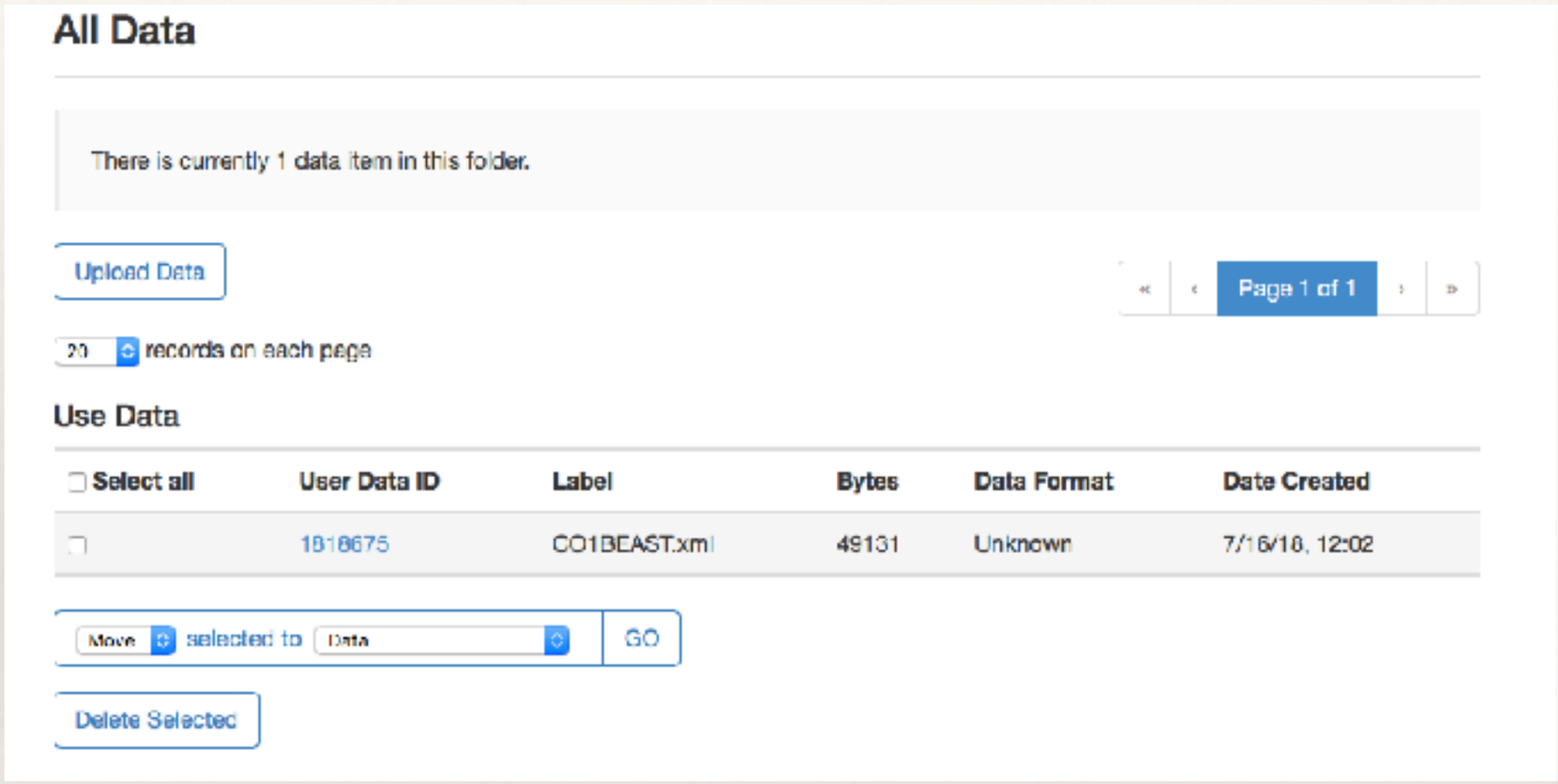
Now go to the MCMC tab to set-up the run. The default is 10,000,000 generations. Click on the arrow next to the tab that says “tracelog”—here, you’ll see that the run will log the posterior every 1,000 generations.



Go the File —> Save —> Save the file as “CO1BEAST.xml”.

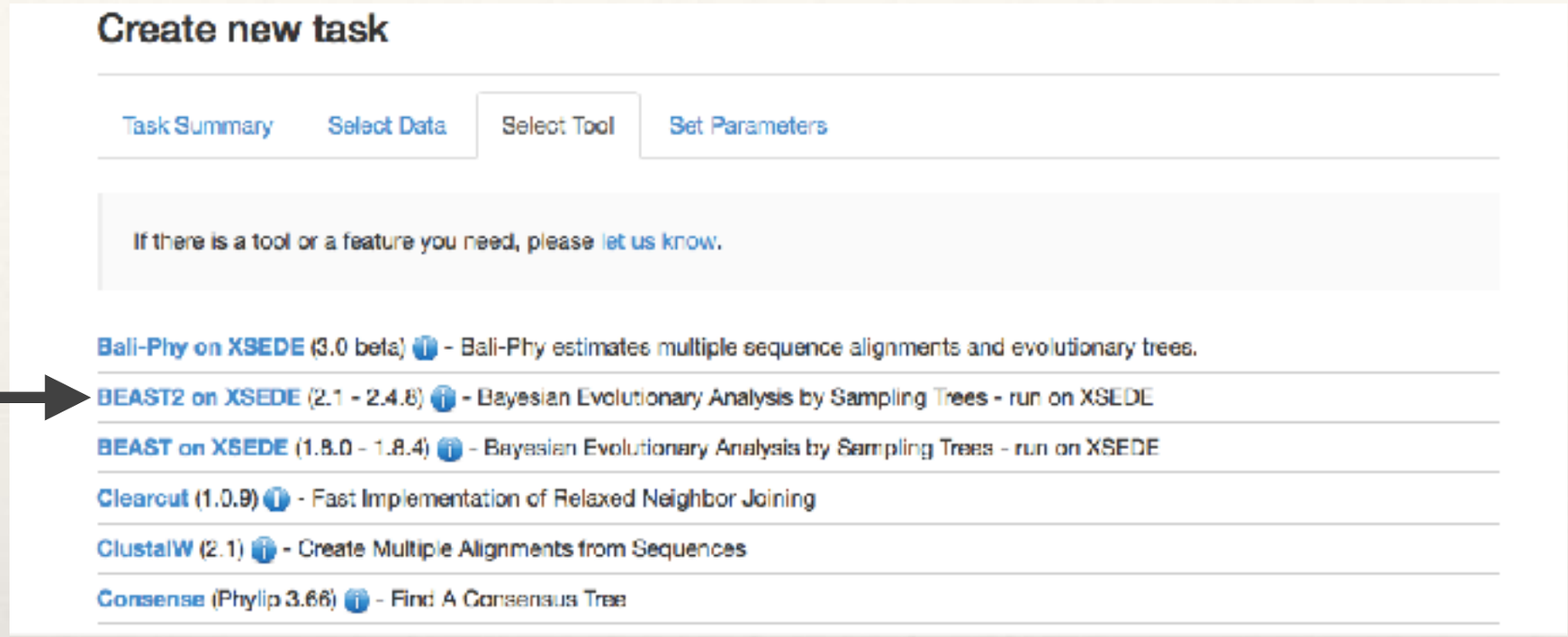
Now, go to https://www.phylo.org, which is the CIPRES portal. Sign-in to your account. Select the tab that says “Create New Folder”, and title it “OSOS”.

The folder should now appear in the left panel with two selectable tabs: Data (0) and Tasks (0). Select Data —> Enter Data —> Choose file —> Select the CO1BEAST.xml file and hit “Save”.



Now, go to the tab on the left that says “Tasks” and then select “Create New Task”.

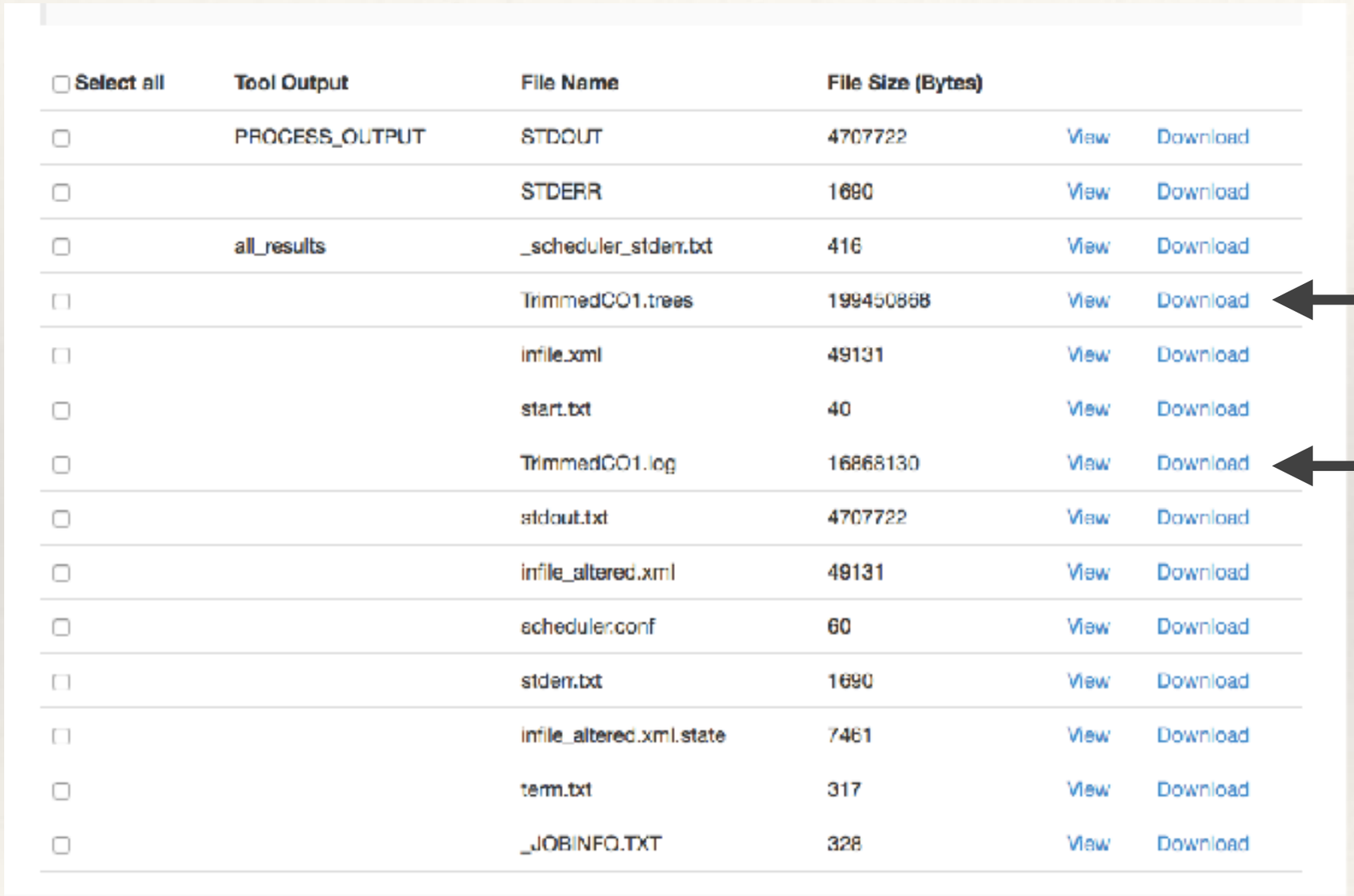
Name the task “CO1BEAST”. Below, choose the data file you just uploaded (should be the only option). Next, select the “Select Tool” button and choose BEAST2 on XSEDE as shown below.



Go to “Input Parameters” and type “**402**” into the tab indicated by the arrow on the left.

Select “Save and Run Task”.

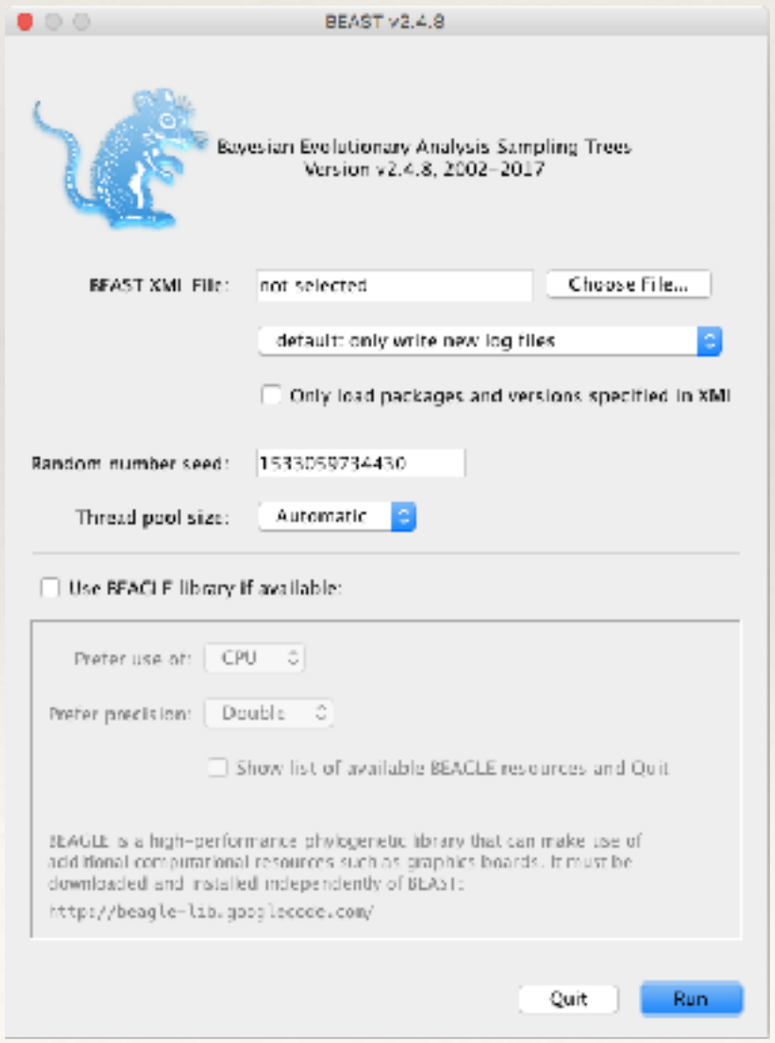
You can choose “View Status” and periodically select the “Refresh Task” button to keep up with the progress of the run. When it is finished, it will say “COMPLETED” beside Status. You will also receive an email stating that the task has terminated.



Select all	Tool Output	File Name	File Size (Bytes)		
	PROCESS_OUTPUT	STDOUT	4707722	View	Download
		STDERR	1690	View	Download
	all_results	_scheduler_stderr.txt	416	View	Download
		TrimmedCO1.trees	199450858	View	Download
		infile.xml	49131	View	Download
		start.txt	40	View	Download
		TrimmedCO1.log	16868130	View	Download
		stdout.txt	4707722	View	Download
		infile_altered.xml	49131	View	Download
		scheduler.conf	60	View	Download
		stderr.txt	1690	View	Download
		infile_altered.xml.state	7461	View	Download
		term.txt	317	View	Download
		_JOBINFO.TXT	328	View	Download

Once completed, download the files ending in “.trees” and “.log”

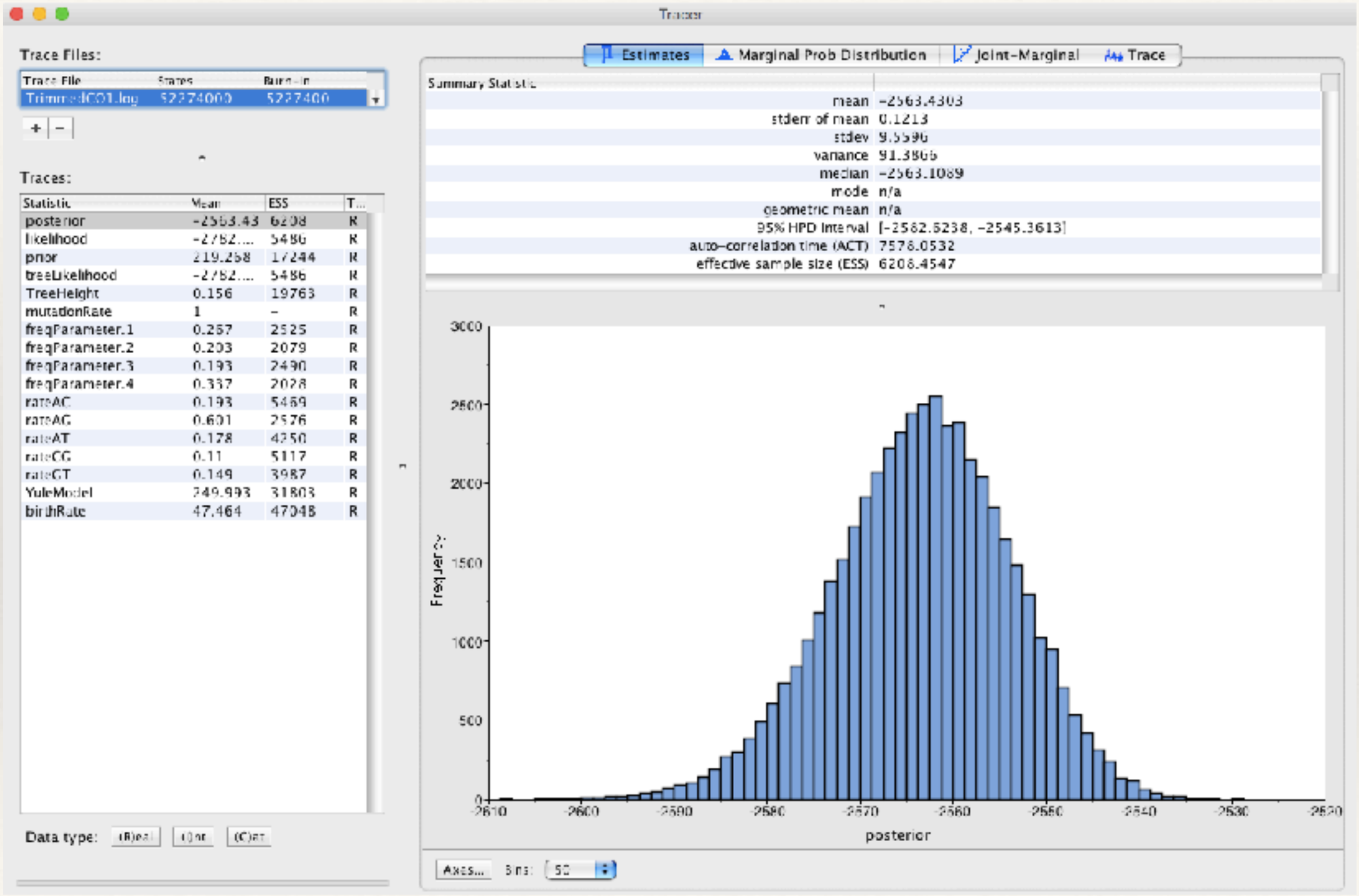
```
      9997000     -2693.5567       738.2       -2940.3613       246.8046 1m9s/Msamples
      9998000     -2689.9170       738.3       -2936.9200       247.0029 1m9s/Msamples
      9999000     -2677.8844       737.9       -2931.1241       253.2396 1m9s/Msamples
     10000000     -2700.4753       738.7       -2948.7471       248.2718 1m9s/Msamples

Operator                                                 Tuning    #accept    #reject      Pr(m)  Pr(acc|m)
ScaleOperator(YuleBirthRateScaler.t:TrimmedCO1)          0.5146      98373     289153     0.0387     0.2538
ScaleOperator(YuleModelTreeScaler.t:TrimmedCO1)          0.7190      78896     307523     0.0387     0.2042
ScaleOperator(YuleModelTreeRootScaler.t:TrimmedCO1)      0.5783      42396     344641     0.0387     0.1095
Uniform(YuleModelUniformOperator.t:TrimmedCO1)                -    2340484    1525729     0.3866     0.6054
SubtreeSlide(YuleModelSubtreeSlide.t:TrimmedCO1)         0.4813       8329    1922195     0.1933     0.0043 Try decreasing size to about 0.241
Exchange(YuleModelNarrow.t:TrimmedCO1)                        -     714790    1220668     0.1933     0.3693
Exchange(YuleModelWide.t:TrimmedCO1)                          -       3337     382768     0.0387     0.0086
WilsonBalding(YuleModelWilsonBalding.t:TrimmedCO1)            -       4407     381641     0.0387     0.0114
DeltaExchangeOperator(FixMeanMutationRatesOperator)   7552.2576     257523          0     0.0258     1.0000 Try setting delta to about 15104.515
DeltaExchangeOperator(FrequenciesExchanger.s:TrimmedCO1)  0.0706       3666       9345     0.0013     0.2818
ScaleOperator(RateACScaler.s:TrimmedCO1)                 0.3798       3573       9031     0.0013     0.2835
ScaleOperator(RateAGScaler.s:TrimmedCO1)                 0.4724       2970      10026     0.0013     0.2285
ScaleOperator(RateATScaler.s:TrimmedCO1)                 0.4027       3219       9632     0.0013     0.2505
ScaleOperator(RateCGScaler.s:TrimmedCO1)                 0.2386       3622       9174     0.0013     0.2842
ScaleOperator(RateGTScaler.s:TrimmedCO1)                 0.3588       3500       9450     0.0013     0.2703

     Tuning: The value of the operator's tuning parameter, or '-' if the operatEnd likelihood: -2700.4753305776576
or can't be optimized.
    #accept: The total number of times a proposal by this operator has been accepted.
```

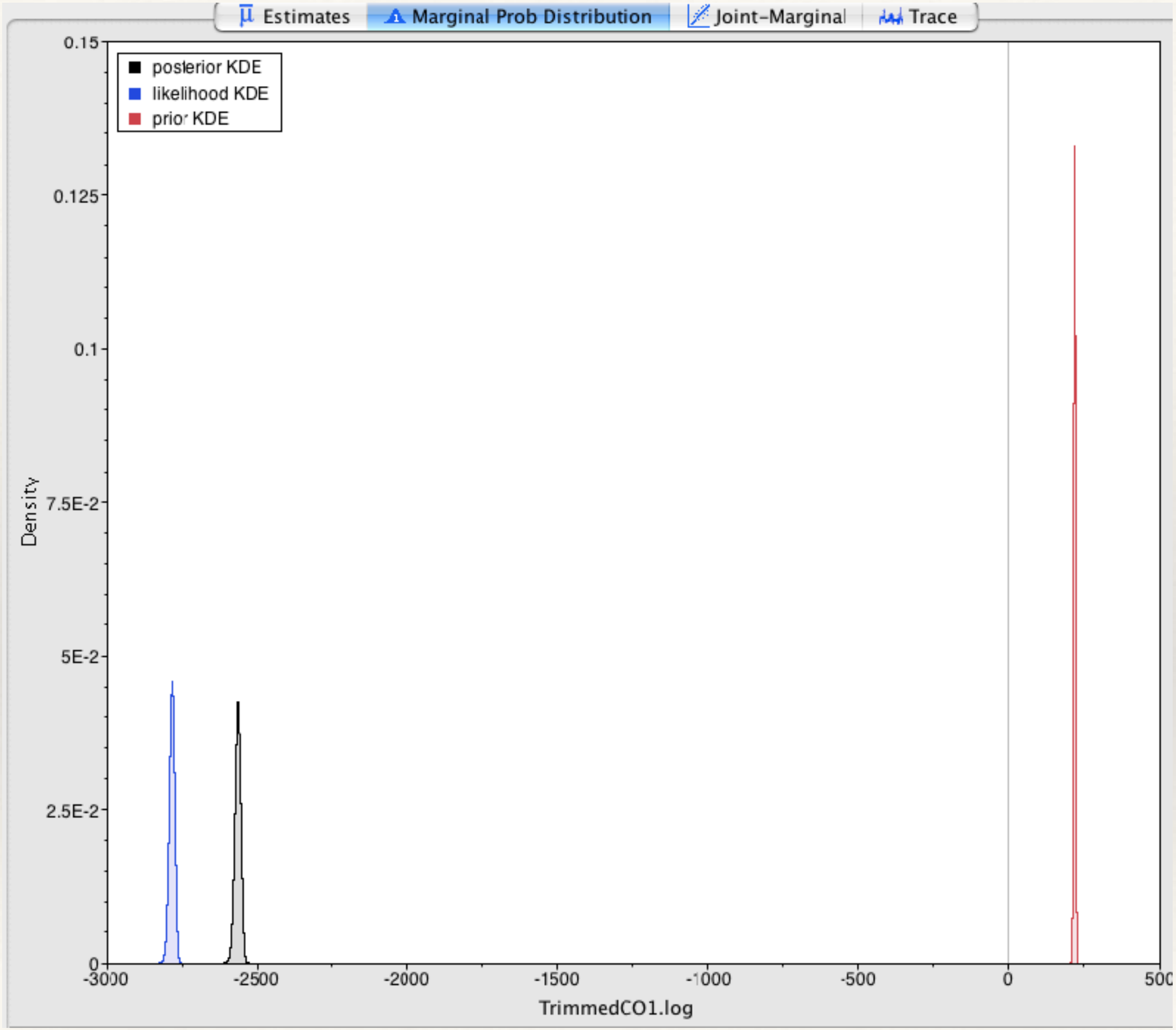


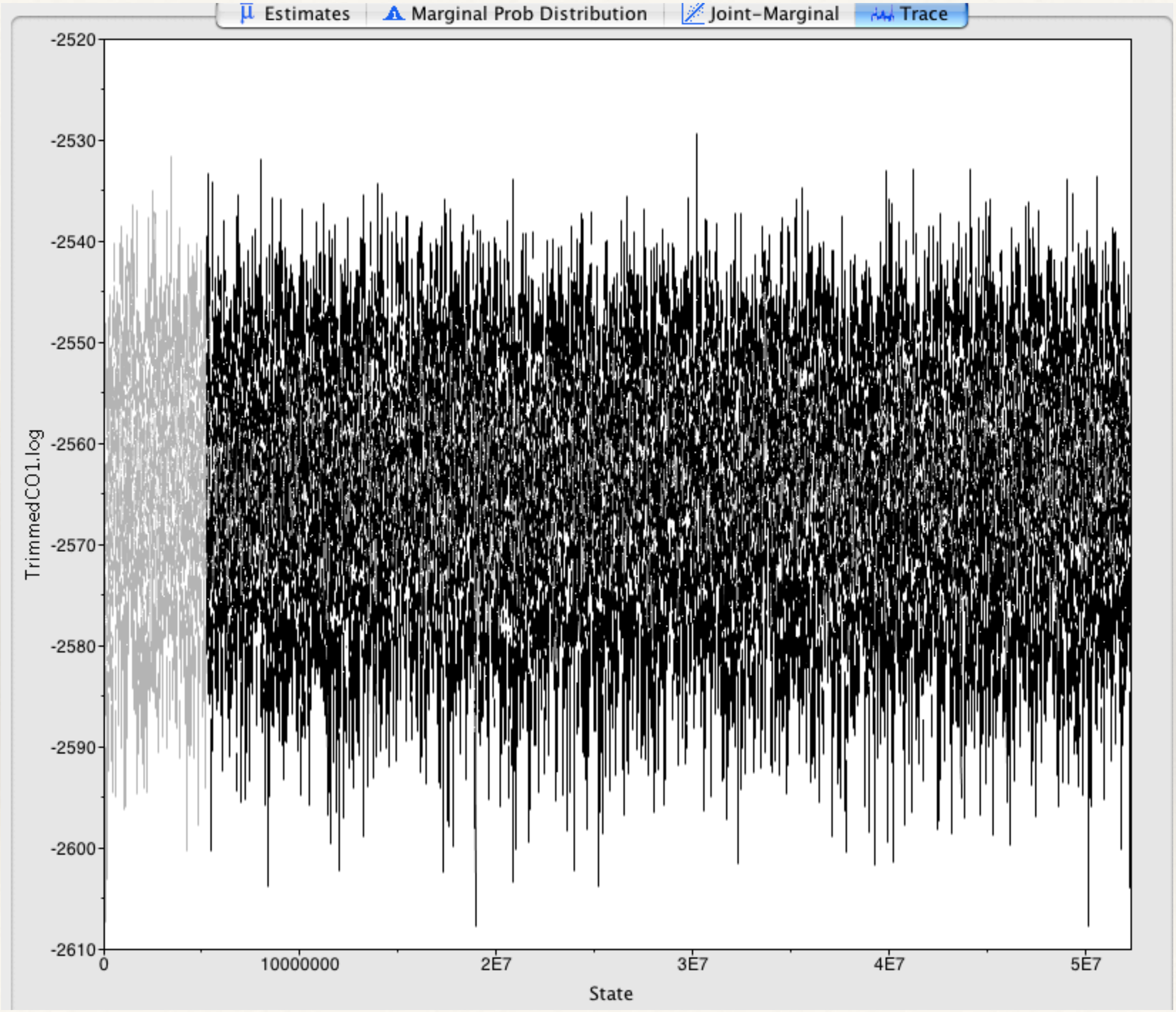
Running BEAST on your own computer will generate a read-out like the one above, allowing you to see each proposal made during the MCMC and how many of them were accepted or rejected.

For high rejections and/or high acceptance, the program will make “suggestions” as to how to even-out the MCMC.

Open TRACER, and import the file “TrimmedCO1.log”.

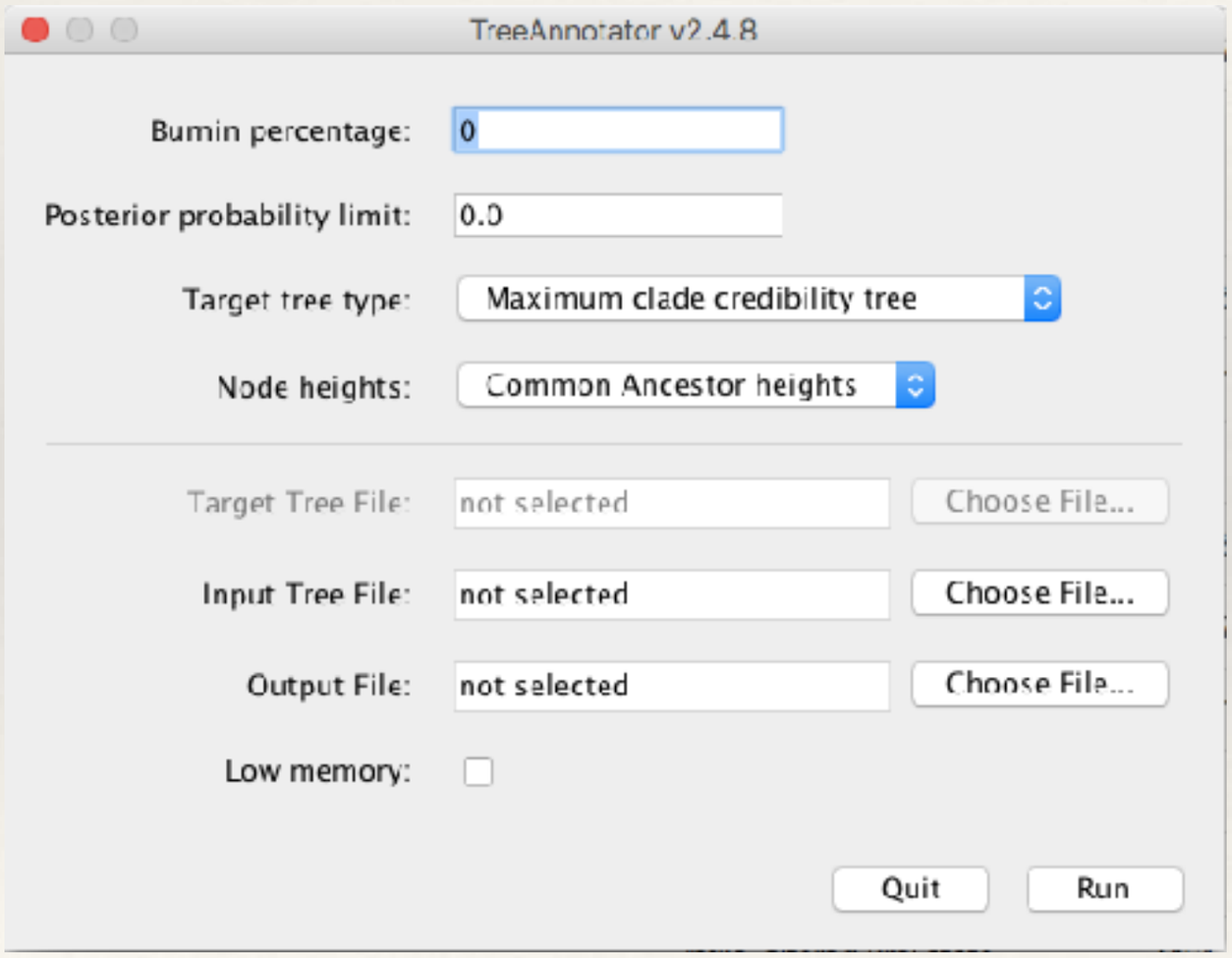


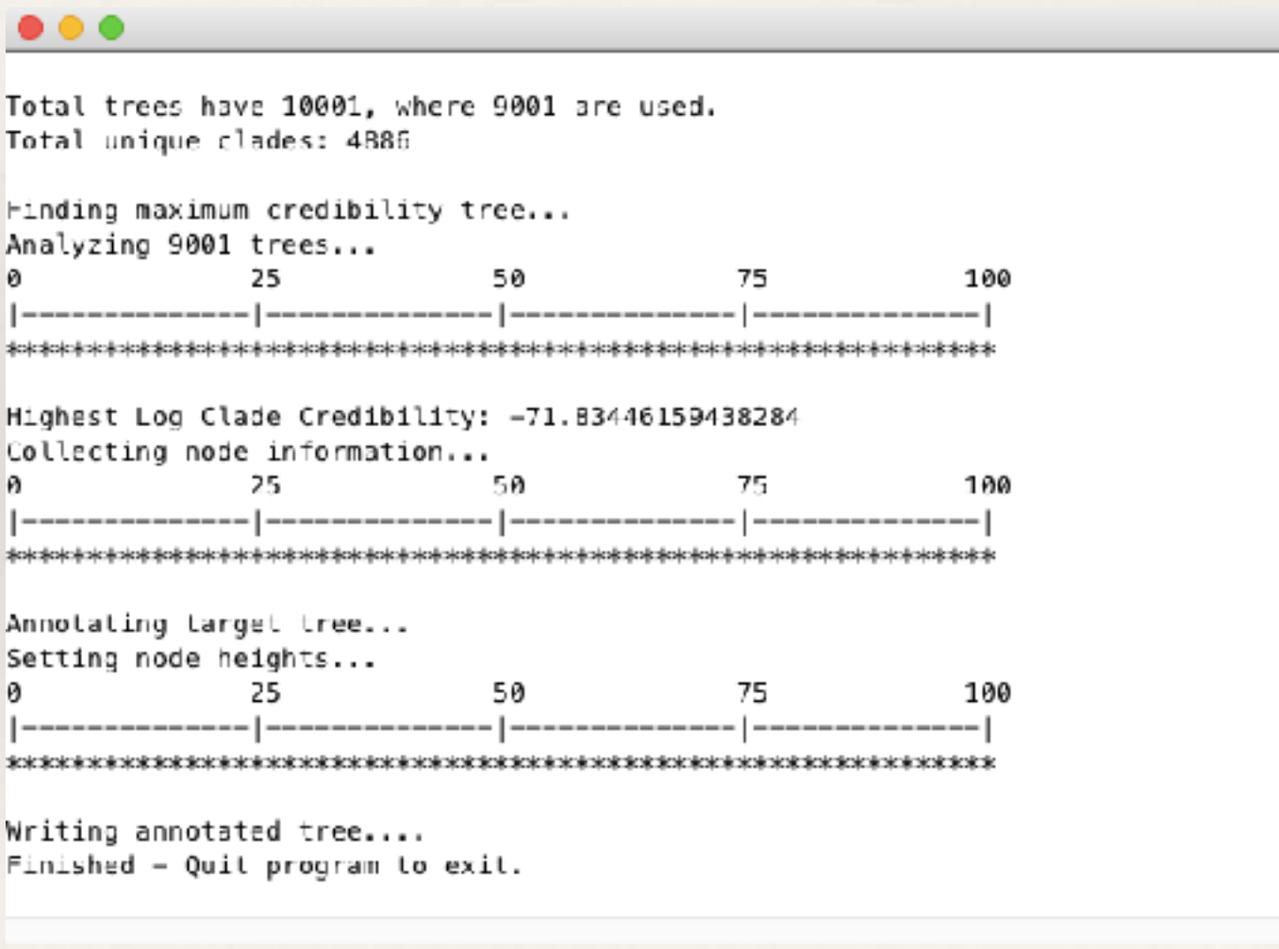

Estimates
Marginal Prob Distribution
Joint-Marginal
Trace
posterior KDE
likelihood KDE
prior KDE
0.15
0.125
0.1
7.5E-2
5E-2
2.5E-2
0
Density
-3000
-2500
-2000
-1500
-1000
-500
0
TrimmedCO1.log


Estimates
Marginal Prob Distribution
Joint-Marginal
Trace
-2520
-2530
-2540
-2550
-2560
-2570
-2580
-2590
-2600
-2610
TrimmedCO1.log
0
10000000
2E7
3E7
4E7
5E7
State

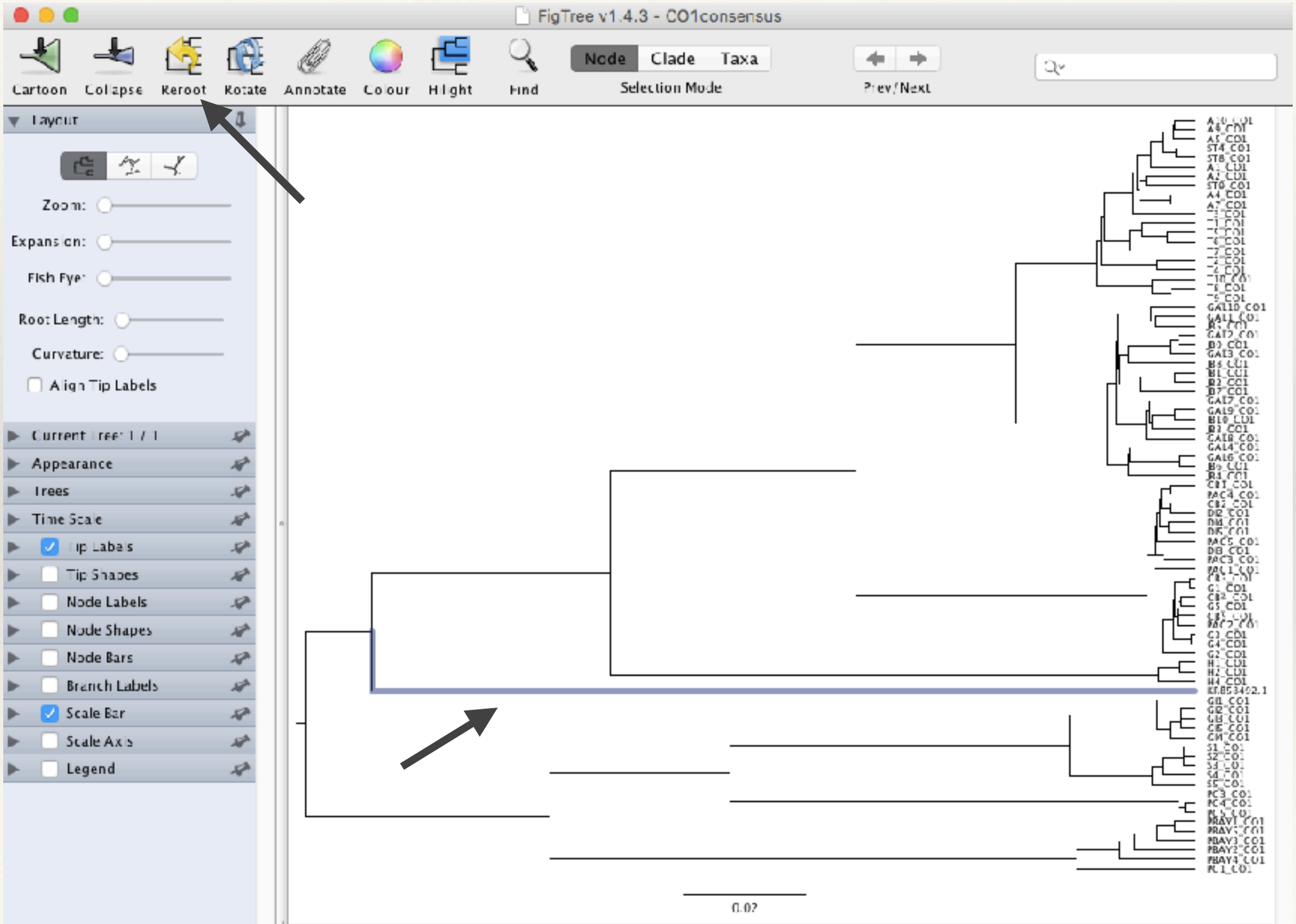
Open the TreeAnnotator app and upload the file “TrimmedCO1.trees”. The Burnin percentage is 10.

Click “Choose File” by the Output File tab and name the new file “CO1consensus”.



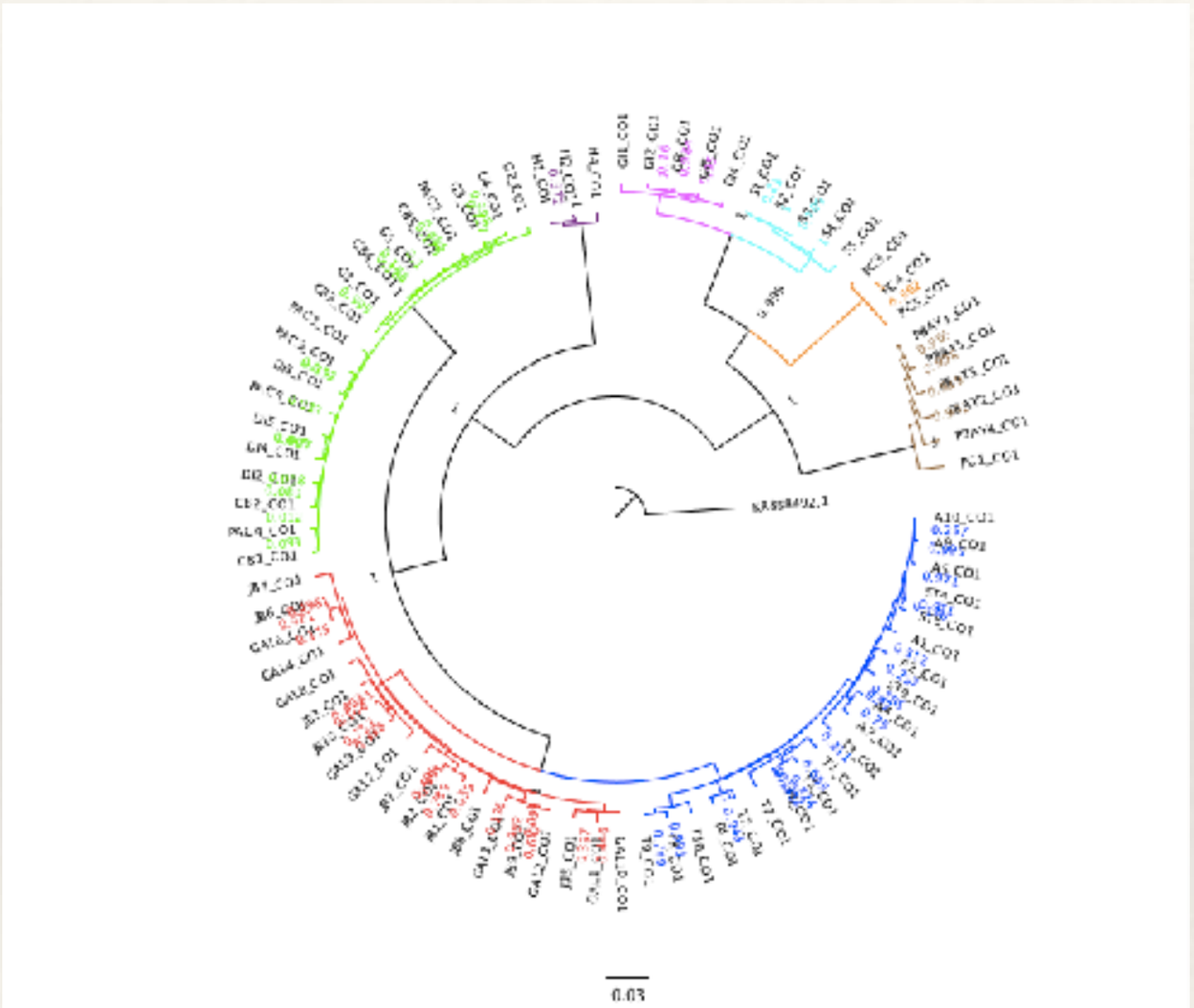


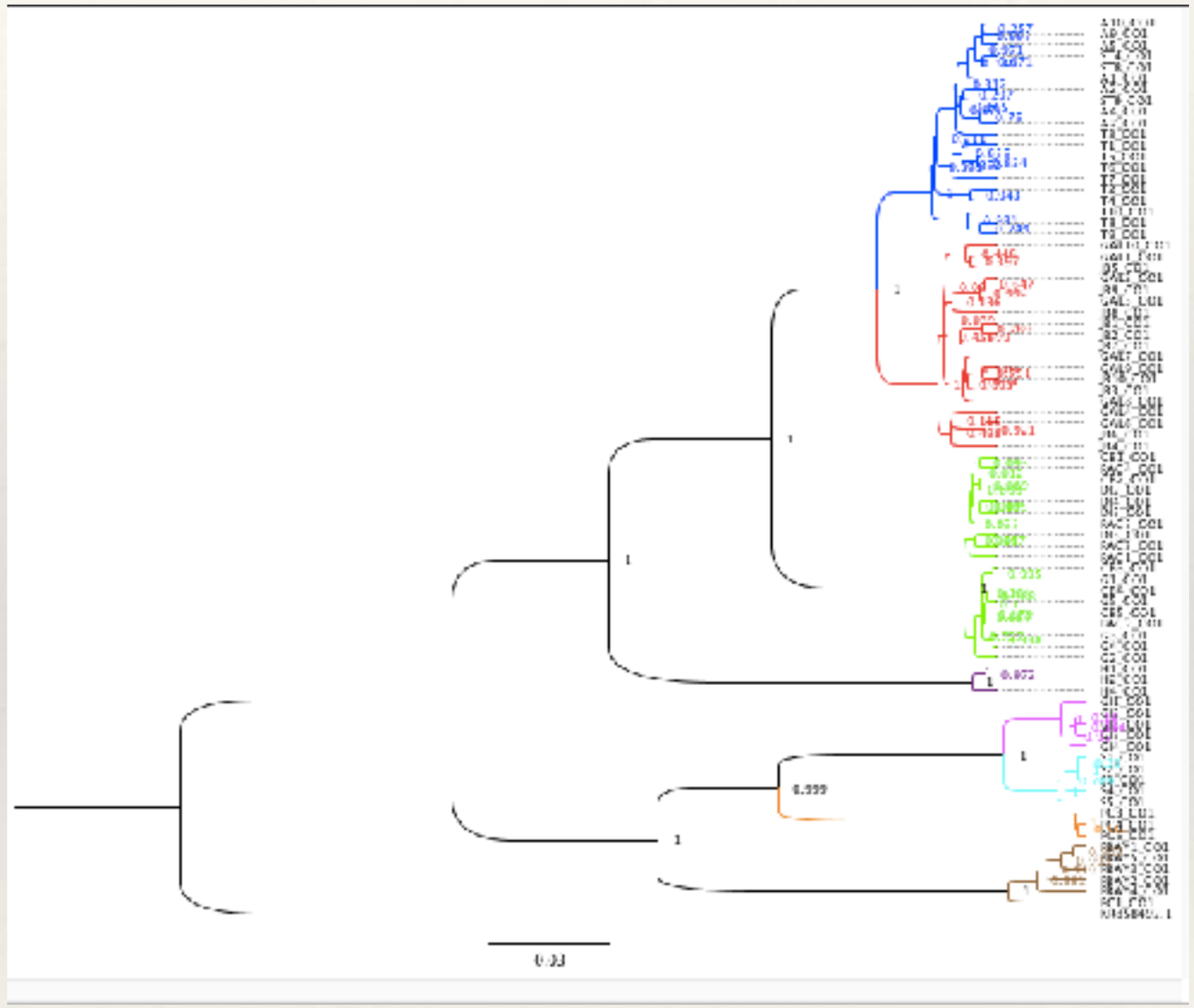
Open FigTree. Go to File —> Open —> choose “CO1consensus.tree”
On the consensus tree, find and select the branch leading to the outgroup (KR858492.1), and select “re-root” at the top.



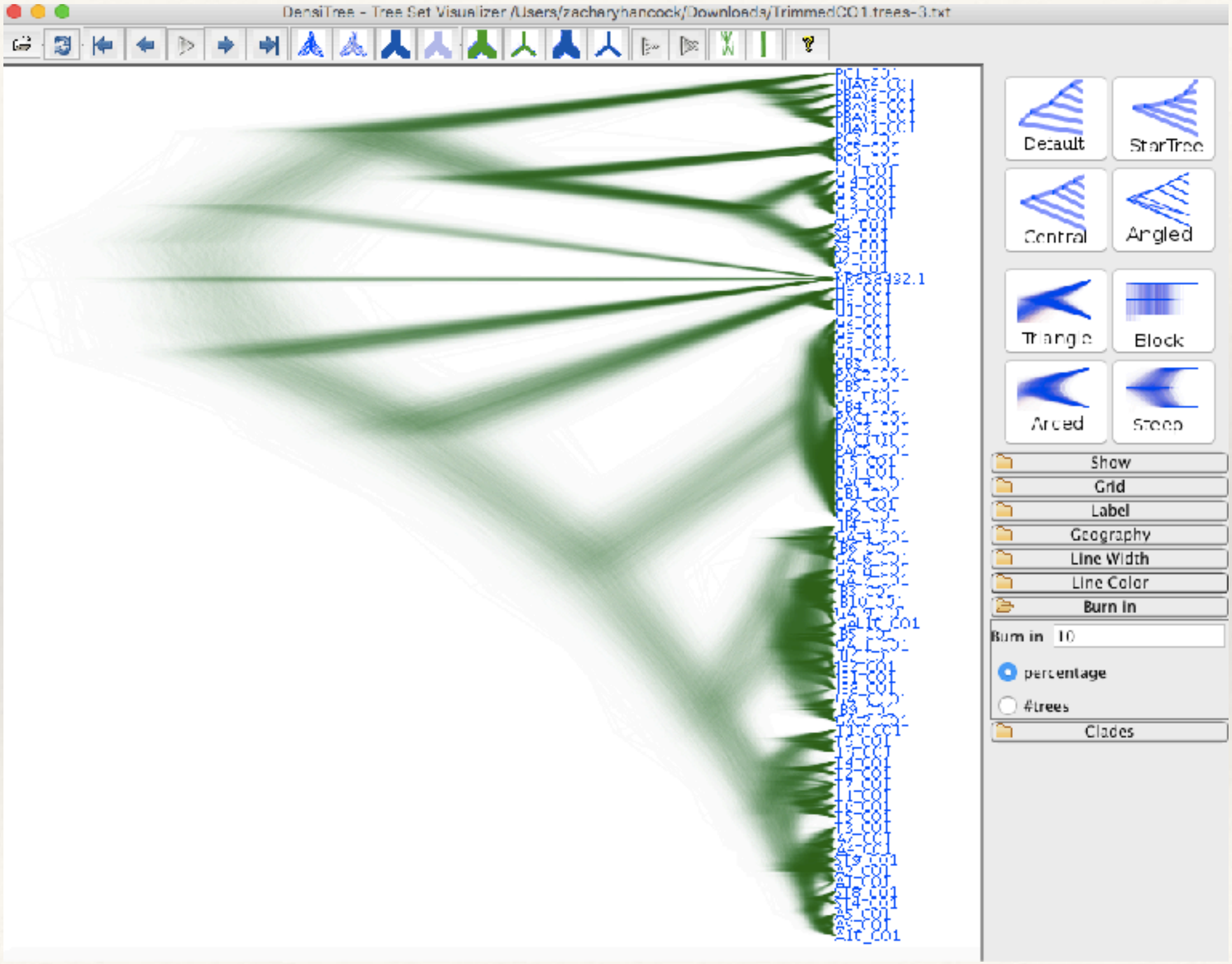
On the left-hand tab, check the box labelled “Node Labels”, and hit the scroll down until you find “posterior”. This will plot the posterior of each node.

You can also play with the shape of the tree, color different clades, etc.





If we want to visualize the full set of trees that were generated, we can do so in DensiTree. Open the DensiTree app, go to File —> Open —> choose "TrimmedCO1.trees"



References

Bouckaert RR et al. (2014) BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Computational Biology, 10(4): e1003537.

Maddison WP, Maddison DR (2018) Mesquite: a modular system for evolutionary analysis. Version 3.40 http://mesquiteproject.org

Swofford DL (2003) PAUP*: Phylogenetic Analysis Using Parsimony (*and other Methods) Version 4. Nature Biotechnology, 18:233–234.

Rambaut A et al. (2013) TRACER v1.6 (http://tree.bio.ed.ac.uk/software/tracer/).

Lewis PO (2018) Phyloseminar lecture series. https://www.youtube.com/watch?v=4PWlnNsfz90&t=2408s.