

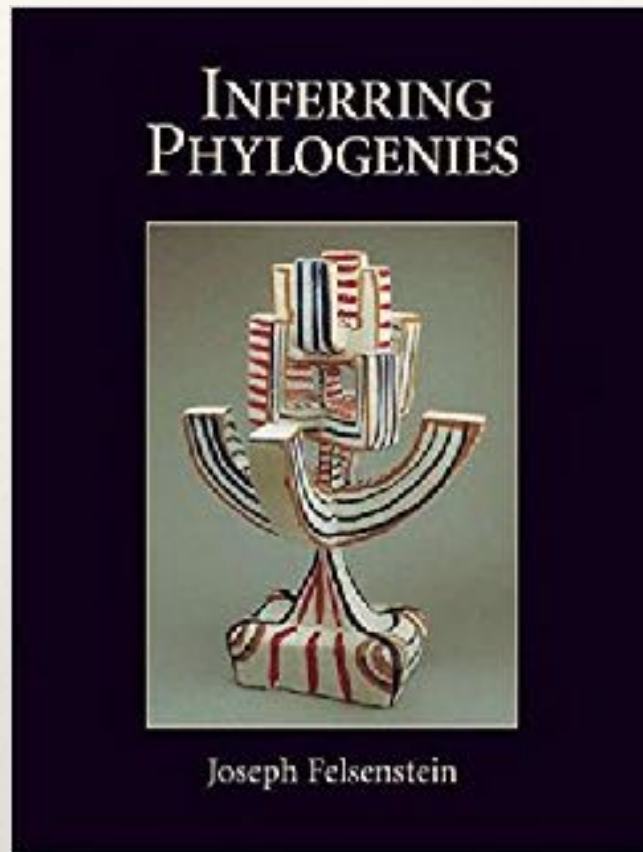


Zach Hancock — PhD student, Ecology & Evolutionary Biology

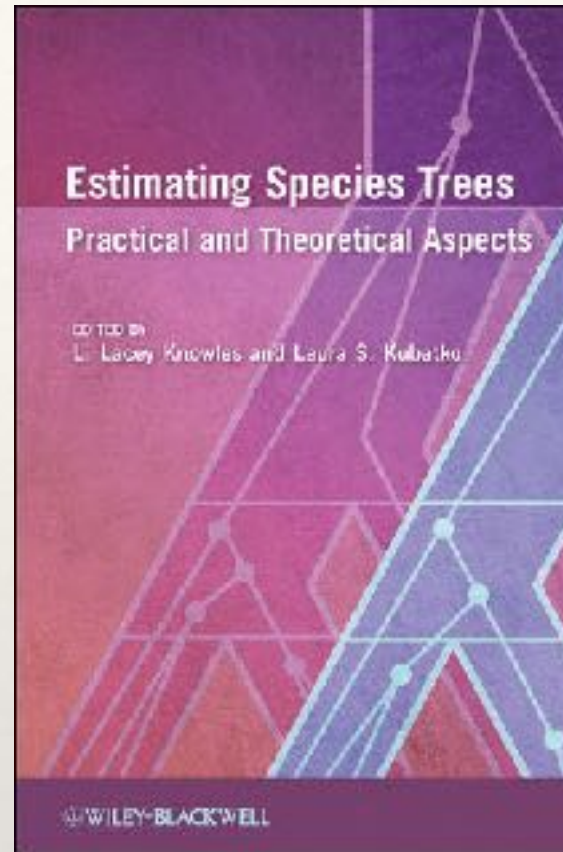
Rousing the BEAST

A molecular phylogenetic
pipeline tutorial

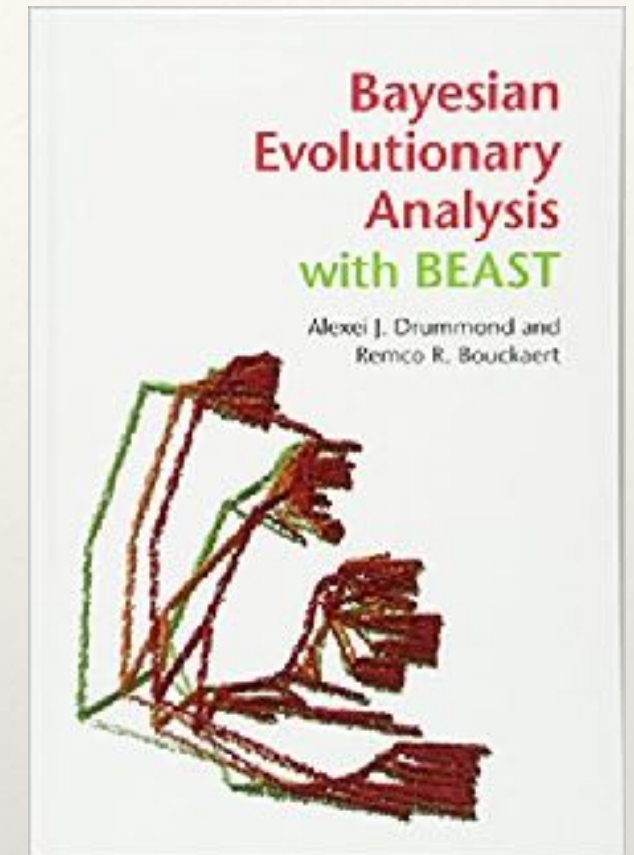
Recommendations



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>)



Beast2

Bayesian evolutionary analysis by sampling trees

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>)

Home »

All submissions are working normally.

- > Codes
- > Requirements
- > Limitations
- > Architecture
- > Known Issues
- > Usage Statistics
- > User Locations
- > Survey Results
- > Publications

The CIPRES Science Gateway V. 3.3

The CIPRES Science Gateway [V. 3.3](#) is a public resource for inference of large phylogenetic trees. It is designed to provide all researchers with access to NSF XSEDE's large computational resources through a simple browser interface. You can now also access these same capabilities programmatically with the [CIPRES REST API](#).

High Performance Parallel Codes for Large Tree Inference and Sequence Alignment on XSEDE:
[RAxML](#); [MrBayes](#); [BEAST](#); [BEAST2](#); [GARLI](#); [MAFFT](#); [DPPDIV](#); [FastTree](#); [jModelTest2](#); [PAUP](#); [ParallelStructure](#); [PartitionFinder2](#); [IQ-Tree](#); and [Migrate-N](#). If you need access to [PhyloBayes](#), please inquire.

Serial Codes for Tree Inference:
[PAUP*](#) (Inference by Parsimony); [Poy](#) (Alignment and Inference);

Serial Codes for Sequence Alignment:
[ClustalW](#); [Contraalign](#); [MUSCLE](#); [PROBCONS](#); [PROBALIGN](#)

[▶ Use the CIPRES Science Gateway](#)

Join the [CIPRES Google Group](#) for questions and problems.

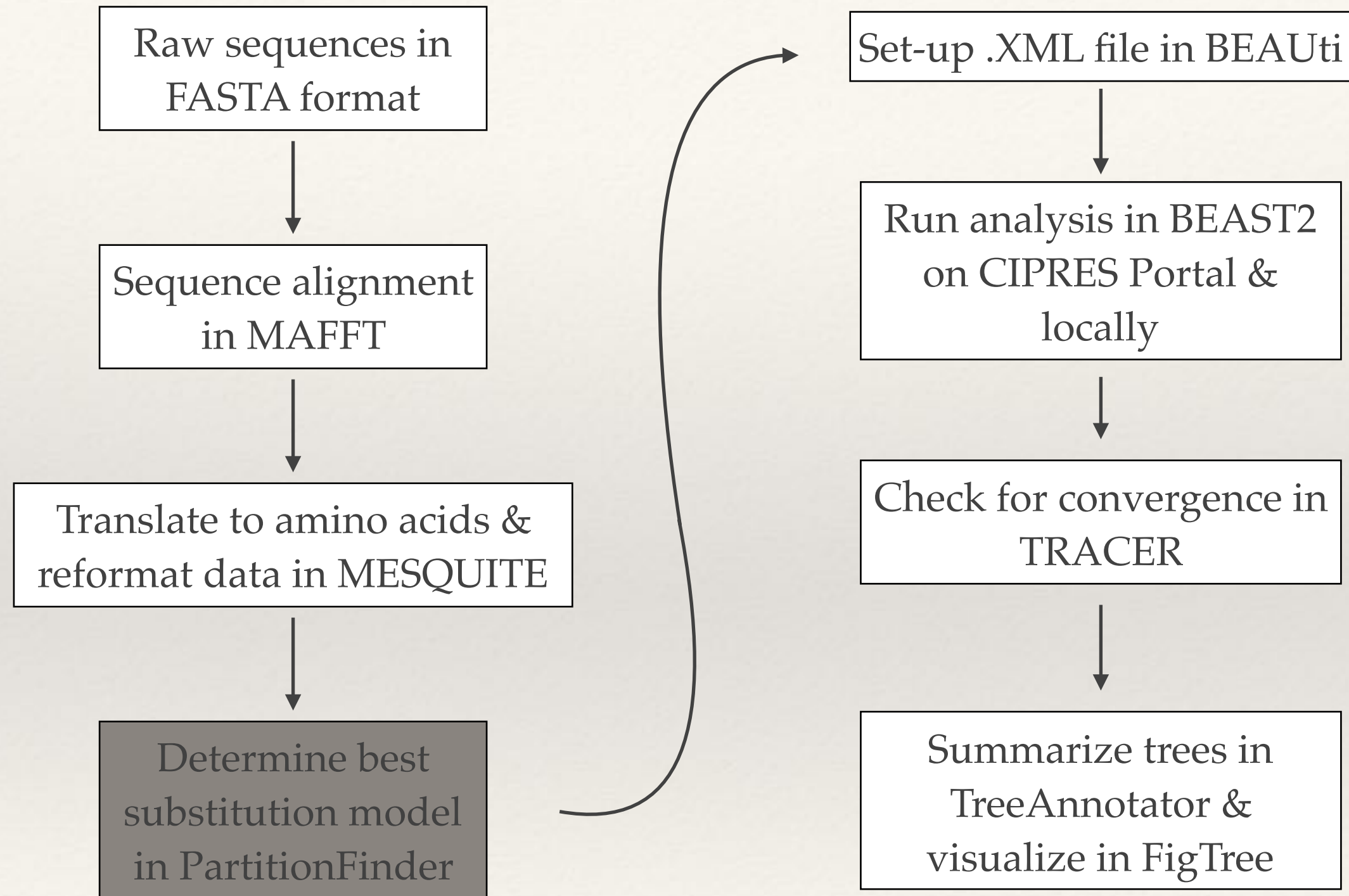
New Users - please register . Or proceed without registration.

Registration

Username	<input type="text"/>
Password	<input type="password"/>
Confirm Password	<input type="password"/>
First Name	<input type="text"/>
Last Name	<input type="text"/>
Email	<input type="text"/>

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*



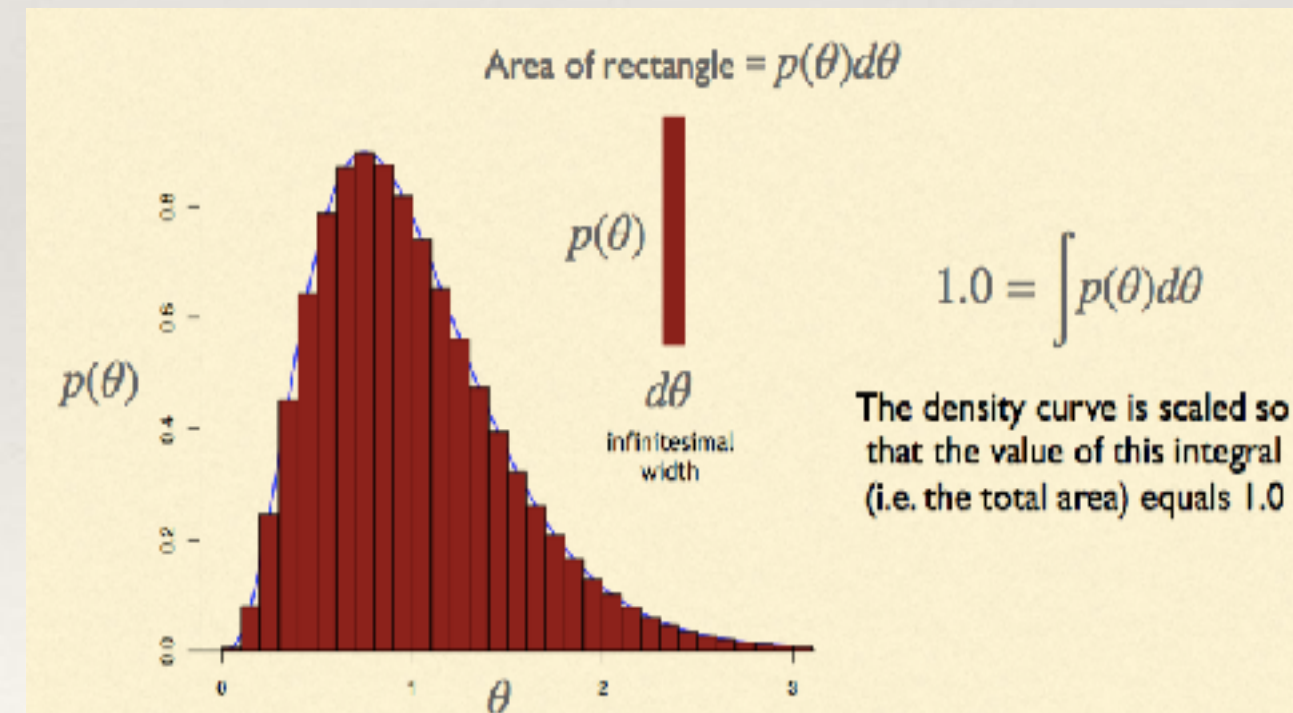
Thomas Bayes (1701–1761)

Likelihood of hypothesis θ **Prior probability of hypothesis θ**

$$\text{Pr}(\theta|D) = \frac{\text{Pr}(D|\theta) \text{Pr}(\theta)}{\sum_{\theta} \text{Pr}(D|\theta) \text{Pr}(\theta)}$$

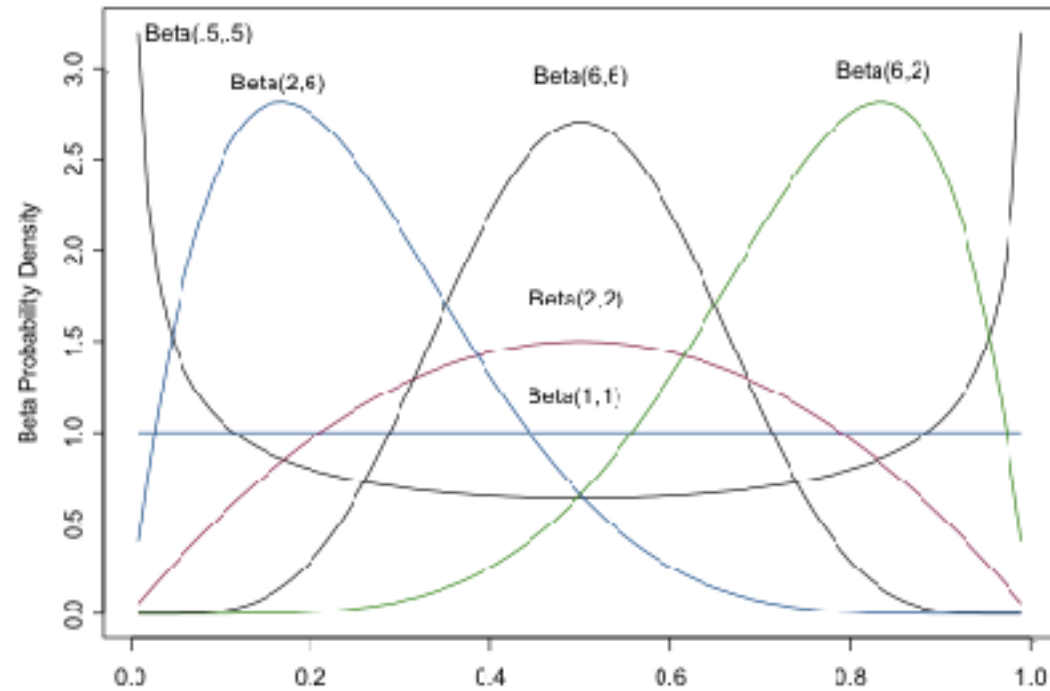
Posterior probability of hypothesis θ **Marginal probability of the data (marginalizing over hypotheses)**

Lewis (2018)

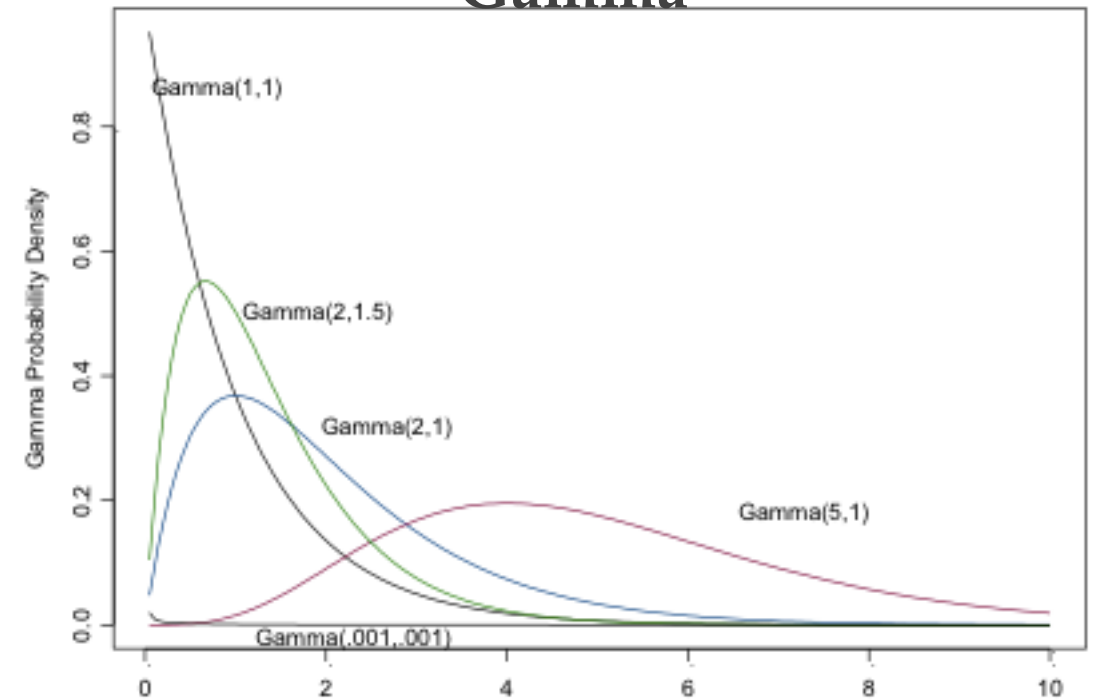


Prior Distributions

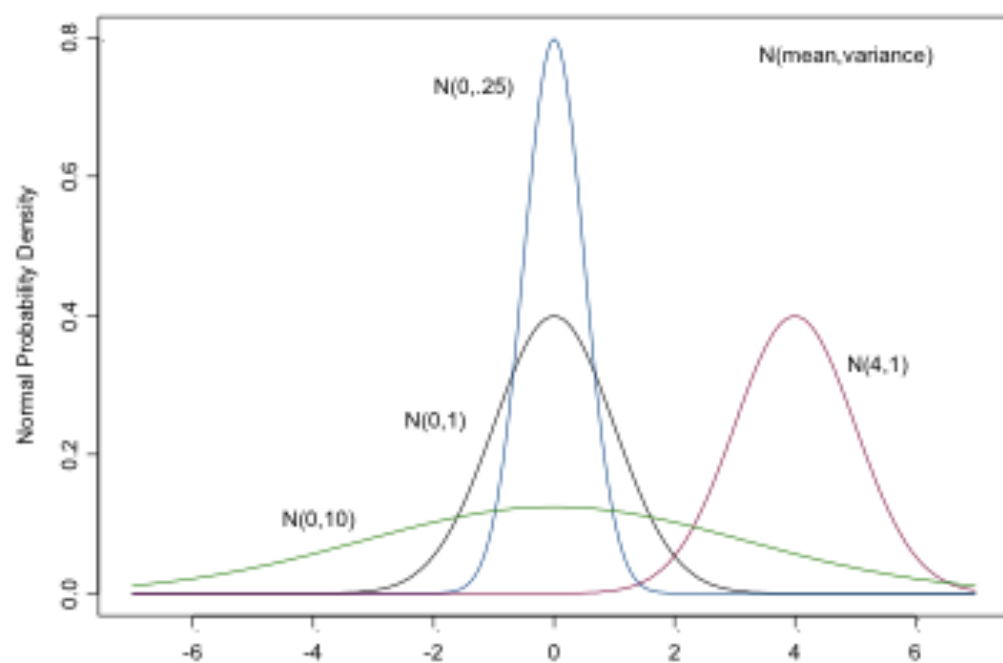
Beta



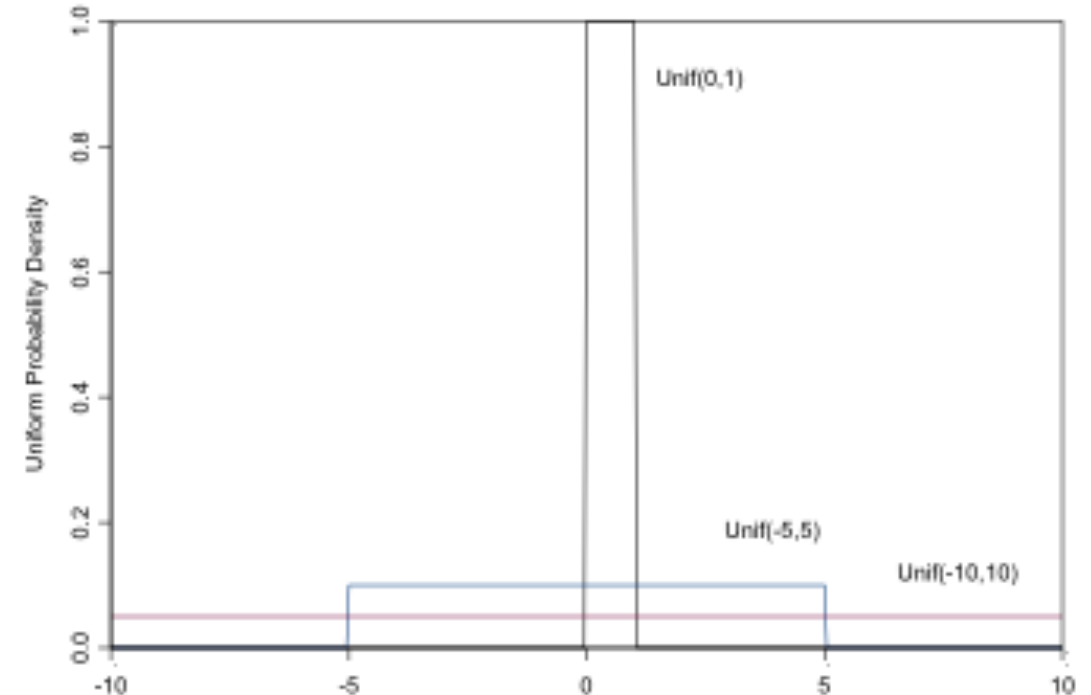
Gamma




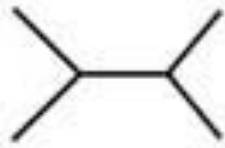
Normal



Uniform



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 ($\sim 10^{78}$)

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

Sampling the Posterior (MCMC)

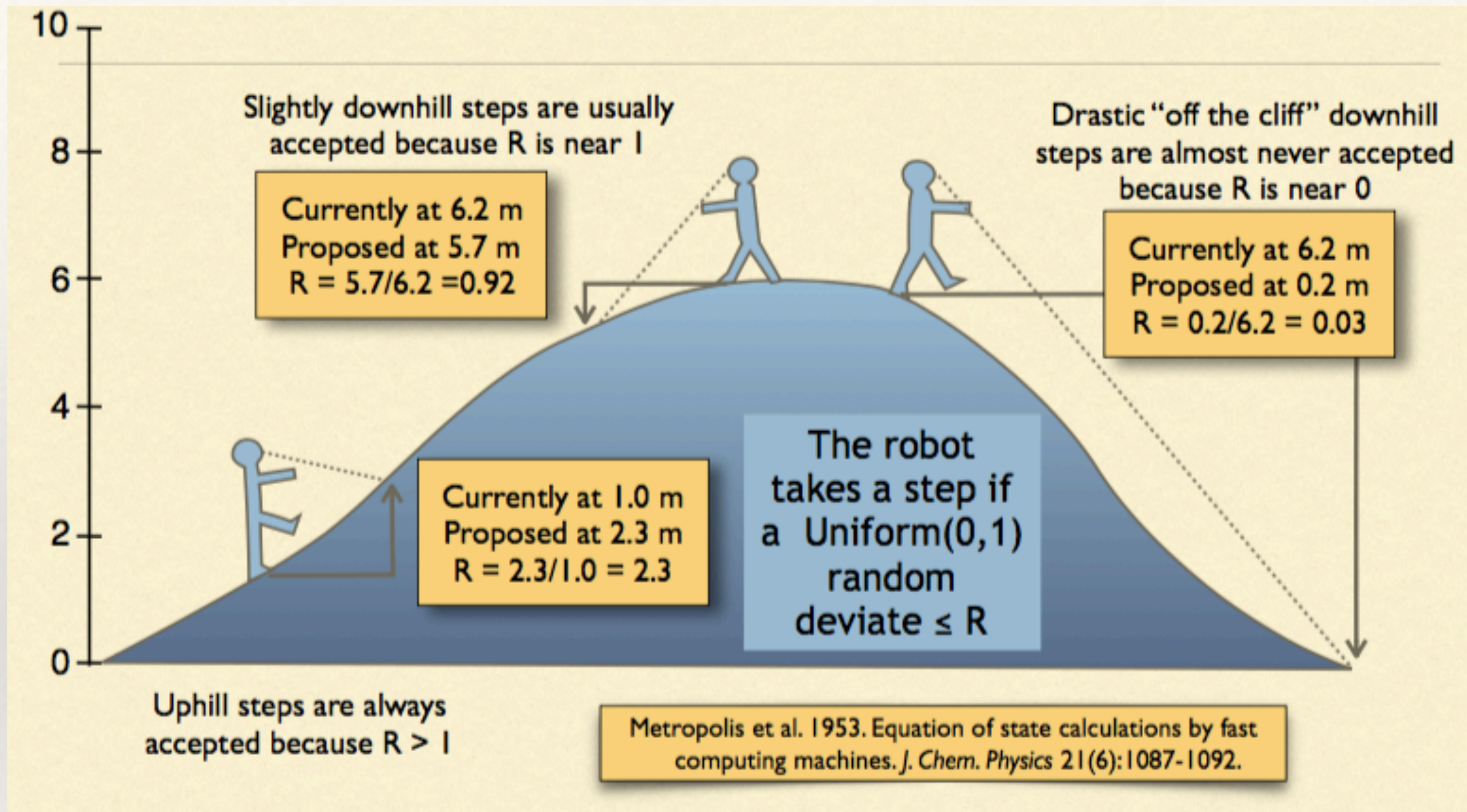
Estimating 2-parameters:

$$p(\theta, \phi | D) = \frac{p(D | \theta, \phi) p(\phi) p(\theta)}{\int_{\theta} \int_{\phi} p(D | \theta, \phi) p(\phi) p(\theta) d\phi d\theta} \longleftarrow \text{The marginal probability can quickly get out of hand}$$

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

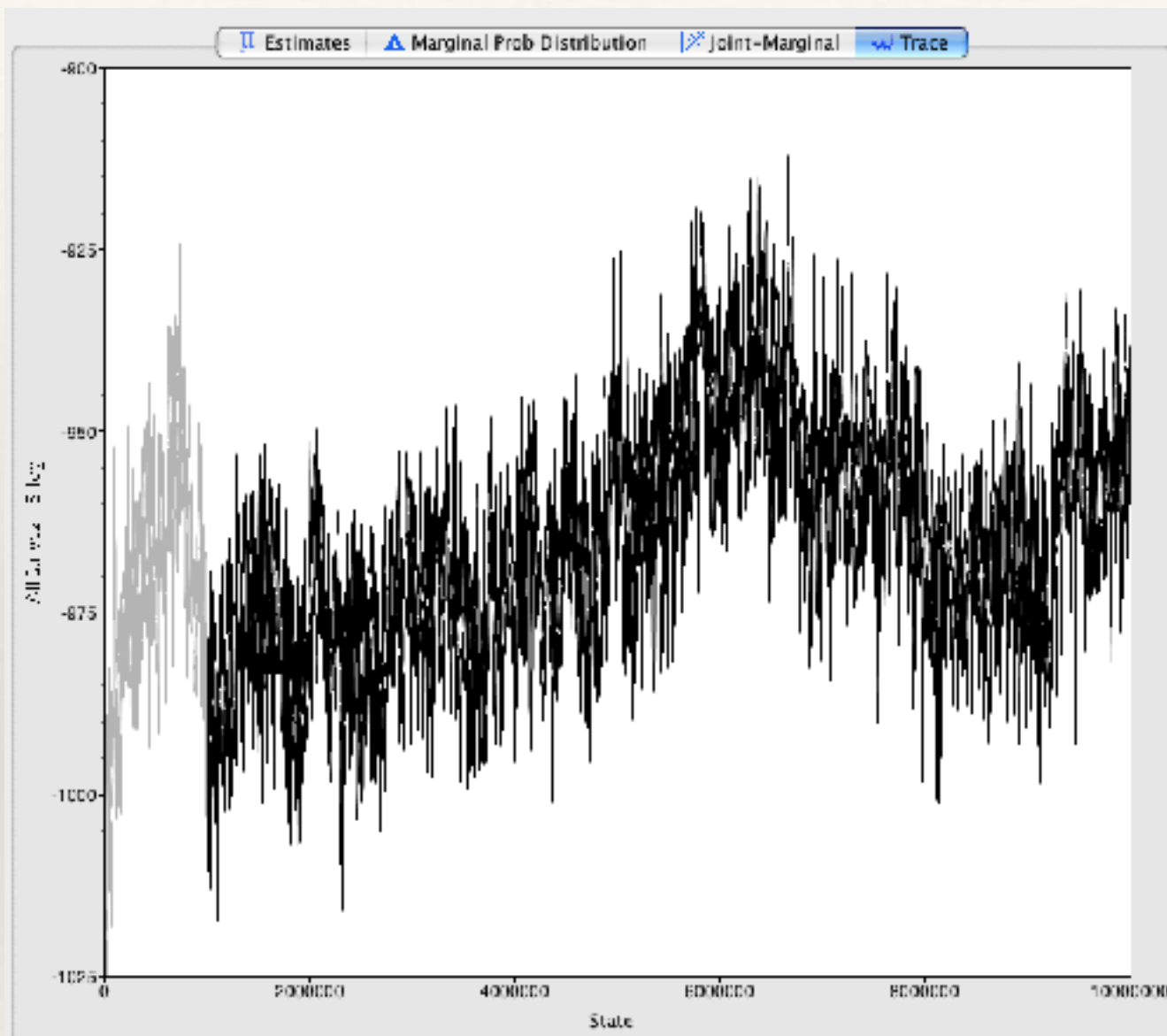
1. Start with a random parameter estimation and tree.
2. Propose a new estimate—is it better or worse?
3. If better, take the step and repeat. If worse, stay where you are and try again.
4. Continue until you have generated a good sample of the posterior.

$$\frac{\frac{p(\theta^* | D)p(\theta^*)}{\cancel{p(D)}}}{\frac{p(\theta | D)p(\theta)}{\cancel{p(D)}}} = \frac{p(\theta^* | D)p(\theta^*)}{p(\theta | D)p(\theta)} \begin{array}{l} \longleftarrow \text{New proposal} \\ \longleftarrow \text{Original proposal} \end{array}$$

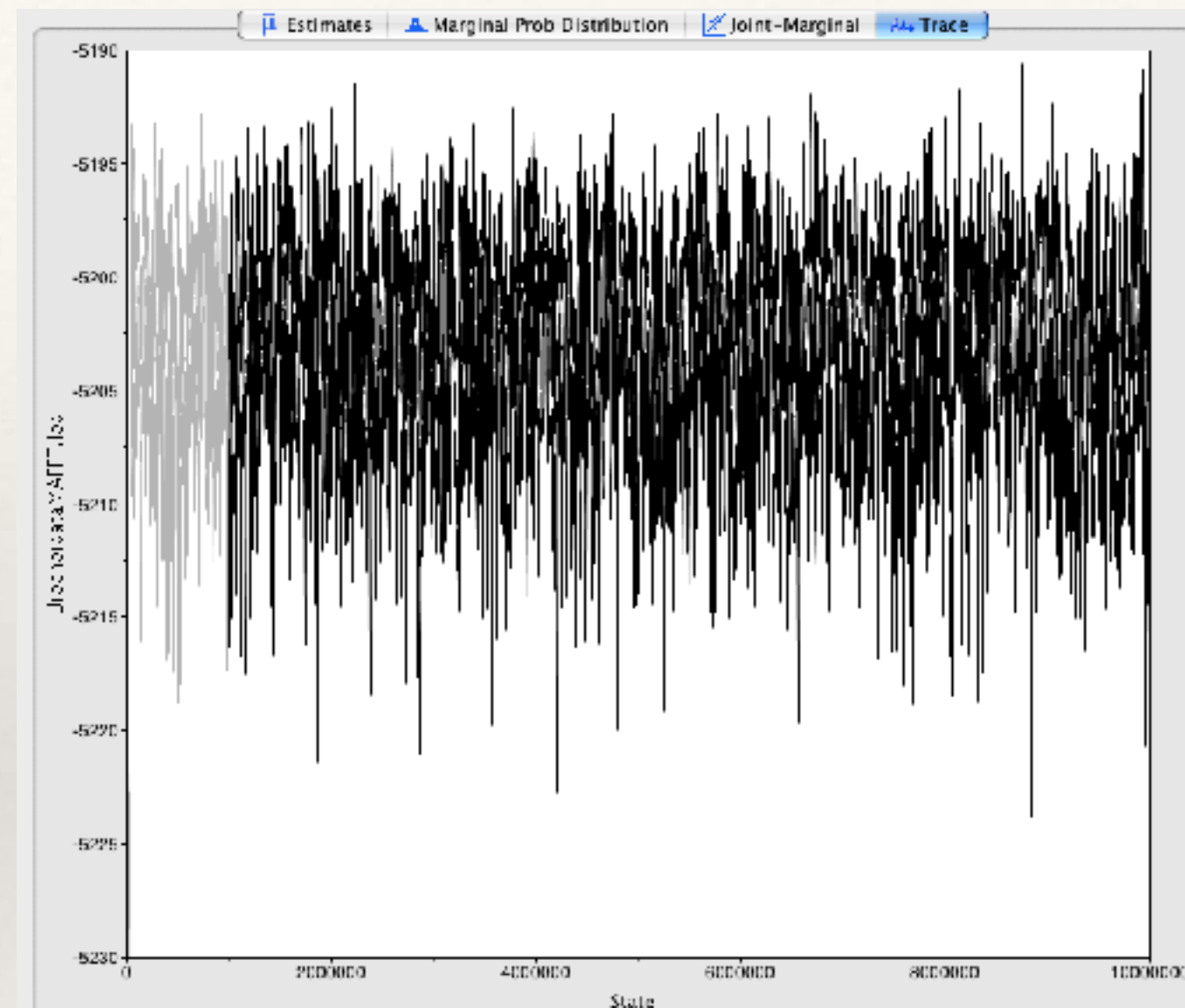


Lewis (2018)

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



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.



The screenshot shows a text editor window titled "RawCO1seqs.txt" with the file path "~/Desktop/OSOSWorkshop/RawCO1seqs.txt". The editor displays a FASTA file with multiple sequences. The first sequence, labeled ">A1_CO1", is highlighted in blue. A red arrow points to the start of this sequence. The sequence itself is a long string of nucleotide bases (A, C, G, T) followed by a newline. The editor interface includes a toolbar with icons for settings, search, and other functions, and a status bar at the bottom indicating "Evaluation Ended".

```
1 >A1_CO1
2 CCGGATCCTTTTGATTTTTGGTCATCCAGAAGTCTATATTTTAATTCTTCCAGCCTTTGGAATAATCTCACACATTGTTAACCAAGAGTCAAGAAAAAAGAAGCCTTCGGCT
3 >A2_CO1
4 CCGGATCCTTTTGATTTTTGGTCATCCAGAAGTCTATATTTTAATTCTTYCAGCCTTTGGAATAATCTCACACATTGTTAACCAAGAGTCAAGAAAAAAGAAGCCTTCGGC
5 >A4_CO1
6 CCGGATCCTTTTGATTTTTGGTCATCCAGAAGTCTATATTTTAATTCTTCCAGCCTTTGGAATAATCTCACACATTGTTAACCAAGAGTCAAGAAAAAAGAAGCCTTCGGCT
7 >A5_CO1
8 CCGGATCCTTTGGTTTTTGGTCATCCAGAAGTCTATATTTTAATTCTTCCAGCCTTTGGAATAATCTCACACATTGTTAACCAAGAGTCAAGAAAAAAGAAGCCTTCGGCT
9 >CB1_CO1
10 GTCATCCTGAGGTATATATTTTAATTCTGCCAGCCTTTGGAATAATTTACACACATTGTCAACCAAGAATCAAGAAAAAAGGAGGCCTTCGGCTCCCTCGGCATGATTTATGCTA
11 >CB2_CO1
12 CCGGATCCTCTTTGATTTTTGGTCATCCTGAGGTATATATTTTAATTCTGCCAGCCTTTGGAATAATTTACACACATTGTCAACCAAGAATCAAGAAAAAAGGAGGCCTTCGGC
13 >CB3_CO1
14 CCGGATCCTTTGATTTTTCGGTCACCCTGAGGTATATATTTTAATTCTGSCAGCCTTTGGAATAATTTACACACATTGTCAACCAAGAATCAAGAAAAAAGGAGGCCTTCGGCTC
15 >CB4_CO1
16 CCGGATCCTTTGGTTTTTGGTCACCCTGAGGTATATATTTTAATTCTGCCAGCCTTTGGAATAATTTACACACATTGTCAACCAAGAATCAAGAAAAAAGGAGGCCTTCGGCT
17 >CB5_CO1
18 CCGGATCTCTCCTGGTTTTTGGTCACCCTGAGGTATATATTTTAATTCTGCCAGCCTTTGGAATAATTTACACACATTGTCAACCAAGAATCAAGAAAAAAGGAGGCCTTCGGC
19 >DI2_CO1
20 TTTTGGTCATCCTGAGGTATATATTTTAATTCTGCCAGCCTTTGGAATAATTTACACACATTGTCAACCAAGAATCAAGAAAAAAGGAGGCCTTCGGCTCCCTCGGCATGATTT
21 >DI3_CO1
```

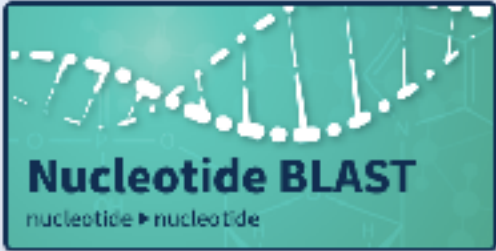
BLAST® Home Recent Results Saved Strategies Help

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)


NEWS
Introducing the BLAST widget - Integrating your BLAST results into NCBI's Genome Data Viewer!
Analyze your BLAST results in a genome browser and compare these results against other genome assembly annotations. Introducing the Genome Data Viewer (GDV) and the BLAST widget.
Tue, 19 Jun 2010 14:00:00 EST [More BLAST news...](#)

Web BLAST

**Nucleotide BLAST**
nucleotide ► nucleotide

blastx
translated nucleotide ► protein

tblastn
protein ► translated nucleotide

**Protein BLAST**
protein ► protein

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

Standard Nucleotide BLAST

[blastn](#) [blastp](#) [blastx](#) [tblastn](#) [tblastx](#)

BLASTN programs search nucleotide databases using a nucleotide query. [more...](#)

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#) [Clear](#)

Or, upload file [Choose File](#) no file selected [?](#)

Job Title

Enter a descriptive title for your BLAST search [?](#)

☐ Align two or more sequences [?](#)

Query subrange [?](#)

From

To

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

BLAST

Sequences producing significant alignments:

Select: [All](#) [None](#) Selected:0

↑ Alignments Download GenBank Graphics Distance tree of results

	Description	Max score	Total score	Query cover	E value	Ident	Accession
<input type="checkbox"/>	Niphargus timavi voucher NA190 cytochrome oxidase subunit I (COI) gene, partial cds; mitochondrial	350	350	93%	3e-82	82%	KR858492.1
<input type="checkbox"/>	Anastrepha fuscata isolate Afur744397 cytochrome oxidase subunit I (COI) gene, partial cds; mitochondrial	320	320	98%	2e-83	80%	KY428379.1
<input type="checkbox"/>	Anastrepha fuscata isolate Afur671665 cytochrome oxidase subunit I (COI) gene, partial cds; mitochondrial	320	320	96%	2e-83	81%	KY428272.1
<input type="checkbox"/>	Homologenus malayensis mitochondrion, complete genome	298	298	94%	1e-76	80%	KJ612407.1
<input type="checkbox"/>	Drosophila pachea cytochrome c oxidase subunit I (COI) gene, partial cds; mitochondrial	298	298	92%	1e-76	80%	KF632609.1

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”.

GenBank

Niphargus timavi voucher NA190 cytochrome oxidase subunit I (COI) gene, partial cds; mitochondrial

GenBank: KR858492.1

[FASTA](#) [Graphics](#) [PopSet](#)

[Help](#)

LOCUS KR858492 592 bp DNA linear INV 09-AUG-2015

DEFINITION Niphargus timavi voucher NA190 cytochrome oxidase subunit I (COI) gene, partial cds; mitochondrial.

ACCESSION KR858492

VERSION KR858492.1

KEYWORDS .

SOURCE mitochondrial Niphargus timavi

ORGANISM [Niphargus timavi](#)
Eukaryota; Metazoa; Bodysozoa; Arthropoda; Crustacea; Malacostraca; Eumalacostraca; Peracarida; Amphipoda; Senticandata; Gammarida; Crangonyctidae; Crangonyctoidea; Niphargidae; Niphargus.

REFERENCE 1 (bases 1 to 592)

AUTHORS Fiser,Z., Altermatt,F., Zakeek,V., Knapic,T. and Fiser,C.

TITLE Morphologically Cryptic Amphipod Species Are 'Ecological Clones' at Regional but Not at Local Scale: A Case Study of Four Niphargus Species

JOURNAL PLoS ONE 10 (7), e0134384 (2015)

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.

Niphargus timavi voucher NA190 cytochrome oxidase subunit I (COI) gene, partial cds; mitochondrial

GenBank: KR858492.1

[GenBank](#) [Graphics](#) [PopSet](#)

>KR858492.1 Niphargus timavi voucher NA190 cytochrome oxidase subunit I (COI) gene,
partial cds; mitochondrial

TCACCCTGAAGTTTATATTTTGATTTTACCTGCTTTTGGCATAATCTCCCATATTGTCAGACAAGAAGCA
GGTAAAAAAGAAACATTCGGGGGCCCTTGGTATAATCTATGCTATATTAGCAATTGGTCTATTAGGGTTTA
TTGTGTGGGCTCACCATATATTTACTGTAGGAATAGATGTGGATACTCGAGCTTATTTACATCTGCTAC
AATAATCATTGCTGTCCCAACAGGTATTAAAGTATTTAGTTGATTAGGTACTCTTCAAGGAGGTAAACTA
TACCTCTCTCCTTCTCTCTTATGAGCTCTTGGGTTTATTTTTTTTATTCACATTAGGAGGTCTAACAGGTA
TTATATTAGCTAATTCATCAATTGATATTATCCTTCATGATACCTATTATGTAGTAGCCATTTTCATTA
TGTGTTATCCATAGGAGCTGTGTTTGGTATTTTGGCTGGATTATTCATTGATTCCCTTTATTCACAGGC
CTTACCCTTAATAATTACTTAGCTAAAATTCATTTTTATCTTATATTTATTGGAGTTAACATAACCTTTT
TCCCCAACATTTTTTAGGTTTAAGGGGTATA

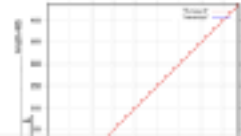
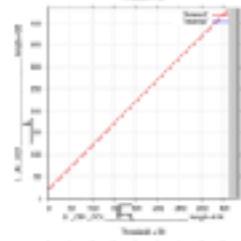
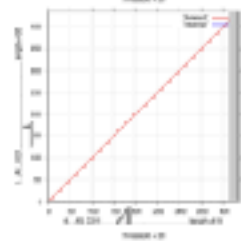
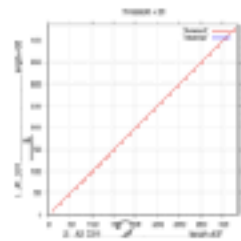
Copy!

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](#)

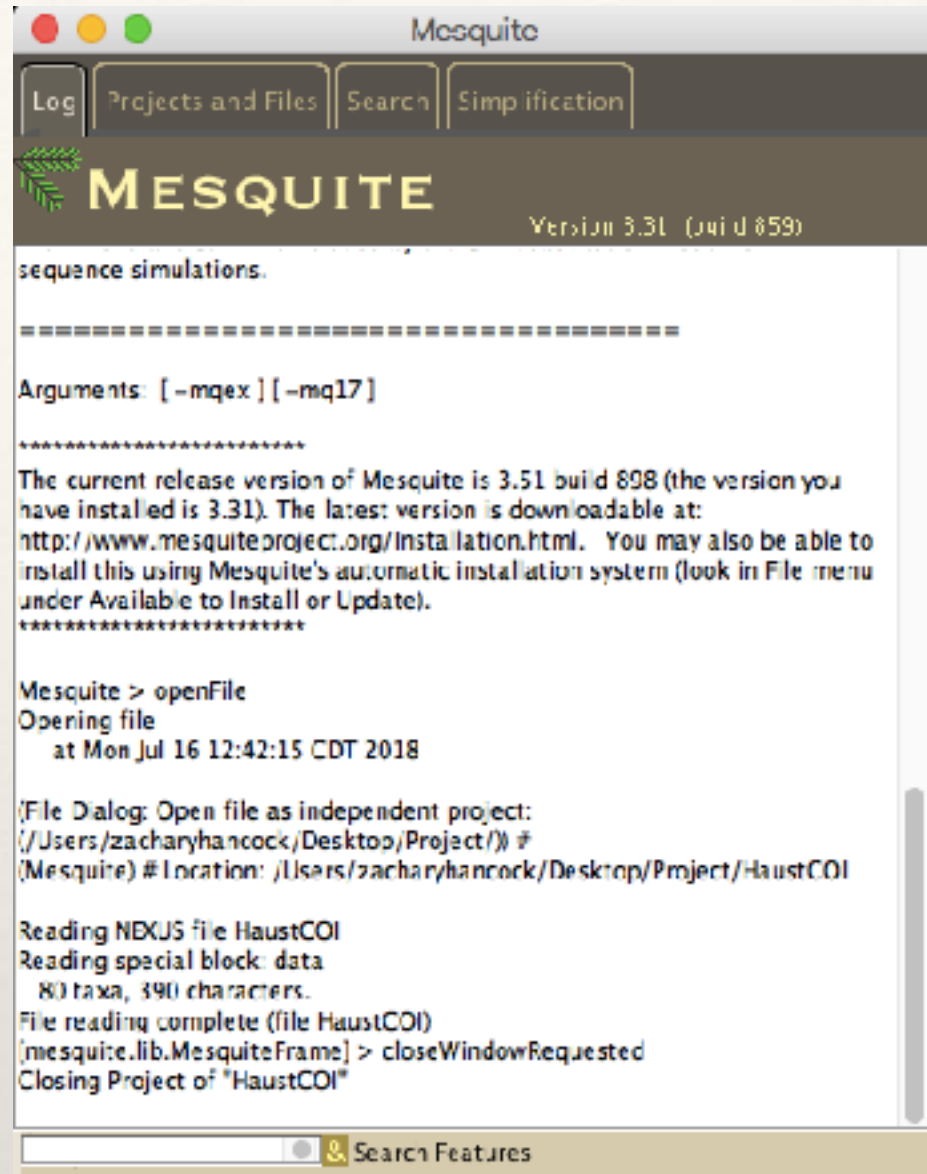
[Phylogenetic tree](#)

MAFFT-L-INS-i Result

CLUSTAL format alignment by MAFFT (v7.402)

```
A1_CO1 -----
A4_CO1 -----
T7_CO1 -----
A5_CO1 -----
A-10_CO1 -----
A-9_CO1 -----
ST-4_CO1 -----
T2_CO1 -----
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 (54-bit word length)
SSE vectorization enabled
SSE3 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 '-'
"Equal" matrix in effect:
  R,r ==> {AC}
  Y,y ==> {CT}
  M,m ==> {AC}
  K,k ==> {GT}
  S,s ==> {CG}
  W,w ==> {AT}
  H,h ==> {ACT}
  B,b ==> {CCT}
  V,v ==> {ACG}
  D,d ==> {ACT}
  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:

Taxon \ Character	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67
1 A1 CO1	G	T	C	A	T	C	C	A	G	A	A	G	T	C	T	A	T	A	T	T	T	T	A	A	T	T	C	T	T	C	C	A	G	C	C	T	T	T	G	G	A	A	T	A	A	T	C	T	C	A	C	A	C	A	T	T	G	T	T	A	A	C	C	A	A	G	A
2 A4 CO1	G	T	C	A	T	C	C	A	G	A	A	G	T	C	T	A	T	A	T	T	T	T	A	A	T	T	C	T	T	C	C	A	G	C	C	T	T	T	G	G	A	A	T	A	A	T	C	T	C	A	C	A	C	A	T	T	G	T	T	A	A	C	C	A	A	G	A
3 T7 CO1	G	T	C	A	T	C	C	A	G	A	A	G	T	C	T	A	T	A	T	T	T	T	A	A	T	T	C	T	T	C	C	A	G	C	C	T	T	T	G	G	A	A	T	A	A	T	C	T	C	A	C	A	C	A	T	T	G	T	T	A	A	C	C	A	A	G	A
4 A5 CO1	G	T	C	A	T	C	C	A	G	A	A	G	T	C	T	A	T	A	T	T	T	T	A	A	T	T	C	T	T	C	C	A	G	C	C	T	T	T	G	G	A	A	T	A	A	T	C	T	C	A	C	A	C	A	T	T	G	T	T	A	A	C	C	A	A	G	A
5 A-10 CO1	G	T	C	A	T	C	C	A	G	A	A	G	T	C	T	A	T	A	T	T	T	T	A	A	T	T	C	T	T	C	C	A	G	C	C	T	T	T	G	G	A	A	T	A	A	T	C	T	C	A	C	A	C	A	T	T	G	T	T	A	A	C	C	A	A	G	A
6 A-9 CO1	G	T	C	A	T	C	C	A	G	A	A	G	T	C	T	A	T	A	T	T	T	T	A	A	T	T	C	T	T	C	C	A	G	C	C	T	T	T	G	G	A	A	T	A	A	T	C	T	C	A	C	A	C	A	T	T	G	T	T	A	A	C	C	A	A	G	A
7 ST-4 CO1	G	T	C	A	T	C	C	A	G	A	A	G	T	C	T	A	T	A	T	T	T	T	A	A	T	T	C	T	T	C	C	A	G	C	C	T	T	T	G	G	R	A	T	M	A	T	C	T	C	A	C	A	C	A	T	T	G	T	T	A	A	C	C	A	A	G	A
8 T2 CO1	G	T	C	A	T	C	C	A	G	A	A	G	T	C	T	A	T	A	T	T	T	T	A	A	T	T	C	T	T	C	C	A	G	C	C	T	T	T	K	G	R	A	T	A	A	T	C	T	C	A	C	A	C	A	T	T	G	T	T	A	A	C	C	A	A	G	A
9 T4 CO1	G	T	C	A	T	C	C	A	G	A	A	G	T	C	T	A	T	A	T	T	T	T	A	A	T	T	C	T	T	C	C	A	G	C	C	T	T	T	G	G	A	A	T	A	A	T	C	T	C	A	C	A	C	A	T	T	G	T	T	A	A	C	C	A	A	G	A
10 ST-6 CO1	G	G	C	A	T	C	C	A	G	A	A	G	T	C	T	A	T	A	T	T	T	T	A	A	T	T	C	T	T	C	C	A	G	C	C	T	T	T	G	G	A	A	T	A	A	T	C	T	C	A	C	A	C	A	T	T	G	T	T	A	A	C	C	A	A	G	A
11 T3 CO1	G	T	C	A	T	C	C	A	G	A	A	G	T	C	T	A	T	A	T	T	T	T	A	A	T	T	C	T	T	C	C	A	G	C	C	T	T	T	G	G	A	A	T	A	A	T	C	T	C	A	C	A	C	A	T	T	G	T	T	A	A	C	C	A	A	G	A
12 T8 CO1	G	T	C	A	T	C	C	A	G	A	A	G	T	C	T	A	T	A	T	T	T	T	A	A	T	T	C	T	T	C	C	A	G	C	C	T	T	T	G	G	A	A	T	A	A	T	C	T	C	A	C	A	C	A	T	T	G	T	T	A	A	C	C	A	A	G	A
13 T10 CO1	G	G	C	A	T	C	C	A	G	A	A	G	T	C	T	A	T	A	T	T	T	T	A	A	T	T	C	T	T	Y	C	C	G	C	C	T	T	T	G	G	C	A	T	A	A	T	C	T	C	A	C	A	C	A	T	T	G	T	T	A	A	C	C	A	A	G	A
14 T5 CO1	G	T	C	A	T	C	C	A	G	A	A	G	T	C	T	A	T	A	T	T	T	T	A	A	T	T	C	T	T	C	C	A	G	C	C	T	T	T	G	G	A	A	T	A	A	T	C	T	C	A	C	A	C	A	T	T	G	T	T	A	A	C	C	A	A	G	A
15 T9 CO1	G	T	C	A	T	C	C	A	G	A	A	G	T	C	T	A	T	A	T	T	T	T	A	A	T	T	C	T	T	C	C	A	G	C	C	T	T	T	G	G	A	A	T	A	A	T	C	T	C	A	C	A	C	A	T	T	G	T	T	A	A	C	C	A	A	G	A
16 A-7 CO1	G	T	C	A	T	C	C	A	G	A	A	G	T	C	T	A	T	A	T	T	T	T	A	A	T	T	C	T	T	C	C	A	G	C	C	T	T	T	G	G	A	A	T	A	A	T	C	T	C	A	C	A	C	A	T	T	G	T	T	A	A	C	C	A	A	G	A
17 A2 CO1	G	T	C	A	T	C	C	A	G	A	A	G	T	C	T	A	T	A	T	T	T	T	A	A	T	T	C	T	T	Y	C	A	G	C	C	T	T	T	G	G	A	A	T	A	A	T	C	T	C	A	C	A	C	A	T	T	G	T	T	A	A	C	C	A	A	G	A
18 T6 CO1	G	T	C	A	T	C	C	A	G	A	A	G	T	C	T	A	T	A	T	T	T	T	A	A	T	T	C	T	T	C	C	A	G	C	C	T	T	T	S	G	R	A	T	A	A	T	C	T	C	A	C	A	C	A	T	T	G	T	T	A	A	C	C	A	A	G	A
19 T1 CO1	G	G	C	A	T	C	C	A	G	A	A	G	T	C	T	A	T	A	T	T	T	T	A	A	T	T	C	T	T	Y	C	A	G	C	C	T	T	T	G	G	A	A	T	A	A	T	C	T	C	A	C	A	C	A	T	T	G	T	T	A	A	C	C	A	A	G	A
20 GALL CO1	G	G	C	A	T	C	C	A	G	A	A	G	T	C	T	A	T	A	T	T	T	T	A	A	T	T	C	T	T	C	C	A	G	C	C	T	T	T	G	G	A	A	T	A	A	T	C	T	C	A	C	A	C	A	T	T	G	T	T	A	A	C	C	A	A	G	A
21 B10 CO1	G	G	C	A	T	C	C	A	G	A	A	G	T	C	T	A	T	A	T	T	T	T	A	A	T	T	C	T	T	C	C	A	G	C	C	T	T	T	G	G	R	A	T	A	A	T	C	T	C	A	C	A	C	A	T	T	G	T	T	A	A	C	C	A	A	G	A
22 B3 CO1	G	G	C	A	T	C	C	A	G	A	A	G	T	C	T	A	T	A	T	T	T	T	A	A	T	T	C	T	T	C	C	A	G	C	C	T	T	T	G	G	R	A	T	A	A	T	C	T	C	A	C	A	C	A	T	T	G	T	T	A	A	C	C	A	A	G	A

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

TrimmedCO1.nex

Project of "TrimmedCO1.nex" Character Matrix Characters "Character Matrix"

Project: TrimmedCO1.nex Add... Taxa (81 taxa) Character Matrix

« List Columns » Window

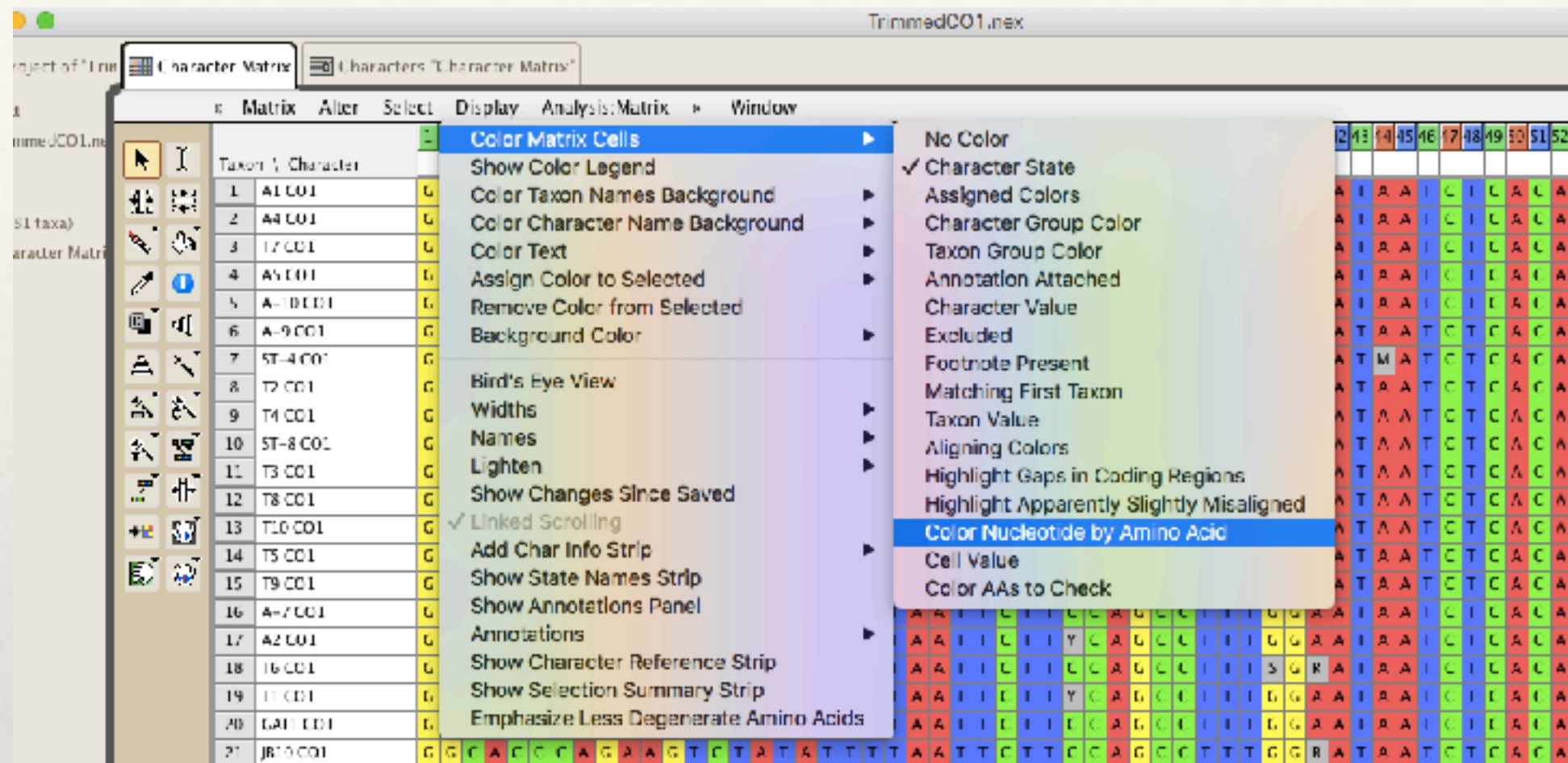
Character	In	Group	Codon Position	Genetic Code
1 Character 1	✓	?	N	Standard
2 Character 2	✓	?	N	Standard
3 Character 3	✓	?	N	Standard
4 Character 4	✓	?	N	Standard
5 Character 5	✓	?	N	Standard
6 Character 6	✓	?	N	Standard
7 Character 7	✓	?	N	Standard
8 Character 8	✓	?	N	Standard
9 Character 9	✓	?	N	Standard
10 Character 10	✓	?	N	Standard
11 Character 11	✓	?	N	Standard
12 Character 12	✓	?	N	Standard
13 Character 13	✓	?	N	Standard
14 Character 14	✓	?	N	Standard
15 Character 15	✓	?	N	Standard
16 Character 16	✓	?	N	Standard
17 Character 17	✓	?	N	Standard
18 Character 18	✓	?	N	Standard
19 Character 19	✓	?	N	Standard

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

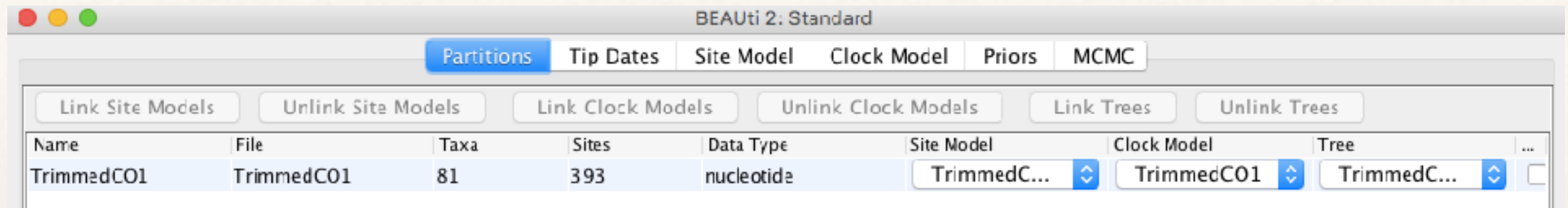
Return to the sequence matrix.

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

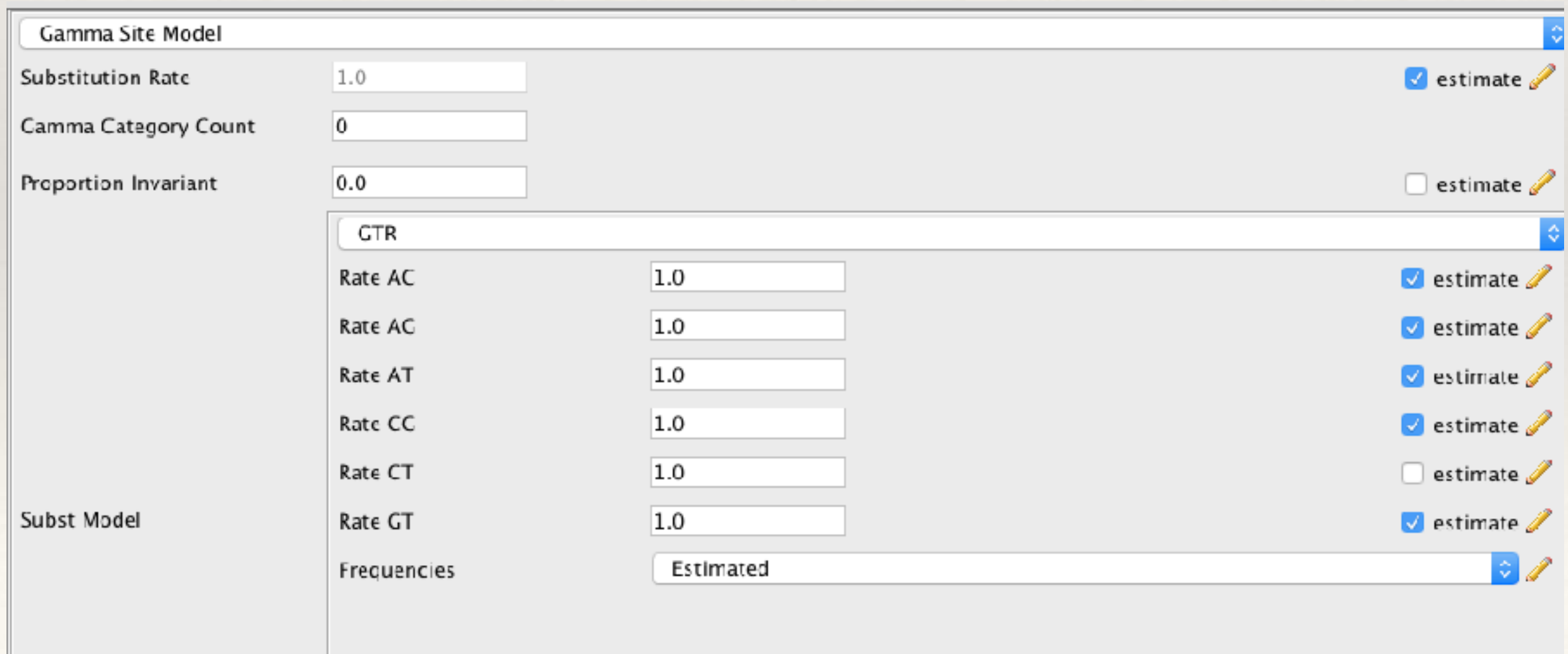


This will change the color of the matrix to match the amino acid coded for. Slide through the matrix to check for black codons—this indicates a premature stop codon (there should not be any!).

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.

BEAUi 2: Standard

Partitions Tip Dates Site Model Clock Model **Priors** MCMC

► Tree.t:TrimmedCO1 Yule Model

► birthRate.t:TrimmedCO1 Gamma initial = [1.0] $[-\infty, \infty]$ Yule speciation process birth rate of partition t:TrimmedCO1

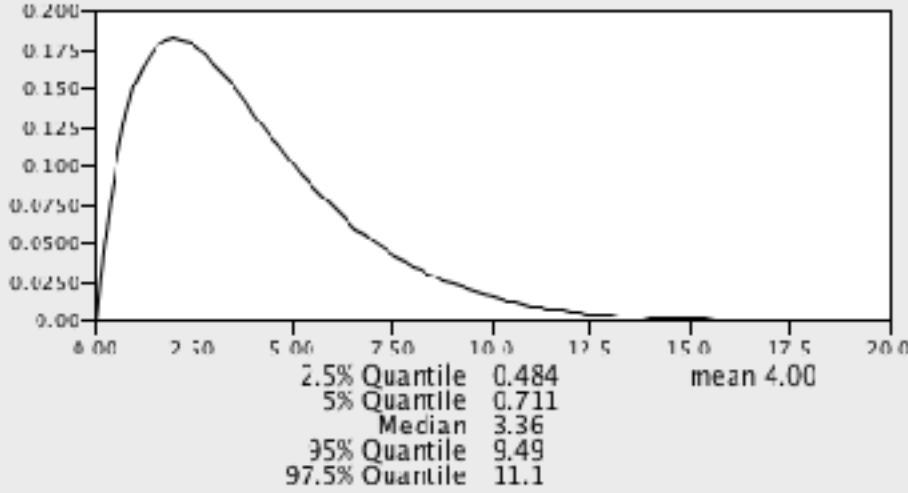
▼ rateAC.s:TrimmedCO1 Gamma initial = [1.0] $[0.0, \infty]$ GTR A-C substitution parameter of partition s:TrimmedCO1

Alpha 2.0 ☐ estimate

Beta 2.0 ☐ estimate

Mode ShapeScale

Offset 0.0



Quantile	Value
2.5% Quantile	0.484
5% Quantile	0.711
Median	3.36
95% Quantile	9.49
97.5% Quantile	11.1

mean 4.00

► rateAG.s:TrimmedCO1 Gamma initial = [1.0] $[0.0, \infty]$ GTR A-G substitution parameter of partition s:TrimmedCO1

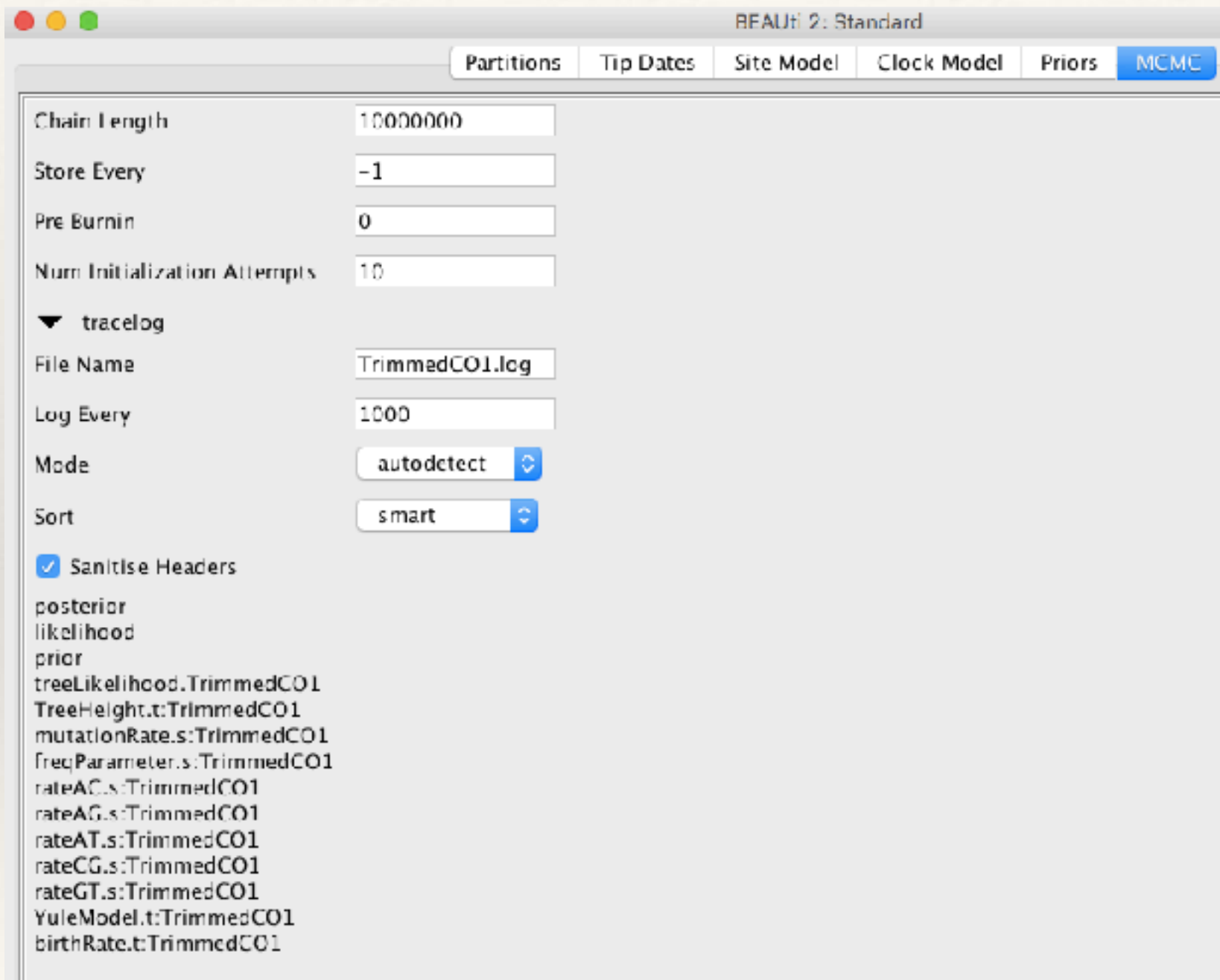
► rateAT.s:TrimmedCO1 Gamma initial = [1.0] $[0.0, \infty]$ GTR A-T substitution parameter of partition s:TrimmedCO1

► rateCG.s:TrimmedCO1 Gamma initial = [1.0] $[0.0, \infty]$ GTR C-G substitution parameter of partition s:TrimmedCO1

► rateGT.s:TrimmedCO1 Gamma initial = [1.0] $[0.0, \infty]$ GTR G-T substitution parameter of partition s:TrimmedCO1

+ Add Prior

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.



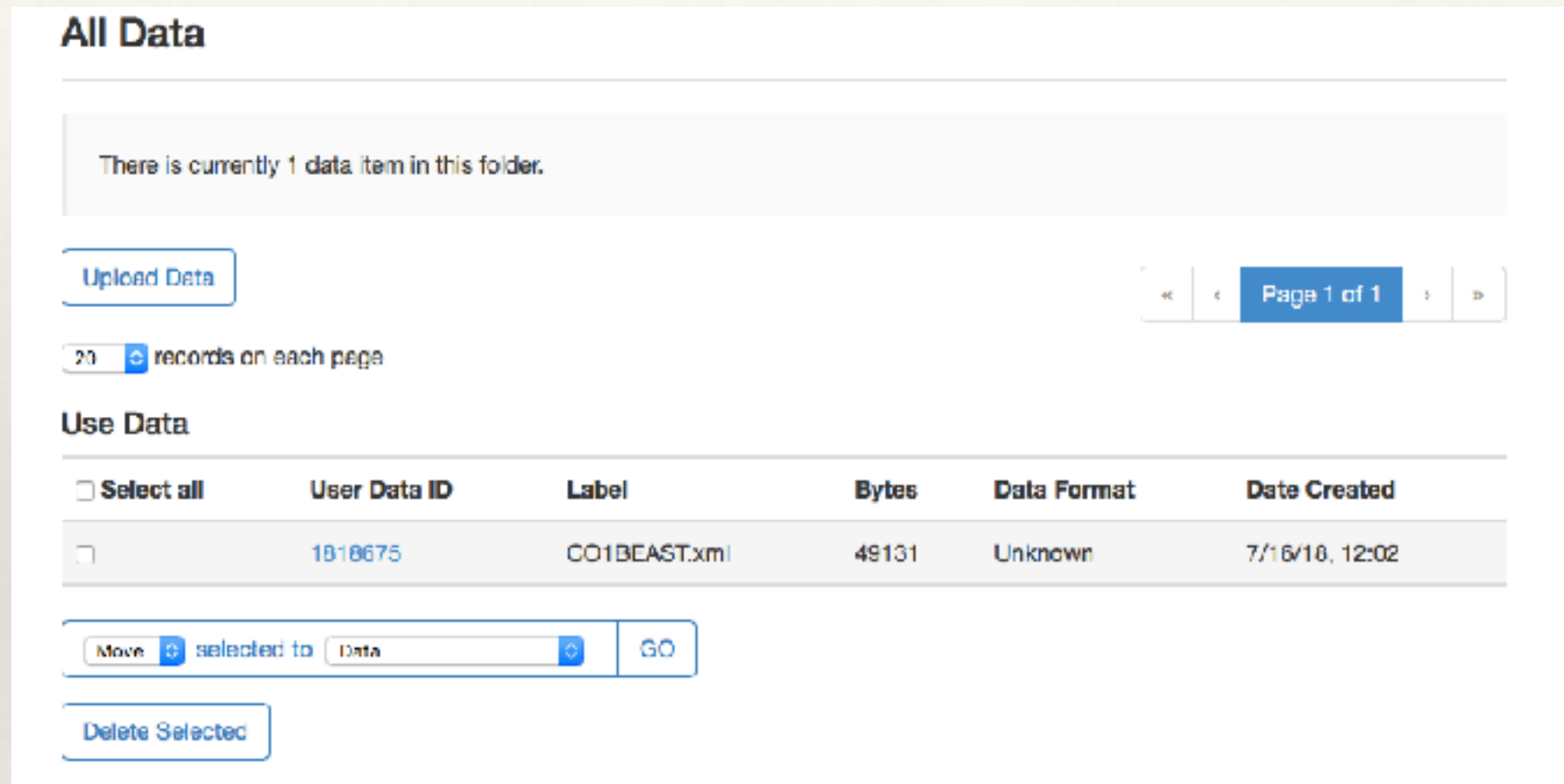
The screenshot shows the BEAUti 2: Standard interface with the MCMC tab selected. The settings are as follows:

Setting	Value
Chain Length	10000000
Store Every	-1
Pre Burnin	0
Num Initialization Attempts	10
tracelog (expanded)	
File Name	TrimmedCO1.log
Log Every	1000
Mode	autodetect
Sort	smart
<input checked="" type="checkbox"/> Sanitise Headers	
posterior	
likelihood	
prior	
treeLikelihood:TrimmedCO1	
TreeHeight.t:TrimmedCO1	
mutationRate.s:TrimmedCO1	
freqParameter.s:TrimmedCO1	
rateAC.s:TrimmedCO1	
rateAG.s:TrimmedCO1	
rateAT.s:TrimmedCO1	
rateCG.s:TrimmedCO1	
rateGT.s:TrimmedCO1	
YuleModel.t:TrimmedCO1	
birthRate.t:TrimmedCO1	

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”.

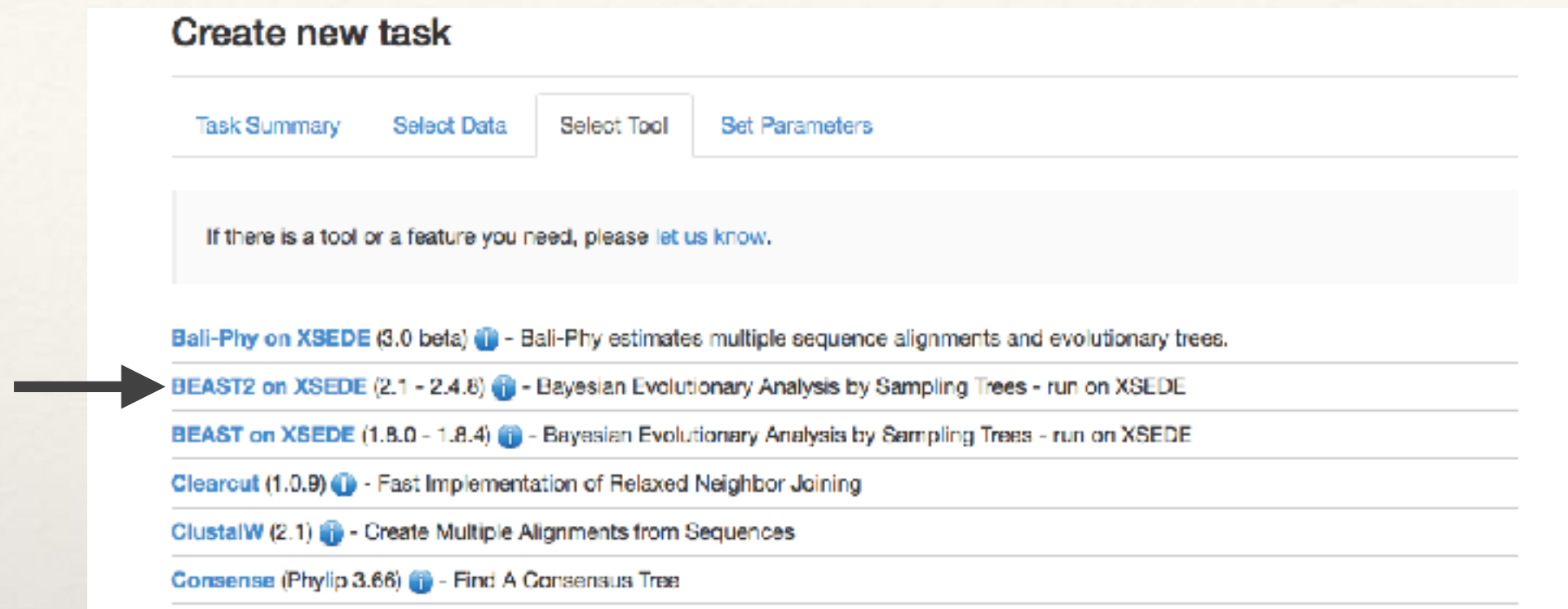


The screenshot shows the 'All Data' interface in the CIPRES portal. At the top, it says 'All Data'. Below that, a message states 'There is currently 1 data item in this folder.' There is an 'Upload Data' button on the left and a pagination control on the right showing 'Page 1 of 1'. Below the message, it says '20 records on each page'. Under the 'Use Data' section, there is a table with the following columns: 'Select all', 'User Data ID', 'Label', 'Bytes', 'Data Format', and 'Date Created'. The table contains one row with the following data: an unchecked checkbox, '1010675', 'CO1BEAST.xml', '49131', 'Unknown', and '7/16/10, 12:02'. Below the table, there is a 'Move' button, a dropdown menu showing 'selected to Data', and a 'GO' button. At the bottom, there is a 'Delete Selected' button.

<input type="checkbox"/> Select all	User Data ID	Label	Bytes	Data Format	Date Created
<input type="checkbox"/>	1010675	CO1BEAST.xml	49131	Unknown	7/16/10, 12:02

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.



Create new task

Task Summary Select Data **Select Tool** Set Parameters

If there is a tool or a feature you need, please [let us know](#).

- [Bali-Phy on XSEDE \(3.0 beta\)](#) ⓘ - Bali-Phy estimates multiple sequence alignments and evolutionary trees.
- [BEAST2 on XSEDE \(2.1 - 2.4.8\)](#)** ⓘ - Bayesian Evolutionary Analysis by Sampling Trees - run on XSEDE
- [BEAST on XSEDE \(1.8.0 - 1.8.4\)](#) ⓘ - Bayesian Evolutionary Analysis by Sampling Trees - run on XSEDE
- [Clearcut \(1.0.9\)](#) ⓘ - Fast Implementation of Relaxed Neighbor Joining
- [ClustalW \(2.1\)](#) ⓘ - Create Multiple Alignments from Sequences
- [Consense \(Phylip 3.66\)](#) ⓘ - Find A Consensus Tree

Simple Parameters

Which BEAST2 Version? *

Maximum Hours to Run (up to 168 hours) *

How many patterns does your data have? (use # chars if you aren't sure) *

My data set is partitioned * ☐

This is a Path Sampling analysis ☐

How many partitions does your data have? *

Do not use Beagle ☐

Specify a seed for this run (by default a random seed is used) ☐

Enter the seed value here

Overwrite existing log files ☐

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.

<input type="checkbox"/> Select all	Tool Output	File Name	File Size (Bytes)		
<input type="checkbox"/>	PROCESS_OUTPUT	STDOUT	4707722	View	Download
<input type="checkbox"/>		STDERR	1690	View	Download
<input type="checkbox"/>	all_results	_scheduler_stderr.txt	416	View	Download
<input type="checkbox"/>		TrimmedCO1.trees	199450868	View	Download
<input type="checkbox"/>		infile.xml	49131	View	Download
<input type="checkbox"/>		start.txt	40	View	Download
<input type="checkbox"/>		TrimmedCO1.log	16868130	View	Download
<input type="checkbox"/>		stdout.txt	4707722	View	Download
<input type="checkbox"/>		infile_altered.xml	49131	View	Download
<input type="checkbox"/>		scheduler.conf	60	View	Download
<input type="checkbox"/>		stderr.txt	1690	View	Download
<input type="checkbox"/>		infile_altered.xml.state	7461	View	Download
<input type="checkbox"/>		term.txt	317	View	Download
<input type="checkbox"/>		_JOBINFO.TXT	328	View	Download

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


9997000	-2693.5567	738.2	-2940.3613	246.8846	1m9s/Msamples
9998000	-2689.9170	738.3	-2936.9200	247.0829	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:TrimmedC01)	0.5146	98373	289153	0.0387	0.2538
ScaleOperator(YuleModelTreeScaler.t:TrimmedC01)	0.7190	78896	307523	0.0387	0.2842
ScaleOperator(YuleModelTreeRootScaler.t:TrimmedC01)	0.5783	42396	344641	0.0387	0.1895
Uniform(YuleModelUniformOperator.t:TrimmedC01)	-	2340484	1525729	0.3856	0.6854
SubtreeSlide(YuleModelSubtreeSlide.t:TrimmedC01)	0.4813	8329	1922195	0.1933	0.0043 Try decreasing size to about 0.241
Exchange(YuleModelNarrow.t:TrimmedC01)	-	714790	1220658	0.1933	0.3693
Exchange(YuleModelWide.t:TrimmedC01)	-	3337	382768	0.0387	0.0086
WilsonBalding(YuleModelWilsonBalding.t:TrimmedC01)	-	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:TrimmedC01)	0.0706	3656	9345	0.0013	0.2818
ScaleOperator(RateACScaler.s:TrimmedC01)	0.3798	3573	9031	0.0013	0.2835
ScaleOperator(RateAGScaler.s:TrimmedC01)	0.4724	2970	10026	0.0013	0.2285
ScaleOperator(RateATScaler.s:TrimmedC01)	0.4027	3219	9632	0.0013	0.2505
ScaleOperator(RateCGScaler.s:TrimmedC01)	0.2386	3622	9124	0.0013	0.2842
ScaleOperator(RateGTScaler.s:TrimmedC01)	0.3588	3500	9450	0.0013	0.2703

Tuning: The value of the operator's tuning parameter, or '-' if the operator's likelihood: -2700.475305776576 or can't be optimized.

#accept: The total number of times a proposal by this operator has been accepted.

BEAST v2.4.8



Bayesian Evolutionary Analysis Sampling Trees
Version v2.4.8, 2002-2017

BEAST XML File:

☐ Only load packages and versions specified in XML

Random number seed:

Thread pool size:

☐ Use BEAGLE library if available:

Prefer use of:

Prefer precision:

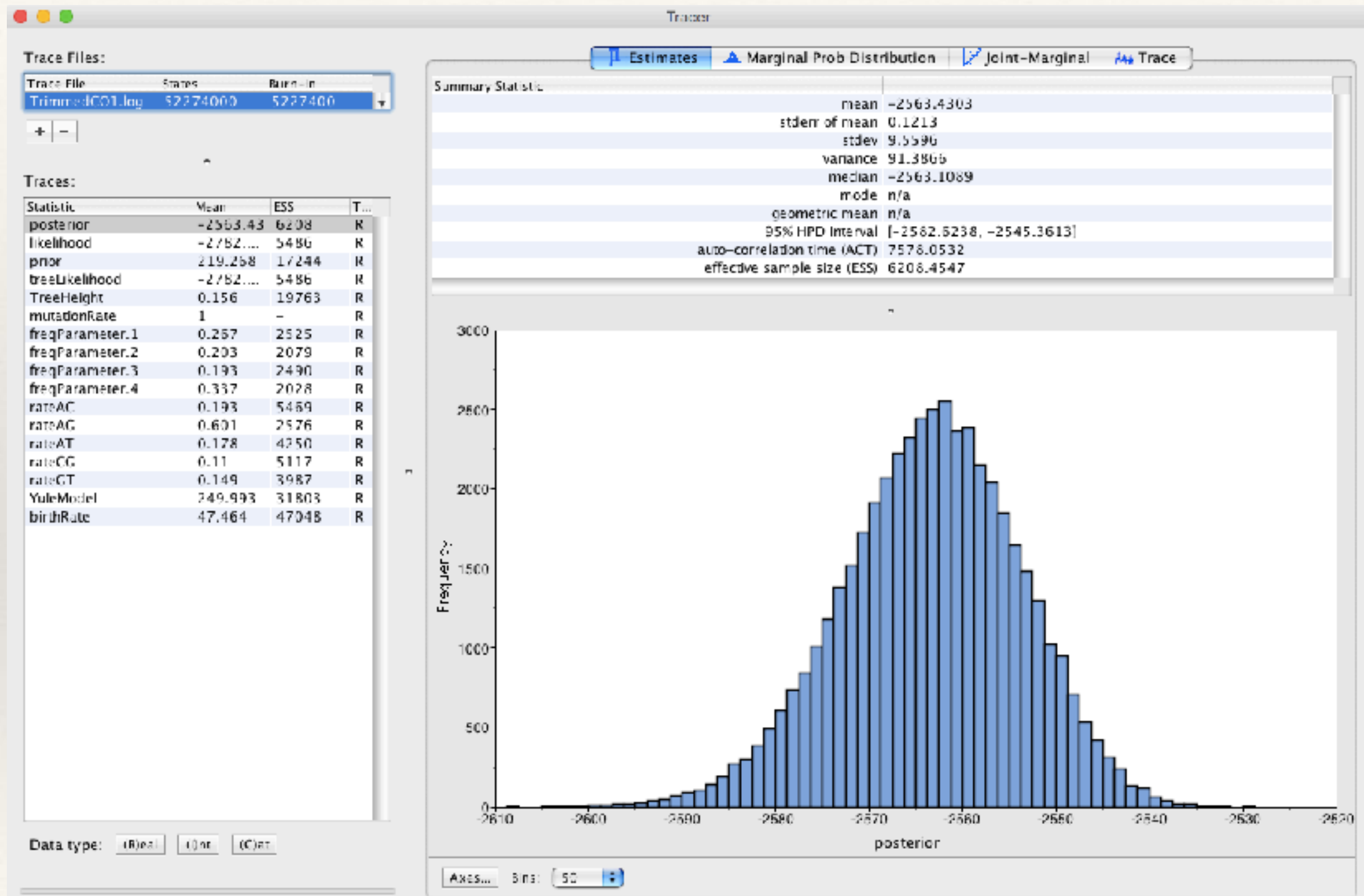
☐ Show list of available BEAGLE resources and Quit

BEAGLE is a high-performance phylogenetic library that can make use of additional computational resources such as graphics boards. It must be downloaded and installed independently of BEAST:
<http://beagle-lib.googlecode.com/>

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”.

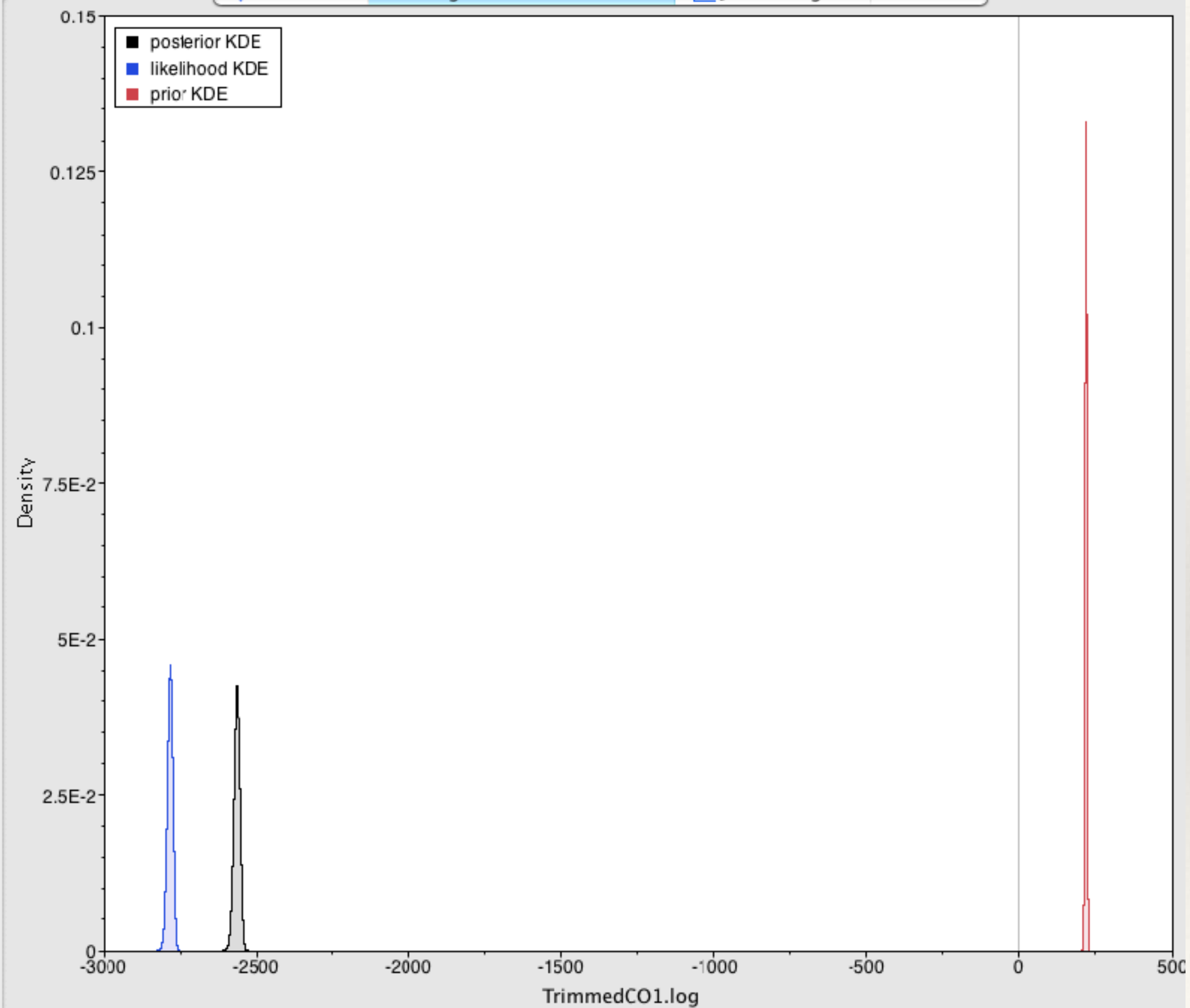


Estimates

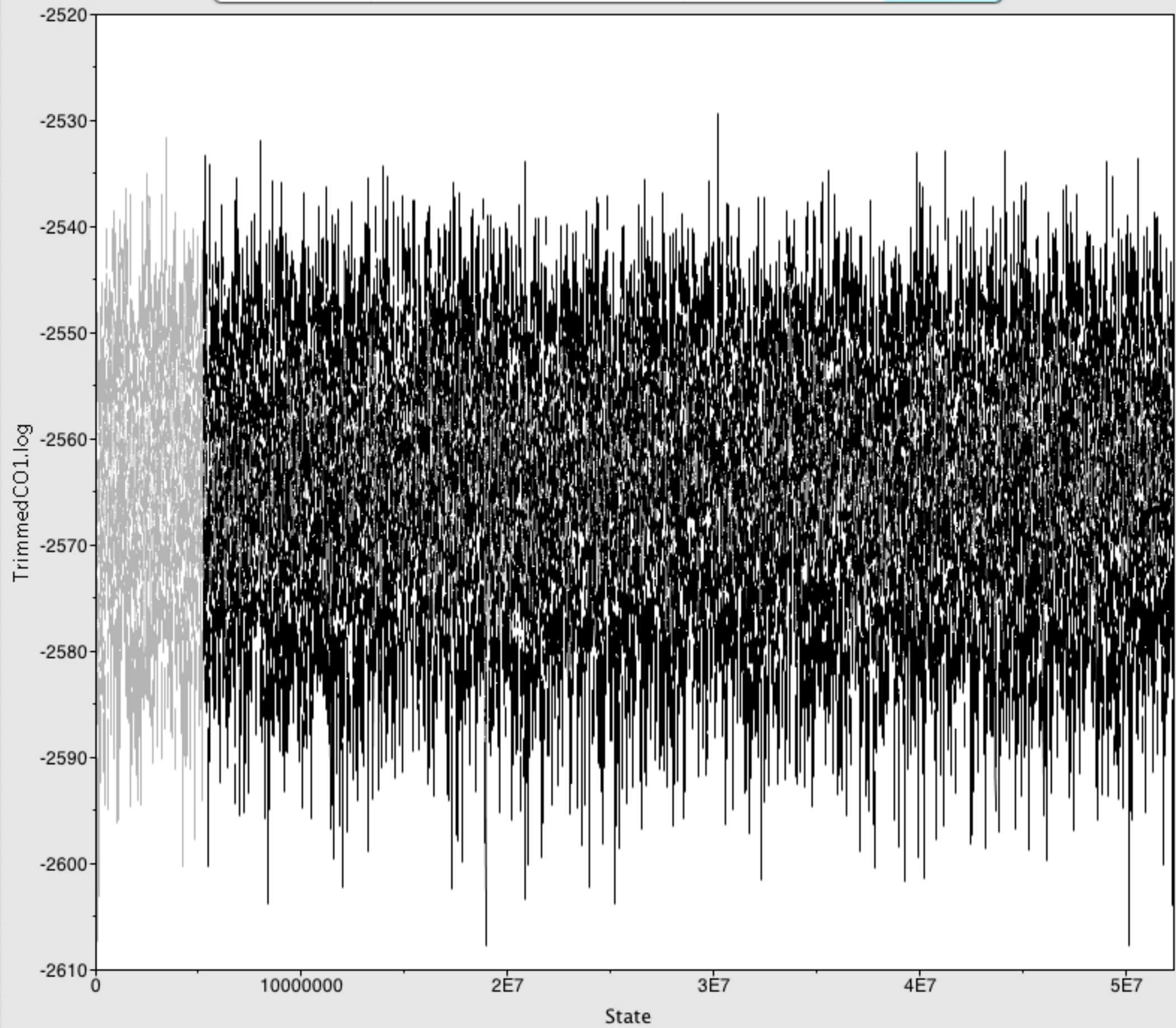
Marginal Prob Distribution

Joint-Marginal

Trace



μ Estimates Δ Marginal Prob Distribution ∇ Joint-Marginal Trace



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”.

TreeAnnotator v2.4.8

Burnin percentage:

Posterior probability limit:

Target tree type:

Node heights:

Target Tree File:

Input Tree File:

Output File:

Low memory: ☐

Total trees have 10001, where 9001 are used.
Total unique clades: 4886

Finding maximum credibility tree...
Analyzing 9001 trees...

0 25 50 75 100
|-----|-----|-----|-----|

Highest Log Clade Credibility: -71.83446159438284
Collecting node information...

0 25 50 75 100
|-----|-----|-----|-----|

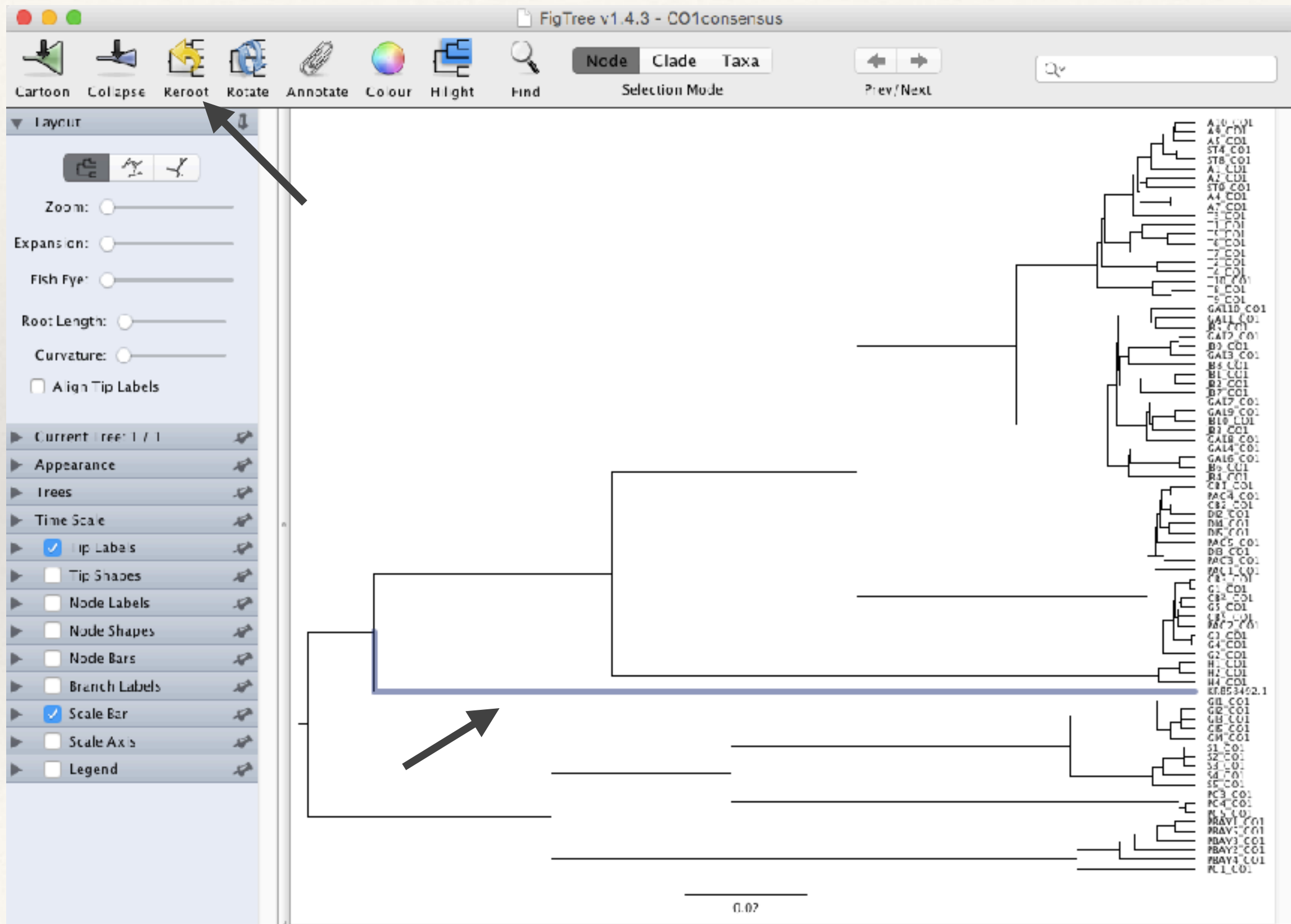
Annotating target tree...
Setting node heights...

0 25 50 75 100
|-----|-----|-----|-----|

Writing annotated tree....
Finished - Quit program to exit.

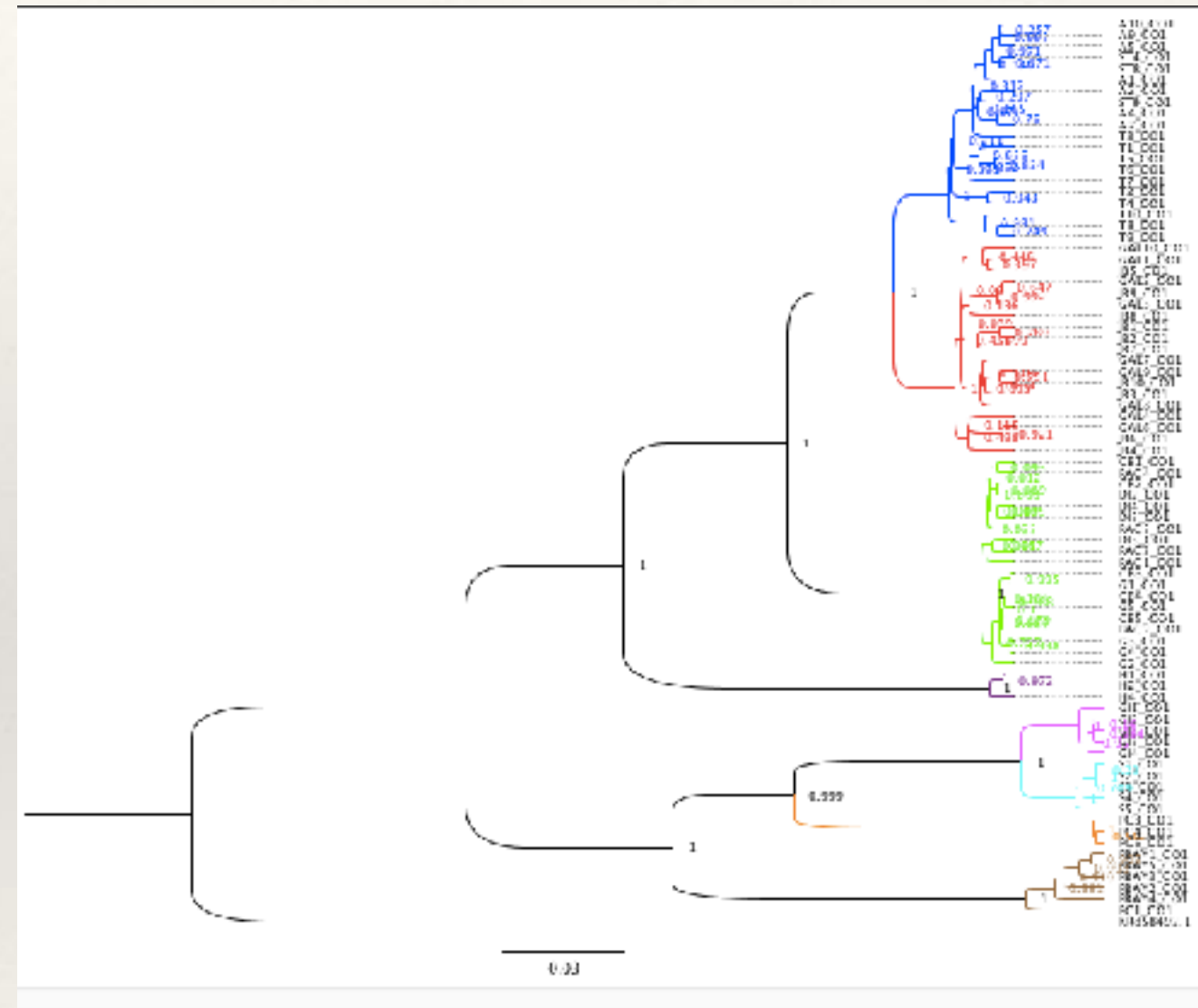
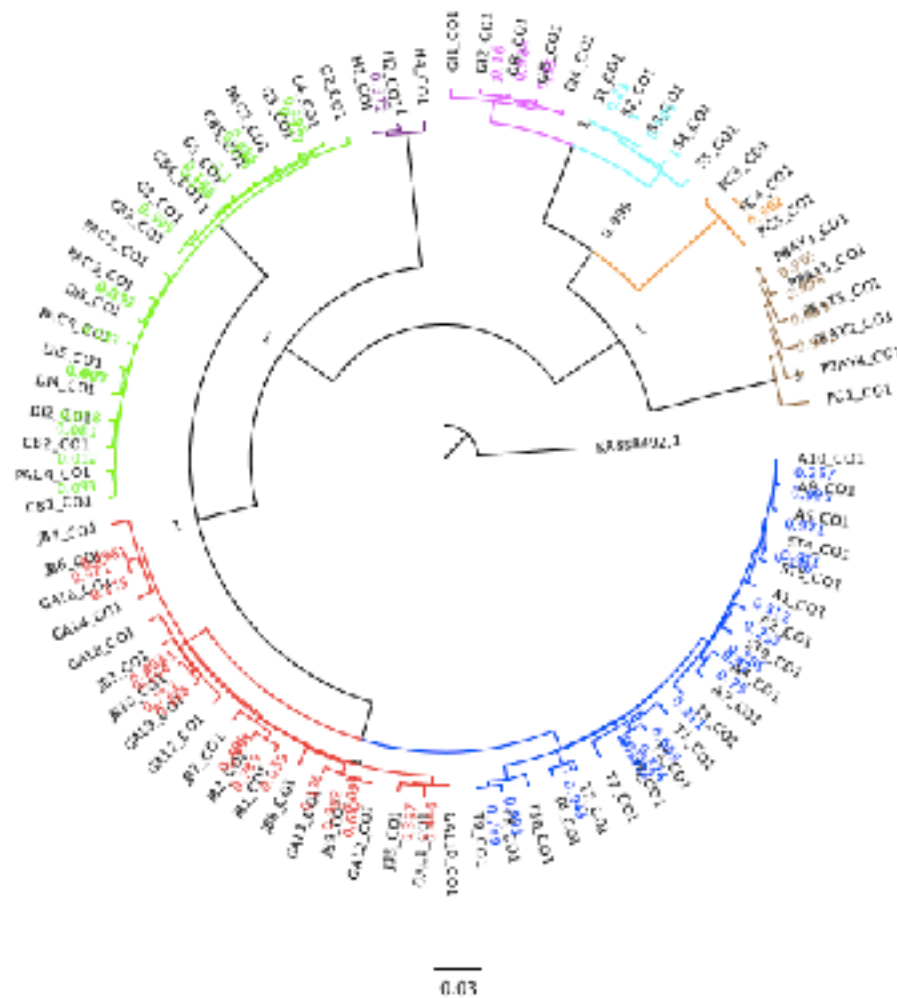
Open FigTree. Go to File —> Open —> choose “CO1consensus.tree”

On the consensus tree, find and select the branch leading to the outgroup (KR858492.1), and select “re-root” at the top.

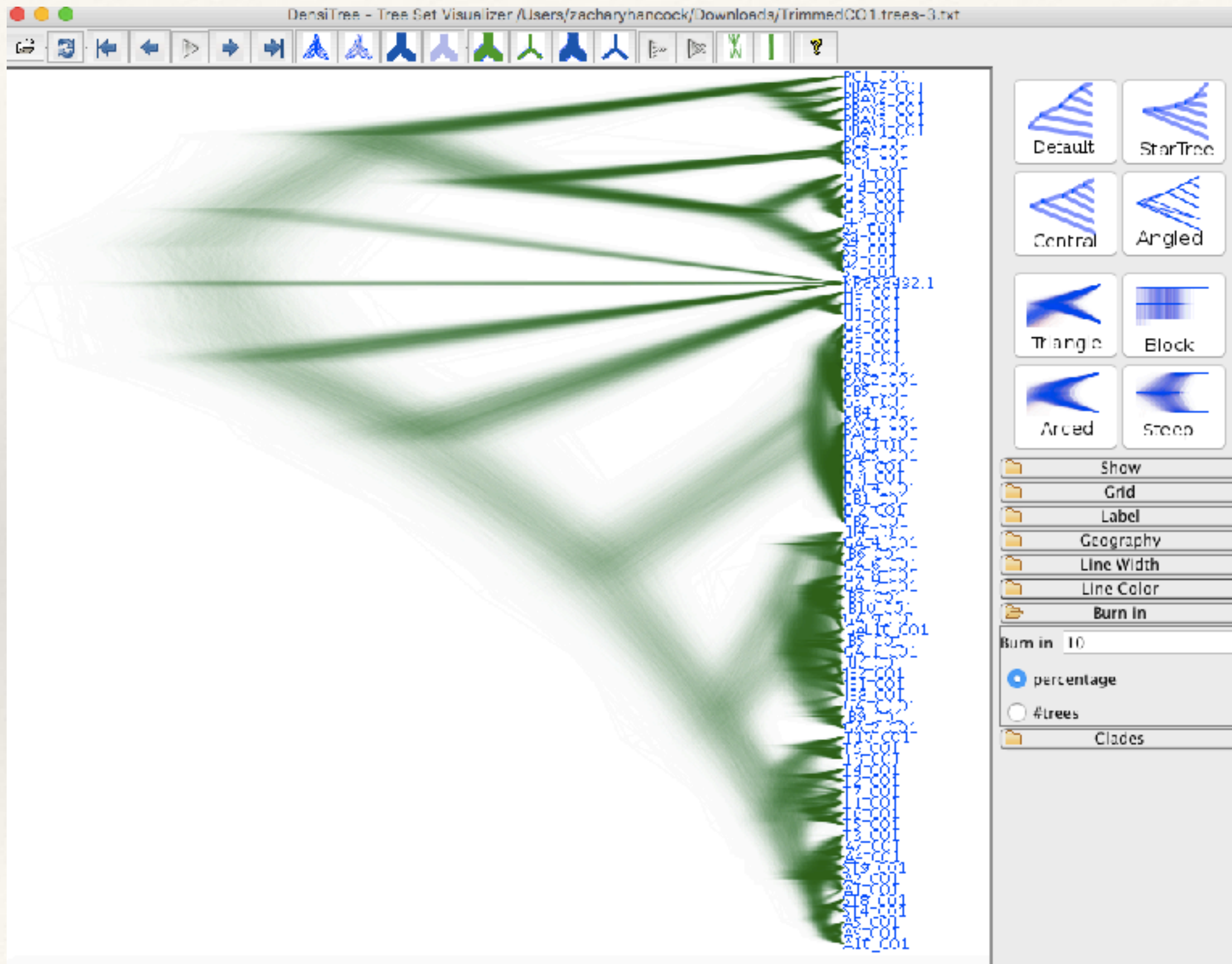


On the left-hand tab, check the box labelled “Node Labels”, and hit the scroll down until you find “posterior”. This will plot the posterior of each node.

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>.