Answers

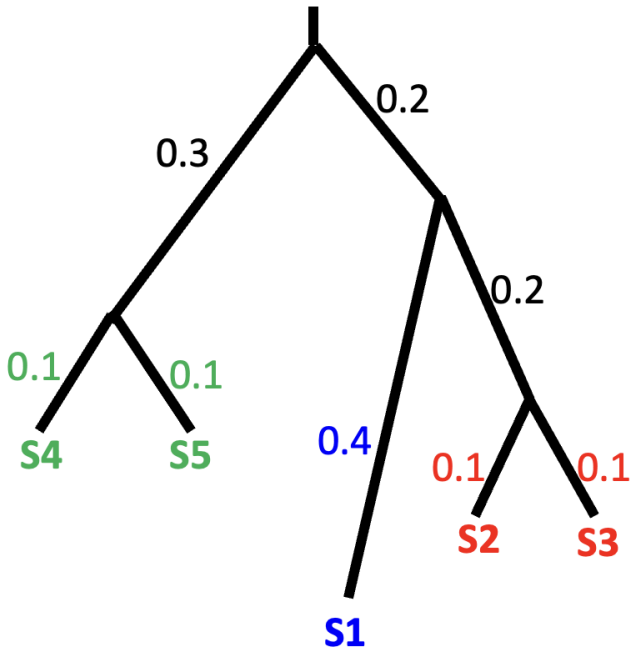
Practical 1

Let's plot and manipulate this tree:

1) represent this tree in Newick format.

2) Save the Newick string in a text file, then open it (File->Open) in **FigTree**

Does the tree look the same? Try to make it look similar (select branch, use "Rotate")!



Newick: "((S1:0.4,(S2:0.1,S3:0.1):0.2):0.2,(S4:0.1,S5:0.1):0.3);"

Practical 2

1) Run MAFFT of the primate-mtDNA-unaligned.fasta file

```
Making a distance matrix ..
    1 / 12
done.

Constructing a UPGMA tree (efffree=0) ...
   10 / 12
done.

Progressive alignment 1/2...
STEP    11 / 11  f
done.

Making a distance matrix from msa..
    0 / 12
done.

Constructing a UPGMA tree (efffree=1) ...
   10 / 12
done.

Progressive alignment 2/2...
STEP    11 / 11  f
done.

disttbfast (nuc) Version 7.520
alg=A, model=DNA200 (2), 1.53 (4.59), -0.00 (-0.00), noshift, amax=0.0
0 thread(s)


Strategy:
 FFT-NS-2 (Fast but rough)
 Progressive method (guide trees were built 2 times.)
```

Practical 2

2) Compare aligned and unaligned files : what's the difference ?

Unaligned

```
>Gorilla
AAGCTTCACCGGCGCAGTTGTTCTTATAATTGCCCACGGACTTACATCATCATTATTATT
CTGCCTAGCAAACTCAAACTACGAACGAACCCACAGCCGCATCATAATTCTCTCTCAAGG
ACTCCAAACCCTACTCCCACTAATAGCCCTTTGATGACTTCTGGCAAGCCTCGCCAACCT
CGCCTTACCCCCCACCATTAACCTACTAGGAGAGCTCTCCGTACTAGTAACCACATTCTC
CTGATCAAACACCACCCTTTTACTTACAGGATCTAACATACTAATTACAGCCCTGTACTC
CCTTTATATATTTACCACAACACAATGAGGCCCACTCACACACCACATCACCAACATAAA
ACCCTCATTTACACGAGAAAACATCCTCATATTCATGCACCTATCCCCCATCCTCCTCCT
ATCCCTCAACCCCGATATTATCACCGGGTTCACCTCCTGTAAATATAGTTTAACCAAAAC
ATCAGATTGTGAATCTGATAACAGAGGCTCACAACCCCTTATTTACCGAGAAAGCTCGTA
AGAGCTGCTAACTCATACCCCCGTGCTTGACAACATGGCTTTCTCAACTTTTAAAGGATA
ACAGCTATCCATTGGTCTTAGGACCCAAAAATTTTGGTGCAACTCCAAATAAAAGTAATA
ACTATGTACGCTACCATAACCACCTTAGCCCTAACTTCCTTAATTCCCCCTATCCTTACC
ACCTTCATCAATCCTAACAAAAAAAGCTCATACCCCCATTACGTAAAATCTATCGTCGCA
TCCACCTTTATCATCAGCCTCTTCCCCACAACAATATTTCTATGCCTAGACCAAGAAGCT
ATTATCTCAAGCTGACACTGAGCAACAACCCAAACAATTCAACTCTCCCTAAGCTT
```

Aligned

```
>Gorilla
aagcttcaccggcgcagttgttcttataattgcccacggacttacatcatcattattatt
ctgcctagcaaactcaaactacgaacgaacccacagccgcatcataattctctctcaagg
actccaaaccctactcccactaatagcccttt gatgacttctggcaagcctcgccaacct
cgccttacccccc accattaacctactaggagagctctccgtactagtaaccacattctc
ctgatcaaacaccacccttttacttacaggatctaacatactaattacagccctgtactc
cctttatatatttaccacaacacaatgaggcccactcacacaccacatcaccaacataaa
accctcatttacacgagaaaacatcctcatattcatgcacctatccccc atcctcctcct
atccctcaaccccgatattatcaccgggttcacctcctgtaaatatagtttaaccaaaac
atcagattgtgaatctgataacagaggctcaca-acccc ttatttaccgagaaagc-tcg
taagagctgctaactcatacccccgtgcttgacaacatggctttctcaactttt aaaggga
taacagctatccattggtcttaggacccaaaaattttggtgcaactccaaataaaagtaa
taactatgtacgctaccataaccaccttagccctaacttccttaattcccccta tcctta
ccaccttcatcaatcctaacaaaaaaagctcatacccccattacgtaaaatctatcgtcg
catccacctttatcatcagcctcttccccacaacaatatttctatgcctagaccaagaag
ctattatctcaagctgacactgagcaacaacccaaacaattcaactctccctaagctt
```

The IMPORTANT difference is that in the aligned file all sequences have the same length, and gap characters "-" have been added

Practical 2

3) Run MAFFT also on the SARS-CoV-2 sequences. Why does it take longer ?

```
nthread = 0
nthreadpair = 0
nthreadtb = 0
ppenalty_ex = 0
stacksize: 8192 kb
generating a scoring matrix for nucleotide (dist=200) ... done
Gap Penalty = -1.53, +0.00, +0.00


Making a distance matrix ..
  101 / 136
done.

Constructing a UPGMA tree (efffree=0) ...
  130 / 136
done.

Progressive alignment 1/2...
STEP   135 / 135  f
done.

Making a distance matrix from msa..
  100 / 136
done.

Constructing a UPGMA tree (efffree=1) ...
  130 / 136
done.

Progressive alignment 2/2...
STEP   135 / 135  f
done.

disttbfast (nuc) Version 7.520
alg=A, model=DNA200 (2), 1.53 (4.59), -0.00 (-0.00), noshift, amax=0.0
0 thread(s)


Strategy:
 FFT-NS-2 (Fast but rough)
 Progressive method (guide trees were built 2 times.)

If unsure which option to use, try 'mafft --auto input > output'.
For more information, see 'mafft --help', 'mafft --man' and the mafft page.

The default gap scoring scheme has been changed in version 7.110 (2013 Oct).
It tends to insert more gaps into gap-rich regions than previous versions.
To disable this change, add the --leavegappyregion option.
```

We now have more (136 vs 11) and longer (30,000 vs 100s of bp) genomes, hence the longer runtime. Higher divergence also leads to longer runtime

Practical 3

1) Run IQ-TREE 2 with GTR substitution model on the alignments we previously generated

```
Computing ML distances based on estimated model parameters...
Computing ML distances took 0.000700 sec (of wall-clock time) 0.000672 sec(of CPU time)
Computing RapidNJ tree took 0.000043 sec (of wall-clock time) 0.000051 sec (of CPU time)
Log-likelihood of RapidNJ tree: -5958.376
--------------------------------------------------------------------
|             INITIALIZING CANDIDATE TREE SET                      |
--------------------------------------------------------------------
Generating 98 parsimony trees... 0.066 second
Computing log-likelihood of 97 initial trees ... 0.045 seconds
Current best score: -5945.484

Do NNI search on 20 best initial trees
Estimate model parameters (epsilon = 0.100)
BETTER TREE FOUND at iteration 1: -5944.643
Iteration 10 / LogL: -5944.665 / Time: 0h:0m:0s
Iteration 20 / LogL: -5944.646 / Time: 0h:0m:0s
Finish initializing candidate tree set (1)
Current best tree score: -5944.643 / CPU time: 0.167
Number of iterations: 20
--------------------------------------------------------------------
|               OPTIMIZING CANDIDATE TREE SET                      |
--------------------------------------------------------------------
Iteration 30 / LogL: -5949.319 / Time: 0h:0m:0s (0h:0m:0s left)
Iteration 40 / LogL: -5944.647 / Time: 0h:0m:0s (0h:0m:0s left)
Iteration 50 / LogL: -5944.651 / Time: 0h:0m:0s (0h:0m:0s left)
Iteration 60 / LogL: -5944.656 / Time: 0h:0m:0s (0h:0m:0s left)
Iteration 70 / LogL: -5944.645 / Time: 0h:0m:0s (0h:0m:0s left)
Iteration 80 / LogL: -5944.728 / Time: 0h:0m:0s (0h:0m:0s left)
Iteration 90 / LogL: -5944.654 / Time: 0h:0m:0s (0h:0m:0s left)
Iteration 100 / LogL: -5944.668 / Time: 0h:0m:0s (0h:0m:0s left)
TREE SEARCH COMPLETED AFTER 102 ITERATIONS / Time: 0h:0m:0s

--------------------------------------------------------------------
|                    FINALIZING TREE SEARCH                        |
--------------------------------------------------------------------
Performs final model parameters optimization
Estimate model parameters (epsilon = 0.010)
1. Initial log-likelihood: -5944.643
Optimal log-likelihood: -5944.643
Rate parameters:  A-C: 10.05648  A-G: 28.54357  A-T: 5.60546  C-G: 2.88589  C-T: 35.28044  G-T: 1.00000
Base frequencies:  A: 0.324  C: 0.304  G: 0.106  T: 0.266
Parameters optimization took 1 rounds (0.001 sec)
BEST SCORE FOUND : -5944.643
Total tree length: 1.535

Total number of iterations: 102
CPU time used for tree search: 0.661 sec (0h:0m:0s)
Wall-clock time used for tree search: 0.481 sec (0h:0m:0s)
Total CPU time used: 0.692 sec (0h:0m:0s)
Total wall-clock time used: 0.520 sec (0h:0m:0s)

Analysis results written to:
  IQ-TREE report:                /Users/demaio/Desktop/presentations/2023/phylogenetics/materials/01-foundations/primate-mtDNA_mafft-aligned_iqtreeGTR.iqtree
  Maximum-likelihood tree:       /Users/demaio/Desktop/presentations/2023/phylogenetics/materials/01-foundations/primate-mtDNA_mafft-aligned_iqtreeGTR.treefile
  Likelihood distances:          /Users/demaio/Desktop/presentations/2023/phylogenetics/materials/01-foundations/primate-mtDNA_mafft-aligned_iqtreeGTR.mldist
  Screen log file:               /Users/demaio/Desktop/presentations/2023/phylogenetics/materials/01-foundations/primate-mtDNA_mafft-aligned_iqtreeGTR.log
```

Distance matrix calculation and NJ

Estimate initial parsimony trees

Optimize parsimony trees with ML

Model optimization

Summary

Output

Practical 3

2) Now let IQ-TREE 2 investigate which model best fits the data (ModelFinder). Does IQ-TREE 2 select a model with rate variation ?

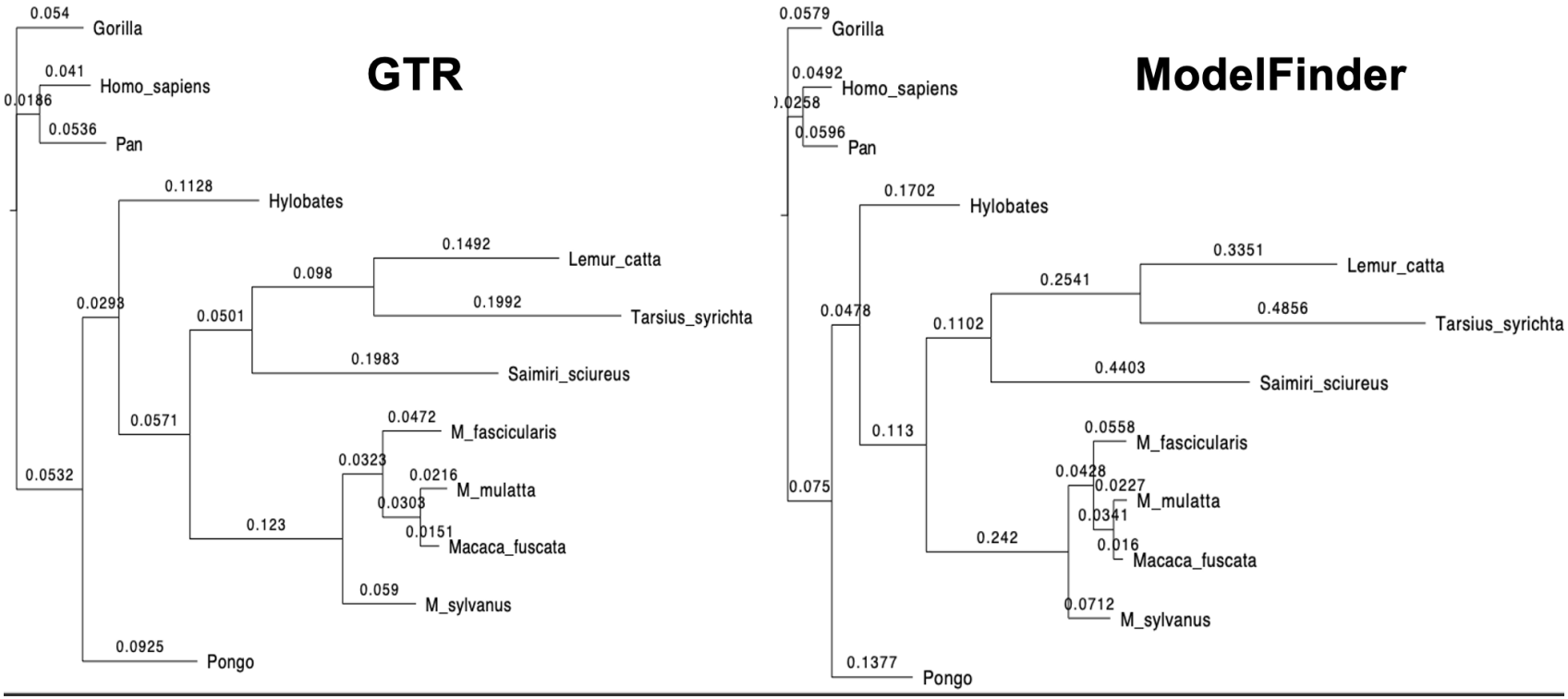
```
NOTE: ModelFinder requires 1 MB RAM!
ModelFinder will test up to 286 DNA models (sample size: 898) ...
 No. Model         -LnL         df  AIC          AICc         BIC
   1 GTR+F         5944.739     29  11947.478    11949.483    12086.683
   2 GTR+F+I       5760.157     30  11580.314    11582.459    11724.319
   3 GTR+F+G4      5721.595     30  11503.190    11505.335    11647.195
   4 GTR+F+I+G4    5722.005     31  11506.011    11508.302    11654.816
   5 GTR+F+R2      5726.477     31  11514.953    11517.244    11663.759
   6 GTR+F+R3      5721.563     33  11509.127    11511.724    11667.532
  16 SYM+G4        5822.223     27  11698.446    11700.184    11828.051
  17 SYM+I+G4      5821.099     28  11698.198    11700.067    11832.603
  29 TVM+F+G4      5721.740     29  11501.480    11503.485    11640.685
  30 TVM+F+I+G4    5722.135     30  11504.270    11506.415    11648.275
  42 TVMe+G4       5871.328     26  11794.656    11796.268    11919.461
  43 TVMe+I+G4     5870.562     27  11795.124    11796.862    11924.729
  55 TIM3+F+G4     5728.695     28  11513.390    11515.258    11647.794
  56 TIM3+F+I+G4   5729.225     29  11516.451    11518.455    11655.656
  68 TIM3e+G4      5896.584     25  11843.168    11844.658    11963.172
  69 TIM3e+I+G4    5895.803     26  11843.605    11845.217    11968.410
  81 TIM2+F+G4     5723.844     28  11503.688    11505.557    11638.093
  82 TIM2+F+I+G4   5724.338     29  11506.676    11508.681    11645.881
  94 TIM2e+G4      5828.207     25  11706.414    11707.904    11826.418
  95 TIM2e+I+G4    5827.271     26  11706.542    11708.154    11831.346
 107 TIM+F+G4      5729.096     28  11514.193    11516.062    11648.598
 108 TIM+F+I+G4    5729.512     29  11517.025    11519.029    11656.230
 120 TIMe+G4       5891.360     25  11832.719    11834.210    11952.723
 121 TIMe+I+G4     5890.105     26  11832.209    11833.821    11957.014
 133 TPM3u+F+G4    5728.704     27  11511.408    11513.146    11641.012
 134 TPM3u+F+I+G4  5729.223     28  11514.446    11516.315    11648.850
 146 TPM3+F+G4     5728.704     27  11511.408    11513.146    11641.012
 147 TPM3+F+I+G4   5729.223     28  11514.446    11516.315    11648.850
 159 TPM2u+F+G4    5723.970     27  11501.940    11503.677    11631.544
 160 TPM2u+F+I+G4  5724.449     28  11504.898    11506.767    11639.303
 172 TPM2+F+G4     5723.964     27  11501.928    11503.666    11631.533
 173 TPM2+F+I+G4   5724.449     28  11504.898    11506.767    11639.303
 185 K3Pu+F+G4     5729.080     27  11512.161    11513.899    11641.765
 186 K3Pu+F+I+G4   5729.504     28  11515.008    11516.877    11649.413
 198 K3P+G4        5944.214     24  11936.428    11937.803    12051.632
 199 K3P+I+G4      5943.350     25  11936.699    11938.190    12056.704
 211 TN+F+G4       5730.635     27  11515.269    11517.007    11644.874
 212 TN+F+I+G4     5731.187     28  11518.374    11520.243    11652.778
 224 TNe+G4        5900.735     24  11849.471    11850.845    11964.675
 225 TNe+I+G4      5899.930     25  11849.861    11851.352    11969.865
 237 HKY+F+G4      5730.635     26  11513.271    11514.883    11638.075
 238 HKY+F+I+G4    5731.191     27  11516.382    11518.120    11645.986
 250 K2P+G4        5952.992     23  11951.985    11953.248    12062.389
 251 K2P+I+G4      5952.671     24  11953.343    11954.717    12068.547
 263 F81+F+G4      6127.054     25  12304.109    12305.600    12424.113
 264 F81+F+I+G4    6124.888     26  12301.777    12303.389    12426.581
 276 JC+G4         6273.172     22  12590.343    12591.500    12695.947
 277 JC+I+G4       6270.751     23  12587.502    12588.765    12697.906
Akaike Information Criterion:           TVM+F+G4
Corrected Akaike Information Criterion: TVM+F+G4
Bayesian Information Criterion:         TPM2+F+G4
Best-fit model: TPM2+F+G4 chosen according to BIC

All model information printed to /Users/demaio/Desktop/presentations/2023/phylogenetics/materials/01-foundations/primate-mtDNA_mafft-aligned_iqtreeModelSelection.model.gz
CPU time for ModelFinder: 0.996 seconds (0h:0m:0s)
Wall-clock time for ModelFinder: 1.005 seconds (0h:0m:1s)
```

IQ-TREE 2 selects a model with rate variation (“G4”)

Practical 3

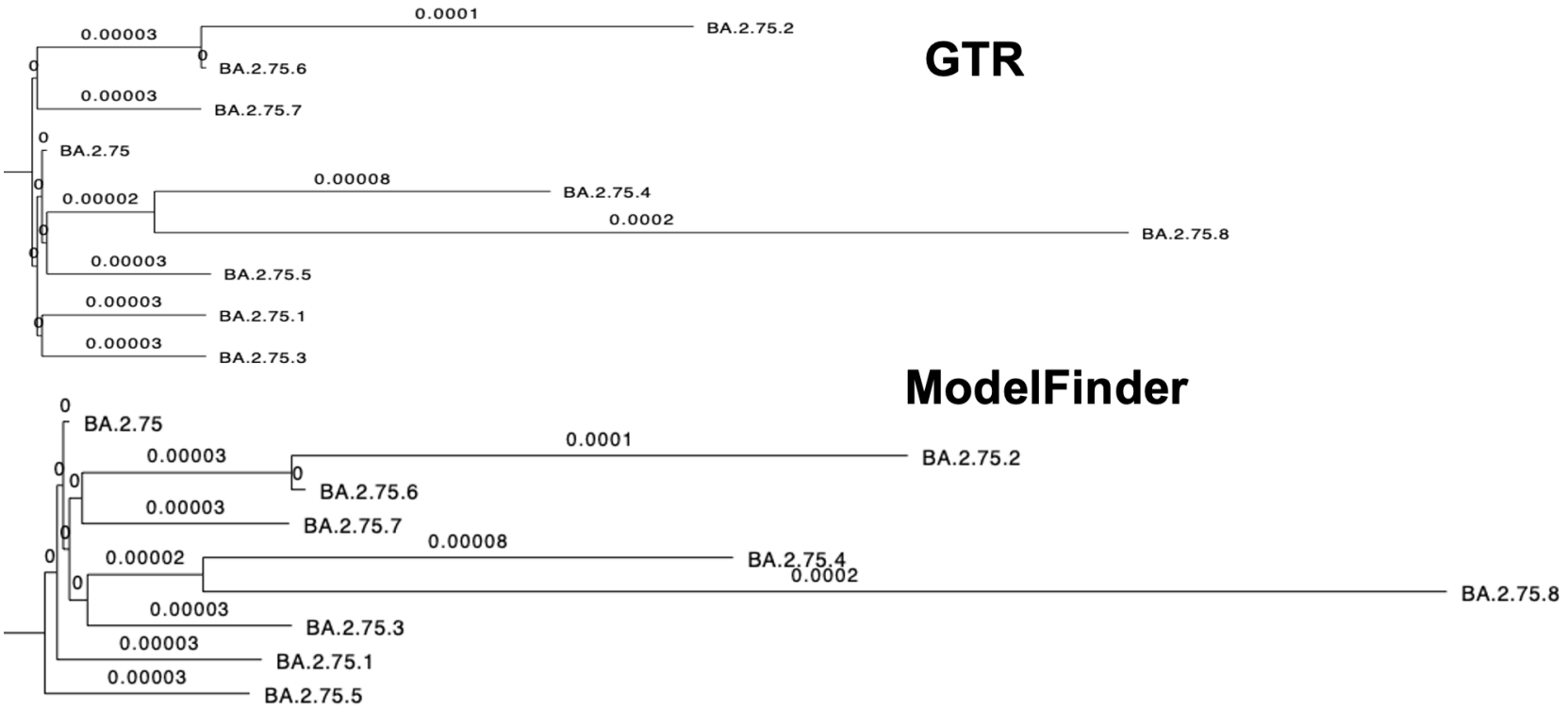
2) Now let IQ-TREE 2 investigate which model best fits the data (ModelFinder). Does the tree differ from before ? How and why ?



IQ-TREE 2 selects a model with rate variation (“G4”)

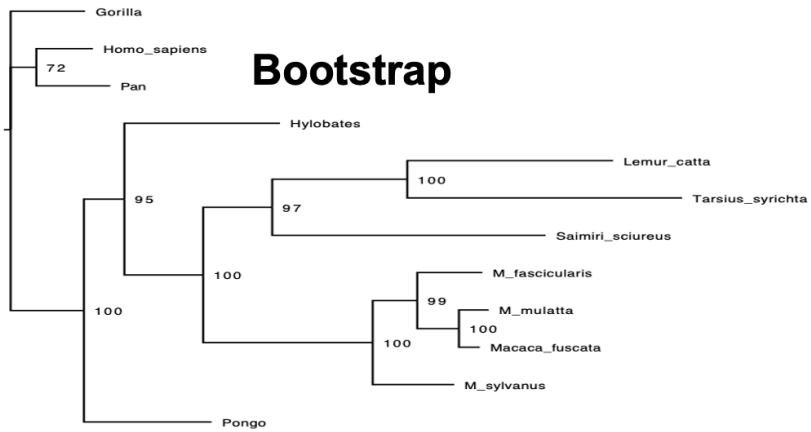
Practical 3

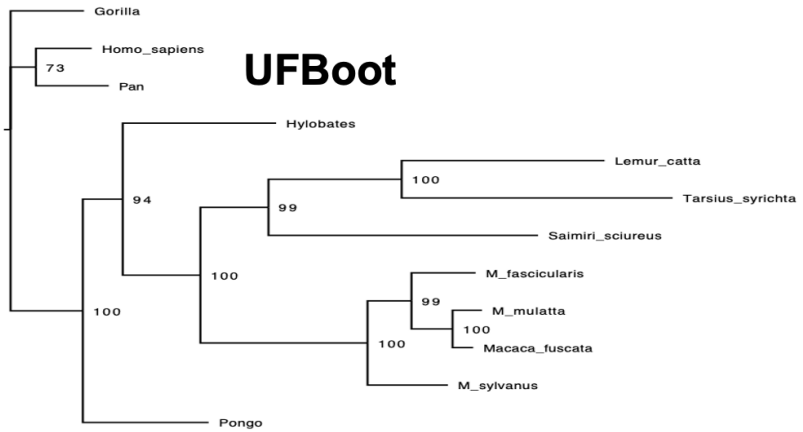
2) Now let IQ-TREE 2 investigate which model best fits the data (ModelFinder). Does the tree differ from before ? How and why ?

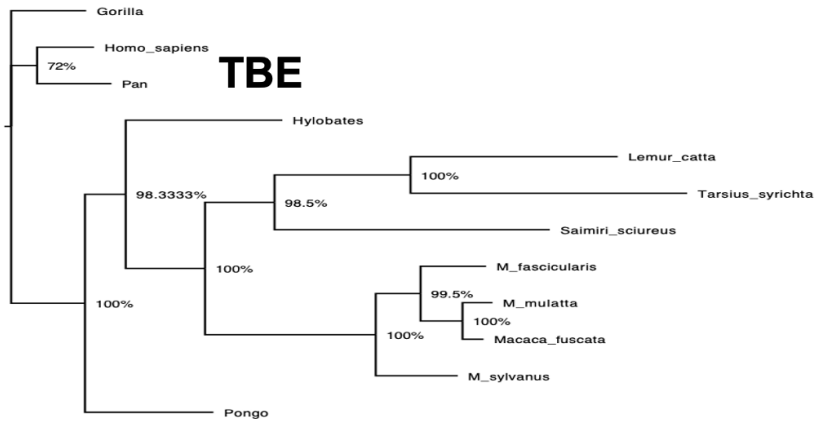


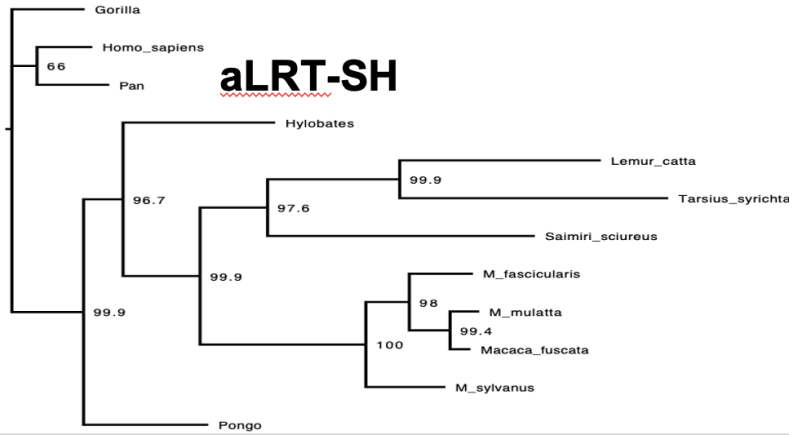
IQ-TREE 2 selects a model with rate variation (“G4”)

Practical 4

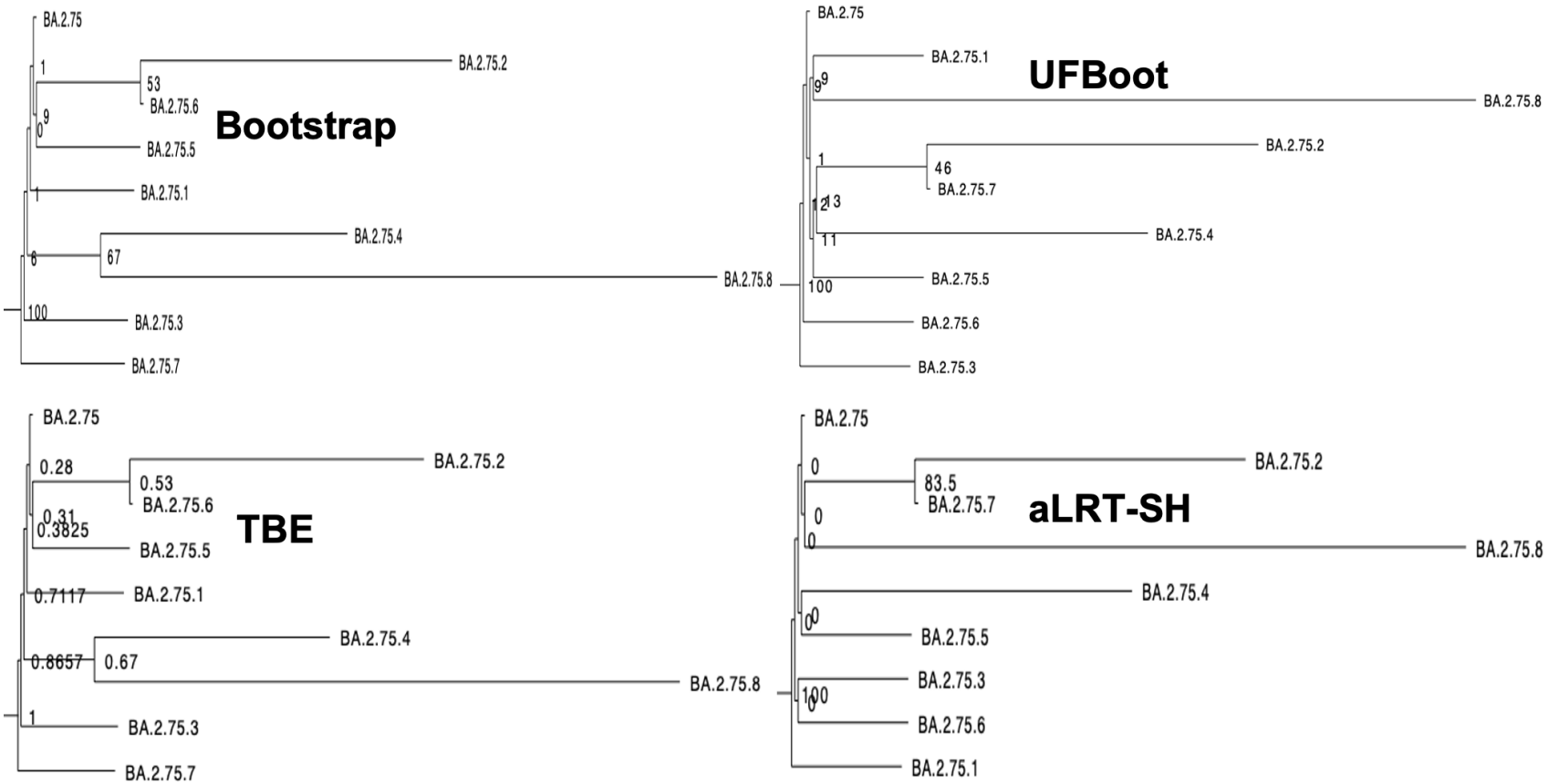

Bootstrap
Gorilla
Homo_sapiens
72
Pan
Hylobates
95
Lemur_catta
100
Tarsius_syrichta
97
Saimiri_sciureus
100
M_fascicularis
99
M_mulatta
100
Macaca_fuscata
100
M_sylvanus
100
Pongo


UFBoot
Gorilla
Homo_sapiens
73
Pan
Hylobates
94
Lemur_catta
100
Tarsius_syrichta
99
Saimiri_sciureus
100
M_fascicularis
99
M_mulatta
100
Macaca_fuscata
100
M_sylvanus
100
Pongo


TBE
Gorilla
Homo_sapiens
72%
Pan
Hylobates
98.3333%
Lemur_catta
100%
Tarsius_syrichta
98.5%
Saimiri_sciureus
100%
M_fascicularis
99.5%
M_mulatta
100%
Macaca_fuscata
100%
M_sylvanus
100%
Pongo

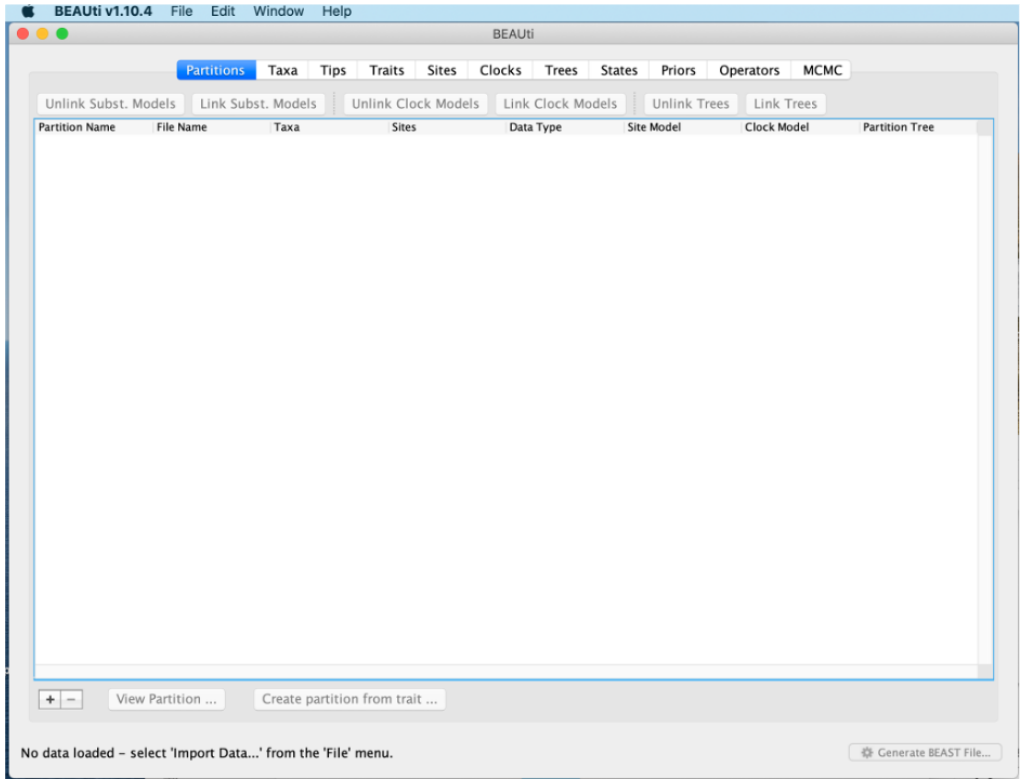

aLRT-SH
Gorilla
Homo_sapiens
66
Pan
Hylobates
96.7
Lemur_catta
99.9
Tarsius_syrichta
97.6
Saimiri_sciureus
99.9
M_fascicularis
98
M_mulatta
99.4
Macaca_fuscata
100
M_sylvanus
99.9
Pongo

Practical 4


BA.2.75
1
53
BA.2.75.2
BA.2.75.6
9
0
BA.2.75.5
Bootstrap
1
BA.2.75.1
BA.2.75.4
6
67
BA.2.75.8
100
BA.2.75.3
BA.2.75.7
BA.2.75
BA.2.75.1
9
9
BA.2.75.8
UFBoot
1
46
BA.2.75.2
BA.2.75.7
1
3
11
BA.2.75.4
100
BA.2.75.5
BA.2.75.6
BA.2.75.3
BA.2.75
0.28
0.53
BA.2.75.2
BA.2.75.6
0.31
0.3825
BA.2.75.5
TBE
0.7117
BA.2.75.1
BA.2.75.4
0.8657
0.67
BA.2.75.8
1
BA.2.75.3
BA.2.75.7
BA.2.75
0
83.5
BA.2.75.2
BA.2.75.7
0
0
BA.2.75.8
aLRT-SH
BA.2.75.4
0
0
BA.2.75.5
100
BA.2.75.3
BA.2.75.6
BA.2.75.1

Practical 5

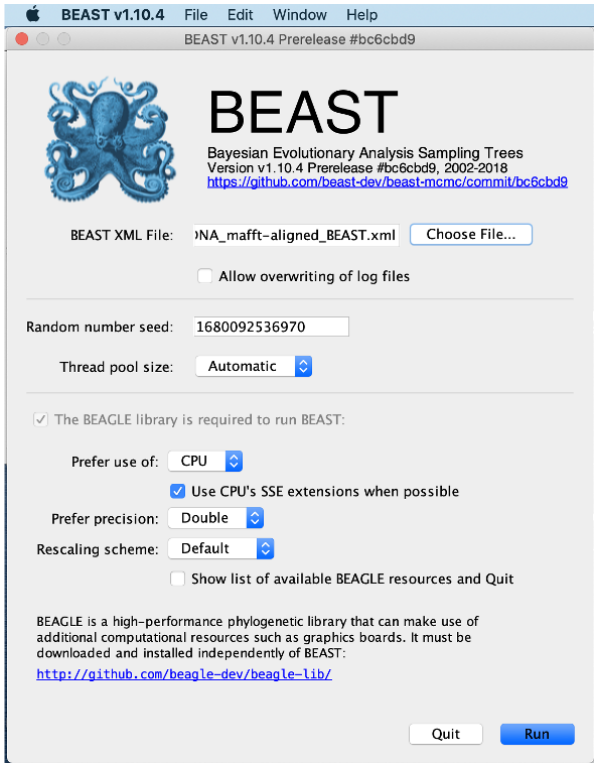
1) Create an input xml file using BEAUti
   - Open BEAUti
   - File->”Import data” ; or, drag alignment file onto BEAUti window
   - Sites-> select GTR substitution model
   - Trees-> Yule process
   - Click “Generate BEAST File”



Practical 5

2) Run the xml file in BEAST

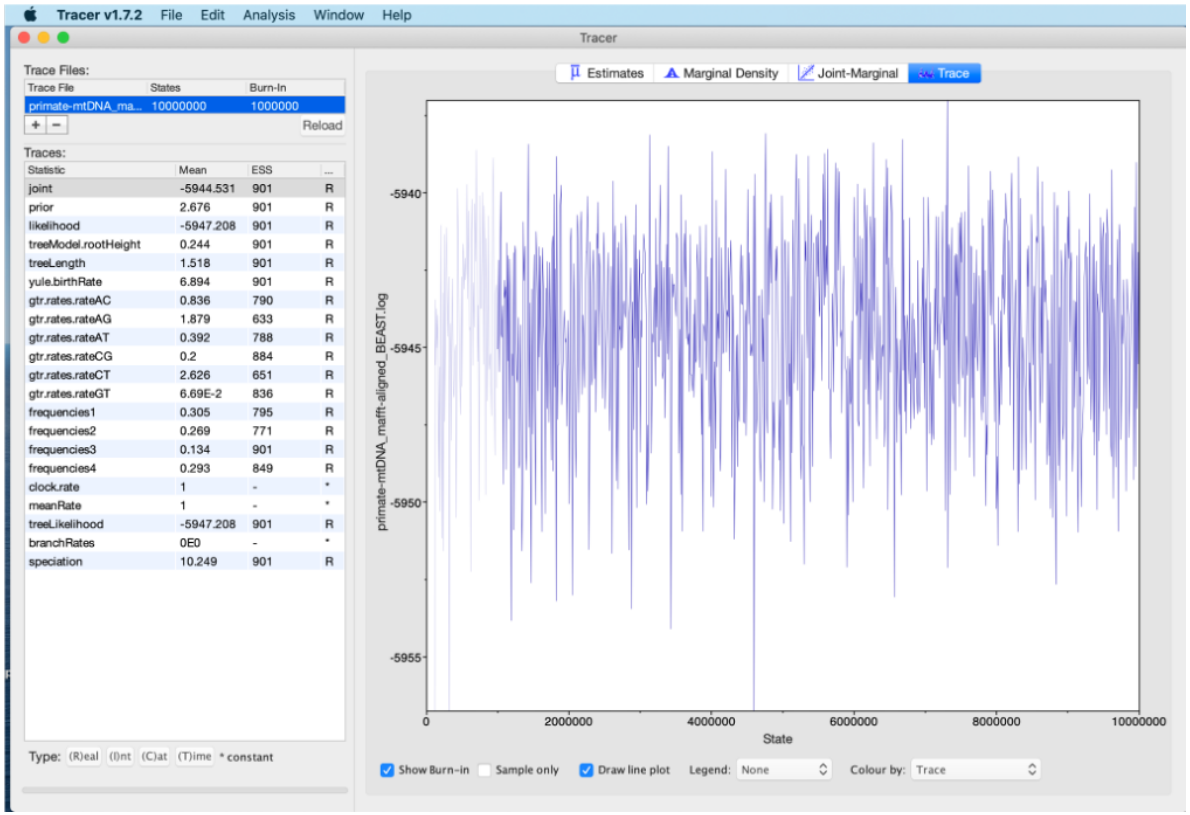
- Give BEAST the .xml file created by BEAUti ; or from the command line : beast primate-mtDNA-mafft-aligned-BEAST.xml
- Click “Generate BEAST File”



Practical 5

3) Analyse the output in Tracer

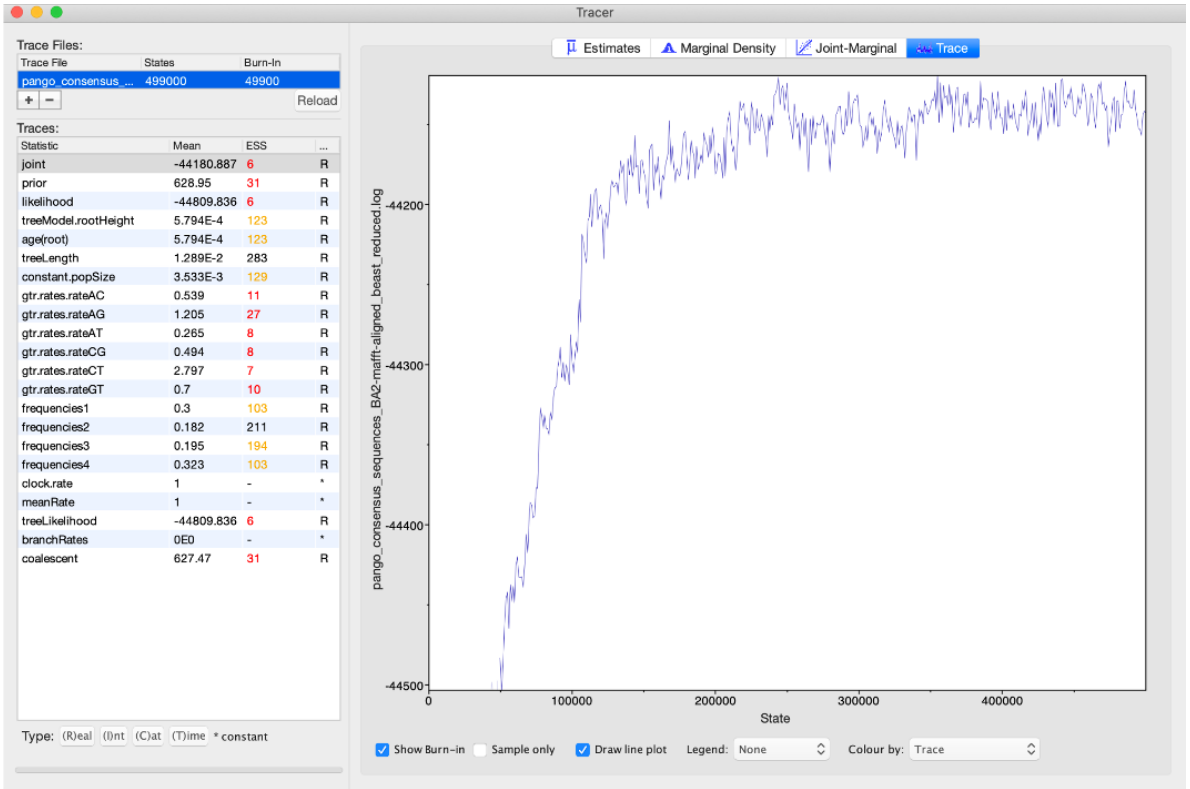
- File->"Import trace file"-> pick .log file created by BEAST ; or just drag it on Tracer the window



Practical 5

3) Analyse the output in Tracer Small ESS values ($<$100, in red) mean that the MCMC needs to run longer

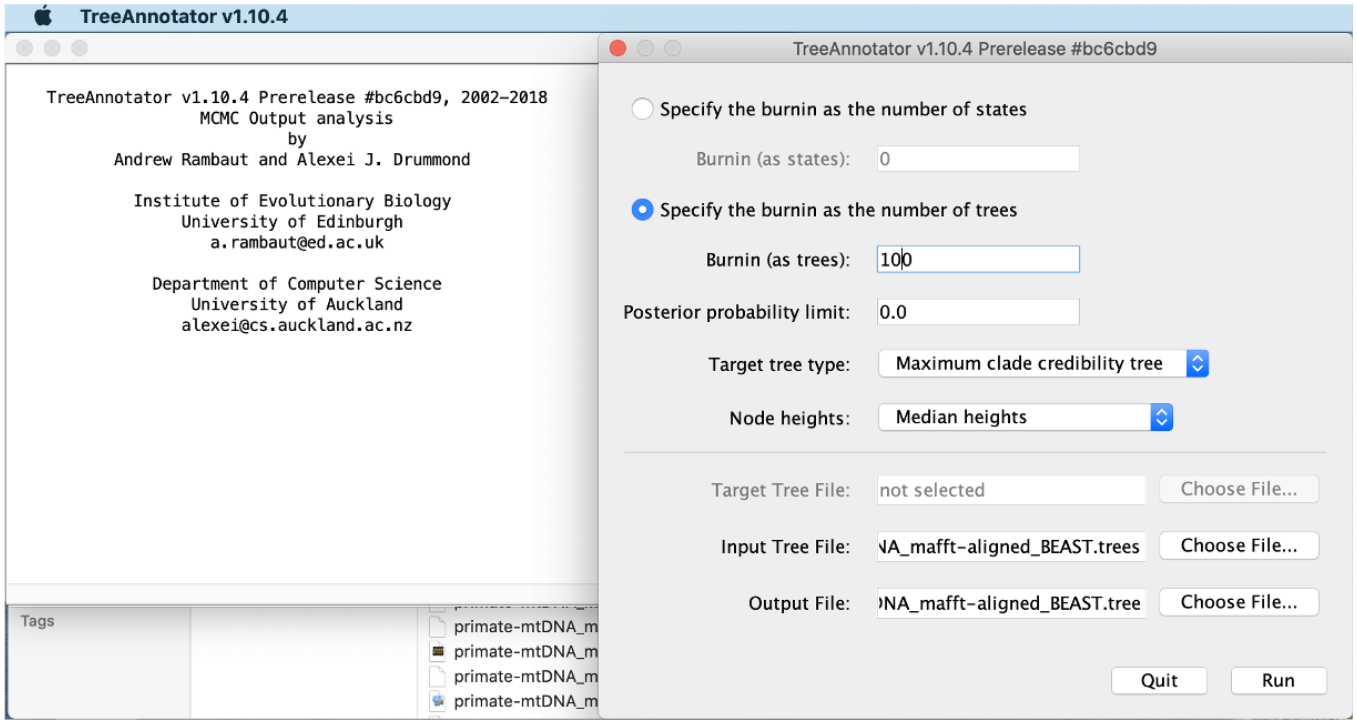
- BEAUti->MCMC-> Length of chain
- File->"Import trace file"-> pick .log file created by BEAST ; or just drag it on Tracer the window



Practical 5

3) Analyse the output in Tracer

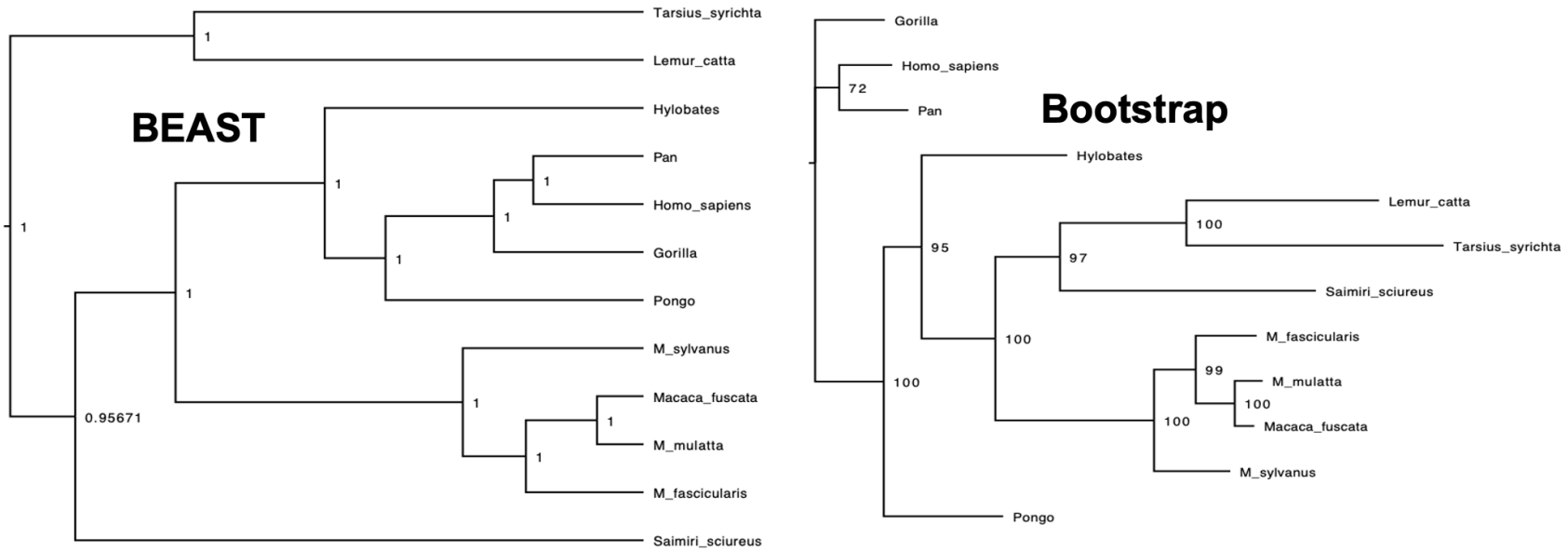
- Pick as input the .trees file created by BEAST. Choose output name



Practical 5

3) Analyse the output in Tracer

- Pick as input the tree file created by TreeAnnotator

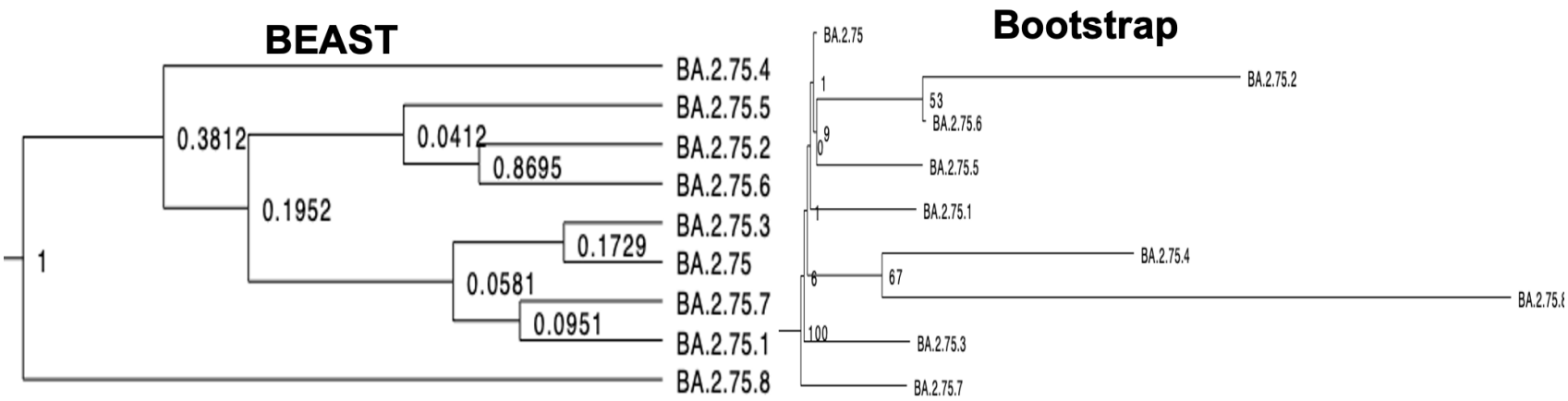


4) Compare to maximum likelihood branch support

Practical 5

3) Analyse the output in Tracer

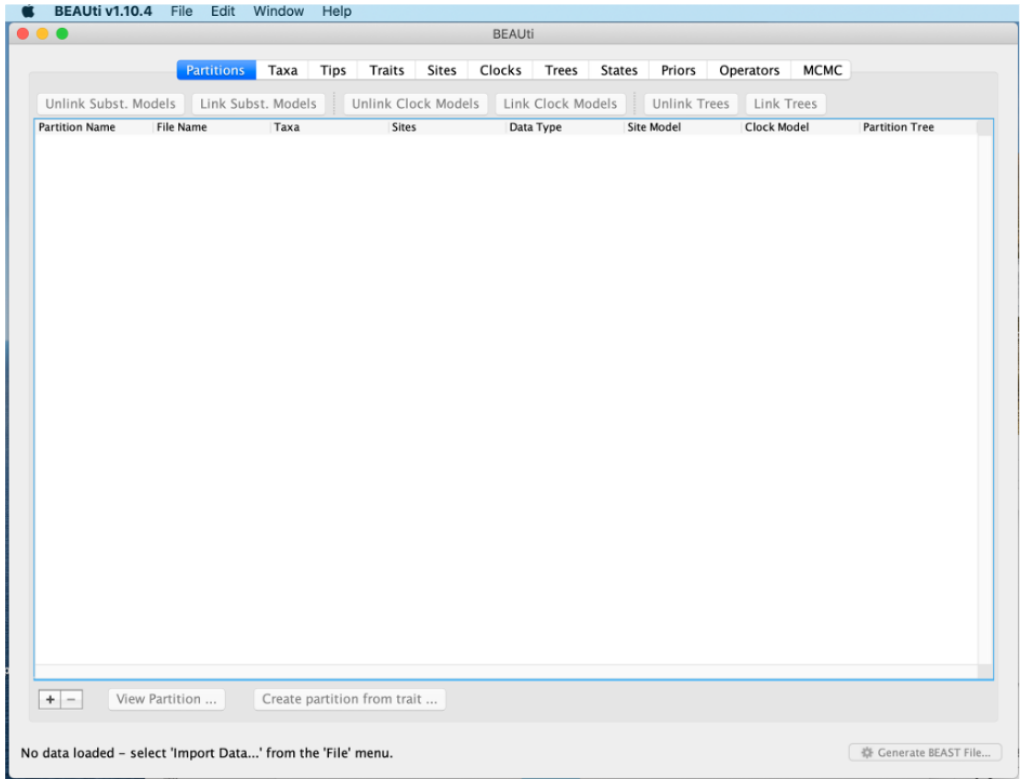
- Pick as input the tree file created by TreeAnnotator



4) Compare to maximum likelihood branch support

Practical 6

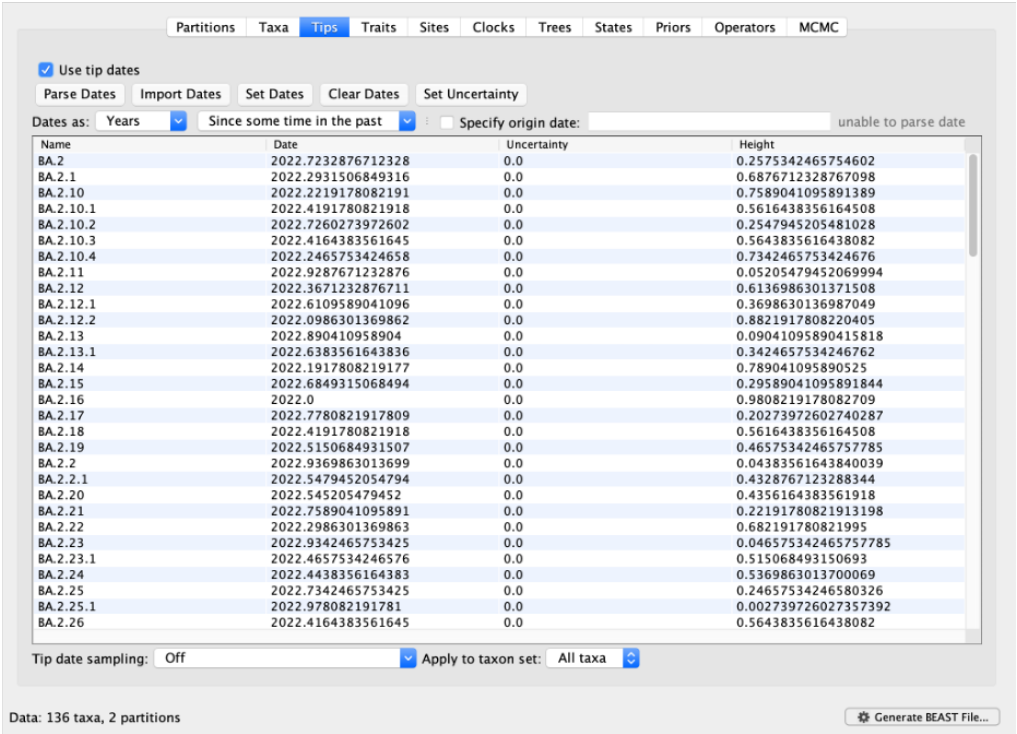
1) Create an input xml file using BEAUti
   - Open BEAUti
   - File->”Import data” ; or, drag alignment file onto BEAUti window
   - Sites-> select GTR substitution model
   - Trees-> Yule process
   - Click “Generate BEAST File”



Practical 6

2) Include time and location data

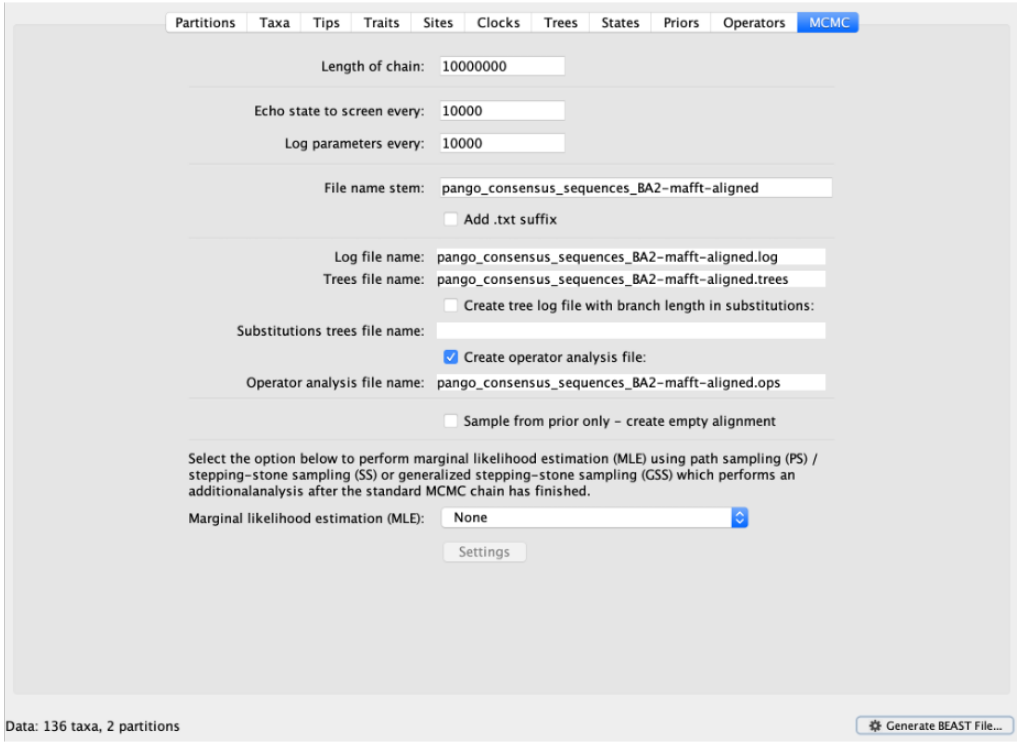
- Tips->”Import dates” ; select given “dates.txt” file
- Traits->”Import traits” ; select given “locations.txt” file
- Traits->”Create partition from trait”



Practical 6

Reduce the number of MCMC steps if you don't want to wait 20 minutes

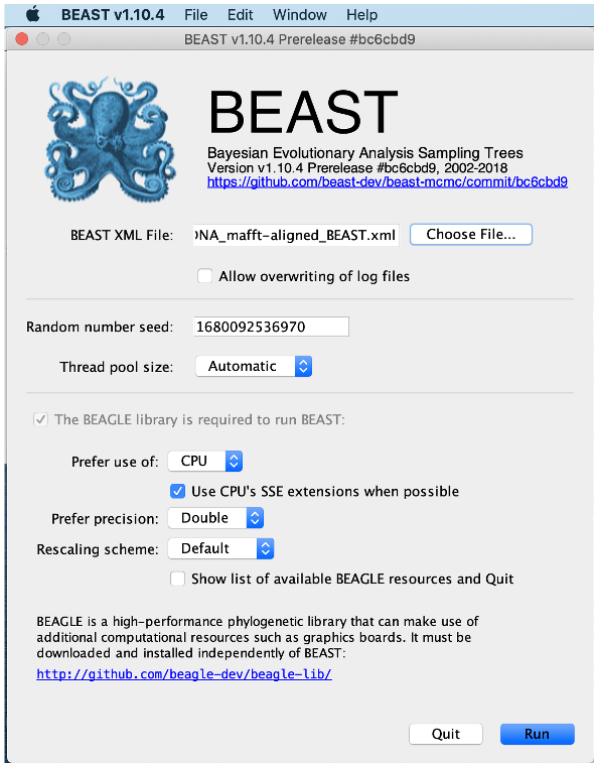
- MCMC->"Length of chain"->1,000,000
- Click "Generate BEAST File"



Practical 6

Now you can run BEAST with the new xml file

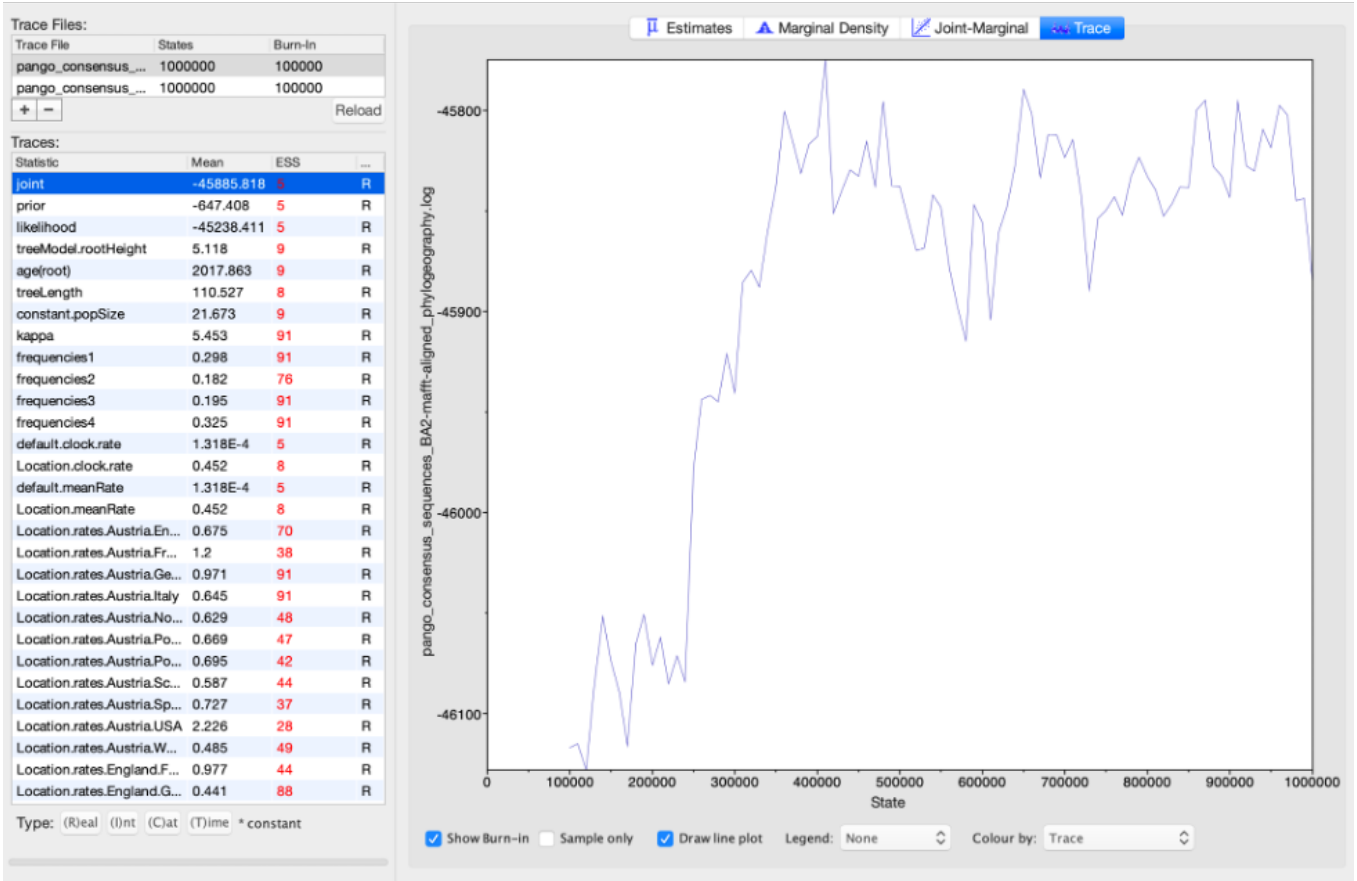
- Give BEAST the .xml file created by BEAUti ; or from the command line : beast primate-mtDNA-mafft-aligned-BEAST.xml
- Click “Generate BEAST File”



Practical 6

Check results in Tracer

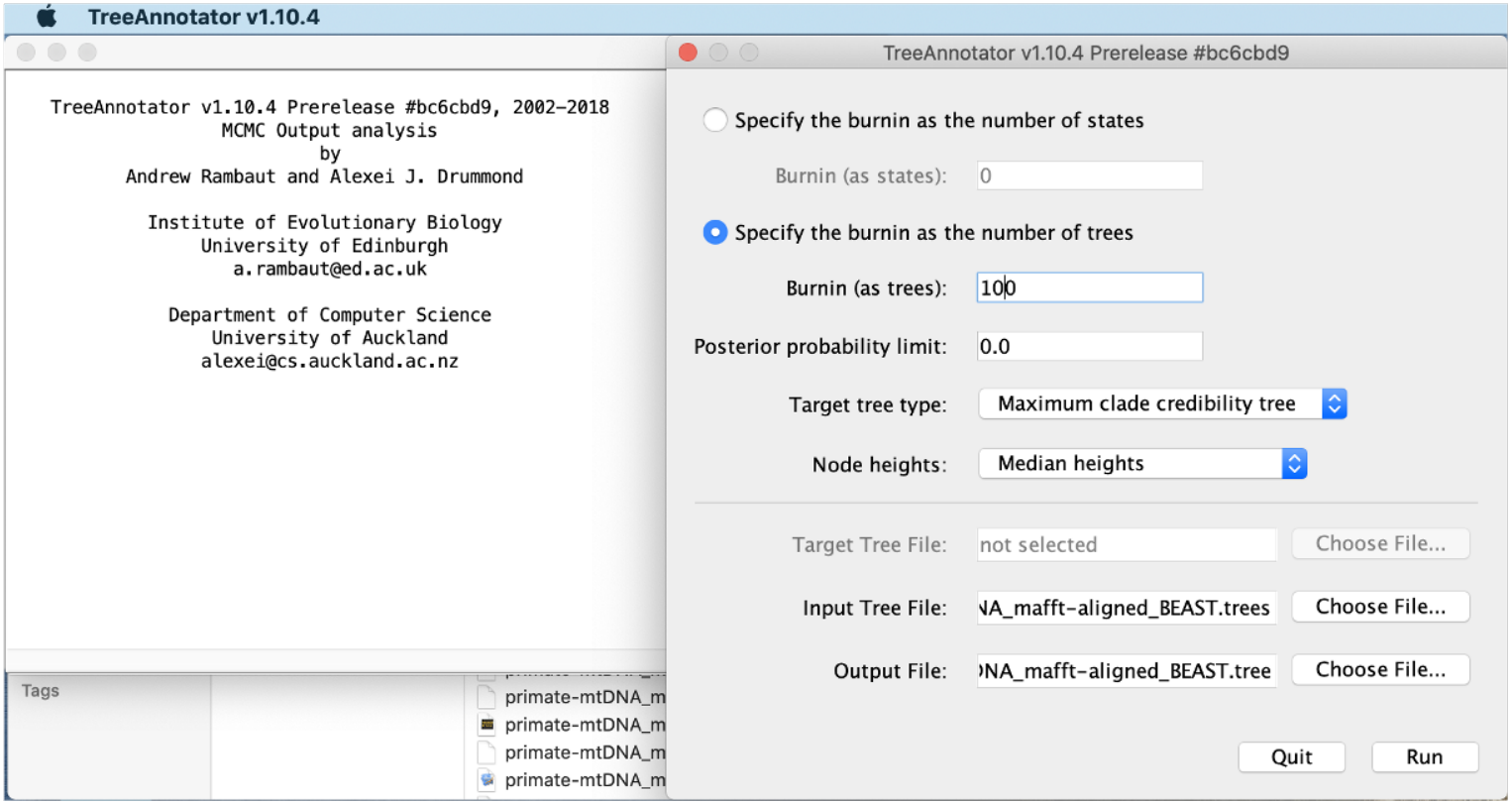
- File->"Import trace file"-> pick .log file created by BEAST ; or just drag it on Tracer the window
- A proper analysis will need to run longer than 1,000,000 MCMC steps



Practical 6

Process the output in TreeAnnotator

- Pick as input the .trees file created by BEAST. Choose output name



Practical 6

Visualize tree in Figtree

- Pick as input the tree file created by TreeAnnotator