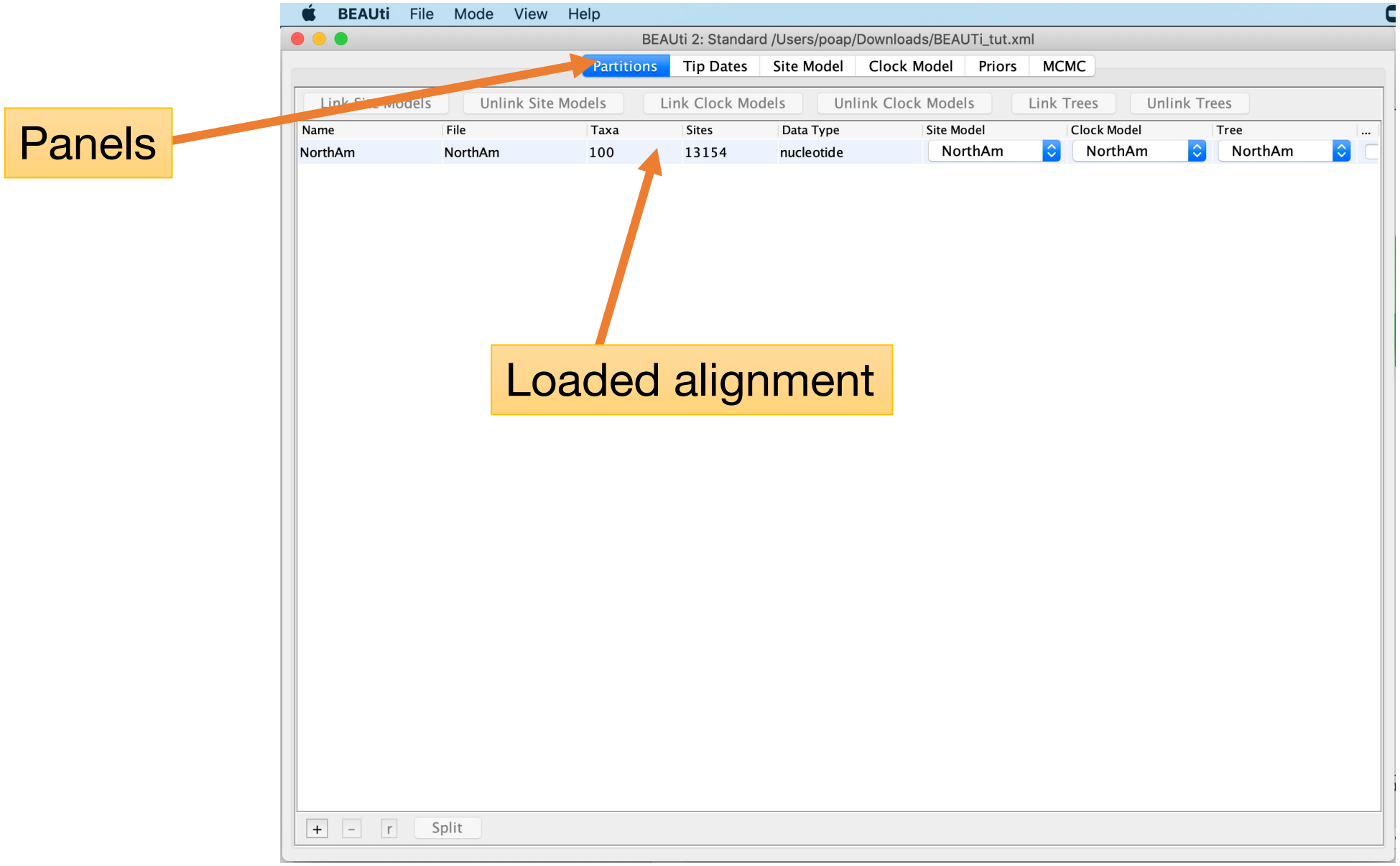
Setting up an analysis in BEAUTi

Ash Porter

Phylogenetics workshop, University of Melbourne, 2021

Getting started with **BEAST**: How to use **BEAUTi**

- BEAUti is a graphical user-interface application to generate XML files to use in BEAST.
- It is the user-friendly way to start BEAST analysis – when you feel more confident, you can edit the XML file directly.
- To begin:
  - Open BEAUTi.
  - Load the alignment file by either dragging in onto the BEAUTi window, or selecting “Import alignment”.
  - We will be using “NorthAm.Nov.fasta”

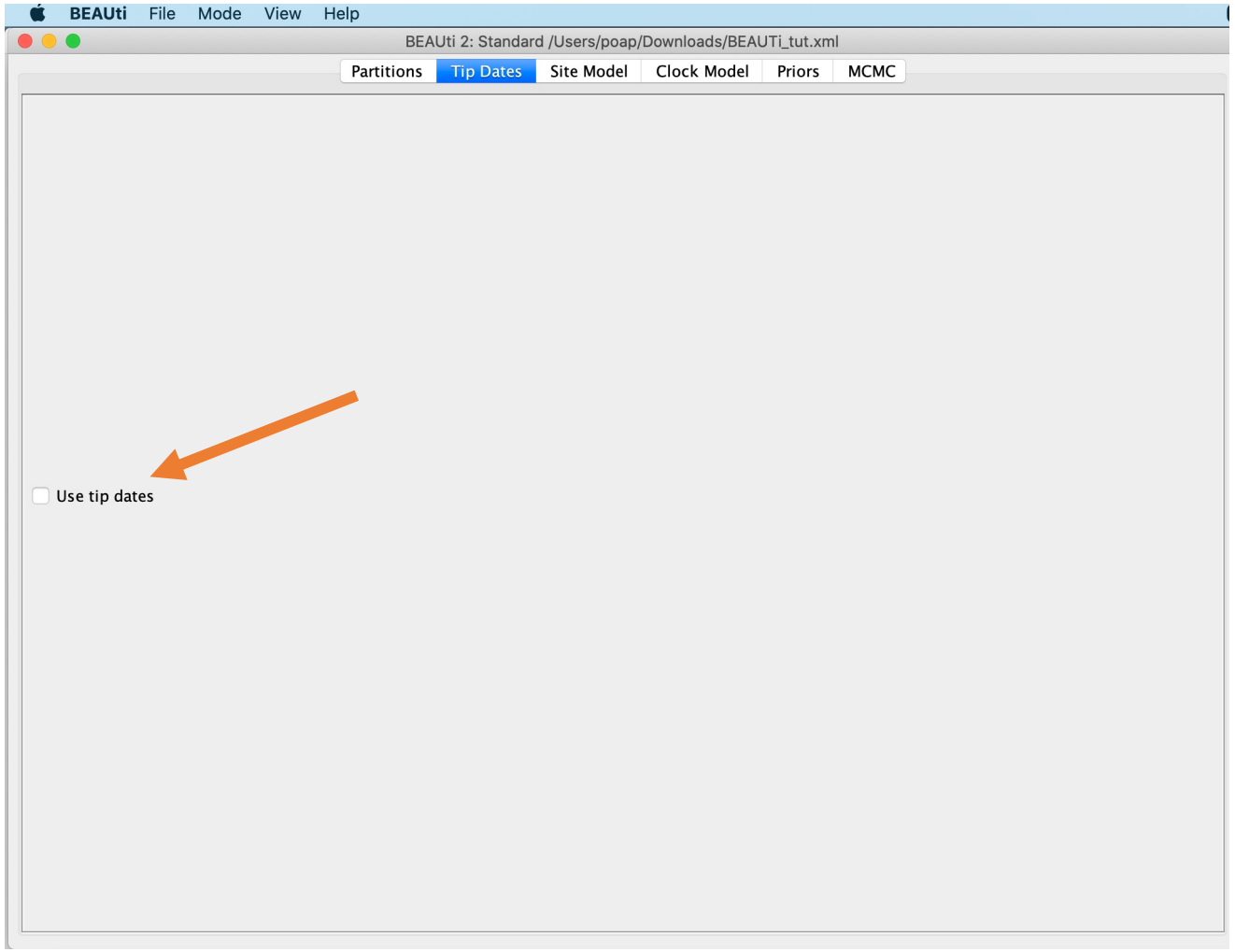

Panels
BEAUti File Mode View Help
BEAUti 2: Standard /Users/poap/Downloads/BEAUTi_tut.xml
Partitions
Tip Dates
Site Model
Clock Model
Priors
MCMC
Unlink Site Models
Link Clock Models
Unlink Clock Models
Link Trees
Unlink Trees
Name
File
Taxa
Sites
Data Type
Site Model
Clock Model
Tree
NorthAm
NorthAm
100
13154
nucleotide
NorthAm
NorthAm
NorthAm
Split
Loaded alignment

Specifying sampling dates

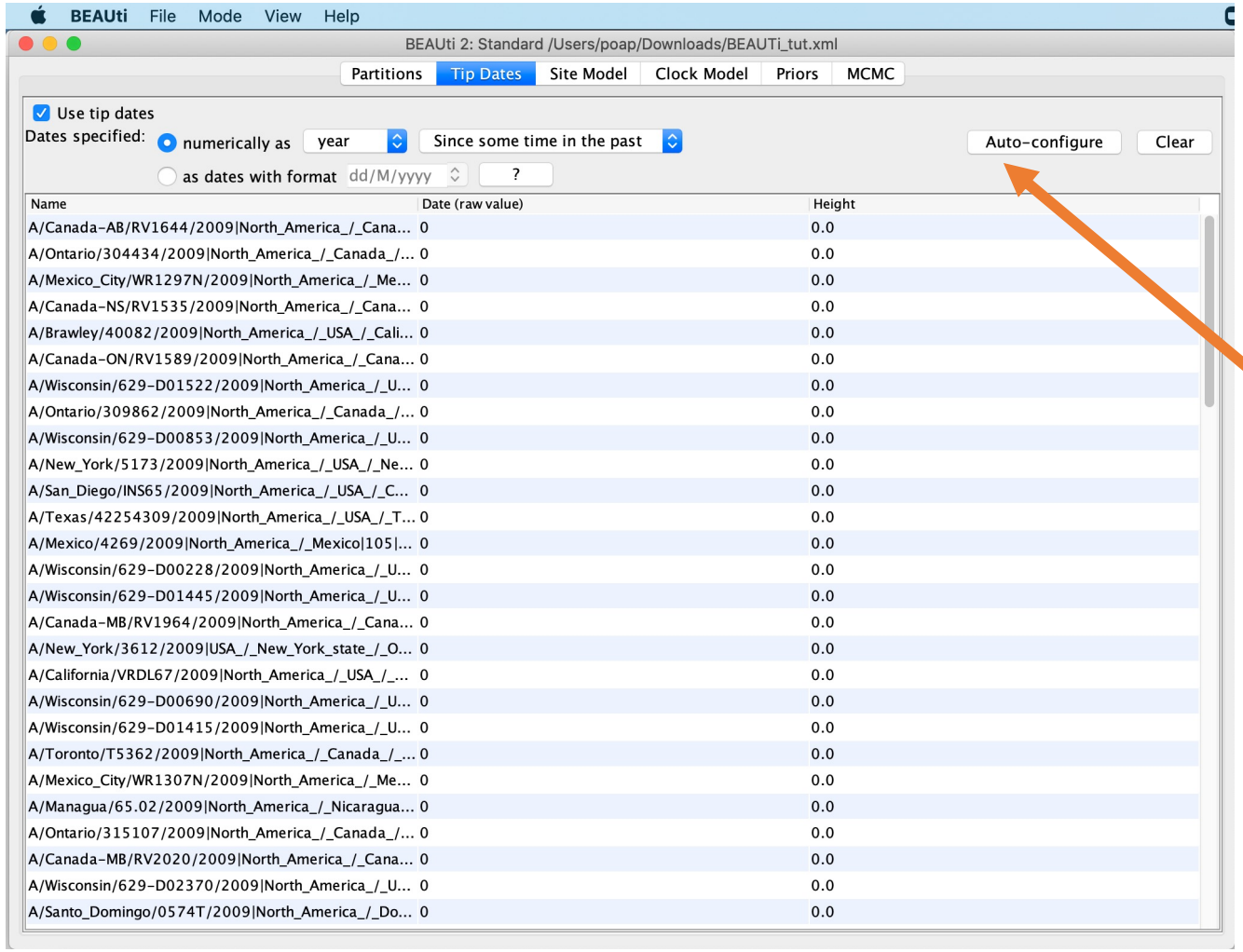
When using sequence data that has been sampled through time:
We can estimate the rates and dates.

- We add tip dates for sequences through the **Tip date** panel
  - Manually (not recommended)
  - Extracting dates from taxon labels
  - Importing dates from a file

Specifying sampling dates

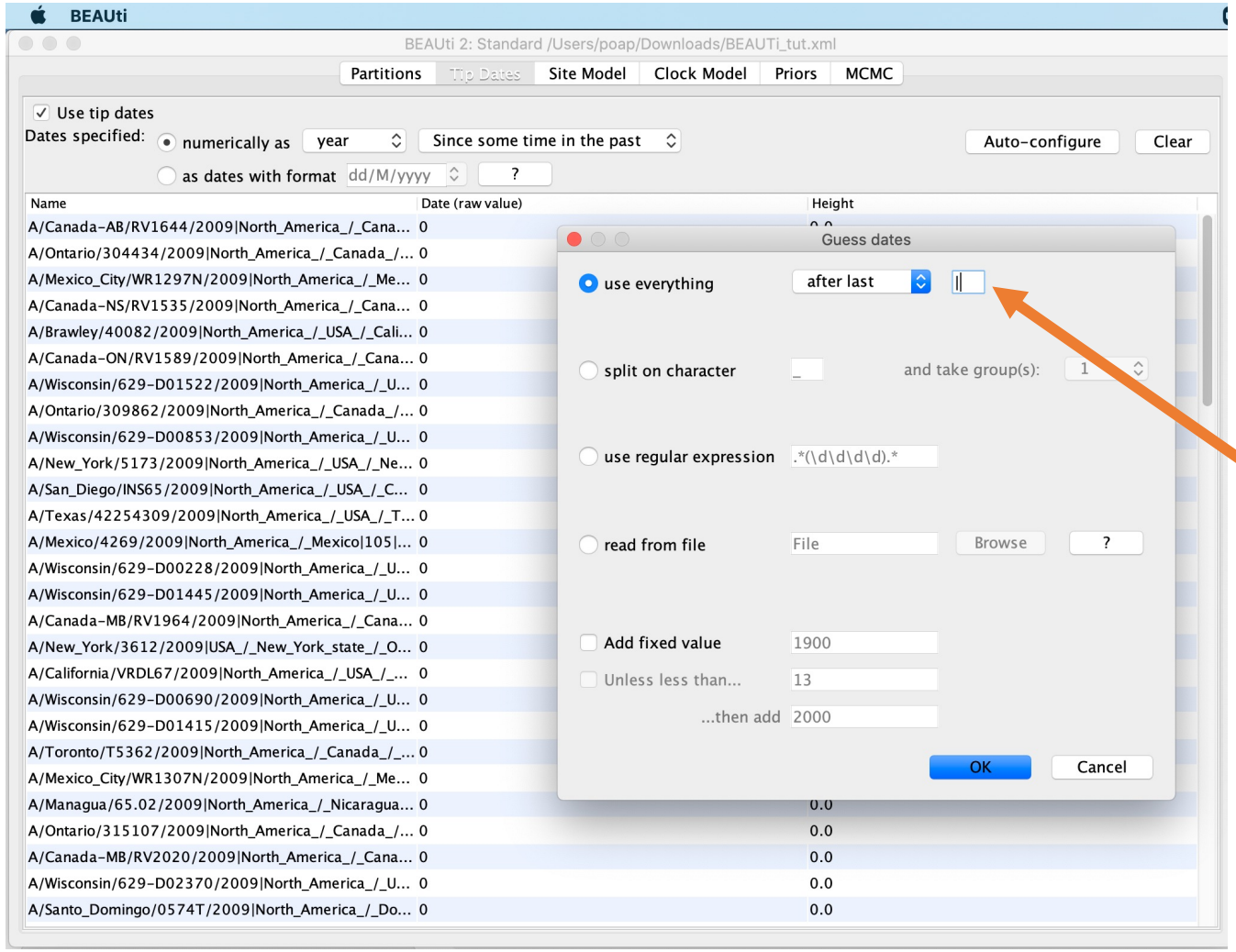


Specifying sampling dates



We are going to extract the dates from the taxon labels.

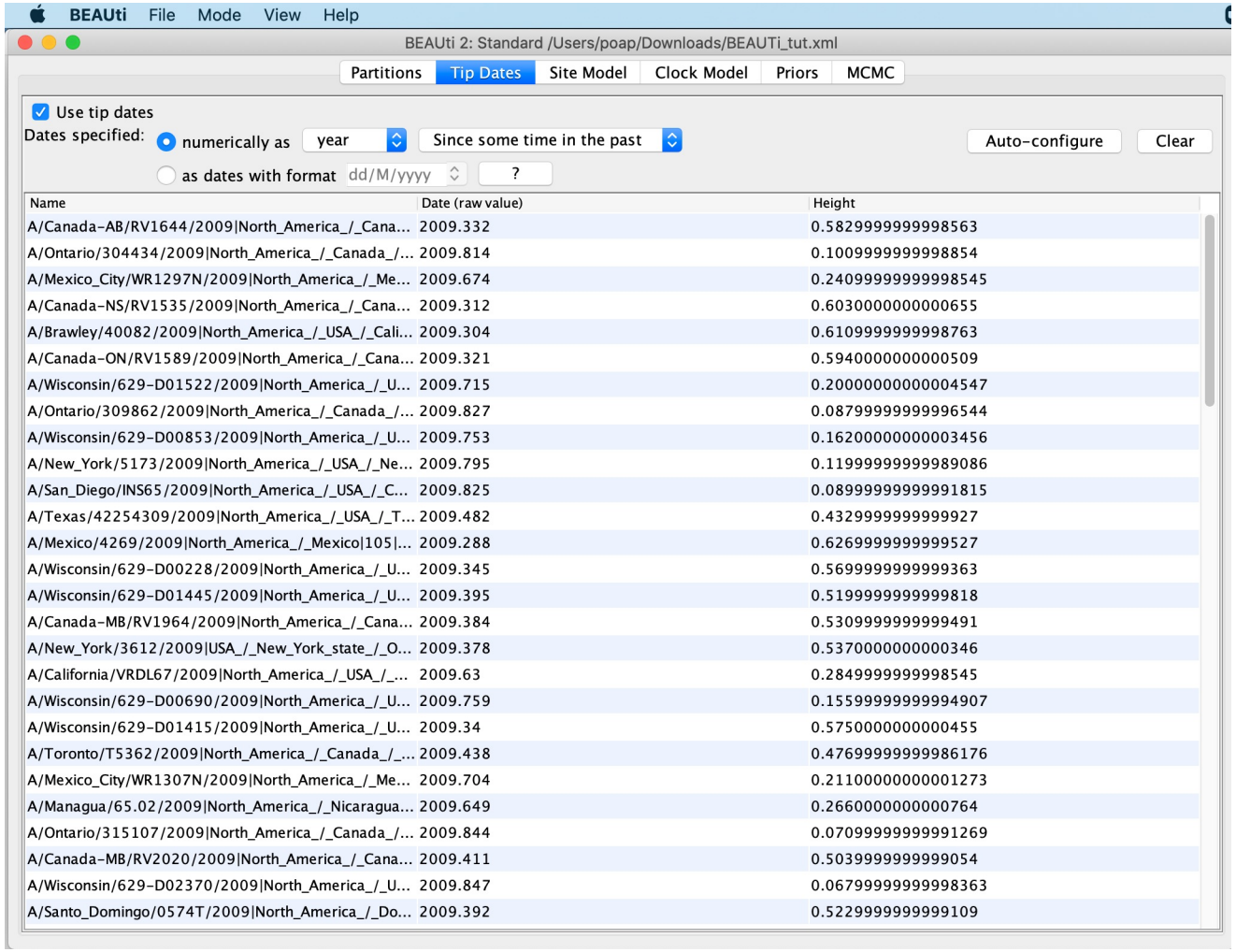
Specifying sampling dates



Want to make this easier?
When setting up your data, make sure that the taxon label has a delimiter before the date!

We can use the delimiter "|" which separates the different information in the taxon label.

Specifying sampling dates



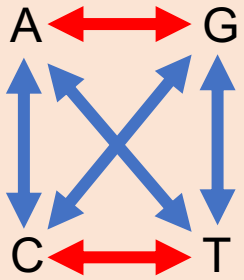
Name	Date (raw value)	Height
A/Canada-AB/RV1644/2009|North_America_/_Cana...	2009.332	0.5829999999998563
A/Ontario/304434/2009|North_America_/_Canada_/...	2009.814	0.1009999999998854
A/Mexico_City/WR1297N/2009|North_America_/_Me...	2009.674	0.24099999999998545
A/Canada-NS/RV1535/2009|North_America_/_Cana...	2009.312	0.6030000000000655
A/Brawley/40082/2009|North_America_/_USA_/_Cali...	2009.304	0.6109999999998763
A/Canada-ON/RV1589/2009|North_America_/_Cana...	2009.321	0.5940000000000509
A/Wisconsin/629-D01522/2009|North_America_/_U...	2009.715	0.20000000000004547
A/Ontario/309862/2009|North_America_/_Canada_/...	2009.827	0.08799999999996544
A/Wisconsin/629-D00853/2009|North_America_/_U...	2009.753	0.16200000000003456
A/New_York/5173/2009|North_America_/_USA_/_Ne...	2009.795	0.11999999999989086
A/San_Diego/INS65/2009|North_America_/_USA_/_C...	2009.825	0.08999999999991815
A/Texas/42254309/2009|North_America_/_USA_/_T...	2009.482	0.4329999999999927
A/Mexico/4269/2009|North_America_/_Mexico|105|...	2009.288	0.6269999999999527
A/Wisconsin/629-D00228/2009|North_America_/_U...	2009.345	0.5699999999999363
A/Wisconsin/629-D01445/2009|North_America_/_U...	2009.395	0.5199999999999818
A/Canada-MB/RV1964/2009|North_America_/_Cana...	2009.384	0.5309999999999491
A/New_York/3612/2009|USA_/_New_York_state_/_O...	2009.378	0.5370000000000346
A/California/VRDL67/2009|North_America_/_USA_/_...	2009.63	0.2849999999998545
A/Wisconsin/629-D00690/2009|North_America_/_U...	2009.759	0.15599999999994907
A/Wisconsin/629-D01415/2009|North_America_/_U...	2009.34	0.5750000000000455
A/Toronto/T5362/2009|North_America_/_Canada_/_...	2009.438	0.47699999999986176
A/Mexico_City/WR1307N/2009|North_America_/_Me...	2009.704	0.21100000000001273
A/Managua/65.02/2009|North_America_/_Nicaragua...	2009.649	0.2660000000000764
A/Ontario/315107/2009|North_America_/_Canada_/...	2009.844	0.07099999999991269
A/Canada-MB/RV2020/2009|North_America_/_Cana...	2009.411	0.5039999999999054
A/Wisconsin/629-D02370/2009|North_America_/_U...	2009.847	0.06799999999998363
A/Santo_Domingo/0574T/2009|North_America_/_Do...	2009.392	0.5229999999999109

Tip!
You don’t have to use decimal dates – if you have your dates in day/month/year format, BEAUTi will automatically calculate this for you.

Site models

Substitution models include…

JC, GTR, and HKY:
- to model base frequencies
- mutation rate for nucleotides



Transitions

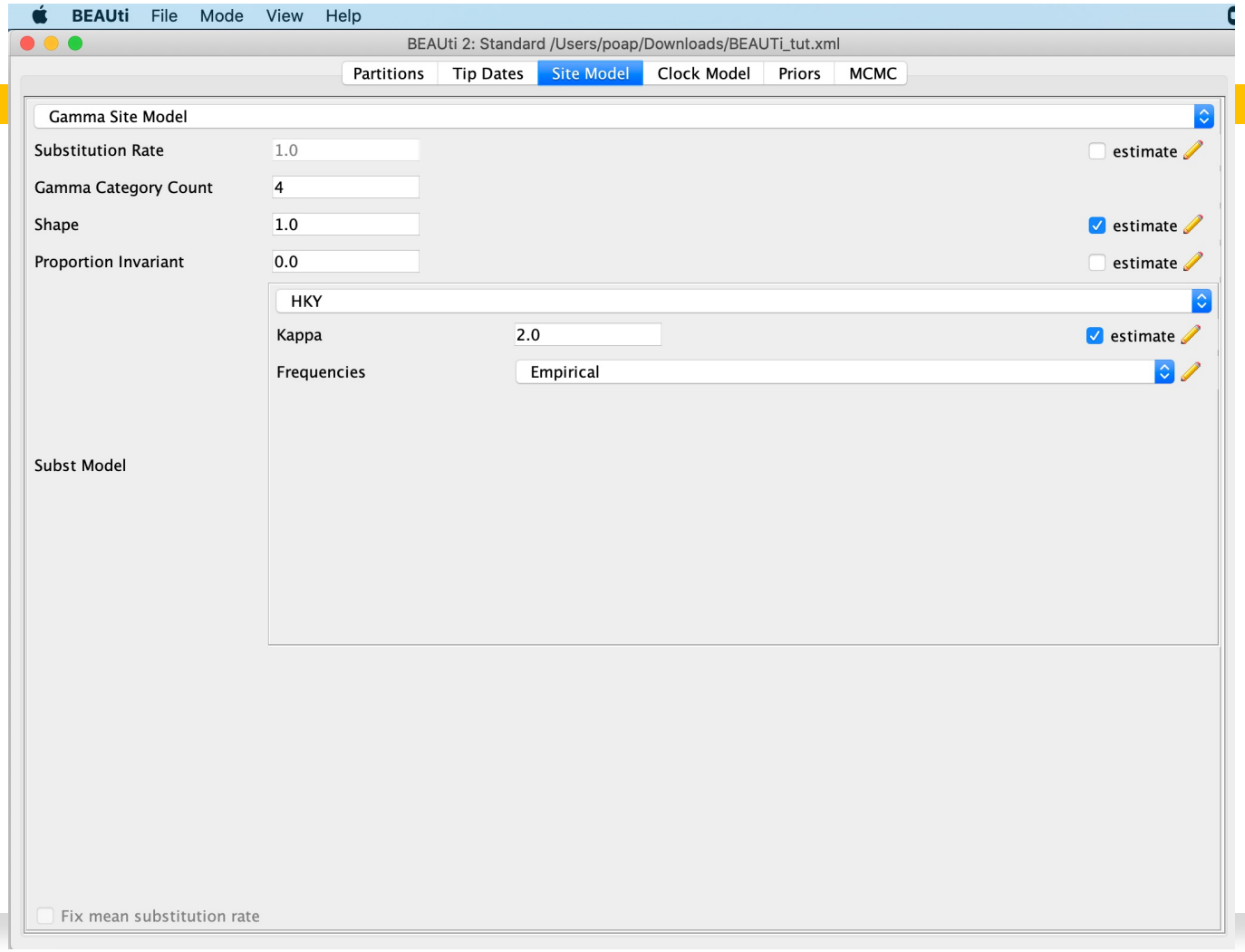
Transversions

Transitions

Want to make this easier?
Selecting **Empirical** from frequencies menu will fix the frequencies to the proportions seen in the data, reducing the number of estimates needed.

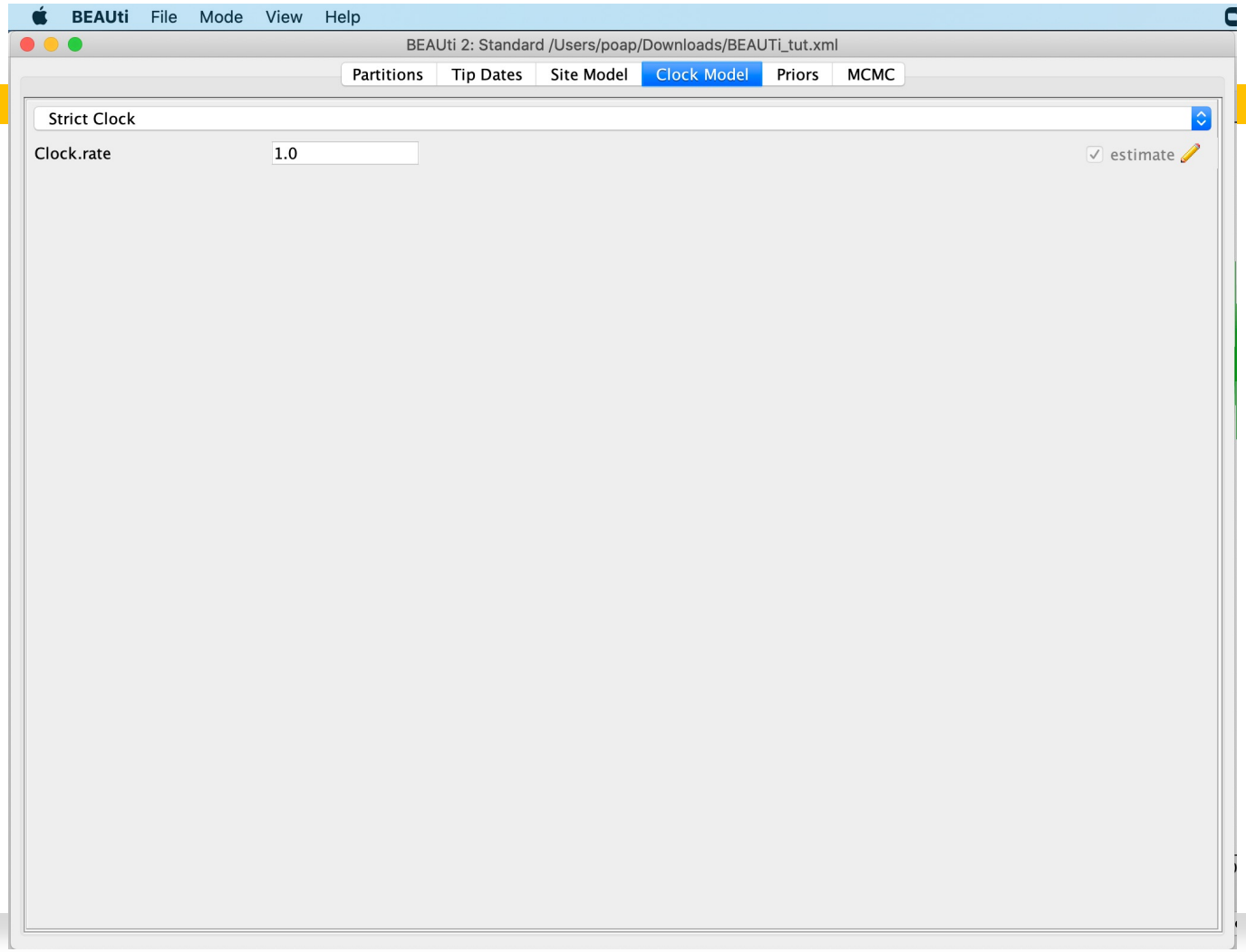
- Can add a substitution model under the **Site model** panel.
  - There are many substitution models to chose from
    - Is your data nucleotides or amino acids?
    - We can estimate the substitution rate, or if there is an accurate estimate already known, we can include it as “prior knowledge”.

Site models



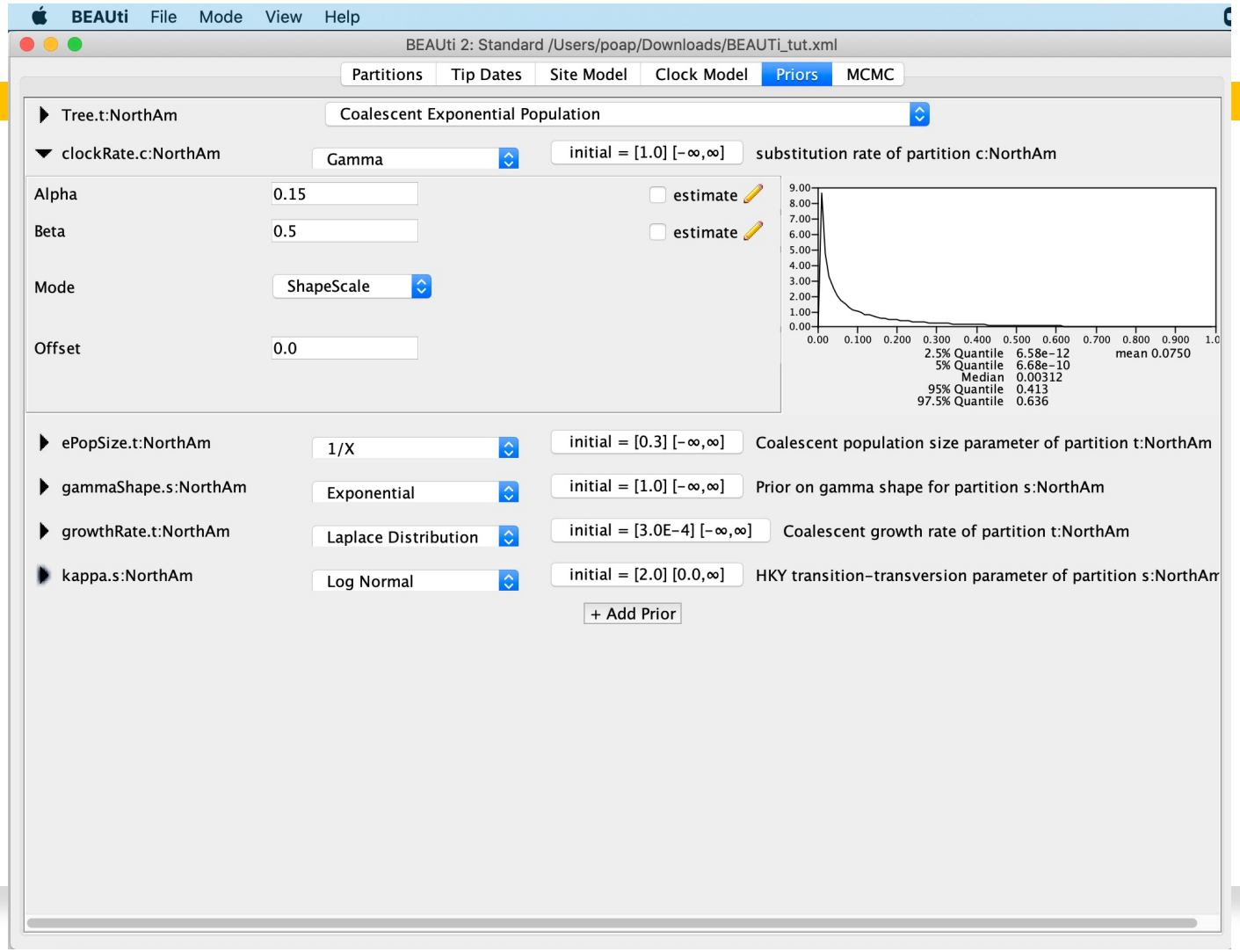
- The Gamma Category count is often set to 4-6, as adding further categories adds little benefit but takes much longer.
  - Gamma shape is often estimated to allow rate variation between sites to be modelled.

Clock models



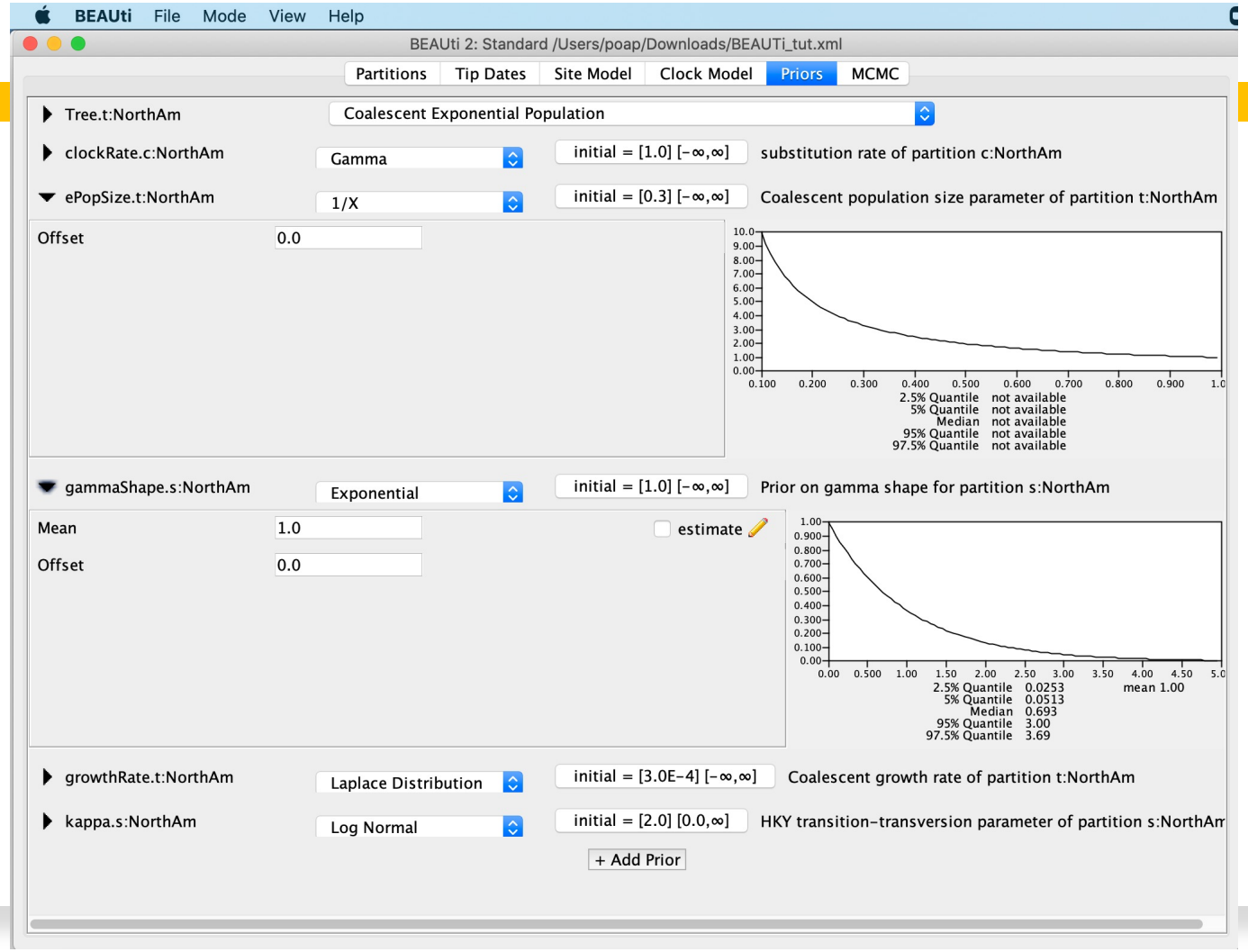
- Can add a clock model under the **Clock** panel

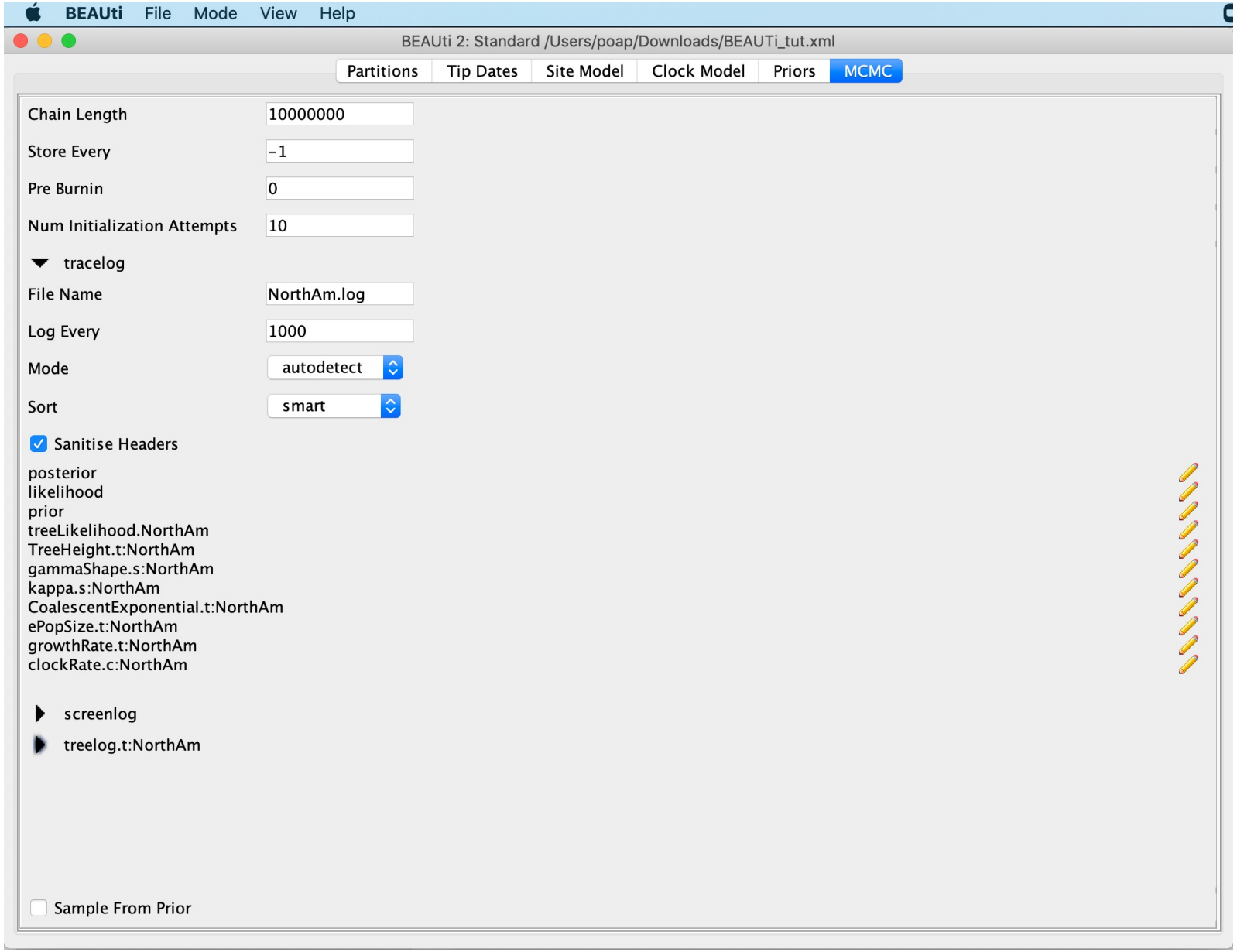
Tree priors

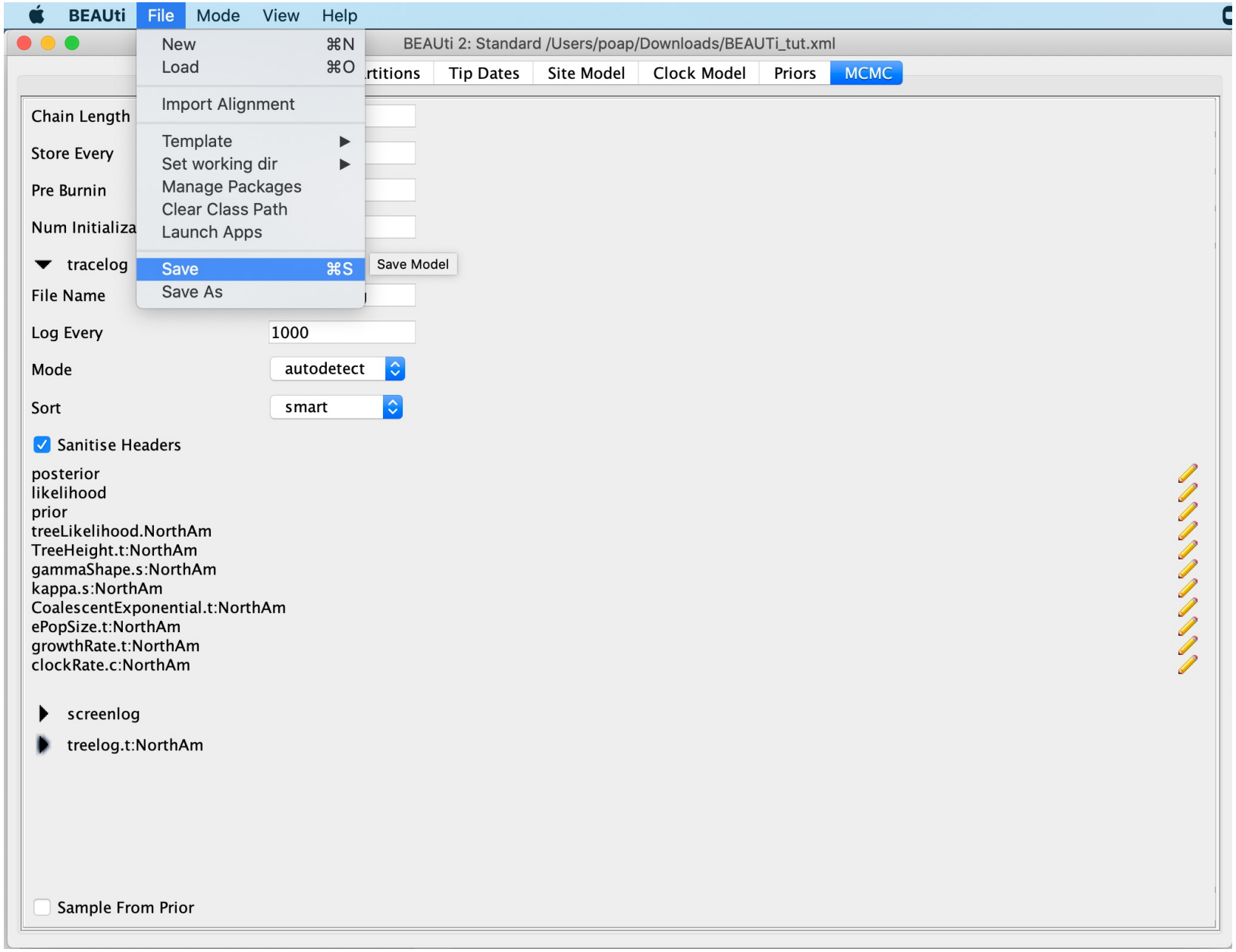


- Can add tree priors under the **Prior** panel

Tree priors

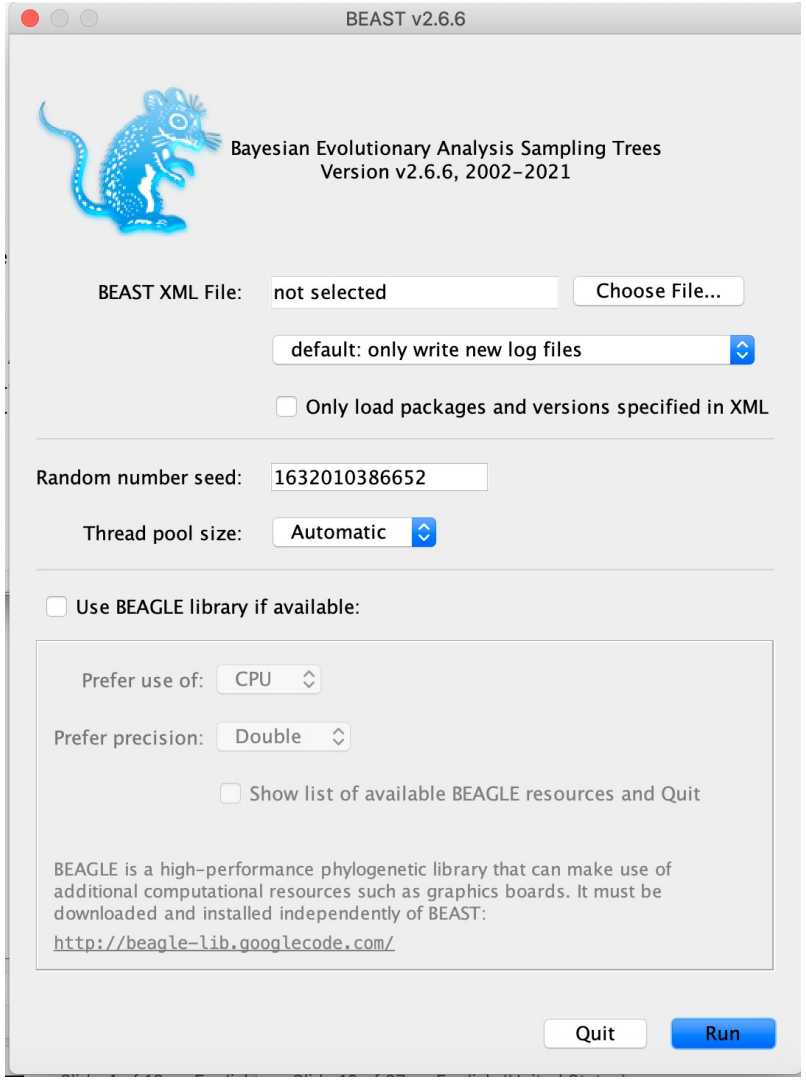

BEAUti File Mode View Help
BEAUti 2: Standard /Users/poap/Downloads/BEAUTi_tut.xml
Partitions Tip Dates Site Model Clock Model Priors MCMC
Tree.t:NorthAm
Coalescent Exponential Population
clockRate.c:NorthAm
Gamma
initial = [1.0] [−∞,∞]
substitution rate of partition c:NorthAm
ePopSize.t:NorthAm
1/X
initial = [0.3] [−∞,∞]
Coalescent population size parameter of partition t:NorthAm
Offset
0.0
2.5% Quantile not available
5% Quantile not available
Median not available
95% Quantile not available
97.5% Quantile not available
gammaShape.s:NorthAm
Exponential
initial = [1.0] [−∞,∞]
Prior on gamma shape for partition s:NorthAm
Mean
1.0
estimate
Offset
0.0
2.5% Quantile 0.0253
5% Quantile 0.0513
Median 0.693
95% Quantile 3.00
97.5% Quantile 3.69
mean 1.00
growthRate.t:NorthAm
Laplace Distribution
initial = [3.0E−4] [−∞,∞]
Coalescent growth rate of partition t:NorthAm
kappa.s:NorthAm
Log Normal
initial = [2.0] [0.0,∞]
HKY transition−transversion parameter of partition s:NorthAm
+ Add Prior


BEAUti File Mode View Help
BEAUti 2: Standard /Users/poap/Downloads/BEAUTi_tut.xml
Partitions
Tip Dates
Site Model
Clock Model
Priors
MCMC
Chain Length
10000000
Store Every
-1
Pre Burnin
0
Num Initialization Attempts
10
tracelog
File Name
NorthAm.log
Log Every
1000
Mode
autodetect
Sort
smart
Sanitise Headers
posterior
likelihood
prior
treeLikelihood.NorthAm
TreeHeight.t:NorthAm
gammaShape.s:NorthAm
kappa.s:NorthAm
CoalescentExponential.t:NorthAm
ePopSize.t:NorthAm
growthRate.t:NorthAm
clockRate.c:NorthAm
screenlog
treelog.t:NorthAm
Sample From Prior


BEAUti
File
Mode
View
Help
New ⌘N
Load ⌘O
Import Alignment
Template
Set working dir
Manage Packages
Clear Class Path
Launch Apps
Save ⌘S
Save As
BEAUti 2: Standard /Users/poap/Downloads/BEAUTi_tut.xml
Tip Dates
Site Model
Clock Model
Priors
MCMC
Chain Length
Store Every
Pre Burnin
Num Initializa
tracelog
Save Model
File Name
Log Every
1000
Mode
autodetect
Sort
smart
Sanitise Headers
posterior
likelihood
prior
treeLikelihood.NorthAm
TreeHeight.t:NorthAm
gammaShape.s:NorthAm
kappa.s:NorthAm
CoalescentExponential.t:NorthAm
ePopSize.t:NorthAm
growthRate.t:NorthAm
clockRate.c:NorthAm
screenlog
treelog.t:NorthAm
Sample From Prior

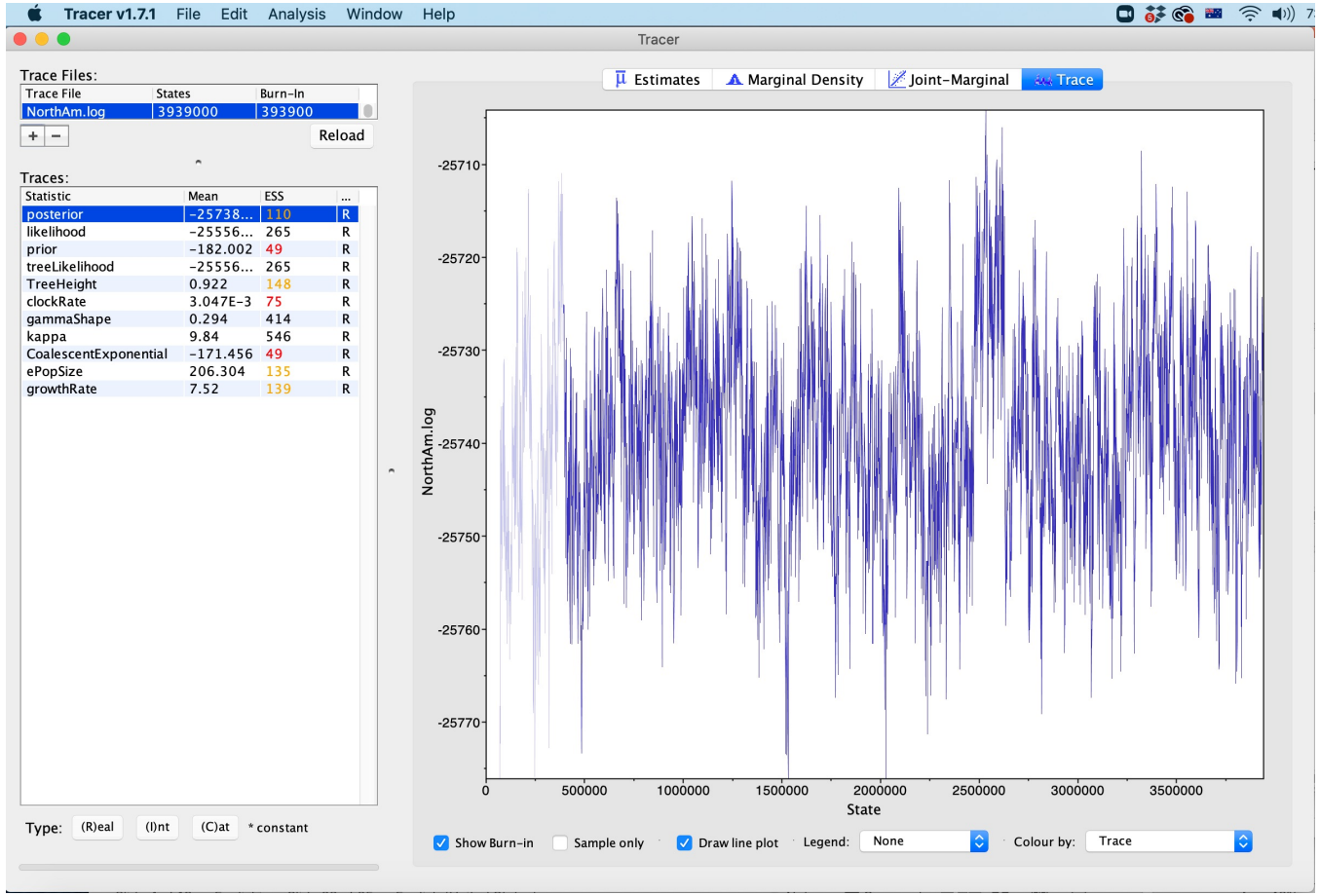
Running BEAST

- Open BEAST2 and choose the xml file
- Hit run!



Viewing the raw trace in Tracer

- Open Tracer and load the .log file



Common stumbling blocks: 1

Using the correct software for alignments is essential, depending on:
- length and number of sequences
- conserved regions of sequences

- The quality of an alignment will affect all downstream analysis, including BEAST! Watch out for:
  - Recombination
  - Poorly aligned regions or ambiguous data
  - Duplicate sequences

Common stumbling blocks: 2

- Setting up your model can be very confusing and overwhelming
- Substitution models should be as simple as possible, while being as “accurate” to the sequences evolutionary process as possible.
- To choose the best clock model, it is wise to look at how your data behaves under a couple of different models (e.g. strict vs relaxed clock).

Common stumbling blocks: 3

- Tree priors
  - Sampling proportions
    - Small sample of individuals from a large population? Coalescent
    - A large dataset where sampling times will be informative? Birth-death

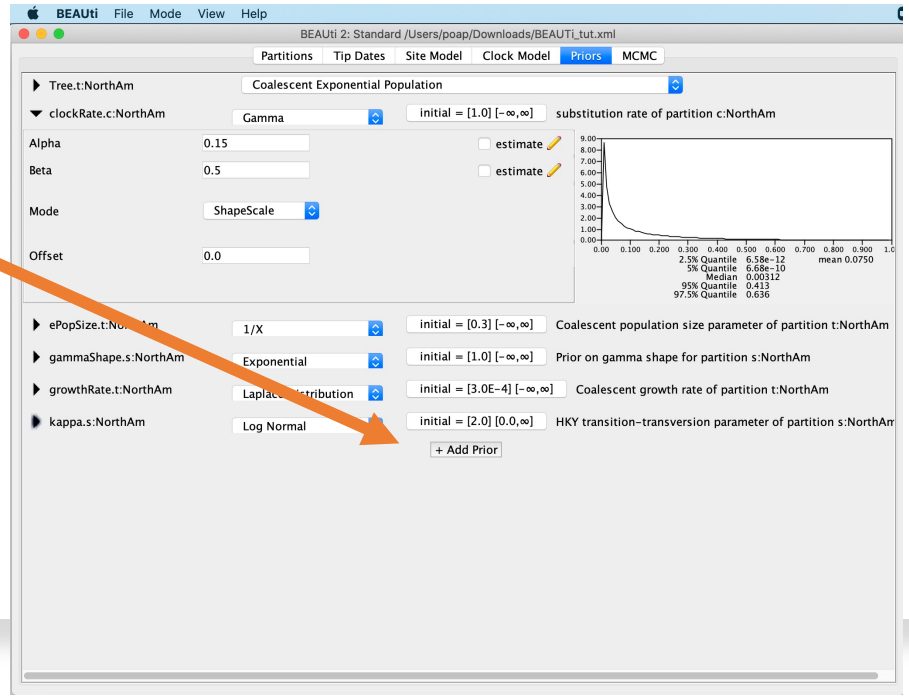
Linking or unlinking models

- Can link (and unlink) models in the **Partitions** panel by selecting the partitions and clicking:
  - “Link Site Models” – links the site models (e.g. GTR)
  - “Link Tree Models” – links the tree models (e.g. Birth-death)
  - “Link Clock Models” – links the clock models (e.g. Strict clock)

Tip date sampling

If the sampling dates are not precise :
We can allow a range of uncertainty around those dates.
Dates will be estimated.

- We add tip date uncertainty for sequences through the **Priors** panel
  - Click “ + add Prior”

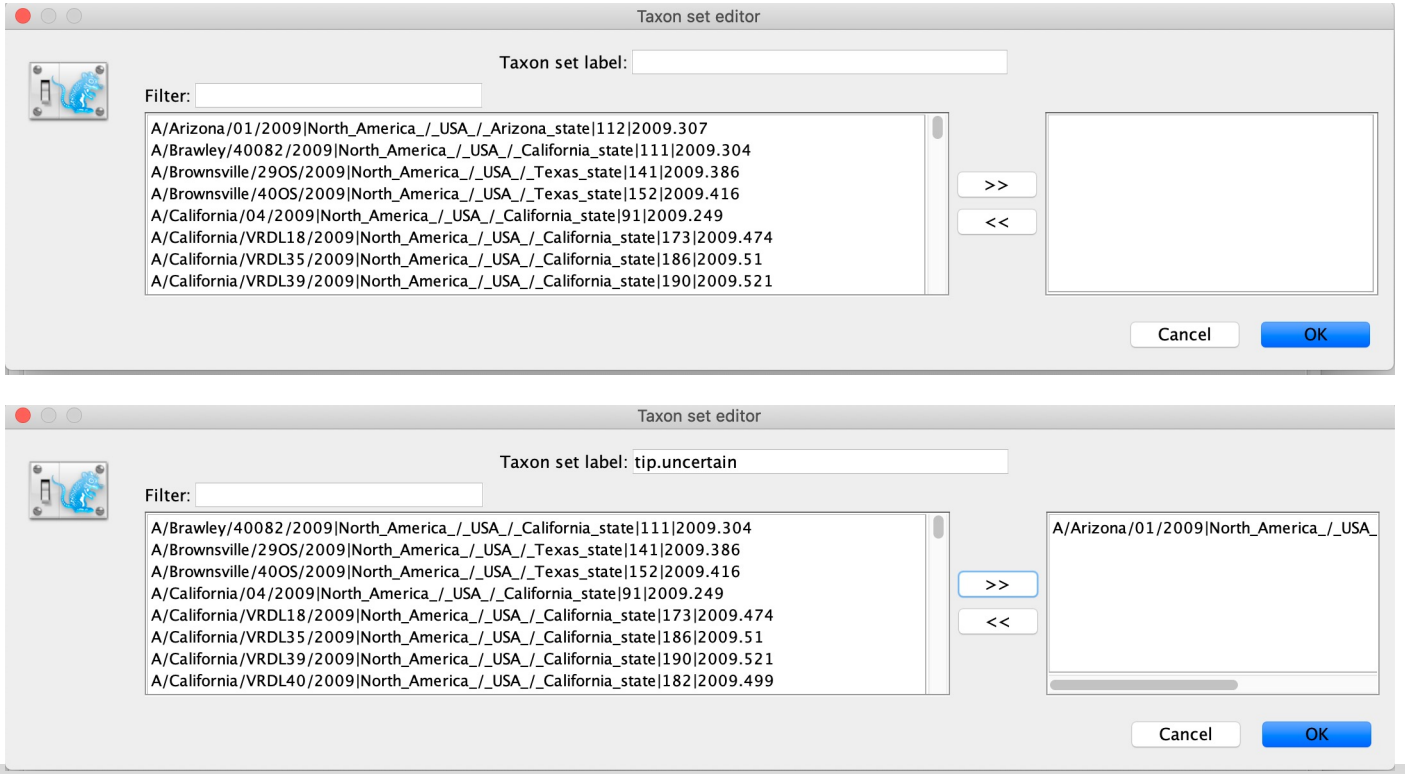


Tip date sampling

If the sampling dates are not precise :

We can allow a range of uncertainty around those dates.

Dates will be estimated.

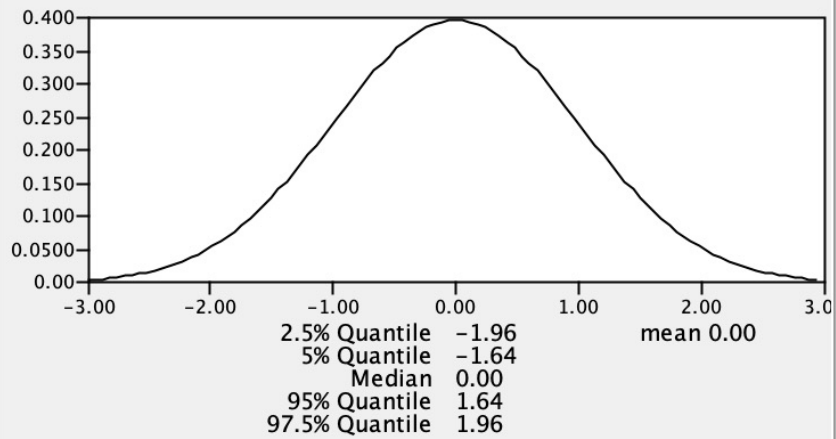


tip.uncertain.prior | Normal | monophyletic | −

Mean 0.0 estimate

Sigma 1.0 estimate

Offset 0.0



Tipsonly

Use Originate

+ Add Prior

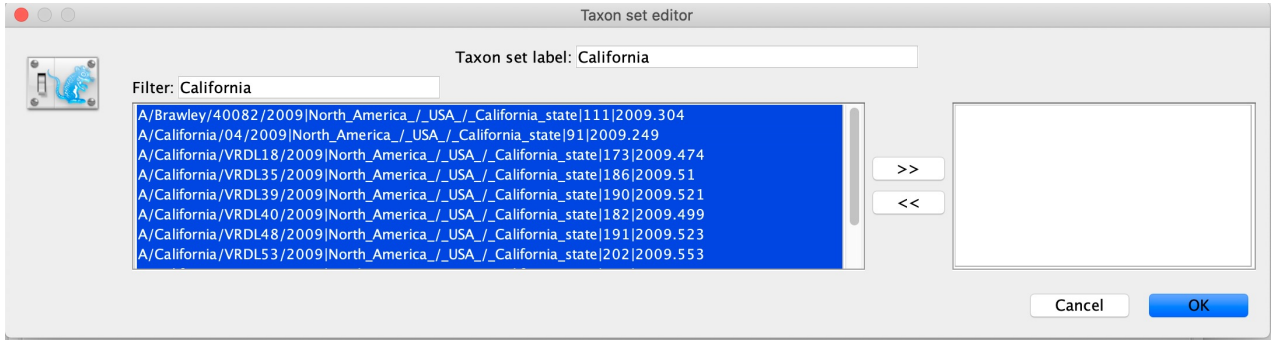
Creating taxon sets

Groups of taxa/sequences that will form a clade within the tree:

- That have a specific rate of evolution,
- You want to apply a specific model,
- Or to specify a node in the tree to calibrate a TMRCA.

Want to make this easier?
When setting up your data, make sure that the taxon label includes something related to its taxon set!

- Can add a taxon set under the **Prior panel**

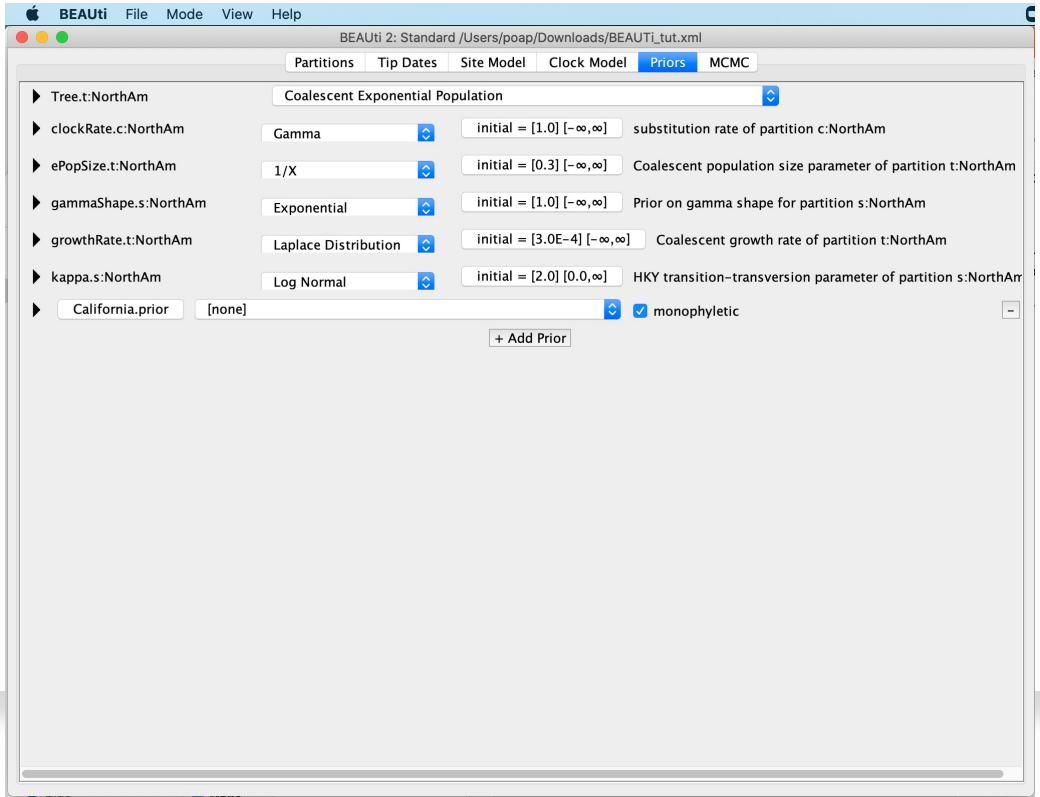


Creating taxon sets

Groups of taxa/sequences that will form a clade within the tree:

- That have a specific rate of evolution,
- You want to apply a specific model,
- Or to specify a node in the tree to calibrate a TMRCA.

Want to make this easier?
When setting up your data, make sure that the taxon label includes something related to its taxon set!



Further reading

- **Online tutorials** at: http://www.beast2.org/tutorials/
- **Papers**
  - Bouckaert R., Vaughan T.G., Barido-Sottani J., Duchêne S., Fourment M., Gavryushkina A., et al. (2019) BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLoS computational biology, 15(4), e1006650.
- **Books**
  - Bayesian Evolutionary Analysis with BEAST – Drummond, Bouckaert (2015)