Analysing Data with BEAST 2:

> Simon J. Greenhill (simon@simon.net.nz).
>
> ARC Centre of Excellence for the Dynamics of Language, Australian National University.
>
> Department of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History.

`BEAST 2` (Bouckaert et al. '14) is one of the most powerful Bayesian Phylogenetic analysis suites around:

> BEAST 2 is a cross-platform program for Bayesian phylogenetic analysis of molecular sequences.
>
> It estimates rooted, time-measured phylogenies using strict or relaxed molecular clock models. It can be used as a method of reconstructing phylogenies but is also a framework for testing evolutionary hypotheses without conditioning on a single tree topology.
>
> BEAST 2 uses Markov chain Monte Carlo (MCMC) to average over tree space, so that each tree is weighted proportional to its posterior probability.
>
> BEAST 2 includes a graphical user-interface for setting up standard analyses and a suit of programs for analysing the results.

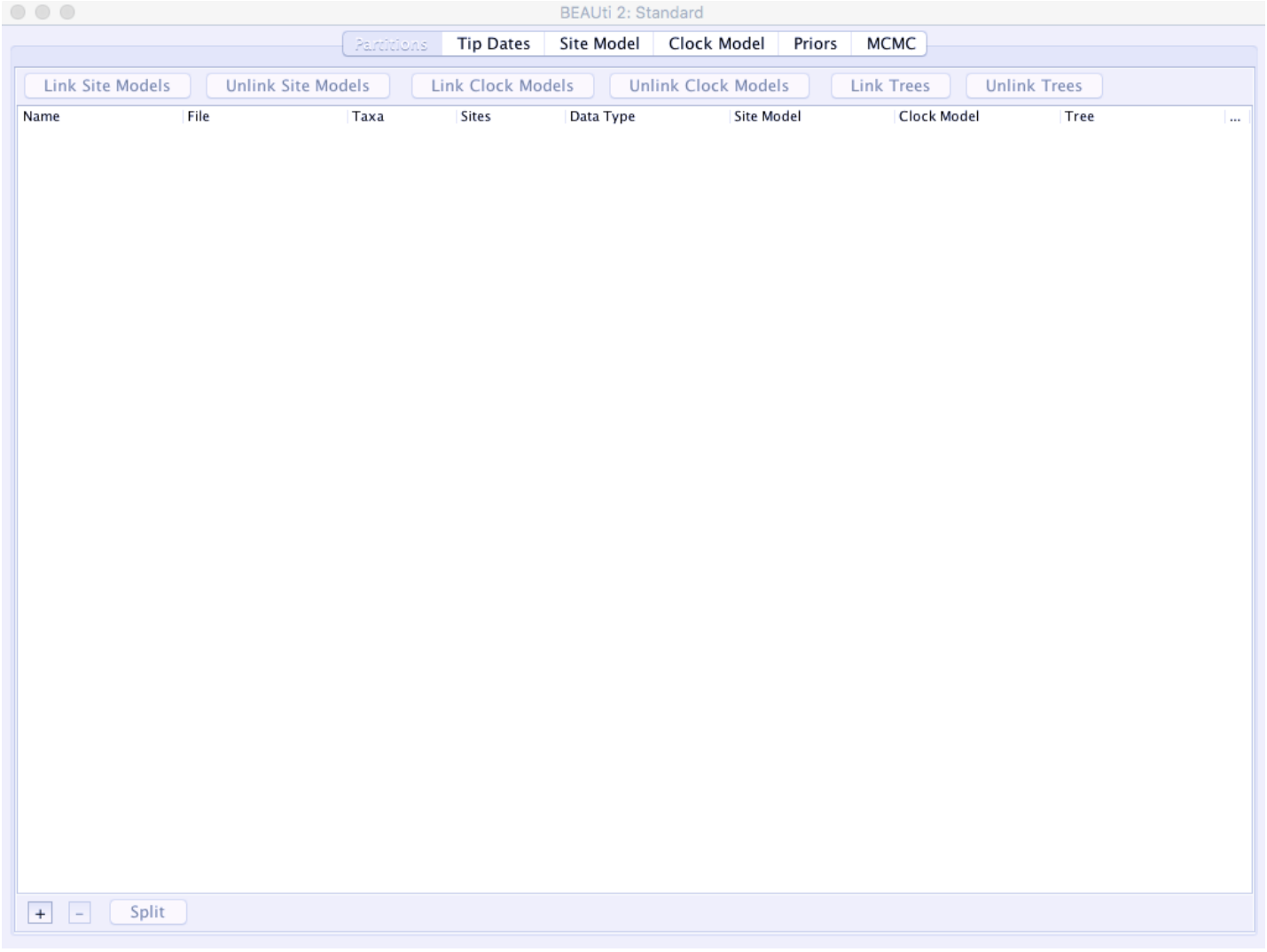
Our workflow:

1. Set up analysis in `BEAUTi`. `BEAUTi` is a graphical user interface that will generate files in the format that `BEAST` wants (XML: https://en.wikipedia.org/wiki/XML).
2. Run the generated XML file in `BEAST`.
3. Examine the analysis log files using `Tracer`
4. Construct summary tree using `TreeAnnotator` and visualise the results using `FigTree` and `Densitree`.

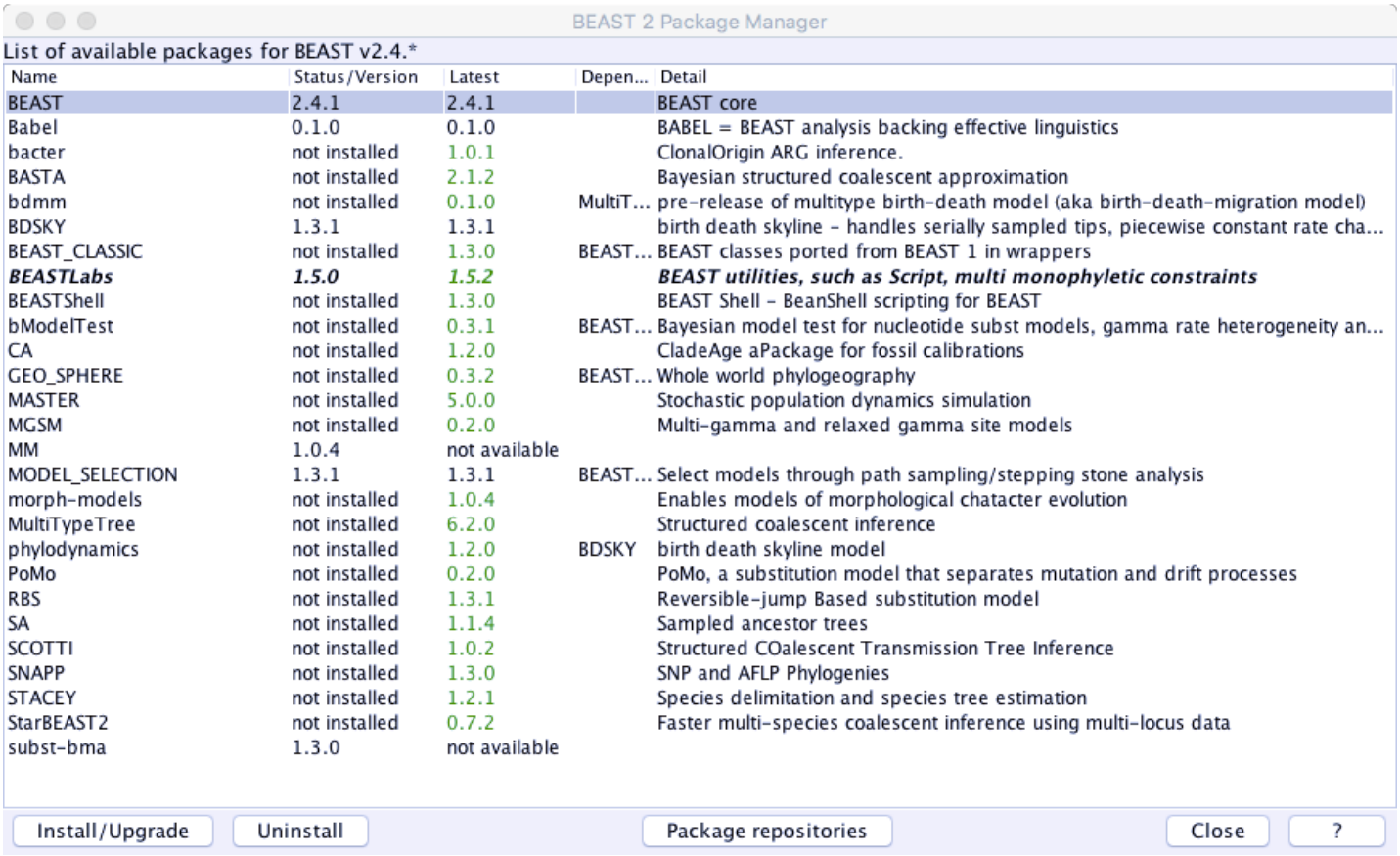
Install packages/addons we need in BEAUTi.

`BEAST` has a package system where new packages can add in extra functionality. This is where all the cool new `BEAST` addins appear. Today we want to install some templates in the `Babel` package that will help us analyse language data.

1. Open `BEAUTi`. It will be whereever you installed `BEAST`. It looks like this, showing you the default `Partitions` tab:



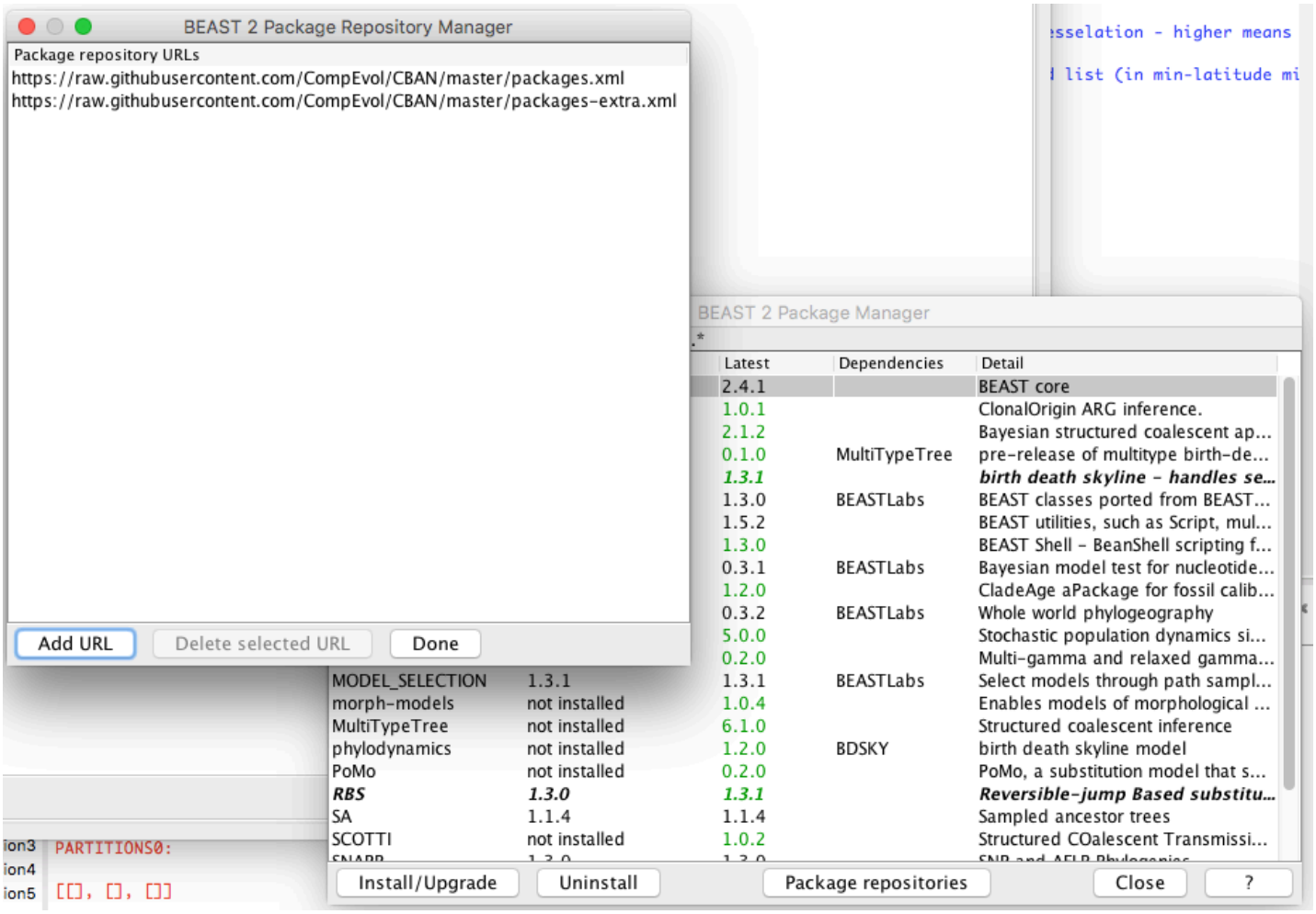
2. From the *File* menu select *Manage Packages*:



3. `Babel` is so new it's not in the general repositories. So, click the *Package Repositories* button down the bottom. Select *Add URL* and enter the following:

```
https://raw.githubusercontent.com/CompEvol/CBAN/master/packages-extra.xml
```

… you should now have something like:



Click *DONE* and then you should see `Babel` in the add-ons list. Select it and click *Install/Upgrade*. This should work.

4. Close BEAUTi and Restart it (to make sure the package contents are loaded).

`[ ] Add the New Package URL.`

`[ ] Make sure you have the Babel package installed.`

Set up Analysis.

Right, let's get started. We're going to set up a very simple analysing the dataset from before (cpacific.nex).

We will run one of the simplest possible analyses to explore the data. Here we're using a Continuous Time Markov Chain Model (CTMC) for binary data. It is essentially the Generalised Time Reversible Model (Tavaré '86) for binary data (Drummond & Bouckaert '15). The CTMC allows cognates to be gained and lost at the same rate, which is probably not correct, but it's simple, and we can relax it later.

1. Open `BEAUTi`.
2. Select a template: `File -> Template -> BinaryCTMC`
3. Add Data: `File -> Add Alignment -> (Choose your file)`

You should see the dataset listed as something like “bin.cpacific” on the `Partitions` tab. Check that nothing weird has happened – it should say the same number of taxa and sites (=characters) that are in the nexus file, and that the data are binary.

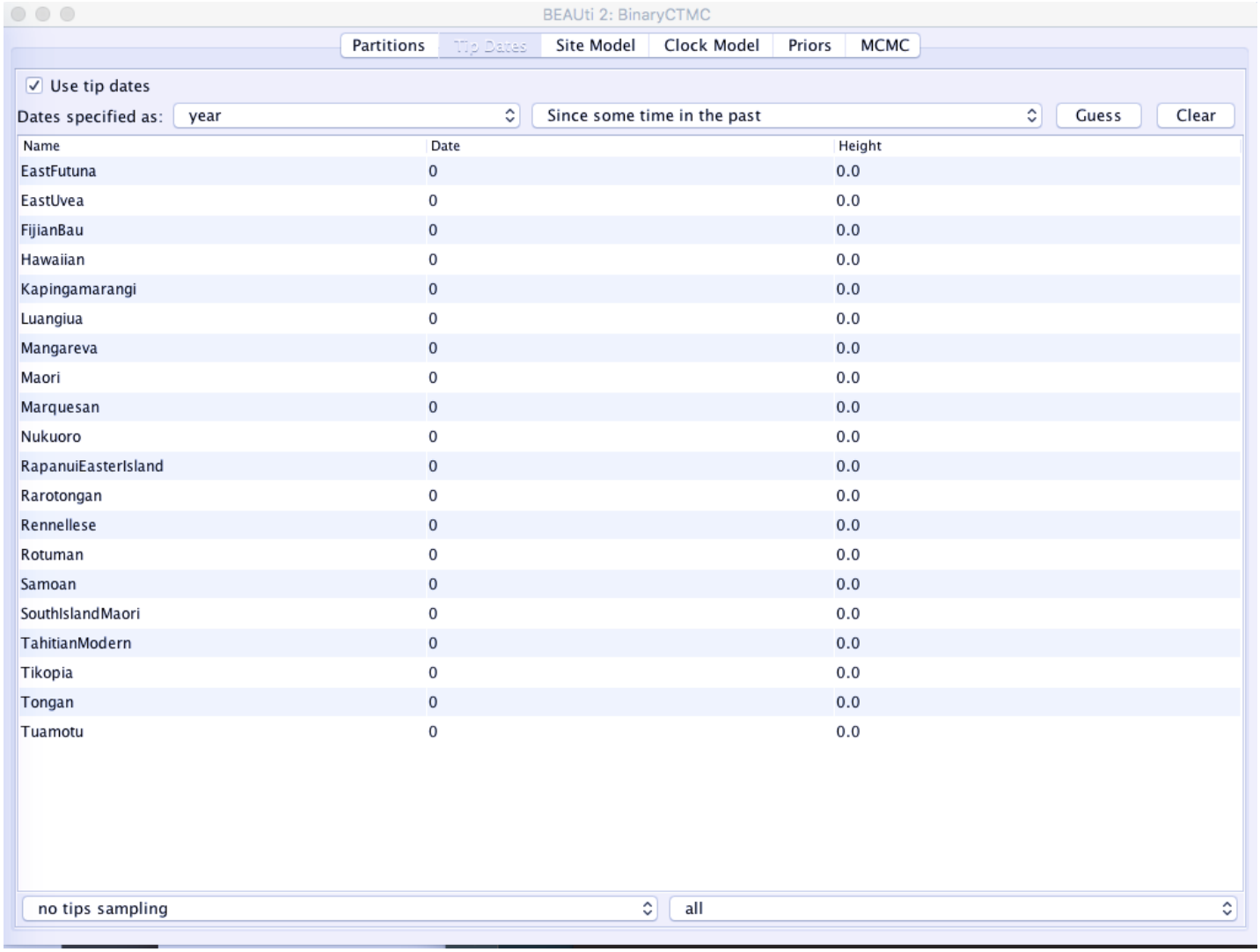
`[ ] Data are loaded into BEAUTi.`

Aside: I’ve corrected for `Ascertainment Bias` for you. If you were to run your own analysis then you will need to do this too. Read the section on “Ascertainment Correction” in the Appendix if you’re interested.

Tip Dates Tab.

Select the `Tip Dates` Tab. If we happened to have extinct languages we could add them here. We don’t, so we won’t use this, but I want you to know it’s there if you need it later.

This is what it looks like. Should be fairly self-explanatory:



`[ ] Admire tip dates tab, but make sure it stays unchecked.`

Site Model Tab.

Select the `Site Model` Tab. Here we specify the "site model", i.e. the model that describes how the characters in our data will evolve.

Using the `Babel` template for `BinaryCTMC` has set most of this up for us as a `CTMC` model.

One thing to note – is the parameter `Gamma Category Count`. BEAUTi has chosen to allow the cognates in our data to vary in rate according to a `gamma` distribution (Yang 1993). This is a useful way to let different cognates evolve at different rates.

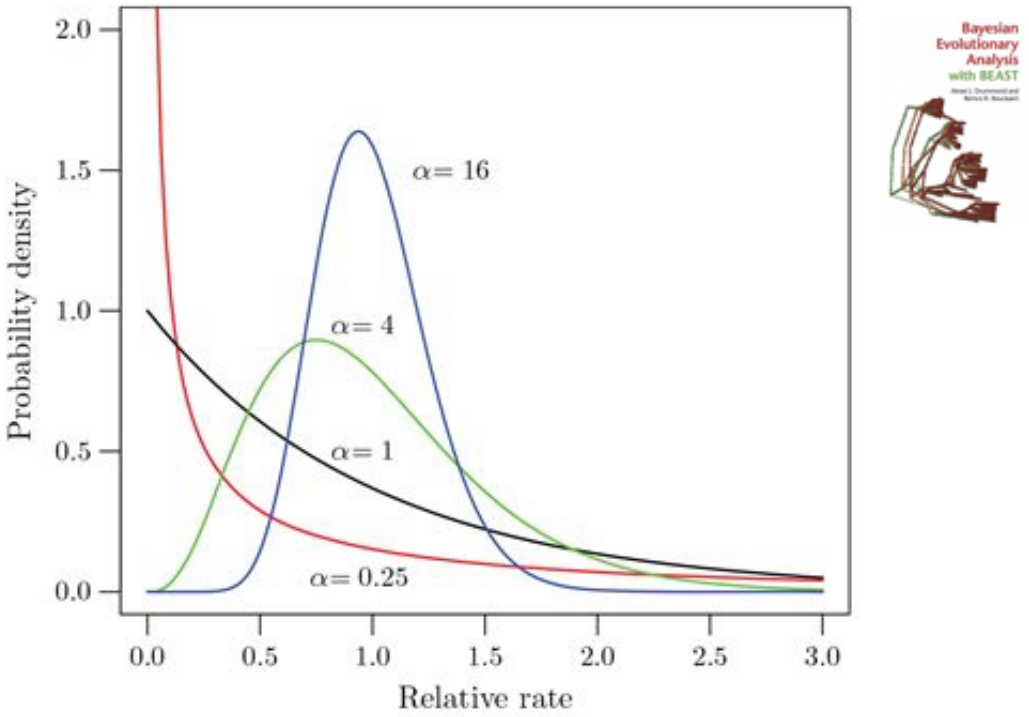


Figure 3.7 The gamma distribution for different values of the shape parameter α with scale parameter set to $1/\alpha$ so that the mean is 1 in all cases.

The gamma distribution (figure from the useful BEAST Book) can take many shapes, all of which are described by a single parameter *α*. For example, if *α* is small there there is lots of variation in rate between different characters (jargon: "high rate heterogeneity"). If *α* is big then there's little difference in rates between our cognate sets.

The Gamma approach to rate heterogeneity is really useful, as we only add one parameter to our model, and we can (a) estimate the *α* parameter from the data, and (b) draw the rates for each cognate from this distribution.

To not overload things the analysis approximates the rate categories by *dividing* the distribution into a

number of categories – here four as specified by `Gamma Category Count`. Four is a good default, you can think of it like dividing that distribution into something like “very slow, slow, medium, fast”. You may want to play around with this if you were doing your own analyses.

One other thing here is the *Proportion Invariant*. This tells `BEAST` to estimate the number of characters that are invariant. If you had a lot of characters that never varied then you might want to select this. However, we won’t because it can interact with the Gamma distribution and the gamma should handle a lot of the invariants anyway by putting them into the slowest rate category.

Clock Model Tab.

Select the `Clock Model` tab. This allows us to specify the clock model that describes how the branches in the tree vary in rates. We have two main options.

A. `Strict Clock`: There are no variations in rates across branches.

B. `Relaxed Clock (Log Normal)`. Variation across branches is autocorrelated, so that neighbouring branches can be more similar than further away (Drummond et al. '06). This works nicely as it allows different lineages to vary but that variation to get larger as languages get more different.

Note that the Relaxed Clock can also be parameterised with an exponential distribution, but the Log Normal one works better almost always (Drummond et al. '06).

We have strong prior beliefs for automatically choosing the `Relaxed Clock` as we know that languages vary in their rates of change. But, remember that we are setting up the simplest analysis and we want to test whether we need to “relax the clock” or not later. So, leave it as the Strict Clock for now.

`[ ] Stick with the Strict Clock for now.`

Priors Tab.

Heres where things get fun.

Remember that BEAST is completely Bayesian, this means we can add other information into the analysis from our 'prior’ beliefs. We can use this to make strong or weak assumptions about pretty much anything in the analysis from the way the trees should look, to the way that any of the other parameters should vary or be constrained, to how the taxa in our analysis should be related.

Choose a Tree Prior:

The Tree Prior describes the process that generates the tree. Here we should set it to a `Yule` process. This is a very simple “pure birth” process which starts with one lineage, waits for some amount of time then splits that lineage into two, and then repeats until we’ve got a tree (Yule ‘24). At any given time slice there’s

a constant probability of a lineage split at any time, and in every slice each lineage has an equal chance of splitting.

The Yule is a good general tree prior to use, unless you have extinct languages in your analysis. If this is the case then you will need to use something more complicated like the `Coalescent Bayesian Skyline` (a nice loose prior based on the Coalescent https://en.wikipedia.org/wiki/Coalescent_theory or a `Birth-Death Skyline` (newer, includes extinction rates), or `Sampled Ancestor` (if you have extinct languages and their descendants (e.g. Latin + French, Italian). These are harder to run and stabilise, so stick with the Yule unless you have to.

`[ ] Make sure that the Tree Prior is set to Yule`

Adding Calibrations:

The main thing we use the prior tab for is to add calibration information. Fortunately, we have good archaeological information about the settlement of Polynesia.

1. New Zealand.

According to Wilmshurst et al. ('11), the settlement of New Zealand can be securely dated to between 1230-1282 A.D.

We have two dialects of NZ Maori in these data: Maori and SouthIslandMaori.

Let's operationalise this calibration like this. If we assume that the `present` is a nice round number like the year 2000 (this makes interpretation easier), then we convert this to before present:

```
2000 - 1230 = 770 years ago
2000 - 1282 = 718 years ago
```

… a good shape for this distribution is a `log normal` as it will allow the lower bound to be tighter than the other side. That is, I'm pretty confident that settlement didn't occur more recently than 718 years, but settlement could have occurred earlier than 770 years ago and language divergence may have slightly preceded that date.

1. click the little `+` symbol.
2. When the dialog pops up, enter `NewZealand` into the field `Taxon set label`
3. Find Maori and S.I. Maori and move them to the right side of the window.
4. Click OK.

`[ ] Done that.`

1. Click the little arrow to the left of that new prior. This will open up a little submenu.
2. Select the check box called `monophyletic` . This will enforce this subgrouping in the analysis. It's a

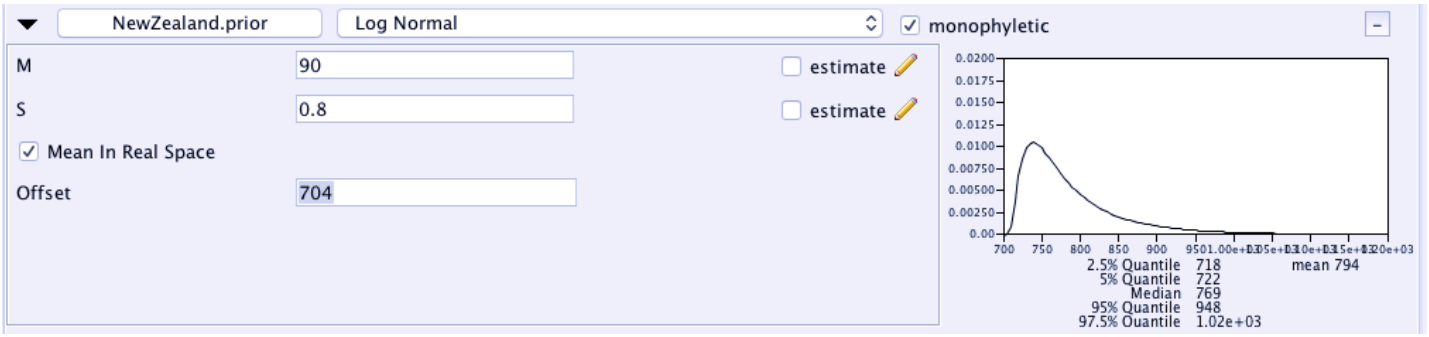
good idea to do this for each calibration, but you should check first that the languages do subgroup correctly.
3. Where it says `[none]`, change that to `Log Normal`.

`[ ] Got it.`

Now we can specify the age range for this subgroup. We want to get the left side of the distribution (the 2.5% quantile) near the earliest age of 718, and have the *median* sitting around the oldest age of 770.

1. tick the box that says `Mean in Real Space` (if you can't think in log units, this will make your life easier).
2. Set the value for M (=μ) to 90.
3. Set the value for S (=σ) to 0.8 (this flattens the distribution a bit.
4. Set the offset to 704.

I got these values by fiddling around a bit. It should look like this:



2. East Polynesian.

Another good calibration is East Polynesian: it's a well-attested linguistic group, and we have good archaeological evidence for when the initial settlement of East Polynesia began.

The ages are a bit controversial between "short" and "long" chronologies e.g.:

```
1025-1121 AD =   975-879 years ago (Wilsmhurst et al. '11)
800-1000 AD    =   1200-1000 years ago (Spriggs '10)
```

The average of these estimates is about 1000 years ago, and they're spread on both sides by about 150-200 years or so. This makes a great candidate for a `Normal` distribution.

Can you do this one? create a new calibration, call it EastPolynesian. Give it a mean of 1000, and a standard deviation of 100.

Add the following languages:

```
Hawaiian, Mangareva, Maori, Marquesan, RapanuiEasterIsland
Rarotongan, SouthIslandMaori, TahitianModern, Tuamotu
```

`[ ] Add East Polynesian clade.`

MCMC Tab.

Select the `MCMC` tab. This tab controls the analysis length and output files etc.

1. `Chain Length` is how long the run will go for (in generations). The default here is 10 million generations. Thankfully we don't need that for these data, so let's change that down to 1 million or you will have to wait 10x longer than everyone else in the room.

`[ ] Change the Chain Length to 1000000 (delete one zero).`

I want you to change the log file names to something sensible (it will make life easier later).

2. Open the `tracelog` section and change the `File Name` to something like "cpacific-ctmc-strict.log".

`[ ] Change the tracelog filename.`

3. Open the `treelog` section and change the `File Name` to something like "cpacific-ctmc-strict.trees".

`[ ] Change the treelog filename.`

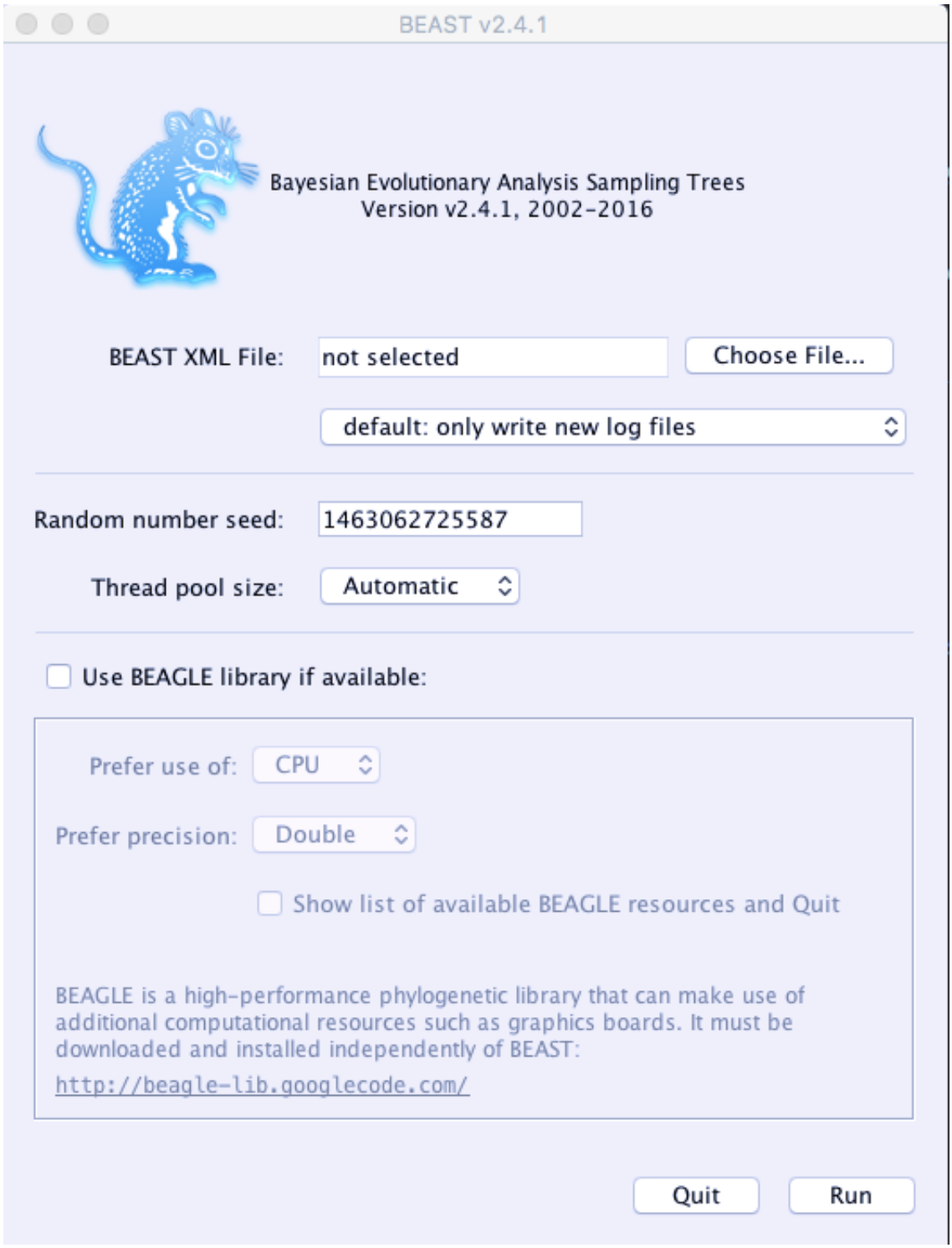
Save the XML file:

1. `File -> Save As`.
2. Make sure you give it a good name e.g. cpacific-ctmc-strict.xml.

`[ ] Make sure you've got the file saved somewhere sensible.`

IMPORTANT: Don't close BEAUTi, just leave it somewhere – we'll use it later to set up another analysis and don't want to have to redo everything.

Run the Analysis:



1. Finally open `BEAST`.
2. Click “Choose File” and feed it the XML file you generated.
3. Click Run.
4. Wait.

`[ ] Run BEAST.`