Temporal tree calibration in BEAST

or... "Help! I'm dating an alien!"

Gytis Dudas[1]

[1]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK

November 18, 2015

Introduction

The BEAST package

The BEAST (Drummond et al., 2012) package contains several programs that make a Bayesian phylogeneticist's life easier:

- **BEAST** is the workhorse of the package. It is what is used to do the actual analyses. BEAST takes XML (Extensible Markup Language) files as input, which contain sequence data, calibrations, evolutionary models, their parameters and desired output. Having an input file for BEAST also means that results can be easily replicated (or easily refuted) by other groups.
- **BEAUti** is what produces XML files via a graphical user interface (GUI). It offers a wide variety of models for multilocus, demographic and temporal inference. Note, however, that many analyses (*e.g.* GLMs, epoch models, *etc.*) that BEAST can do are not available for specification in BEAUti.
- **TreeAnnotator** collects information about the posterior distribution of trees and summarises them onto a single tree.
- **LogCombiner** is used to combine a number of independent MCMC chains into a single file after removing the burn-in from each. More chains are always better when it comes to verifying convergence onto the posterior distribution.
- **TreeStat** is a neat program that allows the recovery of various tree parameters from the posterior distribution of trees. You should use this if you want to recover the posterior distribution of internal branch lengths or tree imbalance statistics and cannot be bothered writing a tree parsing script to extract these parameters.

The entire package is available at: `http://tree.bio.ed.ac.uk/software/beast/`.

Auxiliary programs

In addition to BEAST there are 4 independently distributed programs that are indispensable to BEAST users:

- **BEAGLE** (Ayres et al., 2012) is a general purpose high-performance phylogenetic likelihood computation library. It has been developed to take advantage of parallel computational power, and thanks to its efficiency it has become an indispensable part of BEAST. If something does not work in BEAST a lot of the time it will be because you are not using BEAGLE. Available at `http://beast.bio.ed.ac.uk/beagle`.
- **Tracer** is used to assess convergence of parameters sampled over the course of the MCMC. It takes BEAST log files and reports information about the posterior distribution of each parameter *e.g.* mean, median, 95% highest posterior density intervals, as well as a proxy for convergence, the effective sample size (ESS) of each parameter. Available at: `http://tree.bio.ed.ac.uk/software/tracer/`.
- **FigTree** is one of the most widely used programs to visualise phylogenetic trees. In addition to visualizing trees it also offers the ability to manipulate and explore trees. Available at: `http://tree.bio.ed.ac.uk/software/figtree/`.
- **Path-O-Gen** (optional) offers the ability to perform root-to-tip regressions on phylogenetic trees. By regressing the height (cumulative distance from root) of each tip in a phylogeny against their dates of collection you can see the rate at which substitutions per site accumulate over time. Path-O-Gen works exclusively on tip dates and so is usually used on RNA virus datasets, where sequences collected years apart are sufficiently different to asses this. Available at: `http://tree.bio.ed.ac.uk/software/pathogen/`.

Theory and overview of exercises

This workshop will have two parts, both focused on ways of calibrating molecular phylogenies in BEAST. Calibration involves telling BEAST ***when*** either nodes or tips exist in the tree, and these are then used to arrive at posterior estimates of when (and with what confidence) everything else in the tree existed, given the data and the model.

The first exercise is tip calibration. This is the simplest and arguably the most powerful method of calibration. It involves specifying a date for every tip in the phylogeny, from which point it is only a matter of finding a tree topology and molecular clock rate (and through it the branch lengths) with the highest posterior probability. The power of this type of analysis is derived from the sampling timespan - the greater the timespan between sequences the more signal BEAST has to work with. Unfortunately, the kinds of datasets where full tip calibration is a viable method are rare and for the most part such calibration has been used on RNA virus sequences, although bacterial genomes are now long enough to at least consider it as a potential calibration option.

The second type of calibration involves telling BEAST when a particular node or nodes existed. This is usually based on geological data like estimated emergence dates for geographic barriers or niches, or, more usually, fossils. At least one node in a phylogeny is calibrated by defining a taxon set whose most recent common ancestor (or its stem, *i.e.* its immediate ancestor) receives an informative date prior. The prior is defined by a distribution with appropriate parameters indicating the level of confidence one has in the calibration.

Following the exercises you will also have the basic knowledge of BEAST to be able to use the third "calibration" method, which involves simply giving BEAST a rate prior, which is then used to estimate a phylogeny. This method has a lot of potential problems associated with it, the most obvious one being that an independently acquired and reliable molecular clock rate estimate is required.

At the end of the workshop you will be able to check phylogenies for potentially contaminant sequences, estimate sequence isolation dates and calibrate temporal phylogenies by specifying dates of isolation for sequences or internal node dates.

“Game over, man” - part 1

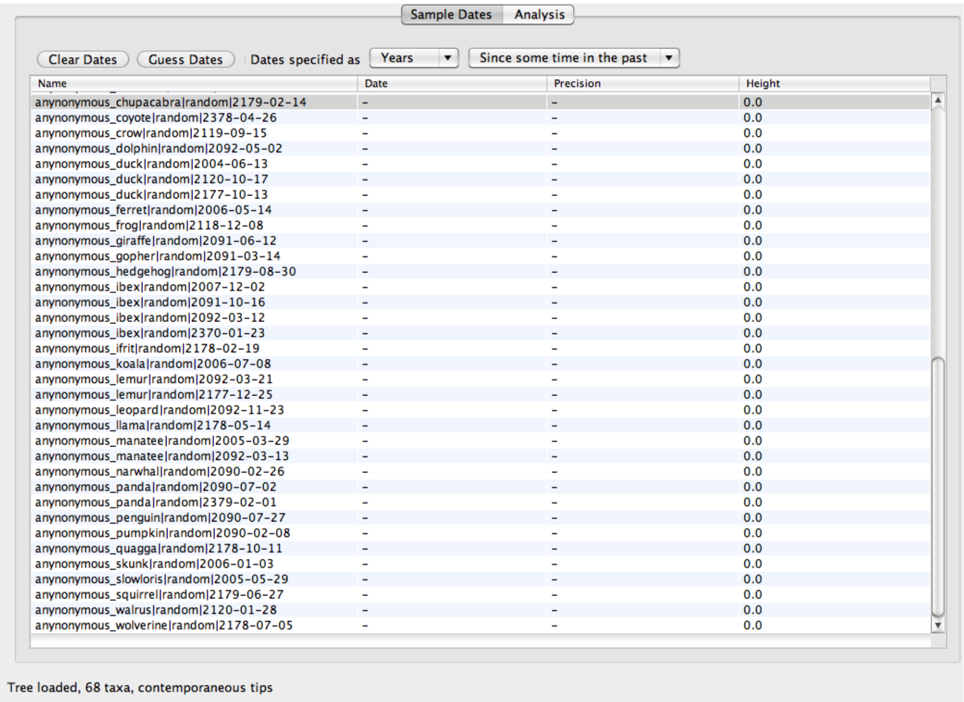
The setting

The Alien film franchise is a series of science fiction films and video games that spans roughly 375 years, if the Predator franchise cross-overs are included. Within the chronology of the joint franchise the events depicted in Alien vs Predator (2004) take place the earliest, in 2004. The last film in the chronology is Alien: Resurrection (1997) set in 2379 and there have been 5 films in between. Despite their usually futuristic setting the films depict phenomena which are of great interest in the present day, such as inter-species contact. We will use this futuristic setting mostly because of the way time passes within the franchise - the films take place decades apart from each other, but also uses technology that seemingly suspends the passage of time (stasis).

We will suppose that in the first 2 films of the franchise, Alien vs Predator (2004) and Aliens vs Predator: Requiem (2007), a rapidly evolving pathogen not unlike an RNA virus is transmitted from one of the extraterrestrial species to humans. Its genome is comprised of 9000 nucleotides that encode, according to the universal genetic code (what are the chances of that?), a polyprotein of 3000 amino acids. The genome does not recombine. Since its introduction this “xenovirus” has been sequenced by the appropriate authorities and we have access to 68 of its sequences isolated over time from the year 2004 all the way to the year 2379.

Checking for outliers with Path-O-Gen

It is always wise to make sure that your data contain enough temporal and genetic information to do a full-scale analysis, but also to make sure no contaminant sequences exist. We can do this by using Path-O-Gen, which performs root-to-tip regressions, where the isolation date of each tip is regressed against the tip’s height in a genetic phylogeny. In the presence of sufficient amounts of time we should observe that older sequences contain fewer mutations than more recent sequences. You should have access to a maximum likelihood phylogeny of the 68 sequences I have prepared earlier: (`https://www.dropbox.com/s/1oolvl4zi14p916/Alien_ML.newick?dl=0`). Once you have found the file you can launch Path-O-Gen v1.4, which will ask you for input. Select the file containing the maximum likelihood phylogeny and click `Open`. Another screen should open up:



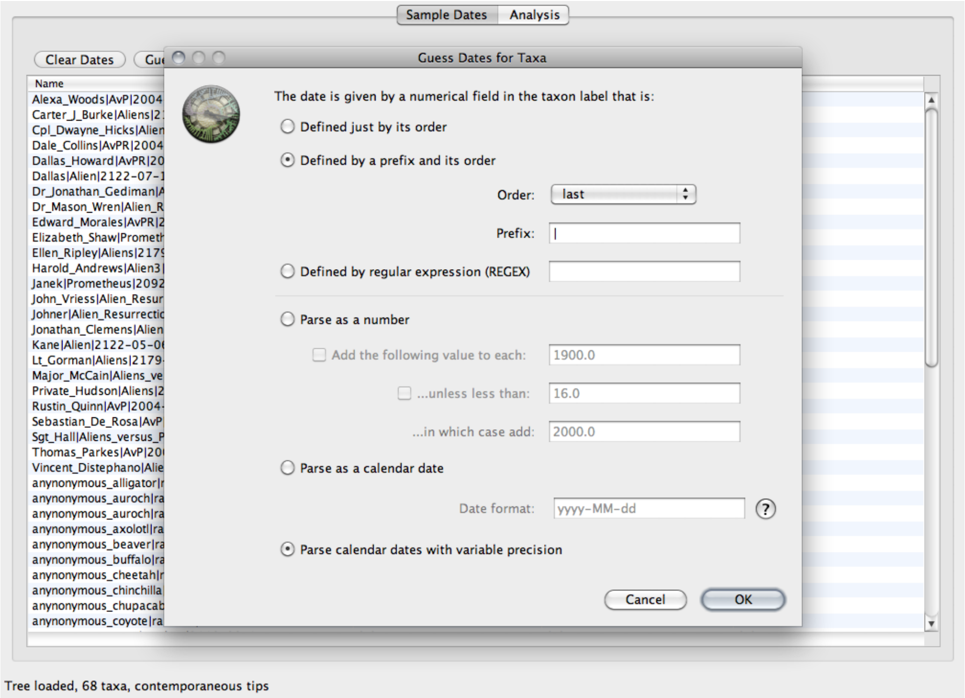
Dates specified as Years ▾ Since some time in the past ▾

<table>
<tr><th>Name</th><th>Date</th><th>Precision</th><th>Height</th></tr>
<tr><td>anynonymous_chupacabra|random|2179-02-14</td><td>-</td><td>-</td><td>0.0</td></tr>
<tr><td>anynonymous_coyote|random|2378-04-26</td><td>-</td><td>-</td><td>0.0</td></tr>
<tr><td>anynonymous_crow|random|2119-09-15</td><td>-</td><td>-</td><td>0.0</td></tr>
<tr><td>anynonymous_dolphin|random|2092-05-02</td><td>-</td><td>-</td><td>0.0</td></tr>
<tr><td>anynonymous_duck|random|2004-06-13</td><td>-</td><td>-</td><td>0.0</td></tr>
<tr><td>anynonymous_duck|random|2120-10-17</td><td>-</td><td>-</td><td>0.0</td></tr>
<tr><td>anynonymous_duck|random|2177-10-13</td><td>-</td><td>-</td><td>0.0</td></tr>
<tr><td>anynonymous_ferret|random|2006-05-14</td><td>-</td><td>-</td><td>0.0</td></tr>
<tr><td>anynonymous_frog|random|2118-12-08</td><td>-</td><td>-</td><td>0.0</td></tr>
<tr><td>anynonymous_giraffe|random|2091-06-12</td><td>-</td><td>-</td><td>0.0</td></tr>
<tr><td>anynonymous_gopher|random|2091-03-14</td><td>-</td><td>-</td><td>0.0</td></tr>
<tr><td>anynonymous_hedgehog|random|2179-08-30</td><td>-</td><td>-</td><td>0.0</td></tr>
<tr><td>anynonymous_ibex|random|2007-12-02</td><td>-</td><td>-</td><td>0.0</td></tr>
<tr><td>anynonymous_ibex|random|2091-10-16</td><td>-</td><td>-</td><td>0.0</td></tr>
<tr><td>anynonymous_ibex|random|2092-03-12</td><td>-</td><td>-</td><td>0.0</td></tr>
<tr><td>anynonymous_ibex|random|2370-01-23</td><td>-</td><td>-</td><td>0.0</td></tr>
<tr><td>anynonymous_ifrit|random|2178-02-19</td><td>-</td><td>-</td><td>0.0</td></tr>
<tr><td>anynonymous_koala|random|2006-07-08</td><td>-</td><td>-</td><td>0.0</td></tr>
<tr><td>anynonymous_lemur|random|2092-03-21</td><td>-</td><td>-</td><td>0.0</td></tr>
<tr><td>anynonymous_lemur|random|2177-12-25</td><td>-</td><td>-</td><td>0.0</td></tr>
<tr><td>anynonymous_leopard|random|2092-11-23</td><td>-</td><td>-</td><td>0.0</td></tr>
<tr><td>anynonymous_llama|random|2178-05-14</td><td>-</td><td>-</td><td>0.0</td></tr>
<tr><td>anynonymous_manatee|random|2005-03-29</td><td>-</td><td>-</td><td>0.0</td></tr>
<tr><td>anynonymous_manatee|random|2092-03-13</td><td>-</td><td>-</td><td>0.0</td></tr>
<tr><td>anynonymous_narwhal|random|2090-02-26</td><td>-</td><td>-</td><td>0.0</td></tr>
<tr><td>anynonymous_panda|random|2090-07-02</td><td>-</td><td>-</td><td>0.0</td></tr>
<tr><td>anynonymous_panda|random|2379-02-01</td><td>-</td><td>-</td><td>0.0</td></tr>
<tr><td>anynonymous_penguin|random|2090-07-27</td><td>-</td><td>-</td><td>0.0</td></tr>
<tr><td>anynonymous_pumpkin|random|2090-02-08</td><td>-</td><td>-</td><td>0.0</td></tr>
<tr><td>anynonymous_quagga|random|2178-10-11</td><td>-</td><td>-</td><td>0.0</td></tr>
<tr><td>anynonymous_skunk|random|2006-01-03</td><td>-</td><td>-</td><td>0.0</td></tr>
<tr><td>anynonymous_slowloris|random|2005-05-29</td><td>-</td><td>-</td><td>0.0</td></tr>
<tr><td>anynonymous_squirrel|random|2179-06-27</td><td>-</td><td>-</td><td>0.0</td></tr>
<tr><td>anynonymous_walrus|random|2120-01-28</td><td>-</td><td>-</td><td>0.0</td></tr>
<tr><td>anynonymous_wolverine|random|2178-07-05</td><td>-</td><td>-</td><td>0.0</td></tr>
</table>

Tree loaded, 68 taxa, contemporaneous tips

Our sequence names are shown on the left. You will observe that each sequence name has 3 fields separated by the pipe character ("|") - the person from which the sequence was isolated (if available), the film in which the character appears and its isolation date. The date fields to the right of sequence names are empty, as are the precision fields. You can edit the 'Date' and 'Precision' fields in Path-O-Gen manually, whilst the 'Height' field tracks how old each sequence is in comparison to the youngest sequence. Date should be in decimal format (*e.g.* January 1, 2016 will be 2016.0), whereas precision dictates how much younger the sequence could potentially be. For example, if you have a dataset where you know the exact year, month and day of isolation for all your sequences but there is one sequence that you know was isolated at some point in 2011 and nothing else you would set its precision to 1.0, meaning the sequence was isolated anywhere between 2011.0 and 2012.0.

We are not interested in converting each date manually into decimal years, so we will set the dates of each sequence and their precision automatically. Click on `Guess Dates` and another dialogue should pop up. The date of each taxon is the last element of the sequence name split by the pipe ("|") character and defined as a calendar date. Click and enter the appropriate options:



Alexa_Woods|AvP|2004
Carter_J_Burke|Aliens|2
Cpl_Dwayne_Hicks|Alien
Dale_Collins|AvPR|2004
Dallas_Howard|AvPR|20
Dallas|Alien|2122-07-1
Dr_Jonathan_Gediman|A
Dr_Mason_Wren|Alien_R
Edward_Morales|AvPR|2
Elizabeth_Shaw|Prometh
Ellen_Ripley|Aliens|2179
Harold_Andrews|Alien3|
Janek|Prometheus|2092
John_Vriess|Alien_Resur
Johner|Alien_Resurrectio
Jonathan_Clemens|Alien
Kane|Alien|2122-05-0
Lt_Gorman|Aliens|2179
Major_McCain|Aliens_ve
Private_Hudson|Aliens|2
Rustin_Quinn|AvP|2004-
Sebastian_De_Rosa|AvP
Sgt_Hall|Aliens_versus_P
Thomas_Parkes|AvP|20
Vincent_Distephano|Alie
anynonymous_alligator|r
anynonymous_auroch|ra
anynonymous_auroch|ra
anynonymous_axolotl|ra
anynonymous_beaver|ra
anynonymous_buffalo|ra
anynonymous_cheetah|r
anynonymous_chinchilla
anynonymous_chupacab
anynonymous_coyote|ra

Guess Dates for Taxa

The date is given by a numerical field in the taxon label that is:

○ Defined just by its order

● Defined by a prefix and its order

Order: last

Prefix: |

○ Defined by regular expression (REGEX)

○ Parse as a number

□ Add the following value to each: 1900.0

□ ...unless less than: 16.0

...in which case add: 2000.0

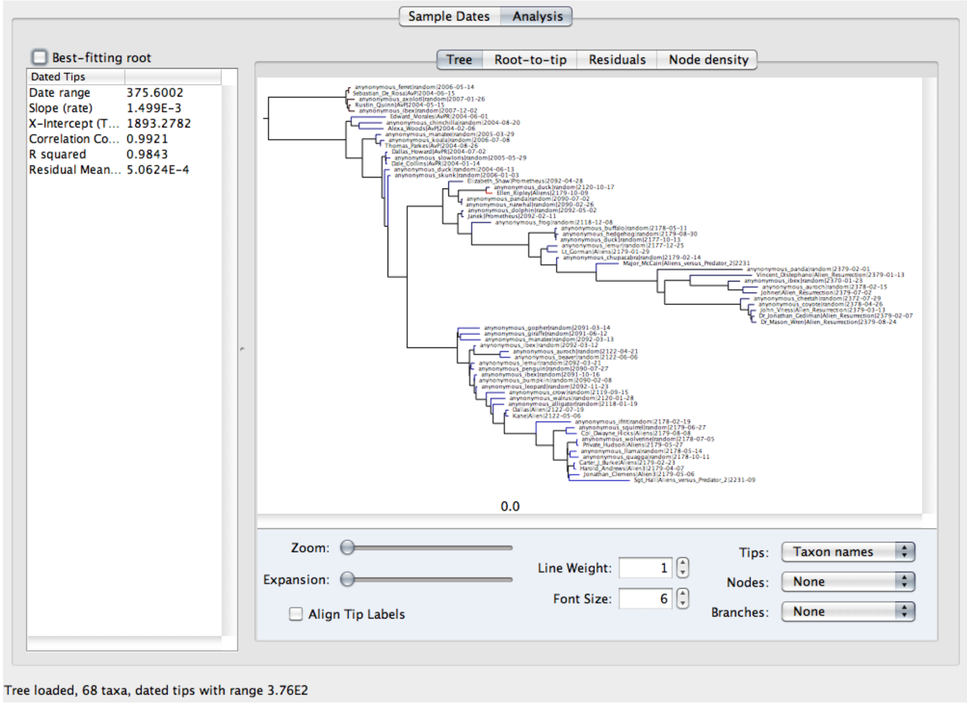
○ Parse as a calendar date

Date format: yyyy-MM-dd

● Parse calendar dates with variable precision

Tree loaded, 68 taxa, contemporaneous tips

We are now ready to do the actual analysis. Click on the `Analysis` tab (next to the `Sample Dates` tab).



☐ Best-fitting root

Dated Tips	
Date range	375.6002
Slope (rate)	1.499E-3
X-Intercept (T...	1893.2782
Correlation Co...	0.9921
R squared	0.9843
Residual Mean...	5.0624E-4

Zoom:

Expansion:

☐ Align Tip Labels

Line Weight: 1

Font Size: 6

Tips: Taxon names

Nodes: None

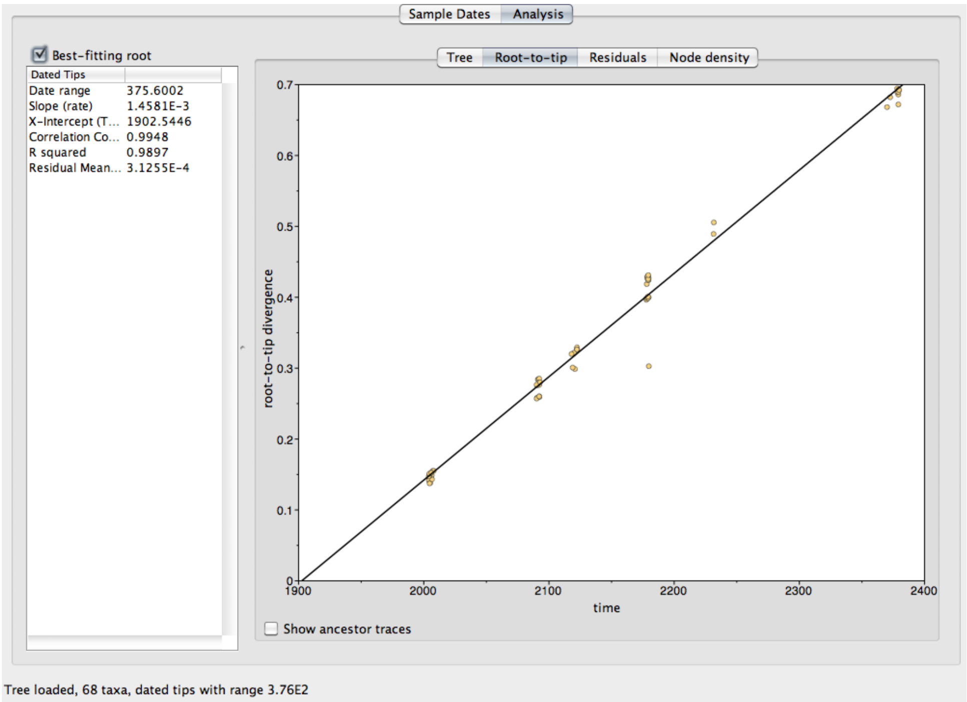
Branches: None

Tree loaded, 68 taxa, dated tips with range 3.76E2

Here you can see several important statistics reported by the current state of the analysis: the timespan of the calibrated tips (Date range, in decimal units), the estimated evolutionary rate (Slope (rate), given as changes per time unit), the most recent common ancestor estimate (X-Intercept, TMRCA) and various statistics from the regression quantifying dispersal - the Correlation Coefficient, R squared and Residual Mean Squared. Our analysis indicates a very good correlation between tip height and their isolation date (r^2=0.9843), but since the tree has been rooted mid-branch we can check whether we can improve the rooting further via a least squares method. Click on the tickbox that says `Best-fitting root`. The r^2 value should have improved to 0.9897, with only a minor shift along the branch connecting the first bifurcation in the tree, a slight reduction in estimated rate and a slightly more recent estimated TMRCA.

Here is a question, though: Path-O-Gen reports various parameters from a linear regression, but not the p-value. Why is that?

Now let's look at the actual root-to-tip regression (click on the `Root-to-tip` tab).



☑ Best-fitting root

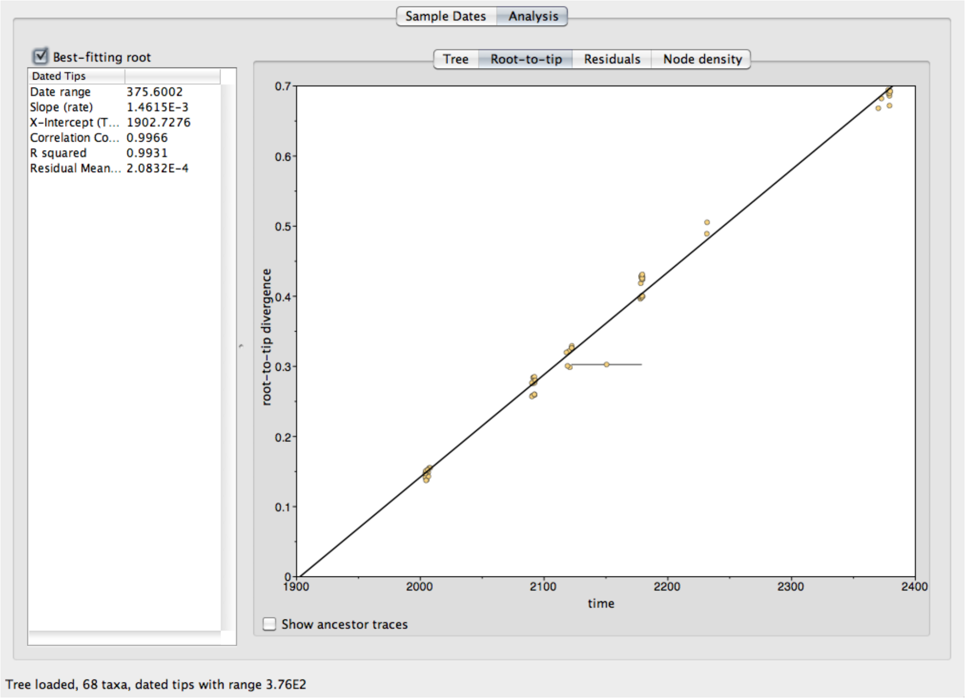
Dated Tips	
Date range	375.6002
Slope (rate)	1.4581E-3
X-Intercept (T...	1902.5446
Correlation Co...	0.9948
R squared	0.9897
Residual Mean...	3.1255E-4

□ Show ancestor traces

Tree loaded, 68 taxa, dated tips with range 3.76E2

The dots in the plot correspond to individual tips. For the most part the correlation seems pretty convincing, except for one point - click and drag to select it and go back to the tree (`Tree` tab). It's a xenovirus sample collected from Ellen Ripley, the date of which indicates that it was collected in the year 2179, when the events depicted in the films Aliens (1986) and Alien^3 (1992) take place. We know that the sample must have been taken immediatelly after Ripley was woken up from stasis, since she was sent off on another xenomorph fighting mission soon after. Could she have been infected with the xenovirus prior to entering stasis? Go back to the `Sample dates` tab, find the sequence labeled "Ellen_Ripley" and change its date (in the `Date` column) to 2122 (when the events depicted in Alien (1979) take place) and precision to 57 (the number of years Ripley was in stasis between the Alien (1979) and Aliens (1986).

Now go back to `Analysis` and click on the `Root-to-tip` tab there:



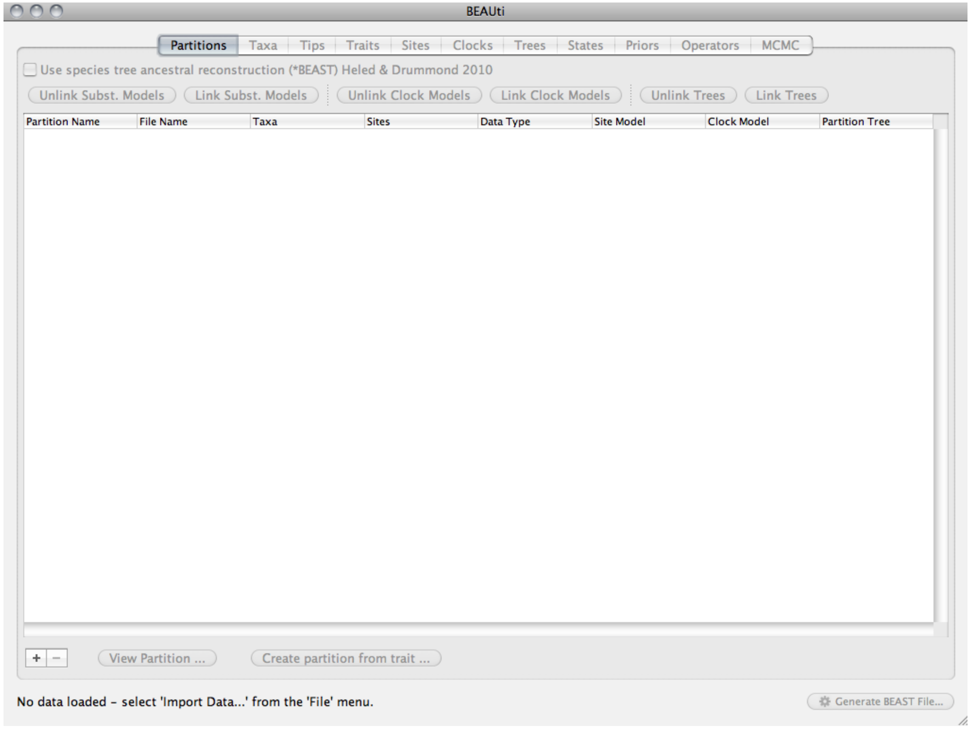
Dated Tips	
Date range	375.6002
Slope (rate)	1.4615E-3
X-Intercept (T...	1902.7276
Correlation Co...	0.9966
R squared	0.9931
Residual Mean...	2.0832E-4

Tree loaded, 68 taxa, dated tips with range 3.76E2

Looks like the authorities failed to account for time Ellen Ripley spent in stasis when indicating the collection date of the sequence. It is highly likely that Ripley was in fact infected during Alien (1979) and when she entered stasis the xenovirus did not replicate until it was sequenced in the year 2179 during Aliens (1986). So even though the xenovirus was isolated in 2179 it has a number of mutations away from the root that is typical of xenoviruses isolated around the year 2122. We will make note of this for future analyses.

Setting up a BEAST XML file

It is time to do the actual analysis. Begin by launching BEAUti.

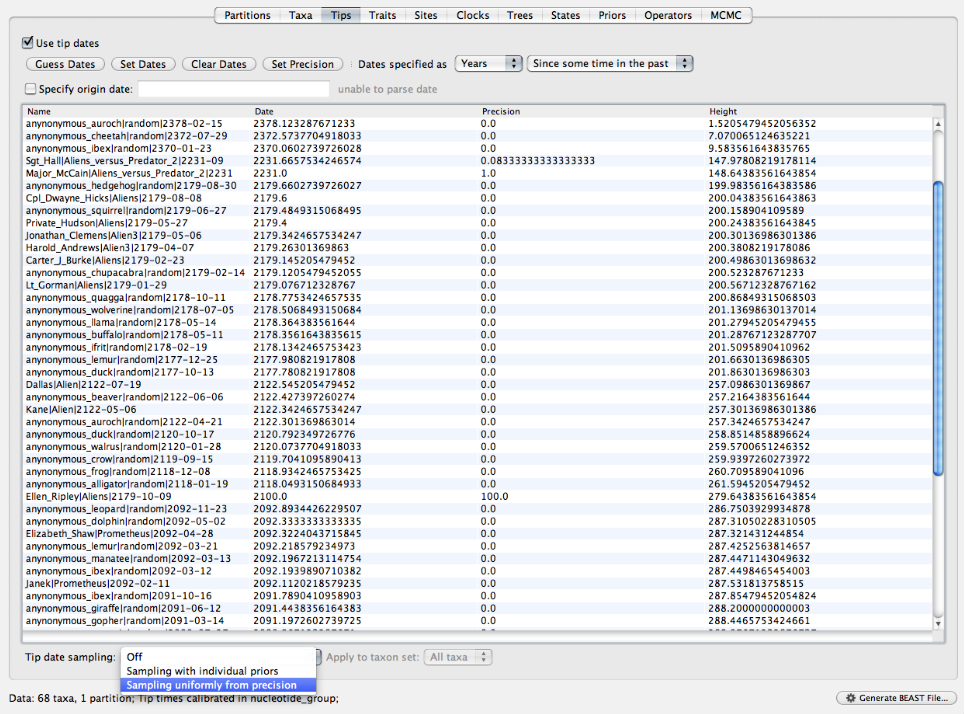


☐ Use species tree ancestral reconstruction (*BEAST) Heled & Drummond 2010

No data loaded – select 'Import Data...' from the 'File' menu.

Then, either go to `Files>Import Date` or press on ‘`+`’ on the bottom left corner. Direct BEAUti to the fasta file that contains our sequences, which is available here: `https://www.dropbox.com/s/lrmjiycqbdefode/Alien_alignment.fasta?dl=0`. Once the sequences are loaded you should have access to all the tabs in BEAUti. Go to the `Tips` tab, tick the `Use tip dates` tickbox and then on `Guess Dates`. Direct BEAUti to the isolation date of each sequence (this should be filled in the same way we did it in Path-O-Gen).

Before we proceed any further remember that we suspect the sequence from Ellen Ripley to be a possible contaminant. We shall set its `Date` to 2100.0 (rather than 2122 to be on the safe side) and `Precision` to 100.0 years and select `Sampling tip dates from precision` from the "Tip date sampling" menu (bottom left).



☑ Use tip dates

Dates specified as: Years; Since some time in the past

☐ Specify origin date: unable to parse date

Name	Date	Precision	Height
anynonymous_auroch|random|2378-02-15	2378.123287671233	0.0	1.5205479452056352
anynonymous_cheetah|random|2372-07-29	2372.5737704918033	0.0	7.070065124635221
anynonymous_ibex|random|2370-01-23	2370.0602739726028	0.0	9.583561643835765
Sgt_Hall|Aliens_versus_Predator_2|2231-09	2231.6657534246574	0.08333333333333333	147.97808219178114
Major_McCain|Aliens_versus_Predator_2|2231	2231.0	1.0	148.64383561643854
anynonymous_hedgehog|random|2179-08-30	2179.6602739726027	0.0	199.98356164383586
Cpl_Dwayne_Hicks|Aliens|2179-08-08	2179.6	0.0	200.04383561643863
anynonymous_squirrel|random|2179-06-27	2179.4849315068495	0.0	200.158904109589
Private_Hudson|Aliens|2179-05-27	2179.4	0.0	200.24383561643845
Jonathan_Clemens|Alien3|2179-05-06	2179.3424657534247	0.0	200.30136986301386
Harold_Andrews|Alien3|2179-04-07	2179.26301369863	0.0	200.3808219178086
Carter_J_Burke|Aliens|2179-02-23	2179.145205479452	0.0	200.49863013698632
anynonymous_chupacabra|random|2179-02-14	2179.1205479452055	0.0	200.523287671233
Lt_Gorman|Aliens|2179-01-29	2179.076712328767	0.0	200.56712328767162
anynonymous_quagga|random|2178-10-11	2178.7753424657535	0.0	200.86849315068503
anynonymous_wolverine|random|2178-07-05	2178.5068493150684	0.0	201.13698630137014
anynonymous_llama|random|2178-05-14	2178.364383561644	0.0	201.27945205479455
anynonymous_buffalo|random|2178-05-11	2178.3561643835615	0.0	201.28767123287707
anynonymous_ifrit|random|2178-02-19	2178.1342465753423	0.0	201.5095890410962
anynonymous_lemur|random|2177-12-25	2177.980821917808	0.0	201.6630136986305
anynonymous_duck|random|2177-10-13	2177.780821917808	0.0	201.8630136986303
Dallas|Alien|2122-07-19	2122.545205479452	0.0	257.0986301369867
anynonymous_beaver|random|2122-06-06	2122.427397260274	0.0	257.2164383561644
Kane|Alien|2122-05-06	2122.3424657534247	0.0	257.30136986301386
anynonymous_auroch|random|2122-04-21	2122.301369863014	0.0	257.3424657534247
anynonymous_duck|random|2120-10-17	2120.792349726776	0.0	258.8514858896624
anynonymous_walrus|random|2120-01-28	2120.0737704918033	0.0	259.5700651246352
anynonymous_crow|random|2119-09-15	2119.7041095890413	0.0	259.9397260273972
anynonymous_frog|random|2118-12-08	2118.9342465753425	0.0	260.709589041096
anynonymous_alligator|random|2118-01-19	2118.0493150684933	0.0	261.5945205479452
Ellen_Ripley|Aliens|2179-10-09	2100.0	100.0	279.64383561643854
anynonymous_leopard|random|2092-11-23	2092.8934426229507	0.0	286.7503929934878
anynonymous_dolphin|random|2092-05-02	2092.3333333333335	0.0	287.31050228310505
Elizabeth_Shaw|Prometheus|2092-04-28	2092.3224043715845	0.0	287.321431244854
anynonymous_lemur|random|2092-03-21	2092.218579234973	0.0	287.4252563814657
anynonymous_manatee|random|2092-03-13	2092.1967213114754	0.0	287.4471143049632
anynonymous_ibex|random|2092-03-12	2092.1939890710382	0.0	287.4498465454003
Janek|Prometheus|2092-02-11	2092.1120218579235	0.0	287.531813758515
anynonymous_ibex|random|2091-10-16	2091.7890410958903	0.0	287.85479452054824
anynonymous_giraffe|random|2091-06-12	2091.4438356164383	0.0	288.2000000000003
anynonymous_gopher|random|2091-03-14	2091.1972602739725	0.0	288.4465753424661

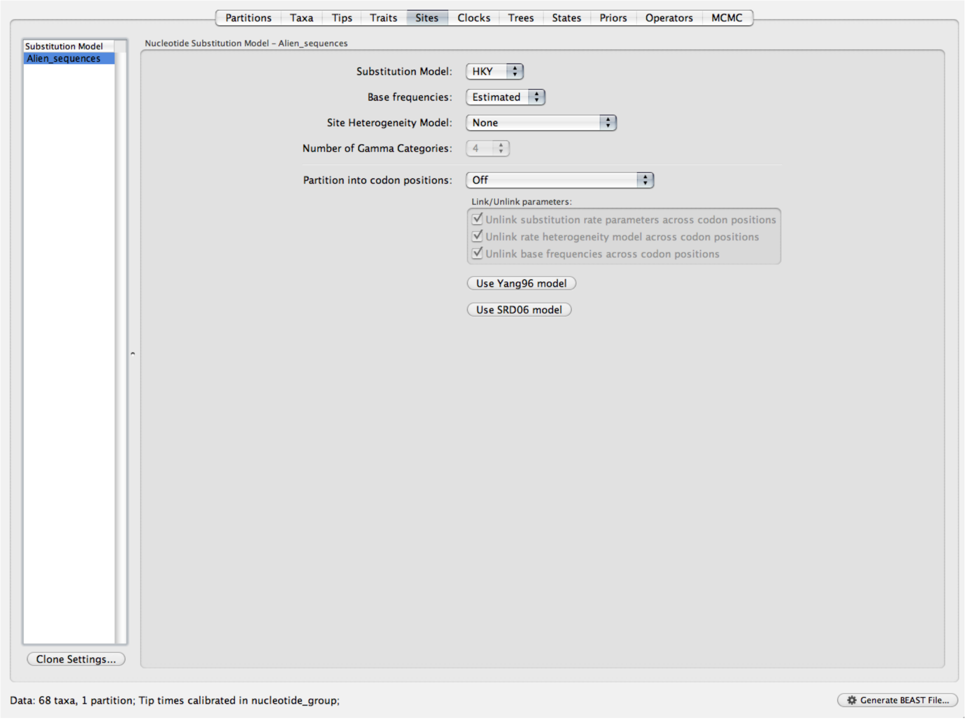
Tip date sampling: Off / Sampling with individual priors / Sampling uniformly from precision

Apply to taxon set: All taxa

Data: 68 taxa, 1 partition; Tip times calibrated in nucleotide_group;

This will sample possible tip dates for each sequence that has non-zero precision, based on the molecular clock rate.

Go to the `Sites` tab. This is where we define the substitution model for each partition. In our case we only have one partition called "Alien_alignment". Click on the `Use SRD06 model` button. This is a combination of data partitioning and nucleotide substitution model. It models substitutions separately for codon positions 1+2 and 3 according to independent HKY+Γ_4 substitution models. This works for our data because all sequences are coding and in frame.



Nucleotide Substitution Model – Alien_sequences

Substitution Model: HKY

Base frequencies: Estimated

Site Heterogeneity Model: None

Number of Gamma Categories: 4

Partition into codon positions: Off

Link/Unlink parameters:

☑ Unlink substitution rate parameters across codon positions

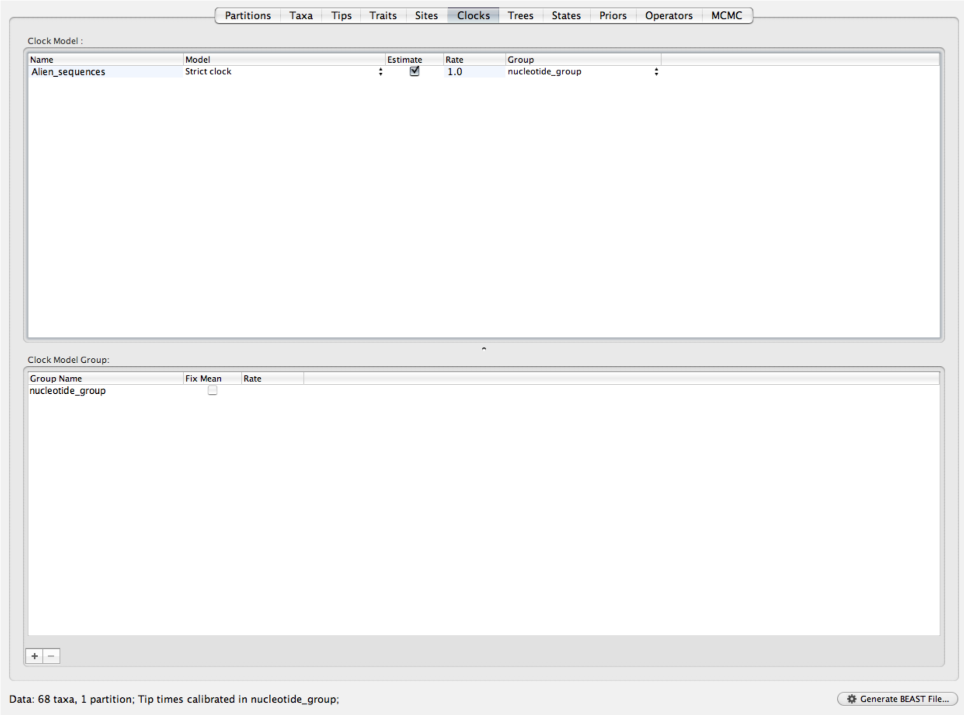
☑ Unlink rate heterogeneity model across codon positions

☑ Unlink base frequencies across codon positions

Data: 68 taxa, 1 partition; Tip times calibrated in nucleotide_group;

If we had non-coding sequences for our xenovirus too we could load them as a separate partition (*i.e.* separate file with the same sequence names) and model substitutions in that according to another HKY+Γ_4 model with `Partition into codon positions` switched off, since they are non-coding.

Next, move on to `Clocks` tab. The molecular clock for our partition ("Alien_alignment") is set to `Strict clock` that is to be estimated by default. We can leave this as is, since Path-O-Gen appeared to indicate a strong clock signal in our data with relatively little deviation from the regression line. If we observed any upward or downward offsets within the data in Path-O-Gen (which could be caused by between-branch rate variation) we would consider using a relaxed lognormal relaxed uncorrelated clock (we will actually use it in the next part of this workshop). The strict clock that we went with estimates a single rate that is applied to every branch in the phylogeny. We also left the `Estimate` tickbox ticked, because we actually have sufficient information to estimate the rate. If instead we knew of an independently derived and reliable rate estimate, but had no other information to calibrate the phylogeny we would untick this, which tells BEAST that the rate should be drawn entirely from the prior.



Clock Model :

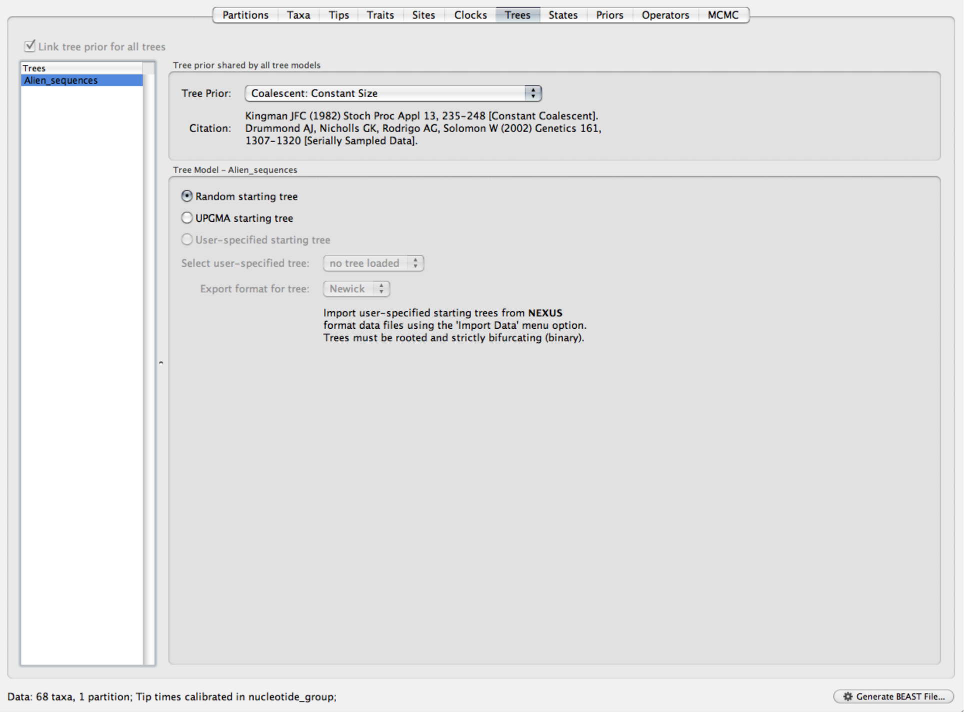
Name	Model	Estimate	Rate	Group
Alien_sequences	Strict clock	☑	1.0	nucleotide_group

Clock Model Group:

Group Name	Fix Mean	Rate
nucleotide_group	□	

Data: 68 taxa, 1 partition; Tip times calibrated in nucleotide_group;

Go to the `Trees` tab. This is where we define the tree prior. Different demographic processes result in differently shaped phylogenies. For example, exponentially growing populations typically have phylogenies with very long external branches and short internal branches. Under the coalescent this is explained as rapid coalescence (short branches) at small population sizes (early) and slow coalescence (long branches) at large population sizes (late). BEAST offers a wide variety of tree priors, but we will stick with the default `Coalescent: Constant Size`. This will assume that our sequences were generated by a population of a constant size, whose size will be estimated over the course of the MCMC.



☑ Link tree prior for all trees

Trees
Alien_sequences

Tree prior shared by all tree models

Tree Prior: Coalescent: Constant Size

Citation: Kingman JFC (1982) Stoch Proc Appl 13, 235-248 [Constant Coalescent].
Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W (2002) Genetics 161, 1307-1320 [Serially Sampled Data].

Tree Model - Alien_sequences

● Random starting tree
○ UPGMA starting tree
○ User-specified starting tree

Select user-specified tree: no tree loaded

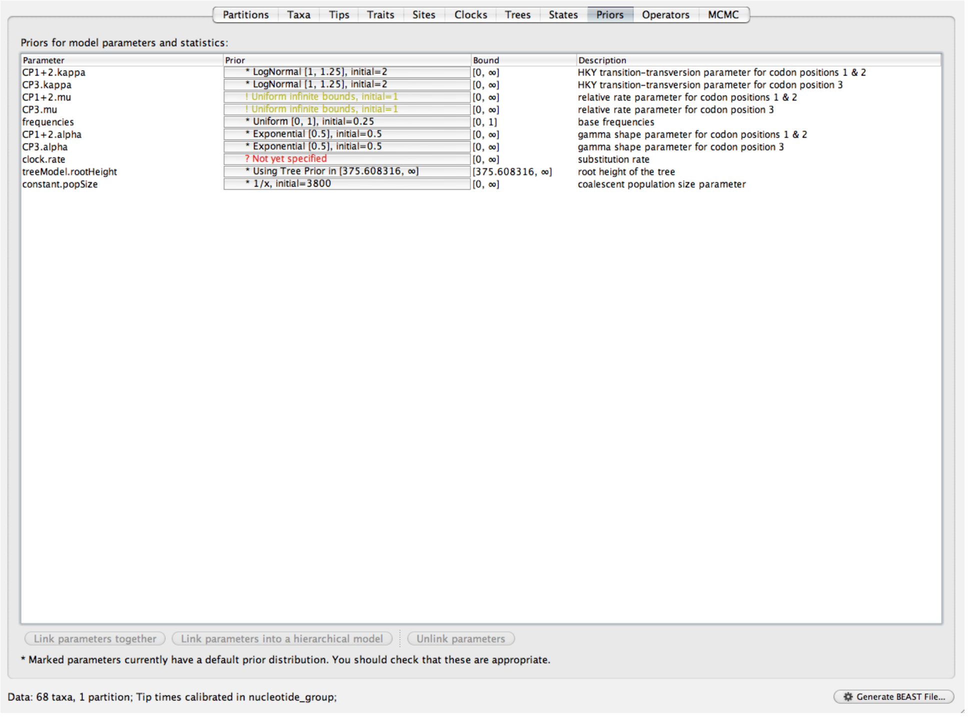
Export format for tree: Newick

Import user-specified starting trees from **NEXUS**
format data files using the 'Import Data' menu option.
Trees must be rooted and strictly bifurcating (binary).

Data: 68 taxa, 1 partition; Tip times calibrated in nucleotide_group;

There are non-coalescent tree priors too, as well as non-parametric coalescent priors which, with sufficient data, are able to estimate the rate of coalescence and thus effective population size (actually N_e×generation time) directly from the data. If we had more than one partition we could also link or unlink tree priors, which would dramatically improve our demographic inference.

Let us move on to the `Priors` tab. This is the most important part of setting up a BEAST analysis when data is not informative. You will notice that the `clock.rate` prior is not specified and highlighted in red. By default you have to specify a molecular clock rate prior manually. Click on `? Not yet specified`. A dialogue box should open where you can specify the family of distributions from which the prior distribution will be sampled, as well as the distribution's parameters. We will select the `CTMC Rate Reference` (Ferreira and Suchard, 2008) option and leave it under the default value. This is an uninformative prior on the molecular clock rate. We could also select a uniform prior or even a gamma distribution with shape 0.001 and scale 100.0, which would place a lot of the probability density on lower rates. Since we are using Bayesian methods we have information coming from the data and from the prior. When there is sufficient information in the data the prior should not influence the final result much, if it does then most of the information is coming from the prior, not the data.



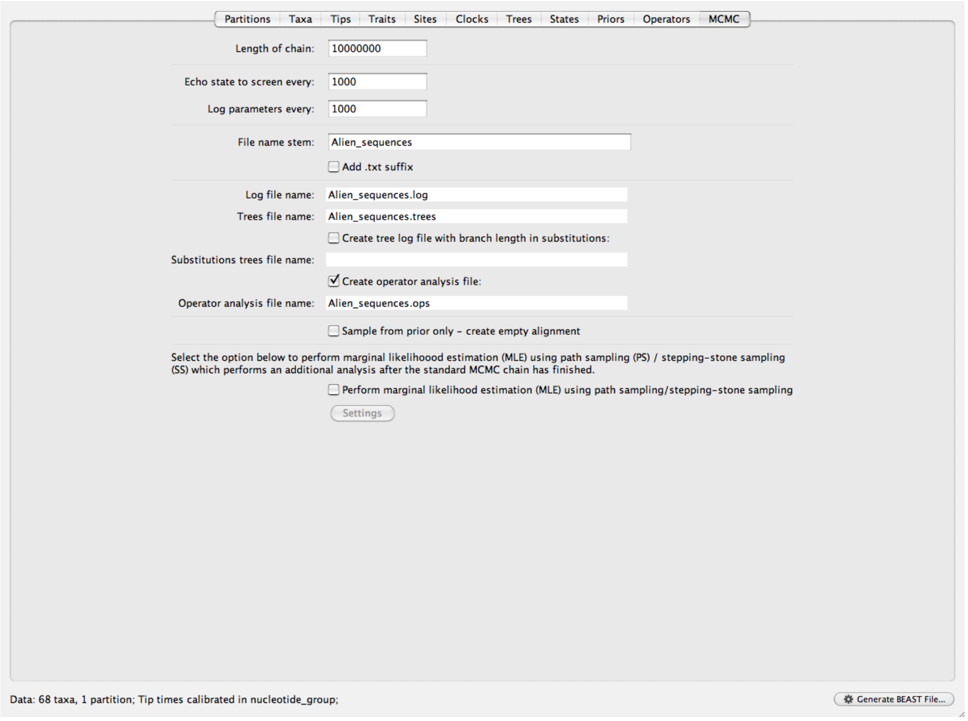
Priors for model parameters and statistics:

Parameter	Prior	Bound	Description
CP1+2.kappa	* LogNormal [1, 1.25], initial=2	[0, ∞]	HKY transition-transversion parameter for codon positions 1 & 2
CP3.kappa	* LogNormal [1, 1.25], initial=2	[0, ∞]	HKY transition-transversion parameter for codon position 3
CP1+2.mu	! Uniform infinite bounds, initial=1	[0, ∞]	relative rate parameter for codon positions 1 & 2
CP3.mu	! Uniform infinite bounds, initial=1	[0, ∞]	relative rate parameter for codon position 3
frequencies	* Uniform [0, 1], initial=0.25	[0, 1]	base frequencies
CP1+2.alpha	* Exponential [0.5], initial=0.5	[0, ∞]	gamma shape parameter for codon positions 1 & 2
CP3.alpha	* Exponential [0.5], initial=0.5	[0, ∞]	gamma shape parameter for codon position 3
clock.rate	? Not yet specified	[0, ∞]	substitution rate
treeModel.rootHeight	* Using Tree Prior in [375.608316, ∞]	[375.608316, ∞]	root height of the tree
constant.popSize	* 1/x, initial=3800	[0, ∞]	coalescent population size parameter

* Marked parameters currently have a default prior distribution. You should check that these are appropriate.

Data: 68 taxa, 1 partition; Tip times calibrated in nucleotide_group;

Next we will move to the MCMC tab. This is where we define the chain length of our MCMC, the sampling frequency and the name the output files will have. By default BEAST is set up to run for 10 million states, sampling every 1 000 states, which gives 10 000 samples from the posterior distribution. MCMC length should be tailored to each dataset to account for model complexity and parameter space, whereas sampling frequency is usually set to give 10 000 samples from the posterior distribution. Ideally at least two independent MCMC chains should be run as well, to ensure that all chains converge onto the same posterior distribution, but for our purposes one will do.



Length of chain: 10000000

Echo state to screen every: 1000

Log parameters every: 1000

File name stem: Alien_sequences

□ Add .txt suffix

Log file name: Alien_sequences.log

Trees file name: Alien_sequences.trees

□ Create tree log file with branch length in substitutions:

Substitutions trees file name:

☑ Create operator analysis file:

Operator analysis file name: Alien_sequences.ops

□ Sample from prior only – create empty alignment

Select the option below to perform marginal likelihoood estimation (MLE) using path sampling (PS) / stepping-stone sampling (SS) which performs an additional analysis after the standard MCMC chain has finished.

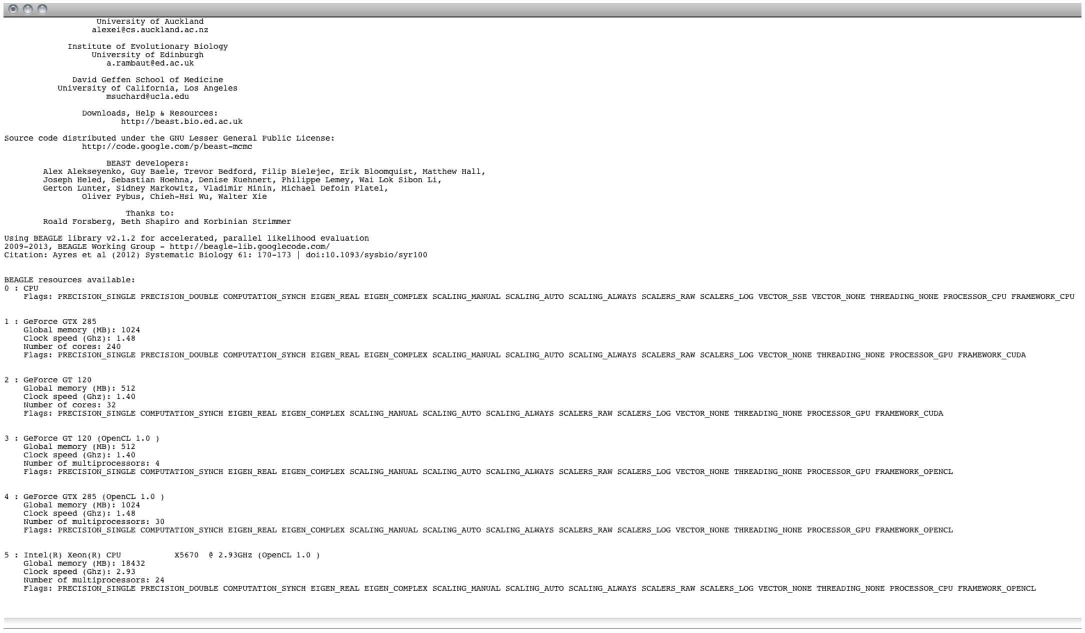
□ Perform marginal likelihood estimation (MLE) using path sampling/stepping-stone sampling

Data: 68 taxa, 1 partition; Tip times calibrated in nucleotide_group;

We will reduce the Length of chain for our purposes down to 3 million, set Echo state to screen every: to 10 000, and leave Log parameters every: at 1 000 as is. This will gives us 3 000 000 / 1 000 = 3 000 samples from the posterior, which should be sufficient. We reduced the frequency of screen updates because we are running a relatively small and simple analysis, which will be relatively fast too. If we left Echo state to screen every: as is at 1 000 BEAST would spend a considerable portion of time sending the updates to screen, which are being logged to file anyway, rather than doing the actual computation. Echoing some parameters to screen is meant to inform where the analysis is right now without having to download the log file (if you're running BEAST on a cluster). A moderately sophisticated BEAST will take hours to run, so it is important to catch any abnormal results early that could be caused by incorrectly formatted sequences or sequence names, wrong models, *etc.*

Once you've changed the sampling/echoing frequencies you can click on Generate BEAST File... at the bottom right corner. Save the XML file somewhere where you can run it and we will be good to go with the actual analysis.

Fire up BEAST. A dialogue box should open. First we'll make sure that you have BEAGLE installed and that it recognizes your hardware. Make sure the `Use BEAGLE library if available` and `Show list of available BEAGLE resources and Quit` tickboxes are ticked, without giving it a file. Press `Run`. The dialogue should scroll through the available resources indexed from 0. In my case it shows 6 available resources which are the CPUs and GPUs on my machine (yours will be different):



```
                University of Auckland
              alexei@cs.auckland.ac.nz

           Institute of Evolutionary Biology
                University of Edinburgh
                   a.rambaut@ed.ac.uk

            David Geffen School of Medicine
         University of California, Los Angeles
                  msuchard@ucla.edu

               Downloads, Help & Resources:
                    http://beast.bio.ed.ac.uk

Source code distributed under the GNU Lesser General Public License:
               http://code.google.com/p/beast-mcmc

                  BEAST developers:
        Alex Alekseyenko, Guy Baele, Trevor Bedford, Filip Bielejec, Erik Bloomquist, Matthew Hall,
        Joseph Heled, Sebastian Hoehna, Denise Kuehnert, Philippe Lemey, Wai Lok Sibon Li,
        Gerton Lunter, Sidney Markowitz, Vladimir Minin, Michael Defoin Platel,
               Oliver Pybus, Chieh-Hsi Wu, Walter Xie

                      Thanks to:
        Roald Forsberg, Beth Shapiro and Korbinian Strimmer

Using BEAGLE library v2.1.2 for accelerated, parallel likelihood evaluation
2009-2013, BEAGLE Working Group - http://beagle-lib.googlecode.com/
Citation: Ayres et al (2012) Systematic Biology 61: 170-173 | doi:10.1093/sysbio/syr100


BEAGLE resources available:
0 : CPU
    Flags: PRECISION_SINGLE PRECISION_DOUBLE COMPUTATION_SYNCH EIGEN_REAL EIGEN_COMPLEX SCALING_MANUAL SCALING_AUTO SCALING_ALWAYS SCALERS_RAW SCALERS_LOG VECTOR_SSE VECTOR_NONE THREADING_NONE PROCESSOR_CPU FRAMEWORK_CPU


1 : GeForce GTX 285
    Global memory (MB): 1024
    Clock speed (Ghz): 1.48
    Number of cores: 240
    Flags: PRECISION_SINGLE PRECISION_DOUBLE COMPUTATION_SYNCH EIGEN_REAL EIGEN_COMPLEX SCALING_MANUAL SCALING_AUTO SCALING_ALWAYS SCALERS_RAW SCALERS_LOG VECTOR_NONE THREADING_NONE PROCESSOR_GPU FRAMEWORK_CUDA


2 : GeForce GT 120
    Global memory (MB): 512
    Clock speed (Ghz): 1.40
    Number of cores: 32
    Flags: PRECISION_SINGLE COMPUTATION_SYNCH EIGEN_REAL EIGEN_COMPLEX SCALING_MANUAL SCALING_AUTO SCALING_ALWAYS SCALERS_RAW SCALERS_LOG VECTOR_NONE THREADING_NONE PROCESSOR_GPU FRAMEWORK_CUDA


3 : GeForce GT 120 (OpenCL 1.0 )
    Global memory (MB): 512
    Clock speed (Ghz): 1.40
    Number of multiprocessors: 4
    Flags: PRECISION_SINGLE COMPUTATION_SYNCH EIGEN_REAL EIGEN_COMPLEX SCALING_MANUAL SCALING_AUTO SCALING_ALWAYS SCALERS_RAW SCALERS_LOG VECTOR_NONE THREADING_NONE PROCESSOR_GPU FRAMEWORK_OPENCL


4 : GeForce GTX 285 (OpenCL 1.0 )
    Global memory (MB): 1024
    Clock speed (Ghz): 1.48
    Number of multiprocessors: 30
    Flags: PRECISION_SINGLE COMPUTATION_SYNCH EIGEN_REAL EIGEN_COMPLEX SCALING_MANUAL SCALING_AUTO SCALING_ALWAYS SCALERS_RAW SCALERS_LOG VECTOR_NONE THREADING_NONE PROCESSOR_GPU FRAMEWORK_OPENCL


5 : Intel(R) Xeon(R) CPU           X5670  @ 2.93GHz (OpenCL 1.0 )
    Global memory (MB): 18432
    Clock speed (Ghz): 2.93
    Number of multiprocessors: 24
    Flags: PRECISION_SINGLE PRECISION_DOUBLE COMPUTATION_SYNCH EIGEN_REAL EIGEN_COMPLEX SCALING_MANUAL SCALING_AUTO SCALING_ALWAYS SCALERS_RAW SCALERS_LOG VECTOR_NONE THREADING_NONE PROCESSOR_CPU FRAMEWORK_OPENCL
```

Close the window (don't save) and start BEAST again. This time click on `Choose File...` and direct it to the XML file we generated earlier. If you installed the BEAGLE library correctly make sure the `Use BEAGLE library if available` tickbox is ticked. If this is your first time running the XML feel free to hit `Run`. If BEAST throws an error it might be because you have run BEAST on the XML previously and killed it. BEAST prevents users from accidentally writing over files that already exist unless it's a conscious decision. You can prevent these errors by either deleting the analysis files (file extensions .log and .trees) that BEAST is trying to overwrite or make sure you tick `Allow overwriting of log files` and then click `Run`. Depending on a lot of variables (but mostly whether or not you run it with BEAGLE) BEAST will take about 30 minutes to complete the run. We will leave it running[1] and prepare another XML file for the next part.

[1] If your BEAST refuses to run the XML you created download the XML I prepared: `https://www.dropbox.com/s/8reua2dpbn2vczb/Alien_analysis.xml?dl=0` and try running that.

“I have come here to chew bubblegum...” - part 1

The setting

They Live! is a satyrical science fiction horror film released in 1988 that has come to enjoy a cult following over the years. The film follows a lone protagonist, called John Nada, who discovers that aliens have been using subliminal messaging to keep people docile and occupied within an economic system that is ruining both the Earth’s climate and the wellbeing of humans. John Nada, however, discovers these plans with the help of a pair of mysterious sunglasses. Similar to the last scenario, these aliens have a cryptic virus, which is transmissible to humans. During their invasion they have been very successful at preventing the transmission of the virus to humans, since a new virus circulating amongst humans is highly conspicuous. After the discovery of the aliens amongst them and their plans, however, humans revolted, causing the containment of the cryptovirus to fail and the virus is now circulating endemically in humans. Since the date of the human revolt is not known precisely, and neither is the date of the first transmission of the cryptovirus to humans, all we can say about it is that it could not have happened earlier than 1988.

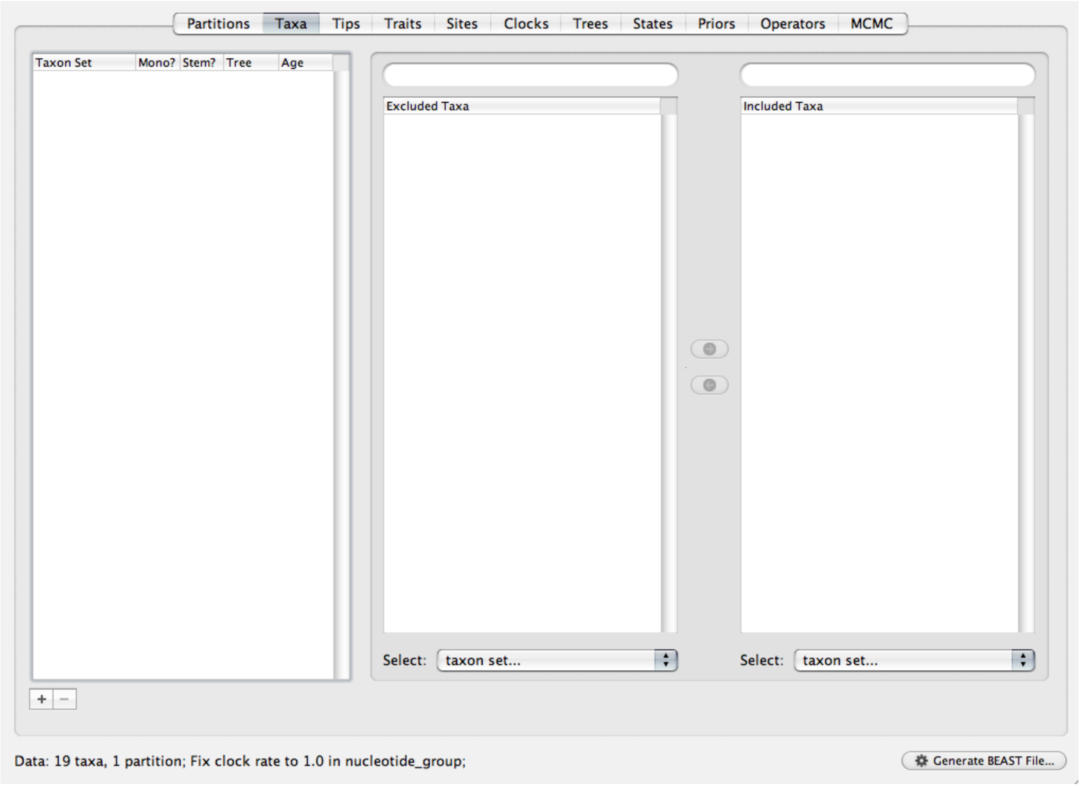
In the year 2231, following successful space exploration, the appropriate authorities have found the home planet of the alien species in They Live! and sequenced their cryptovirus. We also have sequences from the cryptovirus on Earth, where it has been transmitted between humans since the human revolt after the year 1988. Interestingly, there is one lineage of the cryptovirus that is known to have infected John Nada during his trip to the aliens’ base of operation. John Nada’s followers have managed to isolate this cryptovirus and have voluntarily passaged it between themselves all the way to the year 2231, making sure to avoid it escaping into the general population to preserve John’s legacy.

We will be working with 19 sequences, 3 (taxa A-C) from the Earth lineage, 15 (taxa D-R) from the alien lineage and 1 (taxon Y) kindly donated by the followers of John Nada.

Calibrating a phylogeny by specifying node times

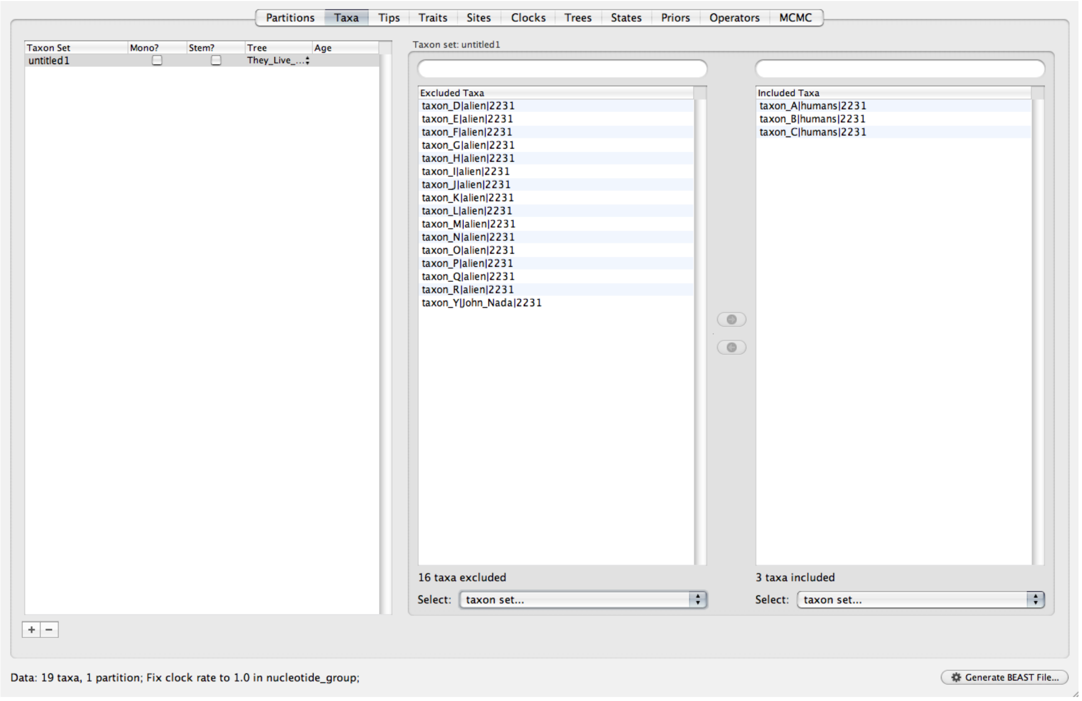
Since all our samples have been sequenced in the year 2231 we can no longer use tip dates to calibrate our phylogeny. However, we can use the timeline of the movie They Live! to calibrate our phylogeny. We (very conveniently) have a lineage that is specific to Earth, which has to have a common ancestor some time post-1988. Similarly, we know that the lineage from John Nada must have been transmitted to him over a similar timescale. Therefore all we need to do is to tell BEAST that the common ancestor of taxa A, B and C, all of which are from Earth, have a common ancestor around (but not before!) 1988. The same applies to John Nada's lineage, which has been allowed to evolve over the 243 or so years since the events that take place in They Live!. In BEAUti this is done by designating a clade by the taxa that comprise it. This is called a taxon set and we can specify additional information about, such as our prior knowledge about whether it is monophyletic (*i.e.* a clade), or whether we want to specify the date of the ancestor (*i.e.* parent node) of the set. First, let's download the alignment: `https://www.dropbox.com/s/31xlnswt56ie5zg/They_Live_alignment.fasta?dl=0`. Then, let's fire up BEAUti.

Import the sequence alignment into BEAUti. Previously we skipped the `Taxa` tab in BEAUti, but this time it is exactly where we need to go:



Data: 19 taxa, 1 partition; Fix clock rate to 1.0 in nucleotide_group;

Click on the '+' at the bottom left of the screen. A new taxon set called 'untitled1' should pop up into the panel on the far left, followed by a list of all taxa in the alignment in the central panel and an empty panel to the left. Select 'Taxon_A—humans—2231', 'Taxon_B—humans—2231' and 'Taxon_C—humans—2231' and press the green arrow pointing to the right. All 3 taxa should now appear in the right panel:



Taxon Set	Mono?	Stem?	Tree	Age
untitled1	□	□	They_Live_...	

Taxon set: untitled1

Excluded Taxa
taxon_D|alien|2231
taxon_E|alien|2231
taxon_F|alien|2231
taxon_G|alien|2231
taxon_H|alien|2231
taxon_I|alien|2231
taxon_J|alien|2231
taxon_K|alien|2231
taxon_L|alien|2231
taxon_M|alien|2231
taxon_N|alien|2231
taxon_O|alien|2231
taxon_P|alien|2231
taxon_Q|alien|2231
taxon_R|alien|2231
taxon_Y|John_Nada|2231

16 taxa excluded

Select: taxon set...

Included Taxa
taxon_A|humans|2231
taxon_B|humans|2231
taxon_C|humans|2231

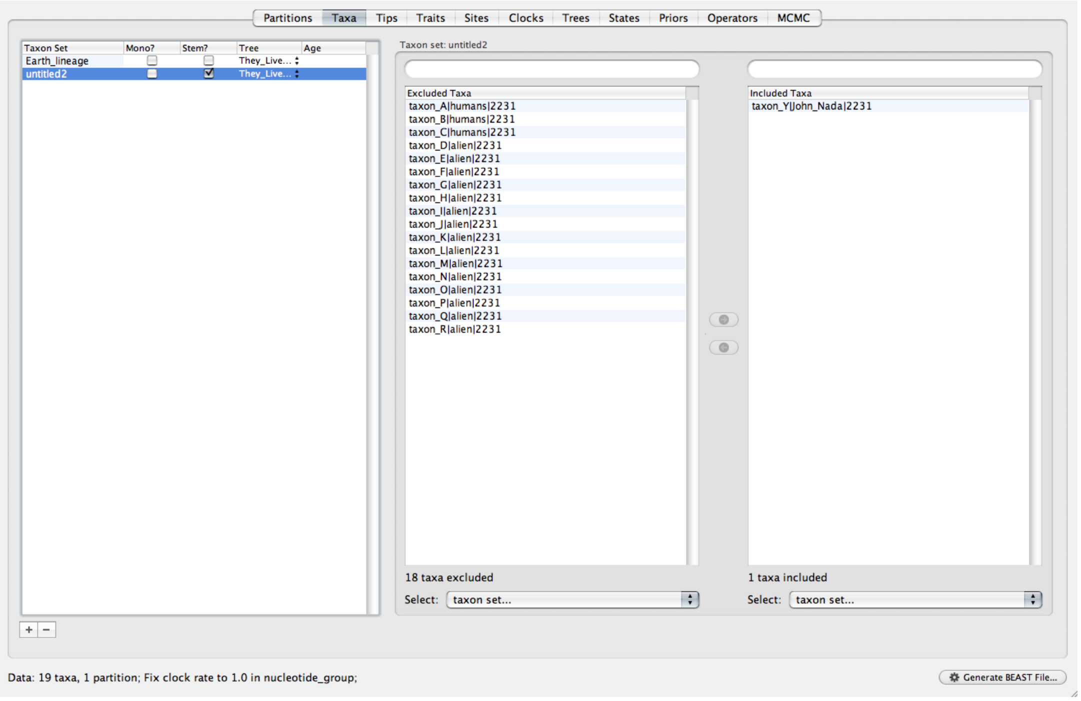
3 taxa included

Select: taxon set...

Data: 19 taxa, 1 partition; Fix clock rate to 1.0 in nucleotide_group;

We have thus specified a taxon set, which we can also define as a clade by ticking the tickbox that asks whether the taxon set is monophyletic (`Mono?`). In our case we have strong suspicions that the Earth lineage is indeed monophyletic, but to be on the safe side we will leave it as is. Specifying a taxon set that is a clade enforces the clade's monophyly throughout MCMC and could be used as a tool in some analyses. Next, double click on 'untitled1' and change the taxon set's name to something more useful, like "Earth_lineage" or whatever you will find useful to help identify it later.

Now we have to define another taxon set, so click on '+' at the bottom left again. Unfortunately we only have one sequence from John Nada's followers, so simply include 'Taxon_Y—John_Nada—2231' as the sole member of the second taxon set. Again, rename the taxon set to something more useful, like "John_Nada_lineage". A single taxon will be monophyletic, so we can leave the `'Mono?'` option unticked, but you can see that we will run into trouble if we want to calibrate the phylogeny using this tip. This is because we know that all of our sequences were collected in 2231 and that we want to calibrate the phylogeny based on when the lineage split from all other lineages. To do this we will specify the date of the 'Taxon_Y—John_Nada—2231' lineage's *ancestor*, not the lineage itself. Tick the `Stem?` tickbox, which tells BEAST that you're interested in the ancestor (stem) of this taxon set.



Taxon Set	Mono?	Stem?	Tree	Age
Earth_lineage	□	□	They_Live...	
untitled2	□	☑	They_Live...	

Taxon set: untitled2

Excluded Taxa
taxon_A|humans|2231
taxon_B|humans|2231
taxon_C|humans|2231
taxon_D|alien|2231
taxon_E|alien|2231
taxon_F|alien|2231
taxon_G|alien|2231
taxon_H|alien|2231
taxon_I|alien|2231
taxon_J|alien|2231
taxon_K|alien|2231
taxon_L|alien|2231
taxon_M|alien|2231
taxon_N|alien|2231
taxon_O|alien|2231
taxon_P|alien|2231
taxon_Q|alien|2231
taxon_R|alien|2231

18 taxa excluded

Select: taxon set...

Included Taxa
taxon_Y|John_Nada|2231

1 taxa included

Select: taxon set...

Data: 19 taxa, 1 partition; Fix clock rate to 1.0 in nucleotide_group;

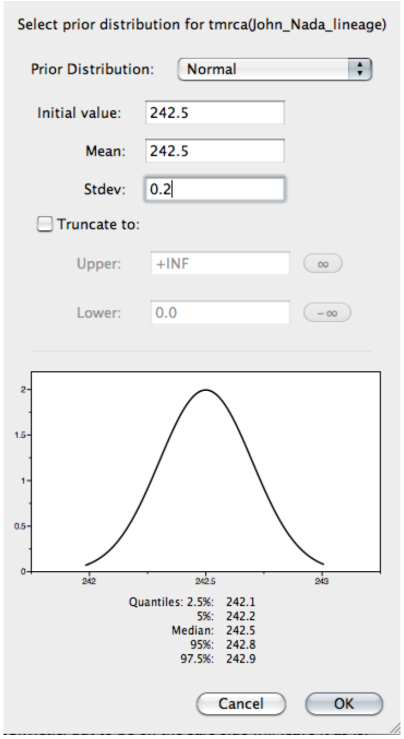
Move onto the `Sites` tab. Turns out the cryptovirus has the exact same genomic structure as the xenovirus from the previous part of the practical (how convenient). All of the genome is coding, so simply select the same SRD06 model (click on `Use SRD06 model`). Can you remember how this model partitions the data and what nucleotide substitution models it uses?

Proceed to the `Clocks` panel. For this analysis we will use an uncorrelated lognormal clock – `Lognormal relaxed clock (Uncorrelated)` (Drummond et al., 2006). This is different from the strict clock we used earlier in that it draws molecular clock rates for each branch from a distribution. This means that every branch receives its own rate over the course of the MCMC (this is what the relaxed bit means). The distribution from which we draw the rates is log-normally distributed (the lognormal part) and we sample its mean and standard deviation over the course of MCMC. Finally, the rates are allocated at random without regard for ancestor-descendent relationships between branches (that's the uncorrelated part). Relaxed clocks are recommended when there is lots of data and suspicions that some lineages might be evolving slower or faster. In real life the tree is probably too small and the use of a relaxed clock would be considered dodgy. Make sure you have ticked the `Estimate?` tickbox.

We will skip the `Trees` tab for now and continue onto the `Priors` tab. You will see that the two taxon sets we have defined earlier are here, at the very top. Specifically, it is their time of most recent common ancestor (TMRCA). If we left the default prior the TMRCAs for both taxon sets would be inferred and reported in the log file. In our case we need to give both these TMRCAs prior distributions to indicate our degree of belief of when they existed.

Let's start with the John Nada lineage (taxon set 2). Click on *Using tree prior next to tmrca(John_Nada_lineage). In the dialogue box that pops up select Normal from Prior distributions. We can now translate our knowledge of when this lineage existed by parameterising a normal distribution. John Nada lived at some point in 1988, which is 243 years before the most recent tip (actually all the tips) exists. In BEAST virtually everything is specified in units of time before the most recent tip. This includes priors.

Given that John Nada lived 243 before the most recent tip we will specify the mean of the normal distribution to be 242.5, that is we believe that the probability distribution of the lineage's ancestor is centered on mid-1988. The lineage that ended up in John Nada could have coalesced with the modern day alien lineage immediately prior to its transmission and thus could be fairly young or could have been a relatively obscure co-circulating lineage that by chance ended up in John Nada and thus is actually a lot older than the events that take place in They Live!. We will assume the former and specify the standard deviation of the distribution to be 0.2, which concentrates virtually all of the probability density on the interval between 242.9 and 242.1 before 2231 (*i.e.* 1988.1 to 1988.9, respectively):



Select prior distribution for tmrca(John_Nada_lineage)

Prior Distribution: Normal

Initial value: 242.5

Mean: 242.5

Stdev: 0.2

□ Truncate to:

Upper: +INF

Lower: 0.0

Quantiles: 2.5%: 242.1
5%: 242.2
Median: 242.5
95%: 242.8
97.5%: 242.9

Let's specify the calibration point for the Earth lineage (taxon set 1). Click on *Using tree prior next to tmrca(Earth_lineage). We will again use a normal distribution on this node's date, except this time we know that the human uprising against the aliens started some time after 1988, but definitely not before then. The mean will be centered on 242.5 again (mid-1988), but we will leave the standard deviation at the default value (1.0) to reflect our uncertainty in the timing of this event. Given our knowledge that the human uprising against aliens (and thus the transmission of the cryptovirus) started after the beginning of 1988 we will truncate our prior distribution. Click on Truncate to:, which will give you access to specify where the prior distribution should be truncated. Remember that virtually everything in BEAST is specified in units of time before the most recent tip, which includes priors. The upper and lower tails of the prior distribution are thus inverted from what you would say from intuition – the upper tail indicates how old the distribution is, whereas the lower tail indicates how recent it is. We will truncate the upper tail of the distribution, that is we will say that the node cannot be older than 243.0 years prior to the most recent tip (this corresponds to 1988.0). We don't need to truncate the lower tail of the distribution, so leave it at 0, which tells BEAST that theoretically the node could be contemporaneous with all tips (it won't be because our prior is quite strong).

Remember that we will also need to specify a rate prior. We will use the CTMC reference prior as before, since it's uninformative.

Let's go back to the `Trees` tab. Although we are happy with saying that the cryptovirus population size is constant through time we have specified two very strong priors on some nodes of the tree. When the initial tree is generated by BEAST as a starting point it will be generated via a neutral coalescent simulation. This is extremely unlikely to generate a tree where the common ancestor of taxa A, B and C exists after 1988 and where the ancestor of taxon Y exists within a narrow time period in 1988. Remember that in Bayesian statistics the posterior distribution is the product of the prior and the likelihood. If you specified the prior to be a truncated distribution with interval [t-1, t+1] you are saying that parameter values outside your distribution (*e.g.* t-2 or t+2) cannot exist. Regardless of what the likelihood says about the probability of the data given a model with a parameter value of t-2 or t+2 (which can be perfectly reasonable) if the prior says it has zero probability, then the posterior will also have zero probability (since posterior probability = prior probability × likelihood).

Similarly, computers have their limitations when it comes to floating point arithmetic. Normal distributions have fairly short tails so the probability of observing a value 6 standard deviations away from the mean will require 9 spaces after zero to be reported with minimal precision (depending on how big the standard deviation is). When probabilities are very low the computer will round them to zero.

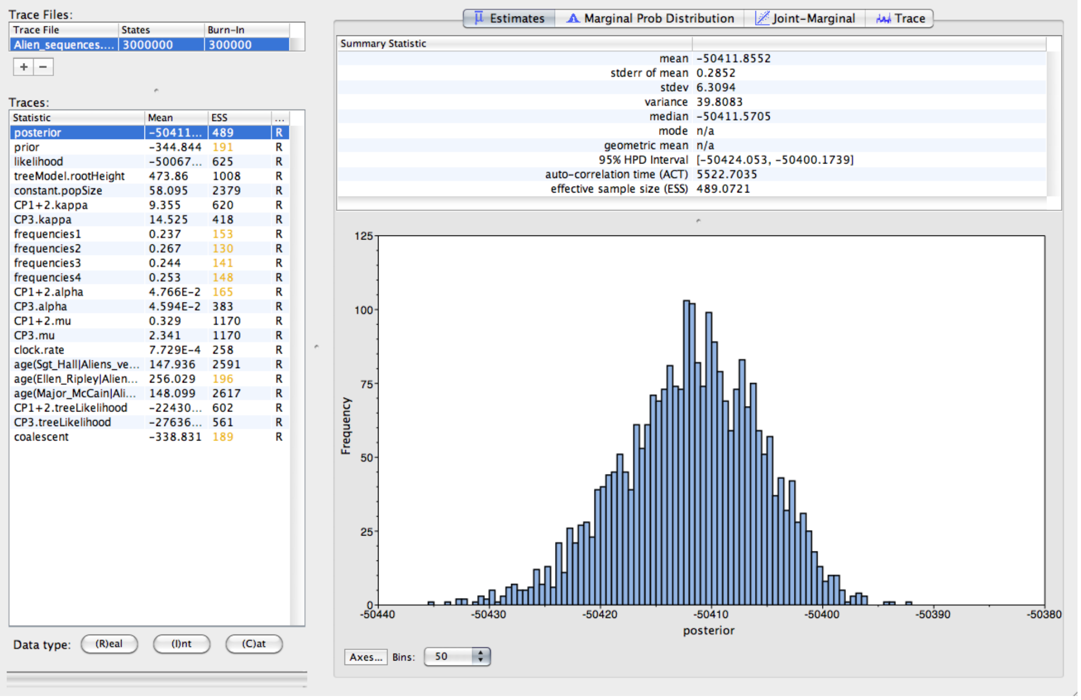
With this in mind we will tell BEAST to start the analysis from a UPGMA (Unweighted Pair Group Method with Arithmetic Mean, an old school tree construction algorithm) tree, so make sure `UPGMA starting tree` is selected in the `Trees` tab. This will ensure that the first tree in MCMC will be somewhere close to the “actual” tree (actual is in quotes here because this is Bayesian phylogenetics). A UPGMA tree is also more likely to fit within the node time constraints we've included.

Finally go to the `MCMC` tab. We will run this analysis for 4 million states, echoing the current state to screen every 10 000 generations and logging parameters every 400 states. How many samples from the posterior will that gives us?

Click on `Generate BEAST File...` to save the XML once you're done.

“Game over, man” - part 2

Once our xenovirus analyses are finished you should set the They Live!-themed cryptovirus analysis to run in BEAST[2]. We will begin the analysis of xenovirus sequences by looking at the log file using Tracer. Fire it up and import the log file (file extension either .log or .log.txt) from the Alien sequences analysis [3]. You should see something like this:



Trace Files:

Trace File	States	Burn-In
Alien_sequences....	3000000	300000

Traces:

Statistic	Mean	ESS	...
posterior	-50411...	489	R
prior	-344.844	191	R
likelihood	-50067...	625	R
treeModel.rootHeight	473.86	1008	R
constant.popSize	58.095	2379	R
CP1+2.kappa	9.355	620	R
CP3.kappa	14.525	418	R
frequencies1	0.237	153	R
frequencies2	0.267	130	R
frequencies3	0.244	141	R
frequencies4	0.253	148	R
CP1+2.alpha	4.766E-2	165	R
CP3.alpha	4.594E-2	383	R
CP1+2.mu	0.329	1170	R
CP3.mu	2.341	1170	R
clock.rate	7.729E-4	258	R
age(Sgt_Hall|Aliens_ve...	147.936	2591	R
age(Ellen_Ripley|Alien...	256.029	196	R
age(Major_McCain|Ali...	148.099	2617	R
CP1+2.treeLikelihood	-22430...	602	R
CP3.treeLikelihood	-27636...	561	R
coalescent	-338.831	189	R

Data type: (R)eal (I)nt (C)at

Summary Statistic	
mean	-50411.8552
stderr of mean	0.2852
stdev	6.3094
variance	39.8083
median	-50411.5705
mode	n/a
geometric mean	n/a
95% HPD Interval	[-50424.053, -50400.1739]
auto-correlation time (ACT)	5522.7035
effective sample size (ESS)	489.0721

The Estimates tab shows the histogram of posterior samples, *i.e.* the posterior distribution itself. Blue bins correspond to where 95% of the posterior samples are, known as the 95% highest posterior density (HPD) interval (this might not be visible for all parameters). In Bayesian statistics it rather straightforwardly says we are 95% sure that the true parameter lies within that region. If you select more than one parameter to display whilst in the Estimates tab it will show their 95% HPDs as lines to help compare HPDs between them.

The Marginal Prob Distribution tab will show the kernel density estimate of the currently selected parameter, which is essentially a smoothed histogram. When selecting multiple values it will display all the distributions as actual distributions, rather than their HPDs like it did in the Estimates tab.

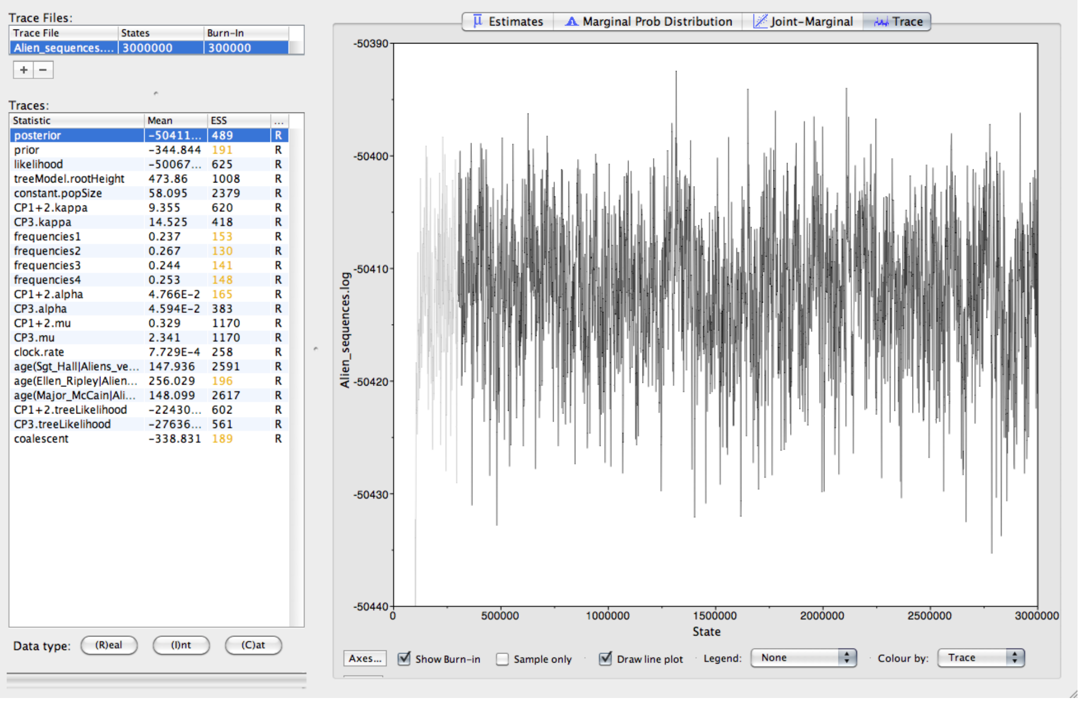
The Joint-Marginal tab will display a scatter plot when exactly two parameters are selected. It plots the value of one parameter against the other which can identify correlations that should not exist or to diagnose what causes problems during an MCMC

[2] If something went wrong and BEAST refuses to run your XML you can download one from here: `https://www.dropbox.com/s/j7toscxxx3948ur/They_Live_analysis.xml?dl=0`

[3] If your copy of BEAST didn't work download the output of the analysis I have run earlier from here: `https://www.dropbox.com/sh/2dvlb2jimxlsdag/AACB9x_eLbyMTxs2pE9Bh0yMa?dl=0`

run. Give it a try by selecting the `posterior` and `likelihood` parameters whilst in the `Joint-Marginal` tab. What do you see? Do you know what's going on here?

Move to the `Trace` tab:



Trace Files:

Trace File	States	Burn-In
Alien_sequences....	3000000	300000

Traces:

Statistic	Mean	ESS	...
posterior	-50411...	489	R
prior	-344.844	191	R
likelihood	-50067...	625	R
treeModel.rootHeight	473.86	1008	R
constant.popSize	58.095	2379	R
CP1+2.kappa	9.355	620	R
CP3.kappa	14.525	418	R
frequencies1	0.237	153	R
frequencies2	0.267	130	R
frequencies3	0.244	141	R
frequencies4	0.253	148	R
CP1+2.alpha	4.766E-2	165	R
CP3.alpha	4.594E-2	383	R
CP1+2.mu	0.329	1170	R
CP3.mu	2.341	1170	R
clock.rate	7.729E-4	258	R
age(Sgt_Hall|Aliens_ve...	147.936	2591	R
age(Ellen_Ripley|Alien...	256.029	196	R
age(Major_McCain|Ali...	148.099	2617	R
CP1+2.treeLikelihood	-22430...	602	R
CP3.treeLikelihood	-27636...	561	R
coalescent	-338.831	189	R

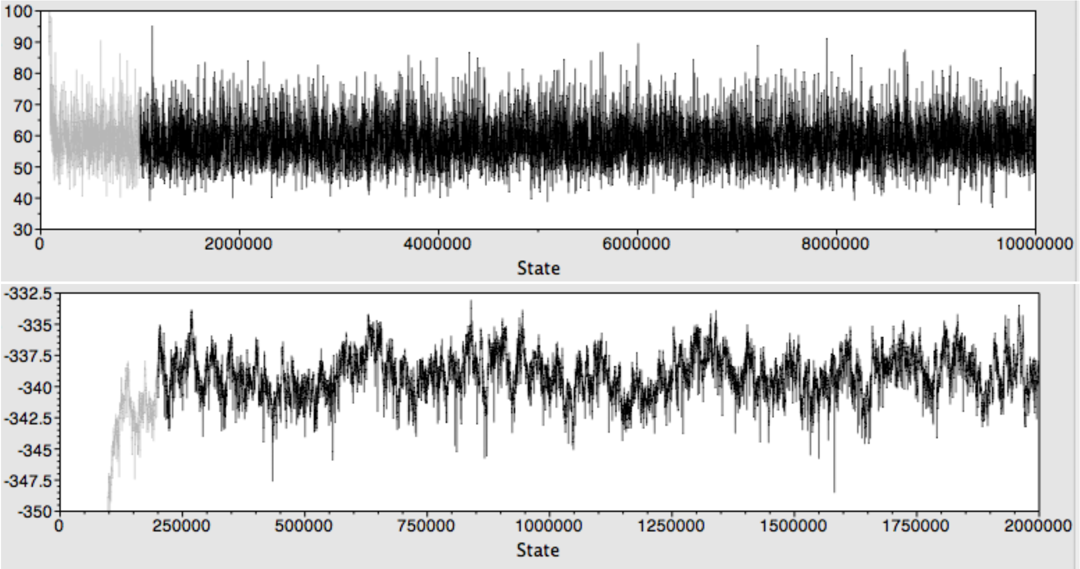
Data type: (R)eal (I)nt (C)at

☑ Show Burn-in □ Sample only ☑ Draw line plot Legend: None Colour by: Trace

This one is arguably the most important tab of all. It plots the value of the parameter that has been logged at each state of the MCMC (MCMC iteration number on the X axis, parameter value on the Y axis). The plot is a continuous line to help identify autocorrelation, which should not exist in a well-constructed MCMC chain. The `Trace` tab is also used to find the burnin (which is 10% of the MCMC chain length by default in Tracer), which is what the values of all parameters that have been sampled on the way from the starting point (which is a very unlikely state) to the posterior distribution (where the chain should have spent most of its time) are called.

Try adjusting the burnin to see how values sampled over the course of MCMC are included or thrown out. If you click through other parameters you will see that not all parameters have been sampled equally. Parameters related to the nucleotide substitution model (kappa, frequencies, alpha) will have relatively long stretches where the value has not been updated for hundreds of states. This is because the MCMC is tuned in such a way that these parameters are changed infrequently (they have lower operator weights in BEAUti), because they can be estimated fairly precisely, but are very costly to compute once changed during MCMC.

By now you must have noticed that each parameter has its mean displayed next to it, but also a parameter called "ESS". This stands for effective sample size and (essentially) tells you how quickly the MCMC trace loses correlation with itself. Highly auto-correlated MCMC chains will have traces that go up and down and have very low ESS values. This can be a problem with the data, but more often than not it is because the operators are not exploring the posterior distribution properly ("poorly mixing MCMC" in MCMC slang), *e.g.* by proposing moves that are too conservative or too liberal. A chain that is mixing well will lose auto-correlation very quickly and will have higher ESS values. Tracer displays ESS values below 200 as yellow and those below 100 as red. These cut-offs are arbitrary but useful. Think of this as trusting a mean derived from 5 values versus 500 values. Compare these two traces:



The trace at the top is of a parameter that has an ESS value of 5215, whereas the one at the bottom only has 86. There is one reliable method for solving low ESS values, which is to run the MCMC for longer or, even better, run the chain multiple times and combine output from different runs. In the example above the bottom trace looks bad because the MCMC was not long enough.

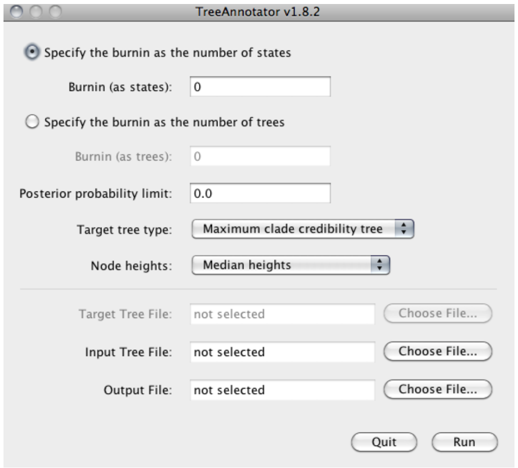
If you want you can download additional log files of our alien analysis from here: `https://www.dropbox.com/sh/2xkvwsv0ekyfx0n/AAAQGsp9FGo9ozbzNYL9x0D0a?dl=0` to see how ESSs change when you import multiple independently run MCMC chains into Tracer. Additionally you can download another chain I have run for 10 million steps instead of 3 million: `https://www.dropbox.com/s/yj64ddg3zlkub4n/Alien_sequences_10M.log.txt?dl=0`. Can you tell why running the MCMC for longer does not narrow down our estimates of all parameters?

Mixing aside, we can start looking at parameters we are actually interested and those are the estimated parameters for the evolutionary rate (`clock.rate`), root height and date estimates for the tips we were interested in. You will remember that in BEAST most things are shown as time units before the most recent tip. In the xenovirus analysis the most recent sequence was collected in 2379 August 24 (2379.6438 in decimal years). The estimated age of Ellen Ripley's sequence in my analyses is centered on 256.1463, which is equivalent to a date of 2123.4975 (95% HPD interval 2120.4774 - 2126.6766) or 2123 July 01 (2120 June 23 - 2126 September 04). This is close to, and the HPDs include, the year 2122 Your estimates should broadly, but not exactly, be the same as mine. Do you know why?

You should also recover the molecular clock rate for this analysis at roughly 7.73×10^{-4} (95% HPDs: 7.42×10^{-4}, 8.02×10^{-4}). The units of this rate are substitutions per site per unit of time. We used decimal years to calibrate our phylogeny and therefore the rate is in substitutions per site per year. Can you tell what the evolutionary rate estimate would have been if we specified the tip dates in decades since 0 C.E.? What about millenia since 0 C.E.? Millions of years?

In terms of finding the right burnin the default burnin of 300 000 (10% of 3 million) did not look too bad, so that's what we will use for the next part, which is recovering a summary phylogeny of all the trees we have sampled over the course of MCMC.

We have sampled 3 000 phylogenies during our MCMC run, but in order to glean any useful information from them we first have to summarise them. This is done using TreeAnnotator, which is distributed with BEAST. Let's fire it up:



● Specify the burnin as the number of states

Burnin (as states): 0

○ Specify the burnin as the number of trees

Burnin (as trees): 0

Posterior probability limit: 0.0

Target tree type: Maximum clade credibility tree

Node heights: Median heights

Target Tree File: not selected

Input Tree File: not selected

Output File: not selected

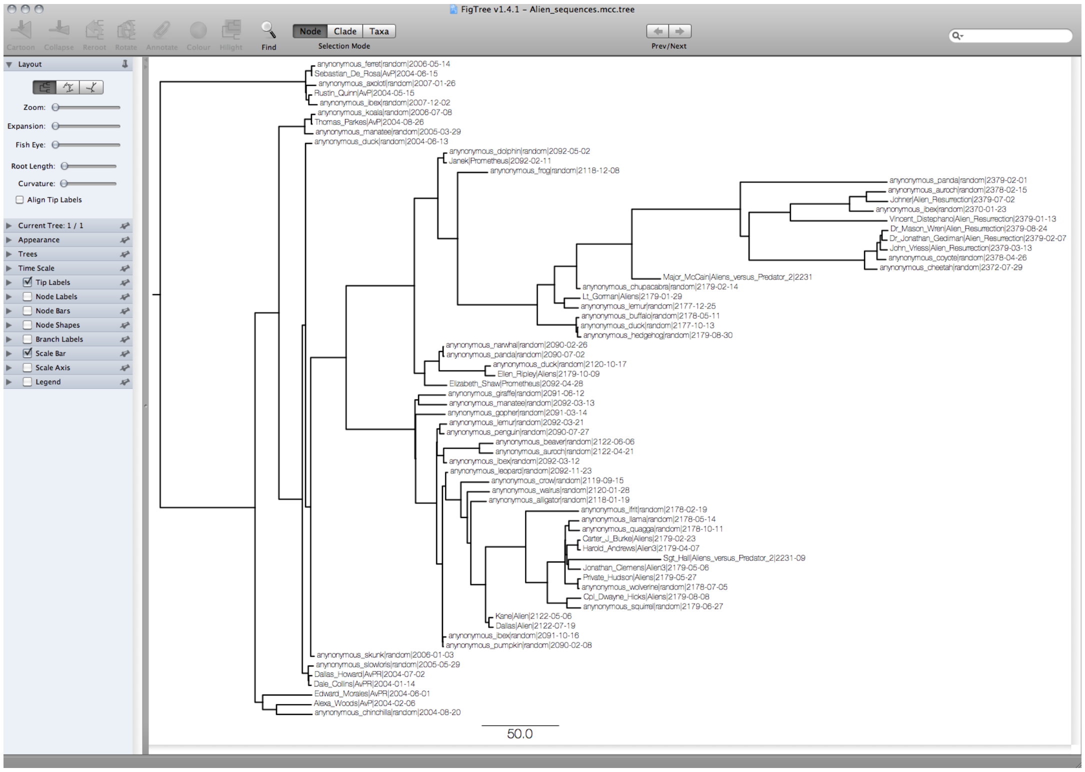
Let's begin by specifying the burnin. We have run our MCMC for 3 million states and sampled states every 1 000 states, which gave us 3 000 samples from the posterior. We also found that a burnin of 300 000 (or 10% of the MCMC) is sufficient, which leaves us with 2 700 samples from the posterior which we can use. Specify the `Burnin (as states):` as 300 000, or if you prefer `Burnin (as trees):` as 300 in TreeAnnotator (up to you). The difference is entirely down to how you count the posterior samples, either as part of a big MCMC chain or as a self-contained collection.

We will leave `Posterior probability limit` at its default value of 0.0 and the `Target tree type` as is at Maximum clade credibility tree (aka MCC tree). It will identify the single most likely phylogeny in our sample and use it as a target, onto which properties of other phylogenies in the sample will be annotated (*e.g.* frequency or posterior probability of clades, rates, branch lengths, *etc.*). This is usually the most sensible option.

Next, select 'Common Ancestor heights' from the `Node heights:` menu. This is because large trees (not applicable in our case) make the tree space difficult to explore and relatively few trees are actually sampled from all possible trees. Conversely, it means that the tree identified by TreeAnnotator as the 'best' tree is far from ideal. If we used the median or the mean heights options on large trees we would see that very frequently some nodes have negative branch lengths, because the 'best' tree has nodes that are much more recent than the rest of the trees in the posterior distribution. The 'Common Ancestor heights' option fixes this 'issue'.

Finally, under `Input Tree File` direct TreeAnnotator to your posterior set of sampled trees (file extension .trees or .trees.txt). For output, under `Output File`, create a file name that you can easily identify. In our lab the convention is to use the same name as the .trees file but with .trees replaced with mcc.tree. Once you're ready press `Run`. It will take TreeAnnotator some time to load all the trees into memory, to find the 'best' tree, to load all the attributes of interest from all the branches of all the other trees and to annotate our target tree.

We can finally have a look at our tree[4]. First, let's start FigTree, then go to `File>Open` and direct it to the MCC tree we just created.

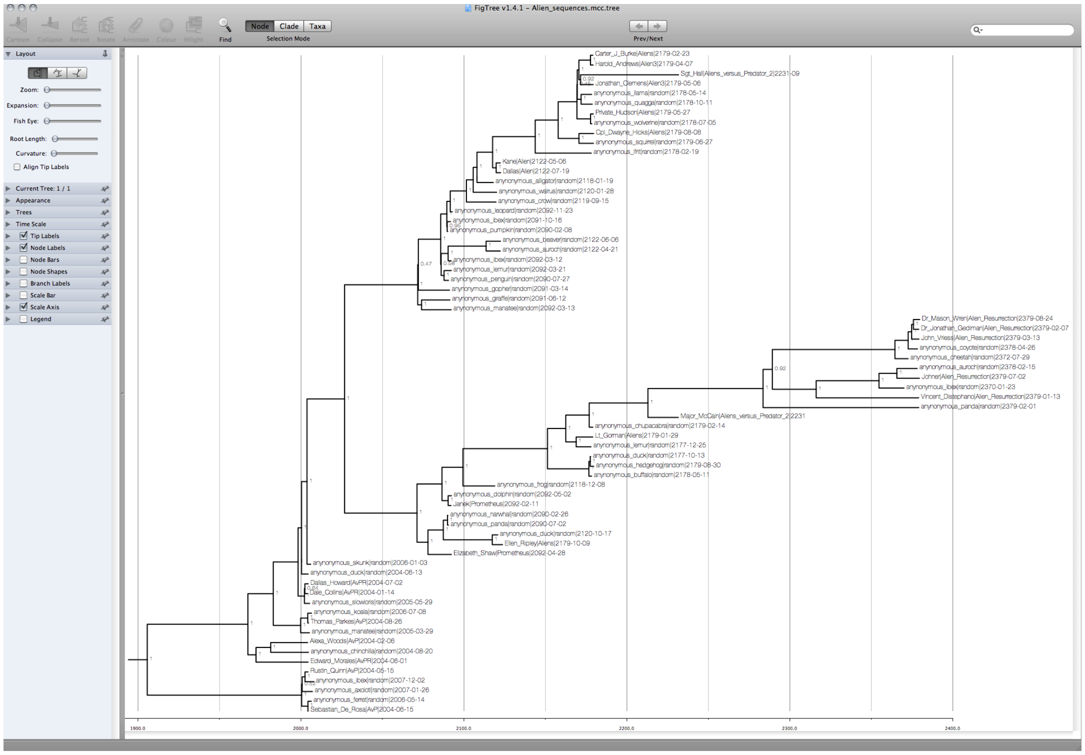

FigTree v1.4.1 – Alien_sequences.mcc.tree
Cartoon
Collapse
Reroot
Rotate
Annotate
Colour
Hilight
Find
Node
Clade
Taxa
Selection Mode
Prev/Next
Layout
Zoom:
Expansion:
Fish Eye:
Root Length:
Curvature:
□ Align Tip Labels
Current Tree: 1 / 1
Appearance
Trees
Time Scale
☑ Tip Labels
□ Node Labels
□ Node Bars
□ Node Shapes
□ Branch Labels
☑ Scale Bar
□ Scale Axis
□ Legend
anynonymous_ferret|random|2006-05-14
Sebastian_De_Rosa|AvP|2004-06-15
anynonymous_axolotl|random|2007-01-26
Rustin_Quinn|AvP|2004-05-15
anynonymous_ibex|random|2007-12-02
anynonymous_koala|random|2006-07-08
Thomas_Parkes|AvP|2004-08-26
anynonymous_manatee|random|2005-03-29
anynonymous_duck|random|2004-06-13
anynonymous_dolphin|random|2092-05-02
Janek|Prometheus|2092-02-11
anynonymous_frog|random|2118-12-08
anynonymous_panda|random|2379-02-01
anynonymous_auroch|random|2378-02-15
Johner|Alien_Resurrection|2379-07-02
anynonymous_ibex|random|2370-01-23
Vincent_Distephano|Alien_Resurrection|2379-01-13
Dr_Mason_Wren|Alien_Resurrection|2379-08-24
Dr_Jonathan_Gediman|Alien_Resurrection|2379-02-07
John_Vriess|Alien_Resurrection|2379-03-13
anynonymous_coyote|random|2378-04-26
anynonymous_cheetah|random|2372-07-29
Major_McCain|Aliens_versus_Predator_2|2231
anynonymous_chupacabra|random|2179-02-14
Lt_Gorman|Aliens|2179-01-29
anynonymous_lemur|random|2177-12-25
anynonymous_buffalo|random|2178-05-11
anynonymous_duck|random|2177-10-13
anynonymous_hedgehog|random|2179-08-30
anynonymous_narwhal|random|2090-02-26
anynonymous_panda|random|2090-07-02
anynonymous_duck|random|2120-10-17
Ellen_Ripley|Aliens|2179-10-09
Elizabeth_Shaw|Prometheus|2092-04-28
anynonymous_giraffe|random|2091-06-12
anynonymous_manatee|random|2092-03-13
anynonymous_gopher|random|2091-03-14
anynonymous_lemur|random|2092-03-21
anynonymous_penguin|random|2090-07-27
anynonymous_beaver|random|2122-06-06
anynonymous_auroch|random|2122-04-21
anynonymous_ibex|random|2092-03-12
anynonymous_leopard|random|2092-11-23
anynonymous_crow|random|2119-09-15
anynonymous_walrus|random|2120-01-28
anynonymous_alligator|random|2118-01-19
anynonymous_ifrit|random|2178-02-19
anynonymous_llama|random|2178-05-14
anynonymous_quagga|random|2178-10-11
Carter_J_Burke|Aliens|2179-02-23
Harold_Andrews|Alien3|2179-04-07
Sgt_Hall|Aliens_versus_Predator_2|2231-09
Jonathan_Clemens|Alien3|2179-05-06
Private_Hudson|Aliens|2179-05-27
anynonymous_wolverine|random|2178-07-05
Cpl_Dwayne_Hicks|Aliens|2179-08-08
anynonymous_squirrel|random|2179-06-27
Kane|Alien|2122-05-06
Dallas|Alien|2122-07-19
anynonymous_ibex|random|2091-10-16
anynonymous_pumpkin|random|2090-02-08
anynonymous_skunk|random|2006-01-03
anynonymous_slowloris|random|2005-05-29
Dallas_Howard|AvPR|2004-07-02
Dale_Collins|AvPR|2004-01-14
Edward_Morales|AvPR|2004-06-01
Alexa_Woods|AvP|2004-02-06
anynonymous_chinchilla|random|2004-08-20
50.0

It looks a bit rough, but we can see more by sorting the tree topology (`Trees>Order nodes`). At this point you might wonder what the outgroup for this analysis is. One of the advantages of molecular clock analyses is that the root is identified as the oldest common ancestor, so the rooted tree you see in FigTree is already rooted. We cannot see *when* the root is exactly, so next we will put our tree on a proper time scale. Click on `Node labels` and expand (click on the triangle) the menu. This allows us to display various statistics we might be interested in at each node of the tree. Try looking at a few of these, especially by looking at the label called 'posterior', which will display the posterior probability of every clade in the tree. If you look at the 'Node ages' option you will see that nodes deep in the tree have bigger numbers than nodes that are more recent. This is because FigTree has only been told the depth of each node in relation to the most recent tip.

To put everything in more interpretable units we have to go to `Timescale` and set `Offset by` to 2379.6438, which is the decimal date of our most recent tip. However, the dates at the nodes are still increasing backwards in time, since node ages are still considered to be specified as going backwards in time (*i.e.* if the offset is 400 years ago, then the node ages are considered as being older than that and therefore the number is higher). Set `Scale factor` to -1.0 instead of 1.0 to specify that time is going forwards, such that all nodes dates are before the date of the most recent tip.

[4] If TreeAnnotator is misbehaving you can download an MCC tree I prepared earlier: `https://www.dropbox.com/s/419543r92ukfhp0/Alien_sequences.mcc.tree?dl=0`

We can make the tree even more legible by turning `Scale Axis` on (make sure you also tick `Reverse axis`) and turning off `Scale Bar`, since the timescale axis is now doing the same job. To finish, set `Node labels` to show the posterior probability. You should have something like this now:



FigTree v1.4.1 – Alien_sequences.mcc.tree

▼ Layout

Zoom:
Expansion:
Fish Eye:
Root Length:
Curvature:
☐ Align Tip Labels

▶ Current Tree: 1 / 1
▶ Appearance
▶ Trees
▶ Time Scale
▶ ☑ Tip Labels
▶ ☑ Node Labels
▶ ☐ Node Bars
▶ ☐ Node Shapes
▶ ☐ Branch Labels
▶ ☐ Scale Bar
▶ ☑ Scale Axis
▶ ☐ Legend

Cartoon
Collapse
Reroot
Rotate
Annotate
Colour
Hilight
Find
Node
Clade
Taxa
Selection Mode
Prev/Next
Carter_J_Burke|Aliens|2179-02-23
Harold_Andrews|Alien3|2179-04-07
Sgt_Hall|Aliens_versus_Predator_2|2231-09
Jonathan_Clemens|Alien3|2179-05-06
anynonymous_llama|random|2178-05-14
anynonymous_quagga|random|2178-10-11
Private_Hudson|Aliens|2179-05-27
anynonymous_wolverine|random|2178-07-05
Cpl_Dwayne_Hicks|Aliens|2179-08-08
anynonymous_squirrel|random|2179-06-27
anynonymous_ifrit|random|2178-02-19
Kane|Alien|2122-05-06
Dallas|Alien|2122-07-19
anynonymous_alligator|random|2118-01-19
anynonymous_walrus|random|2120-01-28
anynonymous_crow|random|2119-09-15
anynonymous_leopard|random|2092-11-23
anynonymous_ibex|random|2091-10-16
anynonymous_pumpkin|random|2090-02-08
anynonymous_beaver|random|2122-06-06
anynonymous_auroch|random|2122-04-21
anynonymous_ibex|random|2092-03-12
anynonymous_lemur|random|2092-03-21
anynonymous_penguin|random|2090-07-27
anynonymous_gopher|random|2091-03-14
anynonymous_giraffe|random|2091-06-12
anynonymous_manatee|random|2092-03-13
Dr_Mason_Wren|Alien_Resurrection|2379-08-24
Dr_Jonathan_Gediman|Alien_Resurrection|2379-02-07
John_Vriess|Alien_Resurrection|2379-03-13
anynonymous_coyote|random|2378-04-26
anynonymous_cheetah|random|2372-07-29
anynonymous_auroch|random|2378-02-15
Johner|Alien_Resurrection|2379-07-02
anynonymous_ibex|random|2370-01-23
Vincent_Distephano|Alien_Resurrection|2379-01-13
anynonymous_panda|random|2379-02-01
Major_McCain|Aliens_versus_Predator_2|2231
anynonymous_chupacabra|random|2179-02-14
Lt_Gorman|Aliens|2179-01-29
anynonymous_lemur|random|2177-12-25
anynonymous_duck|random|2177-10-13
anynonymous_hedgehog|random|2179-08-30
anynonymous_buffalo|random|2178-05-11
anynonymous_frog|random|2118-12-08
anynonymous_dolphin|random|2092-05-02
Janek|Prometheus|2092-02-11
anynonymous_narwhal|random|2090-02-26
anynonymous_panda|random|2090-07-02
anynonymous_duck|random|2120-10-17
Ellen_Ripley|Aliens|2179-10-09
Elizabeth_Shaw|Prometheus|2092-04-28
anynonymous_skunk|random|2006-01-03
anynonymous_duck|random|2004-06-13
Dallas_Howard|AvPR|2004-07-02
Dale_Collins|AvPR|2004-01-14
anynonymous_slowloris|random|2005-05-29
anynonymous_koala|random|2006-07-08
Thomas_Parkes|AvP|2004-08-26
anynonymous_manatee|random|2005-03-29
Alexa_Woods|AvP|2004-02-06
anynonymous_chinchilla|random|2004-08-20
Edward_Morales|AvPR|2004-06-01
Rustin_Quinn|AvP|2004-05-15
anynonymous_ibex|random|2007-12-02
anynonymous_axolotl|random|2007-01-26
anynonymous_ferret|random|2006-05-14
Sebastian_De_Rosa|AvP|2004-06-15
0.92
0.47
0.58
0.95
0.64
1900.0
2000.0
2100.0
2200.0
2300.0
2400.0

“I have come here to chew bubblegum...” - part 2

We will apply the same set of methods to our analysis of cryptovirus sequences. Start with inspecting the log files[5] in Tracer. You will see that in this log file there are parameters that were not in the previous analysis, namely:

- tmrca() statistics, which are the posterior TMRCA heights (*i.e.* years before the most recent tip).
- ucld.mean and ucld.stdev are the posterior distributions of the mean and standard deviation of the lognormal distribution from which branch rates were drawn.
- meanRate is the sum of substitutions divided by the sum of branch lengths and is often the statistic of actual interest, rather than ucld.mean.
- coefficientOfVariation is ucld.mean divided by ucld.stdev and indicates the dispersion of the rate distribution.
- covariance is the correlation between parent and descendent branch rates.

What is the appropriate burnin for this analysis? What is the evolutionary rate estimate that you have recovered? What is the estimated date for the ‘Earth’ lineage of cryptovirus? What is the estimated date of the ancestor of ‘John Nada’s’ lineage?

Using the burnin you just found make an MCC tree in TreeAnnotator. After you are done open the tree in FigTree and place it on a proper timescale. Can you identify the nodes we calibrated? Which branches have the highest rates? Which have the lowest? How old is the root of the tree?

As before, feel free to investigate the behaviour of MCMC by downloading independently run MCMC chains of this analysis from here: `https://www.dropbox.com/sh/rewo2hdy95c65xu/AACjTFWITrYlcE93TQ1kXiCia?dl=0` and a chain that has been run for 10 million states instead of 4 million: `https://www.dropbox.com/s/ovfh8vyneqithty/They_Live_analysis_10M.log.txt?dl=0`. If things went wrong along the way you can download the MCC tree from here: `https://www.dropbox.com/s/hvspbnrishl65um/They_Live_analysis.mcc.tree?dl=0`.

[5] If BEAST refused to run you can download the output of an analysis I did from here: `https://www.dropbox.com/sh/bpa7e46jw81ghkb/AACbpWjtEW-bk9PWjfxMtJuRa?dl=0`.

Congratulations!

You now know the basics of using BEAST.

Summary of analyses

Data	Alien xenovirus	They Live! cryptovirus
Calibration type	tip dates (also inference of unknown dates)	2 nodes
Nucleotide substitution model	SRD06 (1st+2nd codon positions under HKY+Γ_4, 3rd codon positions under HKY+Γ_4)	SRD06 (1st+2nd codon positions under HKY+Γ_4, 3rd codon positions under HKY+Γ_4)
Clock type	Strict	Relaxed uncorrelated lognormal
Demographic model (tree prior)	Constant population size	Constant population size
Priors set	CTMC reference prior on rate	Normal distribution priors on clade dates, CTMC reference prior on rate
Number of sequences	68	19
MCMC chain length	3 million	4 million
MCMC sampling frequency	1 000	400
Number of samples from the posterior distribution	3 000	10 000

References

Ayres DL, Darling A, Zwickl DJ, et al. (12 co-authors). 2012. BEAGLE: An Application Programming Interface and High-Performance Computing Library for Statistical Phylogenetics. Systematic Biology. 61:170–173.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4:e88.

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Molecular Biology and Evolution. 29.

Ferreira MAR, Suchard MA. 2008. Bayesian analysis of elapsed times in continuous-time markov chains. Canadian Journal of Statistics. 36:355–368.