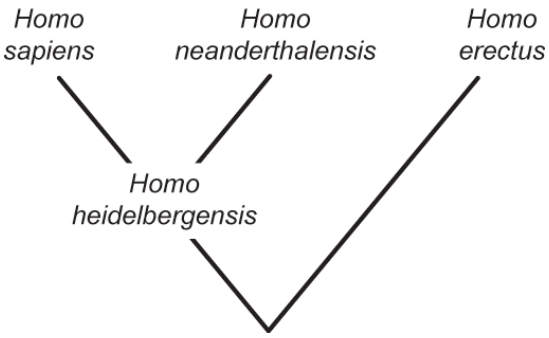
Practical 2.1: Denisovan Hominin

Bayesian Analysis and Molecular Dating Using *BEAST*

In 2008, a phalanx (finger bone) from an unidentified hominin was excavated in Denisova Cave, in the Altai Mountains of Siberia. The cave had already yielded evidence of episodic human occupation stretching back to >125,000 yr ago.

Using radiocarbon dating, the age of the new specimen was estimated at 30,000 to 48,000 yr. This means that the individual existed at a time when Modern Humans (*Homo sapiens*) and Neanderthals (*Homo neanderthalensis*) lived together across Eurasia, before the extinction of Neanderthals about 40,000 yr ago.

Prior to the appearance of Modern Humans and Neanderthals, there was another human species that was widespread across Eurasia: *Homo erectus*. There is evidence that *Homo erectus* migrated out of Africa about 1.9 million yr ago, possibly surviving in Indonesia up to 100,000 yr ago. In contrast, genetic and archaeological evidence suggests that Modern Humans expanded out of Africa ~50,000 yr ago to colonise Eurasia.



The current view of the hominin phylogeny is that Modern Humans and Neanderthals are sister species, having shared an ancestor, possibly *Homo heidelbergensis*, about half a million years ago. *Homo erectus* is a more distant relative. About 6–7 million yr ago, the lineage leading to all of these *Homo* species diverged from the lineage leading to the two chimpanzees (*Pan troglodytes* and *Pan paniscus*). Together, these species form a group called “Hominini”.

Researchers from the Max Planck Institute for Evolutionary Anthropology, Leipzig, sequenced the mitochondrial genome of the Denisovan phalanx. Sequencing DNA from ancient hominins is a major undertaking. The DNA is highly degraded because of post-mortem damage, and the low concentration of authentic endogenous DNA means that contamination from other sources is a serious risk. However, these challenges have now largely been overcome using high-throughput sequencing techniques.

In this practical, you will investigate the Denisovan hominin by performing a Bayesian phylogenetic analysis. The analysis will allow you to elucidate the relationship of the mysterious individual to other hominins and to estimate the evolutionary timescale.

Section A: Bayesian phylogenetic analysis of hominin relationships

Before you begin, check that you have recent versions of the following software:

- *BEAST 2* package (including *BEAUti*, *BEAST*, and *TreeAnnotator*)
- *Tracer*
- *FigTree*

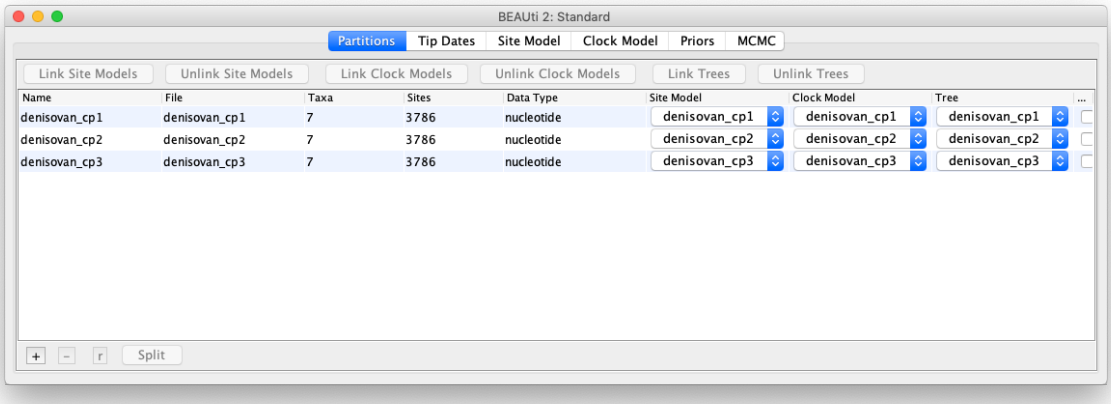
The three data files for this section of the practical exercise are **denisovan_cp1.nex**, **denisovan_cp2.nex**, and **denisovan_cp3.nex**. Each of these files contains a DNA sequence alignment in 'Nexus' format, which is a type of text format that is used by a range of phylogenetics software. The sequence alignments are the 1st, 2nd, and 3rd codon positions of the 13 mitochondrial protein-coding genes of 7 hominids: (i) Denisovan hominin; (ii) Neanderthal (*Homo neanderthalensis*); (iii) Modern Human (*Homo sapiens*); (iv) Common Chimpanzee (*Pan troglodytes*); (v) Pygmy Chimpanzee (*Pan paniscus*); (vi) Gorilla (*Gorilla gorilla*); and (vii) Orangutan (*Pongo pygmaeus*).

This exercise will use the Bayesian phylogenetic software *BEAST 2*. The program is quite complex and requires a detailed input file in XML format. However, these input files can be readily created with the user-friendly program *BEAUti*. There are four parts to the analysis:

- Creating an input file using *BEAUti*
- Bayesian phylogenetic analysis using *BEAST*
- Allowing rate variation across sites
- Processing the output using *Tracer*, *TreeAnnotator*, and *FigTree*

Creating an input file using *BEAUti*

- Open the program *BEAUti*. The purpose of this software is to create a working input file for *BEAST*. The first step is to load the sequence data into the program. Select "Import Alignment" from the "File" menu and open the 3 alignment files, **denisovan_cp1.nex**, **denisovan_cp2.nex**, and **denisovan_cp3.nex**. Alternatively, you can drag and drop the data files into the *BEAUti* window.
- You should now be in the **Partitions** tab of *BEAUti*. The window will display some of the characteristics of the data that you have loaded. You can see that each alignment contains 7 taxa and has 3786 aligned nucleotides.



There are various options in this window relating to data partitioning. Partitioning allows us to apply separate evolutionary models to different parts of the sequence alignment.

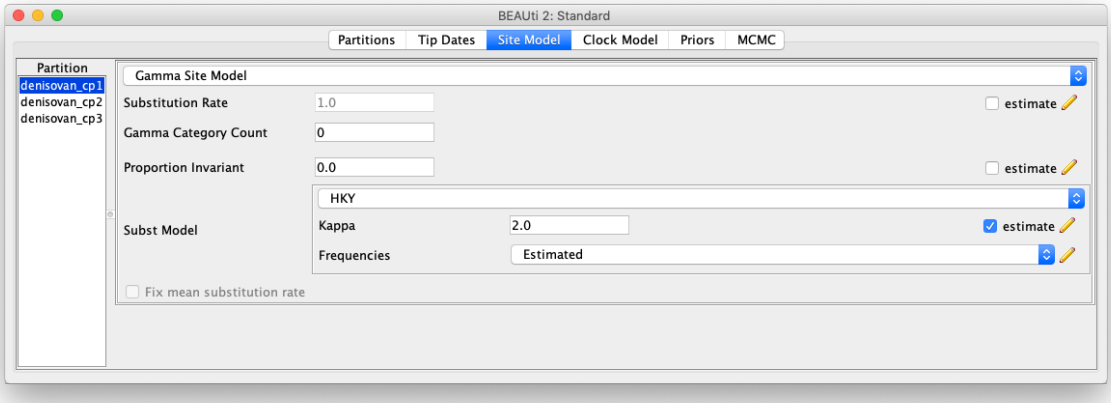
By default, *BEAUti* has assigned separate substitution (site) models, clock models, and trees to each of the 3 codon positions. This is indicated by the 3 different names that are listed under each of the corresponding columns. For this analysis we do indeed want separate substitution models and clock models for the 3 codon positions. However, we can safely assume that the 3 codon positions all share the same phylogenetic tree, because they are all linked in the non-recombining mitochondrial genome.

Select the 3 sequence alignments in this window and click on the button “Link Trees”. You can now see that the trees for the 3 codon positions all have the same name “denisovan_cp1”.

- Skip the **Tip Dates** tab. Normally, this section would allow us to include the ages of the Denisovan hominin (30,000–48,000 years) and Neanderthal (38,790 years) in the analysis. For computational reasons, however, we will skip this step. This is also partly because the ages of the sequences are negligible compared with the overall timeframe spanned by the tree.
- Go to the **Site Model** tab. Here we choose the nucleotide substitution model. In the current analysis, we shall use the HKY model of nucleotide substitution for each of the 3 codon positions. The HKY model allows transitions (A↔G and C↔T) to have a different rate from transversions (A↔C, A↔T, G↔C and G↔T). To specify this model, select “HKY” in the “Subst Model” box. Leave the default starting value of 2.0 for “Kappa”, which represents the ratio of transitions to transversions. This parameter will be estimated in the analysis. Leave the “Frequencies” as “Estimated”, which means that we will be estimating the frequencies of the four nucleotides.

  Now select all of the data partitions in the window on the left of the tab. You will now see an option to “Clone from denisovan_cp1”, which means that *BEAUti* is giving us the option of copying the substitution model across to the other 2 codon positions.

  Click “OK” to assign HKY models to the remaining 2 codon positions. Now each of the 3 codon positions has its own HKY substitution model.



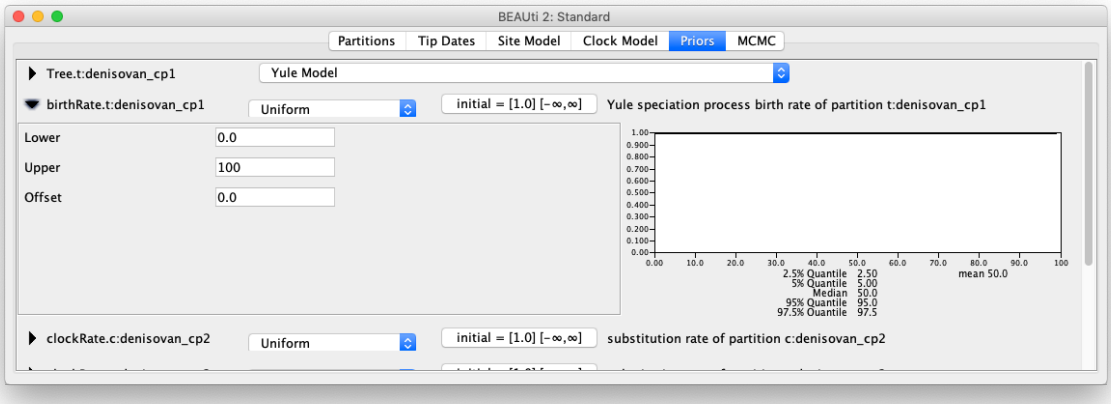
- Go to the **Clock Model** tab. Here we need to choose the type of molecular clock that we want to use in our analysis. Note that even though we are not estimating the timescale in this particular analysis, *BEAST* requires a clock model to be chosen. *BEAST* only infers rooted trees. Here we will use a "strict clock" model for each of the 3 codon positions. For each codon position, leave the "Clock.rate" at the default value of 1.0.

  The clock model that we have set up here is a separate strict clock for each of the 3 codon positions. This allows each codon position to evolve at its own rate. Specifically, the substitution rate of the 1st codon position is fixed to a value of 1.0, but the relative rates at the 2nd and 3rd codon positions are estimated in the analysis.

- Go to the **Priors** tab. Here we need to choose the prior distribution for each of the parameters in our analysis, including the tree. In the drop-down menu to the right of "Tree.t:denisovan_cp1", there are various models that can be used to generate a prior distribution for the tree. In the current analysis, we are dealing with sequences from different species, which means that we need to use one of the speciation models. The "Coalescent" models are only appropriate for population-level analyses. Choose the simplest speciation model, which is the Yule process. This is a pure-speciation model in which all lineages have an equal chance of splitting into two descendent lineages.

  If you click on the black arrow to the left of "Tree.t:denisovan", you will see the options for the Yule model. Leave the starting value for the birth rate at the default value of 1.0.

  Have a look at the other priors in this tab. There is a uniform prior for the birth rate (=speciation rate) of the Yule model, as well as a lognormal prior for the kappa parameter (the ratio of transitions to transversions in the HKY model) for each of the 3 codon positions. We should change the uniform prior for the birth rate so that its upper bound is finite. Click on the black triangle next to "birthrate.t:denisovan" and change the value of "Upper" to 100.



- Go to the **MCMC** tab. Here we need to specify how long we want to spend on drawing samples from the posterior distribution using Markov chain Monte Carlo (MCMC) simulation. Remember that we want to estimate the posterior distributions of the parameters and the tree. However, these cannot be obtained directly. Instead, we can draw samples from the posterior distribution using an appropriately designed MCMC simulation. By plotting these samples, we can gain an approximation of the posterior distribution. To keep the analysis fairly short, change the "Chain Length" to 5,000,000. Click on the black triangle next to "tracelog" and change the "File name" of the log file to **denisovan_hky.log**. Now click the black triangle next to "treelog.t:denisovan_cp1" and change the "File name" of the trees file to **denisovan_hky.trees**.

- Now go to the File menu and select "Save As". Save the file as **denisovan_hky.xml** on your computer's desktop or your current working directory (wherever you want *BEAST* to write the output files. This should produce a file in XML format, which can be read as an input file for *BEAST*. Keep *BEAUti* open because we will want to change some settings later.

Bayesian phylogenetic analysis using *BEAST*

- Open the program *BEAST* and choose the XML file that you created above. Click on the "Run" button in the bottom right of the window.
- While the analysis is in progress, *BEAST* will continually write to three files. The .log file contains samples from the posterior distribution of model parameters, while the .trees file contains samples from the posterior distribution of trees. The .state file records the current MCMC state in case you would like to resume sampling from this state.
- The analysis will take a few minutes, depending on the speed of your computer. While you are waiting, proceed to the next part below.

Allowing rate variation across sites

Now we will use a more complex substitution model, in which we allow the evolutionary rate to vary across sites in the sequence alignment.

Q. *Consider a data set that has evolved with considerable rate variation across sites. What are the potential consequences of failing to account for this?*

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

- Go back to *BEAUti* and click on the **Site Model** tab. Next to "Gamma Category Count", enter a value of 4. This changes the substitution model to the HKY+G model. The "+G" part of the name of the model means that we are allowing different sites in the alignment to have different rates, and that we are assuming that these rates follow a gamma distribution. In practice, we are using a discrete gamma distribution with 4 rate categories here.

  If you have already closed *BEAUti*, go back to the first part of this exercise and set everything up in the same way (except for the site model).

Now select all of the data partitions in the window on the left of the tab. Copy the substitution model across to the remaining 2 codon positions. Now each of the 3 codon positions has its own HKY+G substitution model.

- Go to the **MCMC** tab and change the "File name" for the trace log to **denisovan_hkyg.log**. Now click the black triangle next to "treelog.t:denisovan" and change the "File name" of the trees file to **denisovan_hkyg.trees**.
- Go to the File menu and select "Save As". Save this file as **denisovan_hkyg.xml**.
- Run *BEAST* using the new input file.

While you are waiting for your analysis to finish, try answering these questions about Bayesian phylogenetics.

Q. *How do we choose the prior distribution for each parameter in the analysis?*

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

Q. *Would it be appropriate to use estimates from our data set to inform our choice of prior distributions? Why or why not?*

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

Q. *Given that it is not possible to obtain the posterior distribution directly, what method can we use to estimate the posterior distribution?*

. . . . . .

. . . . . .

. . . . . .

Processing the output using *Tracer*, *TreeAnnotator*, and *FigTree*

- Open the program *Tracer*, click on "Import Trace File" in the "File" menu, and import the .log files from your two *BEAST* analyses.
- You can inspect the characteristics of the posterior distributions of parameters. The first thing to check is that the effective sample sizes (ESSs) of all of our sampled parameters are greater than 200. This indicates that we have drawn enough samples to be able to produce a reliable estimate of the posterior distribution of each parameter. The effective sample size is smaller than the actual number of samples because the samples drawn from the MCMC are not entirely independent of each other. If any ESS values are below ~200, it means that we need to run the MCMC analysis for a larger number of steps. If this is the case, ignore it for the purposes of this practical exercise.
- In addition, we want to draw our samples only from the stationary distribution. For this reason, we normally discard the first ~10% of samples. This is known as the 'burn-in' phase. By default, *Tracer* excludes the first 10% of your samples when calculating the mean and other statistics. Typically we would want to run the analysis multiple times to check for consistency between runs, but we will ignore this for the purposes of this practical exercise.
- Let us have a look at some of the parameter estimates for the analysis using the HKY+G model. Specifically, look at the parameters "clockRate.2" and "clockRate.3". These represent the relative substitution rates at codon positions 2 and 3, respectively, compared with codon position 1 (which has a rate of 1.0).

Q. *What are the relative substitution rates at codon positions 2 and 3? Is this consistent with our expectations of the molecular evolutionary process, given what we know about how codons encode amino acids?*

. . . . . .

. . . . . .

. . . . . .

. . . . . .

- Now we want to compare the fit of the two substitution models to our data. This can be done using the Bayes factor, which is a comparison of the marginal likelihoods of the two models. To do this properly, we would need to run an additional analysis (e.g., stepping-stone sampling). However, we can get a very rough idea of the relative fit of the two models by looking at their likelihoods in *Tracer*. Select both of the files in the top-left box of *Tracer*. In the "Traces" window in the bottom left, select "likelihood".
- You will see that there is a huge difference in likelihoods between the two models. We are doing this by visual inspection here, but keep in mind that this is not normally recommended! In practice, we should estimate the marginal likelihoods using a computationally intensive method such as stepping-stone sampling.

Q. *Which model has the higher likelihood?*

. . . . . .

. . . . . .

- Open the program *TreeAnnotator*. This program is used to process the .trees file from *BEAST*. It reads all of the sampled trees and summarises the information in the form of a single tree.
- In the box next to "Burnin percentage", enter the value "10". This means that we are discarding the first 10% of our samples as burn-in. For the "Input Tree File", click "Choose File" and select the .trees file produced by the *BEAST* analysis using the HKY+G model. For the "Output File", click "Choose File" and select the directory where you want to save the output file from *TreeAnnotator*. Give the output file the name **denisovan_hkyg.tre** and click "Run".
- Open the program *FigTree* and use it to view the file **denisovan_hkyg.tre** produced by *TreeAnnotator* in the previous step. The summary tree for your Bayesian phylogenetic analysis will be displayed. You can play around with the settings and *FigTree* will display some of the information associated with the tree. Try clicking on the three symbols in the "Layout" menu. From left to right, these allow you to display the tree as a rooted tree, circular tree, and unrooted tree, respectively.

  We are mainly interested in two features of the tree. First, we want to see where the Denisovan hominin has been placed. Check the box next to "Node Labels" and select "posterior" in the drop-down menu next to "Display". This will label the nodes of the tree with posterior probabilities, which indicate the support for each of the groupings represented in the tree.

Q. *Where has the Denisovan hominin been placed in the phylogenetic tree? What is the posterior probability of the grouping of the Denisovan hominin with the other two humans?*

. . . . . .

. . . . . .

. . . . . .

Q. *Is the Denisovan hominin a Modern Human, a Neanderthal, or neither?*

. . . . . .

. . . . . .

. . . . . .

. . . . . .

If you have spare time, you might want to try answering these more challenging questions about MCMC analysis.

Q. *When drawing samples during the MCMC analysis, we chose to log parameters every 1000 steps. Why did we not want to log the parameters at every single step?*

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

Q. *What can we do to reduce the number of steps that need to be discarded as 'burn-in'? That is, how can we help the Markov chain to reach the stationary distribution more quickly?*

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

Q. *Sometimes the Markov chain fails to find some of the peaks in the landscape, such that our samples do not provide a good representation of the posterior distribution. How can we tell whether this is the case or not?*

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

Section B: Bayesian molecular dating

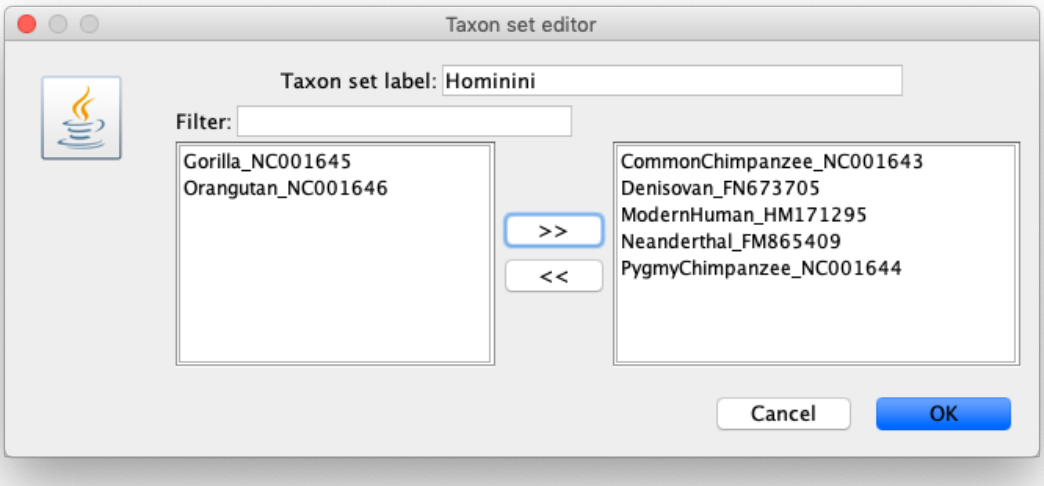
In this section you will conduct further analyses of the hominin data set in order to estimate the evolutionary timescale. There are four parts to the analysis:

- Creating an input file using *BEAUti*
- Bayesian phylogenetic analysis using *BEAST*
- Using a relaxed-clock model
- Processing the output using *Tracer*, *TreeAnnotator*, and *FigTree*

Creating an input file using *BEAUti*

- Open the program *BEAUti*. Select "Import Alignment" from the "File" menu and open the alignment file **denisovan.nex**. Alternatively, you can drag and drop the data file into the *BEAUti* window. This file contains the 13 concatenated protein-coding genes from the mitochondrial genomes of the 7 hominids that we analysed in Section A. For this section of the practical, we will simplify things by not partitioning the data. In other words, we will apply a single substitution model and single clock model to the whole sequence alignment.
- Go to the **Site Model** tab. Select the HKY+G model of nucleotide substitution, with 4 categories for gamma-distributed rates across sites.
- Go to the **Clock Model** tab. Select the "Strict Clock" model, which assumes that all lineages evolve at the same rate.
- Go to the **Priors** tab. Select the "Yule Model" for the tree prior and change the uniform prior for the birth rate so that it ranges from 0 to 100.
- Now we want to define groups of taxa that might be of interest. In the current analysis, we are interested in one of the nodes of the tree that can be used for age calibration. We can define a group of taxa by clicking on the "+ Add Prior" button and adding an "MRCA prior".

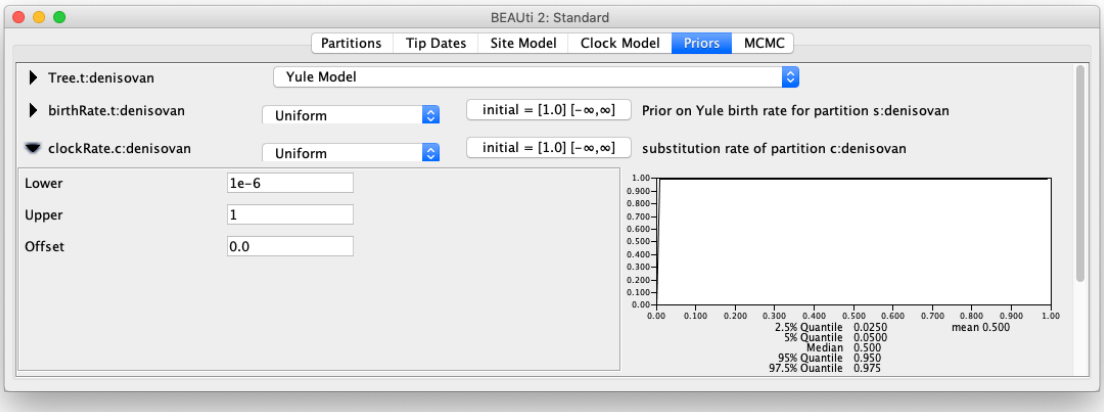
  For the taxon set label, type "Hominini". In this taxon set, we want to include the three humans (Modern Human, Neanderthal, and Denisovan) and two chimpanzees (Common Chimpanzee and Pygmy Chimpanzee). Select these taxa and click on the rightward-pointing arrows to include them in the taxon set. Then click "OK".



- Now we want to specify the prior distribution for the age of this taxon set. We believe that the ancestor of humans and chimpanzees existed about 6.5 million years ago, most likely around 6–7 million years ago. We can use this information to choose the parameters of a normal distribution. In the drop-down menu to the right of "Hominini.prior", select "Normal".

  We now need to set the parameters of this normal prior. Click on the black triangle on the left and change the parameters of the normal distribution so that the mean is 6.5 and the standard deviation (sigma) is 0.2551. Note that we are giving the dates in Myr. Take the time to have a look at the distribution and its features.

- The next step is to set the priors for the substitution rate. Published work suggests that mitochondrial substitution rates in animals all fall within the range of 10^{-6} to 1 substitutions/site/Myr. We can use these values to place uniform priors on the substitution rates. Note that the value of 10^{-6} is entered as "1e-6", as shown in the figure below.



- Go to the **MCMC** tab. To keep the analysis fairly short, choose a "Chain Length" of 10,000,000. Choose a file name of **denisovan_strictclock.log** for the trace log and a file name of **denisovan_strictclock.trees** for the tree log.
- Save the file as **denisovan_strictclock.xml** on your computer's desktop or your current working directory (wherever you want BEAST to write the output files). Keep *BEAUti* open because we will want to change the clock model later.

Bayesian phylogenetic analysis using *BEAST*

- Open the program *BEAST* and run an analysis using the XML file that you created above. The analysis will take a few minutes. While you are waiting, proceed to the next part below.

Using a relaxed-clock model

Now we will use a more complex clock model that allows a distinct rate along each branch of the tree.

- Go back to *BEAUti* and click on the **Clock Model** tab. Select the "Relaxed Clock Log Normal". This implements the uncorrelated lognormal relaxed clock, which assumes that the branch rates are drawn from an underlying lognormal distribution.

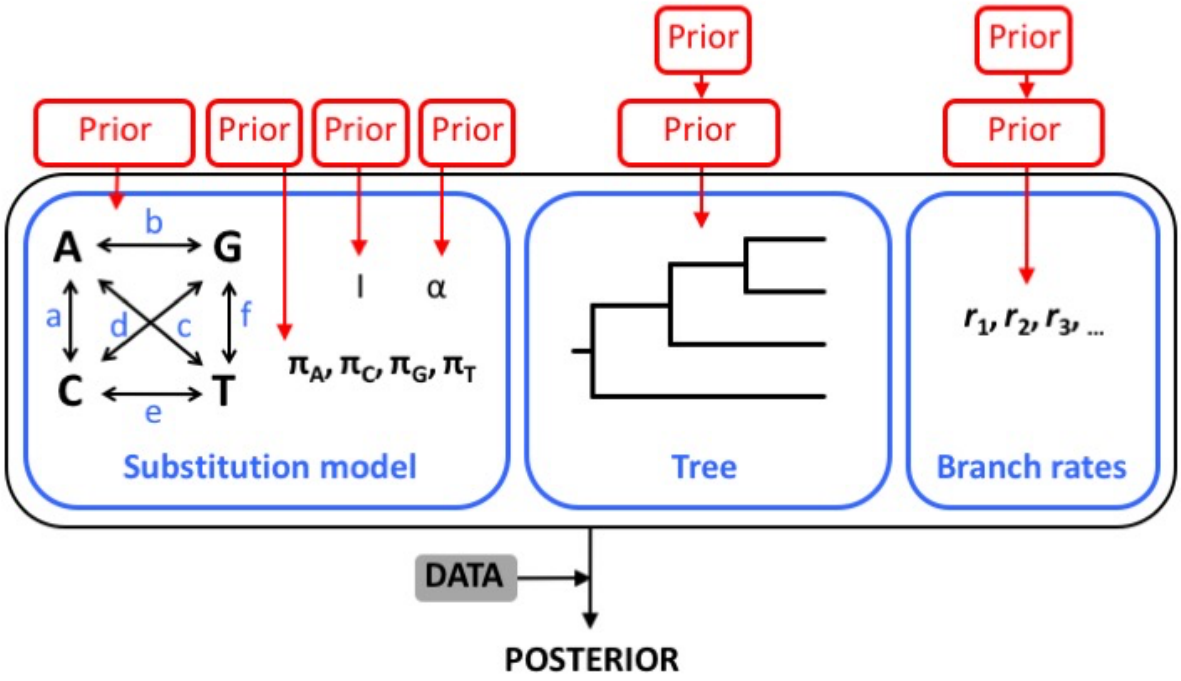
  If you have already closed *BEAUti*, go back to the first part of this section (Section B) and set everything up in the same way (except for the clock model).
- Go to the **Priors** tab and put a uniform prior of 10^{-6} to 1 substitutions/site/Myr on the mean rate ("ucldMean.c:denisovan").
- In the **MCMC** tab, change the file name for the trace log to **denisovan_relaxedclock.log**. Then change the file name of the tree log to **denisovan_relaxedclock.trees**.
- Save the file as **denisovan_relaxedclock.xml**. Keep *BEAUti* open for now, because it will be useful for answering the question below (which relates to the settings that you have used for your analysis).
- Run *BEAST* using the new XML file. If it looks as though the analysis will take too long, you can simply stop the analysis and instead work with the output files that have been provided ("pre-cooked runs").

While you are waiting for your analysis to run, try annotating the diagram below to show which models and priors you are using in your analysis. To work out these details, you will mainly need to look at the **Priors** tab in *BEAUti*.

To get you started, have a look at the alpha parameter in the substitution model. This is the shape parameter of the gamma distribution for rate variation across sites. In the **Priors** tab, you will see that we have used an exponential prior distribution for this parameter.

Note that the diagram below shows 6 parameters (*a* to *f*) for the pairwise exchange rates of the substitution model. These are the rates of change between pairs of nucleotides. In the analyses here, we are using the HKY substitution model which uses a different parameter, kappa, to represent the ratio of transitions to transversions.

Another hint is that we are not assuming a proportion of invariable sites in this analysis. So there is no "I" parameter in the model being used in the current analysis.



Processing the output using *Tracer*, *TreeAnnotator*, and *FigTree*

- Open the program *Tracer*, click on "Import Trace File" in the "File" menu, and import the two .log files from your *BEAST* analyses.
- Check that the effective sample sizes (ESSs) of all of the sampled parameters are greater than 200. This indicates that we have drawn enough samples to be able to produce a reasonable estimate of the / distribution of each parameter. If any ESS values are below ~200, it means that we need to run the MCMC analysis for a greater number of steps. If this is the case, ignore it for the purposes of this practical. You can also try changing the amount of burn-in to see if you can increase the ESS values.

Q. *Look at the results from your analysis based on the strict-clock model. What are the mean and 95% HPD interval (=95% credibility interval) for the estimate of the age of Hominini, which is given by* mrcatime(Hominini)*? Does this match the prior distribution that we assigned to it?*

. . . . . .

. . . . . .

. . . . . .

- Now go to the results from the relaxed-clock model and have a look at the estimate of "rate.coefficientOfVariation". This is the coefficient of variation of branch rates, which is calculated as the standard deviation of branch rates divided by their mean. It provides a measure of rate variation across branches in the tree, where a value of 0 indicates a strict clock. Have a look at the posterior distribution.

Q. *Does the posterior distribution for the coefficient of variation (of branch rates) bump against zero? If not, what does this indicate about the degree of rate variation across branches?*

. . . . . .

. . . . . .

. . . . . .

. . . . . .

- Open the program *TreeAnnotator*. In the box next to "Burnin percentage", enter the value "10". This means that we are discarding the first 10% of our samples as burn-in. For the "Input Tree File", select the .trees file produced by the *BEAST* analysis using the relaxed-clock model. For the "Output File", select the directory where you want to save the output file from *TreeAnnotator*. Give the output file the name **denisovan_relaxedclock.tre** and click "Run".

- Open the program *FigTree* and use it to view the file **denisovan_relaxedclock.tre** produced by *TreeAnnotator* in the previous step.

  Here we are mainly interested in the evolutionary timescale. In the "Node Labels" box, select "height" in the drop-down menu next to "Display". This will label the nodes of the tree with the mean posterior ages in Myr. You can also view the 95% HPD intervals by selecting "height_95%_HPD" in the drop-down menu next to "Display".

Q. *What are the mean and 95% HPD (credibility) interval for the estimate of the age of the split between the Denisovan hominin and the other two humans?*

. . . . . .

. . . . . .

- In the "Appearance" section, colour the branches by "rate_median". Play around with the colour gradient to show low rates in blue and high rates in red.

Q. *Which branches of the tree appear to have a high evolutionary rate?*

. . . . . .

. . . . . .

- Now use *TreeAnnotator* to process the .trees file from the *BEAST* analysis using the strict-clock model. View the tree in *FigTree*.

Q. *Does the date estimate for the split between the Denisovan hominin and the other two humans differ between the two clock models?*

. . . . . .

. . . . . .

Q. *Do you see any differences in the 95% credibility intervals of the date estimates made using the strict and relaxed clocks? Why might this be the case?*

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

Q. *We calibrated the dating analysis using a normal distribution for the age of Hominini. However, the lognormal distribution is often regarded as the most suitable parametric distribution for summarising fossil information. Why might this be the case?*

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

Q. *How might we be able to improve the precision of our molecular date estimates?*

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .