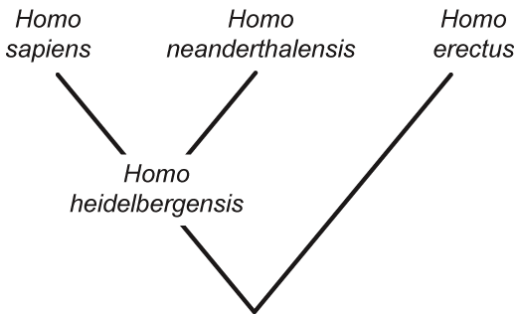
# Practical 2: A mysterious hominin from Siberia

## Background

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 25,000 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 DNA means that contamination from other sources is a serious risk. However, these challenges are being 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:

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

The data file, **denisovan.nex**, contains a DNA sequence alignment in 'Nexus' format. These are the concatenated DNA sequences 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*).

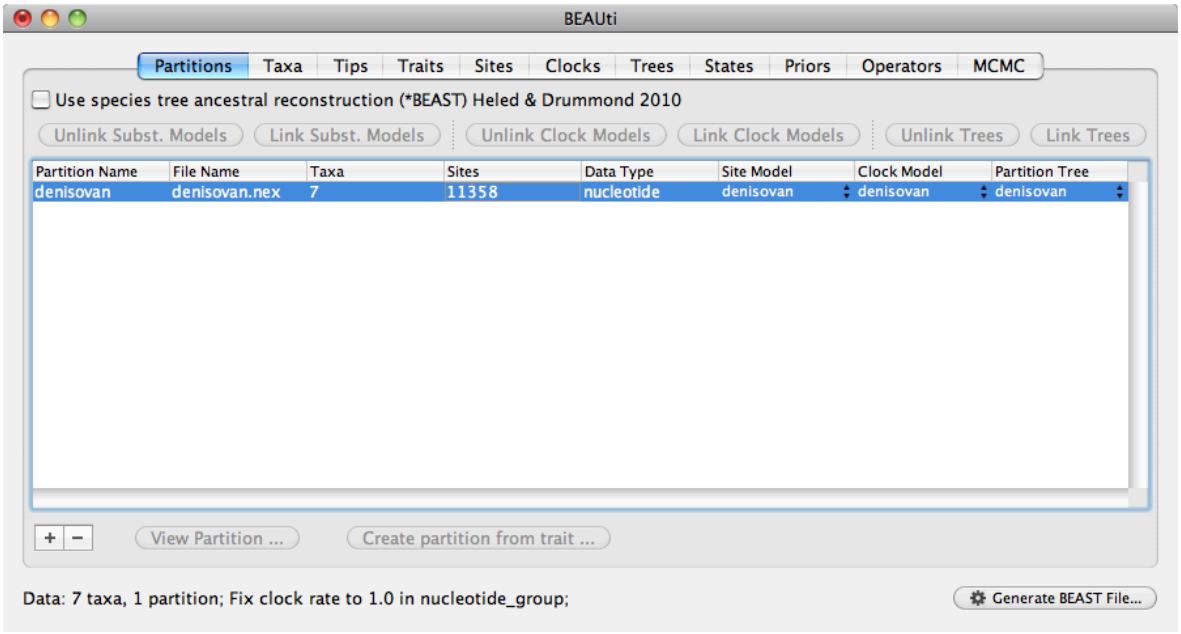
This exercise will use the Bayesian phylogenetic software *BEAST*. The program is quite complex and requires a detailed input file in XML format. However, these input files can be readily created with the user-friendly program *BEAUti*. There are four parts to the analysis:

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

## Creating an input file using *BEAUti*

a) 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 alignment file, **denisovan.nex**. Alternatively, you can drag and drop the data file into the BEAUti window.

a) You should now be in the **Partitions** section of *BEAUti*. The window will display some of the characteristics of the data that you have loaded. For example, we can see that the alignment contains 7 taxa and has 11,358 aligned nucleotides.



There are various options in this window, but these pertain to 'partitioned' analyses in which we divide the alignment into distinct subsets. Partitioning allows separate evolutionary models to be applied to different parts of the sequence alignment. For example, if we have multiple genes in our alignment, then we might wish to assign a different evolutionary model to each gene. In the current analysis, we will keep it simple and assume that a single evolutionary model is sufficient for the whole alignment.

b) Skip the **Taxon Sets** section. This section allows us to define groups of sequences that might be of interest, but we do not need to do this for the current analysis.

c) Skip the **Tips** and **Traits** sections. Normally, the **Tips** 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.

d) Go to the **Sites** section. Here we choose the nucleotide substitution model. In the current analysis, we shall use the HKY model of nucleotide substitution. 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" for "Substitution Model", "Estimated" for "Base Frequencies", and "None" for "Site Heterogeneity Model".

e) Go to the **Clocks** section. 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, BEAST requires a clock model to be chosen. BEAST only infers rooted trees. Here we will use a "strict clock" model, which is a very simple model that assumes that all lineages evolve at the same rate.

f) Go to the **Trees** section. Here we need to choose the prior distribution for the tree in our analysis. In the drop-down menu next to "Tree Prior", 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. For this analysis, we shall choose the simplest speciation model, which is the Yule process. This is a pure-birth model in which all lineages have an equal chance of splitting into two descendent lineages. Choose the "Speciation: Yule Process" model.

g) Skip the **States** section. Note that this section contains an option to choose a model of post-mortem DNA damage, which might be appropriate for the ancient DNA sequences in this data set. But we will ignore this option in this analysis.

h) Go to the **Priors** section. Here we need to choose prior distributions for the various parameters in the analysis. The default choices can be left as they are.

i) Skip the **Operators** section. This section lists the mechanisms for proposing changes to the tree and parameter values during the MCMC analysis. The default settings are fine.

j) Go to the **MCMC** section. 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 distributions using an appropriately designed MCMC simulation. By plotting these samples, we can then gain an approximation of the posterior distribution. To keep the analysis fairly short, choose a "Length of Chain" of 5,000,000. Choose a value of 500 for both "Echo state to screen every" and "Log parameters every". Choose the names of your output files by typing a desired name into the field next to "File name stem". Something like "denisovan_hky" should be fine (where "hky" denotes the HKY substitution model). Uncheck the box next to "Create operator analysis file".

k) Now click on "Generate BEAST file" in the bottom-right corner of the *BEAUti* interface. When the new window appears, click on "Continue" and save the file as **denisovan_hky.xml** in the current directory. 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*

a) Open the program *BEAST*. Open the file that you created above. Uncheck the box next to "Use BEAGLE library if available" and click on the "Run" button.

b) While the analysis is in progress, *BEAST* will continually write to two 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.

c) The analysis will take about 5 to 10 min, depending on 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?*

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

a) Go back to *BEAUti* and click on the **Sites** section. Next to "Site Heterogeneity Model", select "Gamma". 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. Change "Number of Gamma Categories" to 6.

   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 substitution model).

b) Change the "File name stem" to "denisovan_hkyg" and save the file under a name that is different from your previous file. Something like **denisovan_hkyg.xml** should be fine.

c) If your computer has more than one processor, run *BEAST* using the new input file. Otherwise, you should wait until your analysis from the previous part is done before starting the second *BEAST* analysis.

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*

a) Open the program *Tracer*, click on "Import Trace File" in the "File" menu, and import the .log files from your two *BEAST* analyses.

b) 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.

c) 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.

d) Now we want to compare the fit of the two substitution models to our data. This can be done using Bayes factors, which is a comparison of the marginal likelihoods of the two models. To calculate the marginal likelihoods, we will use a quick (but not very accurate) method known as the harmonic-mean estimator.

   Select both of the files in the top-left box of *Tracer*. Go to the "Analysis" menu and select "Model Comparison". For "Analysis type", select "harmonic mean". Change the number of bootstrap replicates to "100" and click "OK".

e) The Bayes factor now needs to be interpreted. The natural log of the Bayes factor is displayed. Values of ln(BF) can be interpreted as follows: 1-3 positive support, 3-5 strong support, >5 decisive support.

**Q**. *What is the ln Bayes factor of the HKY+G model compared with the HKY model? What level of support does this indicate?*

. . . . . .

. . . . . .

f) Open the program *TreeAnnotator*. This program is used to process the .trees file from *BEAST*. It reads all of the 1,000 sampled trees and summarises the information in the form of a single tree.

g) In the box next to "Burnin (as trees)", enter the value "1000". This means that we are throwing out the first 1000 samples (10%) because we are regarding them 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".

h) 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 questions about the analysis.

**Q**. *When drawing samples during the MCMC analysis, we chose to log parameters every 500 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:

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

## Creating an input file using *BEAUti*

a) 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.

b) Go to the **Taxon Sets** section. Here we can define groups of sequences 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. Taxon sets can be created by clicking on the "+" symbol in the bottom-left of the *BEAUti* window.

   Create a taxon set and call it "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 green arrow to put them in the "Included Taxa" window.

   Do not check the boxes in the columns "Mono?" or "Stem?".

c) Skip the **Tips** and **Traits** sections.

d) Go to the **Sites** section. Select the HKY+G model of nucleotide substitution, with 6 categories for gamma-distributed rates across sites. This was the preferred model from Section A of this practical.

e) Go to the **Clocks** section. Select the "strict clock" model, which assumes that all lineages evolve at the same rate.

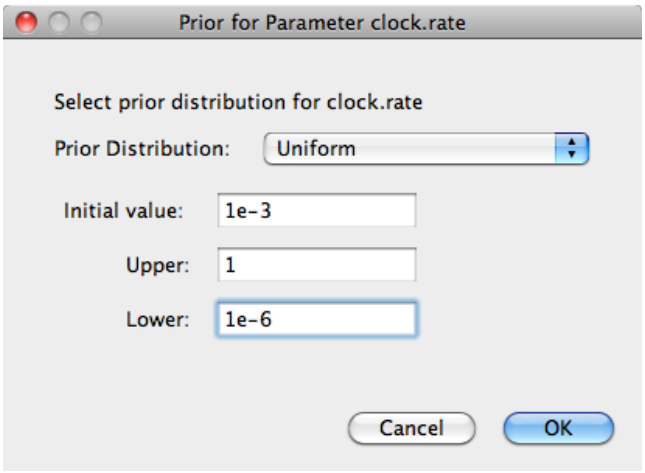
f) Go to the **Trees** section. Choose the "Speciation: Yule Process" model.

g) Skip the **States** section.

h) Go to the **Priors** section. Here we need to choose prior distributions for the various parameters in the analysis. Most of the default choices can be left as they are. However, we need to change the prior distribution for "tmrca(Hominini)", which represents the age of the group that we specified in the **Taxon Sets** section. We want to include our prior knowledge about the age of this group, which we have obtained from the fossil record.

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. Click on the "Using Tree Prior" box next to "tmrca(Hominini)" and change the prior to a "Normal Distribution" with a mean of 6.5 and a standard deviation of 0.2551. Use an initial value of 6.5. Note that we are giving the dates in Myr. Take the time to have a look at the distribution and its features.

We also need to specify the prior distribution for the substitution rate (clock.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 a uniform prior on clock.rate. Note that the value of $10^{-6}$ is entered as "1e-6", as shown in the figure on the right. Additionally, we need to give a starting value for clock.rate that falls within this range, such as $10^{-3}$ substitutions/site/year.




i) Skip the **Operators** section. The default settings are fine.

j) Go to the **MCMC** section. To keep the analysis fairly short, choose a "Length of Chain" of 5,000,000. Choose a value of 500 for both "Echo state to screen every" and "Log parameters every". Choose the names of your output files by typing a desired name into the field next to "File name stem". Something like "denisovan_strictclock" should be fine. Uncheck the box next to "Create operator analysis file".

k) Now click on "Generate BEAST file" in the bottom-right corner of the *BEAUti* interface. When the new window appears, click on "Continue" and save the file as **denisovan_strictclock.xml** in the current directory. Keep *BEAUti* open because we will want to change the clock model later.

## Bayesian phylogenetic analysis using *BEAST*

a) Open the program *BEAST*. Open the file that you created above. Uncheck the box next to "Use BEAGLE library if available" and click on the "Run" button.

b) While the analysis is in progress, *BEAST* will continually write to two 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.

c) The analysis will take about 5 to 10 min. 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.

a) Go back to *BEAUti* and click on the **Clocks** section. Select the "Uncorrelated relaxed clock". 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).

b) Go to the Priors tab and put a uniform prior of $10^{-6}$ to 1 substitutions/site/Myr on the mean rate (ucld.mean). Use a starting value of $10^{-3}$, as done above for the strict clock.

c) Change the "File name stem" to "denisovan_relaxedclock" and save the file under the name **denisovan_relaxedclock.xml** to distinguish it from your other XML files.

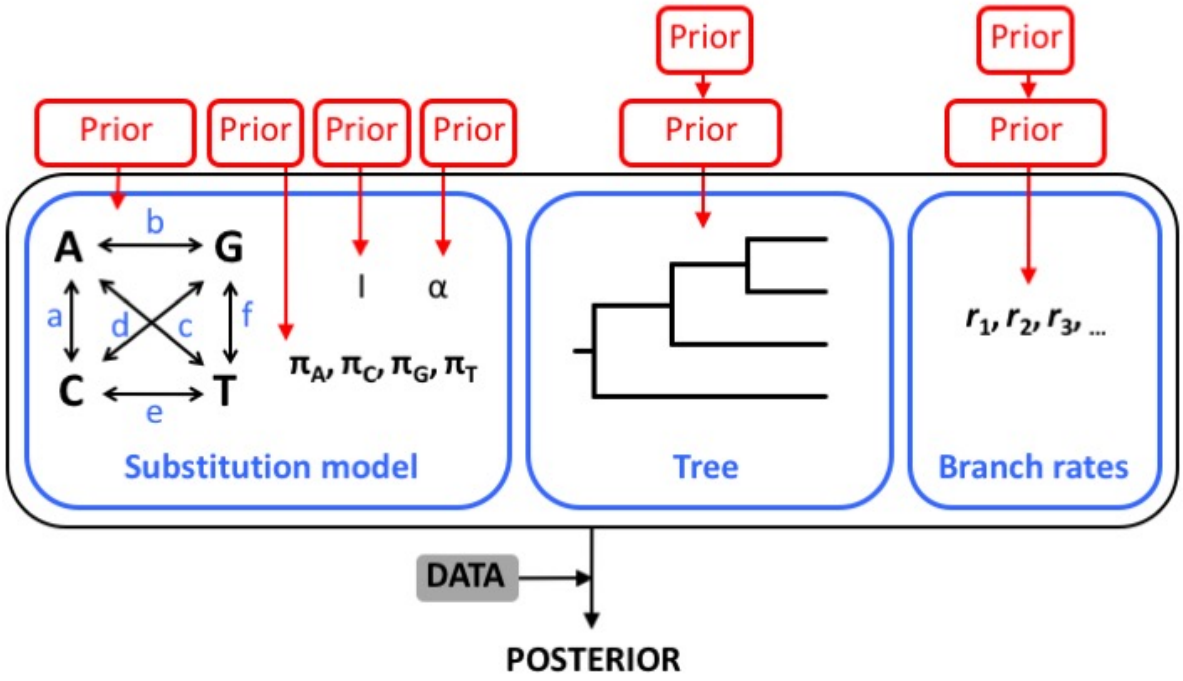
d) If your computer has more than one processor, run *BEAST* using the new input file. Otherwise, you should wait until your analysis using the strict clock is done before starting the second *BEAST* analysis.

e) It is likely that the relaxed-clock analysis will take too long to run (>30 min). If this is the case, you can simply stop the analysis and instead use the output files that have been provided ("pre-cooked runs").

While you are waiting for your analysis to run, try annotating the diagram on the next page 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 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.




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

a) Open the program *Tracer*, click on "Import Trace File" in the "File" menu, and import the two .log files from your *BEAST* analyses.

b) 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 posterior 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.

**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* tmrca(Hominini)*? Does this match the prior distribution that we assigned to it?*

. . . . . .

. . . . . .

. . . . . .

c) Now we want to compare the fit of the two clock models to our data. As in Section A, we will calculate the Bayes factor based on marginal likelihoods obtained using the harmonic-mean estimator.

Select both of the files in the top-left box of *Tracer*. Go to the "Analysis" menu and select "Model Comparison". For "Analysis type", select "harmonic mean". Change the number of bootstrap replicates to "100" and click "OK".

d) The Bayes factor now needs to be interpreted. The natural log of the Bayes factor is displayed. Values of ln(BF) can be interpreted as follows: 1-3 positive support, 3-5 strong support, >5 decisive support.

**Q**. *What is the ln Bayes Factor of the relaxed-clock model compared with the strict-clock model? What level of support does this indicate?*

. . . . . .

. . . . . .

e) In the main Tracer window, have a look at the estimate of "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?*

. . . . . .

. . . . . .

. . . . . .

. . . . . .

**Q**. *In terms of rate variation across branches, did you obtain similar conclusions from the Bayes factor and from inspection of the coefficient of variation?*

. . . . . .

. . . . . .

. . . . . .

. . . . . .

f) Open the program *TreeAnnotator*. In the box next to "Burnin (as trees)", enter the value "1000". For the "Input Tree File", click "Choose File" and select the .trees file produced by the *BEAST* analysis using the relaxed-clock 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_relaxedclock.tre** and click "Run".

g) 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 estimated 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?*

. . . . . .

. . . . . .

h) In the "Appearance" section, colour the branches by "rate". 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?*

. . . . . .

. . . . . .

i) 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?*

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .