# ***10kTrees* - Exercise #1**

## **Using *10kTrees* to Download a Consensus Tree and Tree Block of Multiple Trees**

The goal of this worked exercise is to introduce the *10kTrees* website. Additional exercises will enable you to view the trees in several different programs, add additional species to a tree block, and run a comparative analysis across the tree block.

### ***10kTrees* and Primate Phylogeny**

Phylogeny is now considered an essential component of comparative research, yet the true evolutionary history of a group of organisms is never known with certainty. This is true both in terms of topological relationships (i.e., the branching structure of relationships), and in terms of the lengths of branches that connect the nodes on the tree. Some researchers have used phylogenetic uncertainty as a justification for *not* controlling for phylogeny, or for setting the branches to be equal. Doing so makes a number of assumptions that can affect the outcome of the analysis (Martins and Garland 1991; Diaz-Uriarte and Garland 1998). Instead, it is better to incorporate the current level of uncertainty for specific nodes and branch lengths (Huelsenbeck 2000; Pagel and Lutzoni 2002). Indeed, different trees have been shown to produce different results in comparative analyses (Lutzoni et al. 2001).

The *10kTrees* website is a new web resource for phylogeny that allows users to incorporate phylogenetic uncertainty into their analyses in a principled way (Arnold et al. 2010). Currently, we provide phylogenetic trees for primates. In the near future, we will add additional mammalian orders, such as carnivorans, artiodactyls, and perissodactyls. The approach uses major developments in phylogenetics research to control for phylogenetic uncertainty (Pagel and Lutzoni 2002). These *Bayesian methods* provide a way to sample a set of trees in proportion to their posterior probabilities using Markov chain Monte Carlo (MCMC). This allows researchers to run analyses on an entire set of trees rather than using a single tree; thus, results are no longer conditioned on a single tree being correct. Using these methods, we can begin to understand the degree to which phylogenetic uncertainty affects the results of comparative studies – maybe it has no effect, but we can't say for sure one way or the other, yet…

With the *10kTrees* website, users can download up to 10,000 phylogenies sampled from a Bayesian tree inference. To be honest, this is probably way more trees than you will ever need for an analysis – but it is nice to know you can obtain this many trees if you need to do so. The website provides a variety of options, and it is designed so that it can be easily updated as new data or methods become available. Thus far, we have produced two versions of our Bayesian phylogeny for Primates, and we are already working on the third version. We recommend using the latest version, and it is important to state in the paper which version is used. However, if your dataset includes species found only in Version 1 (but not in Version 2), then it would be best to use the earlier version of 10kTrees. Version 3 will include species that were included in either previous version, plus additional species.

We request that people cite the paper describing the *10kTrees* approach and website (Arnold et al. 2010).

## Getting to know *10kTrees*

Let's begin by going to the *10kTrees* website: http://10ktrees.fas.harvard.edu/.

Take a few minutes to click through the main tabs at the top of the page. Notice that when you click on "Trees," an option appears for "Primates." Click on primates, and explore the headers. You are now on the "Primates" part of the website, which is indicated by a color change and a different menu. Note that under 'Dataset,' the full genetic dataset is available, and we provide graphical representations of the consensus tree in various formats. We also provide an availability matrix showing which genes are available for which species. For example, you can look at the availability matrix to see that 5 genes were available for *Cercopithecus diana* (and you can look at the columns to see exactly which ones). You can also get a full list of species, and the final aligned dataset. We are aiming for a totally open-source project, and you should feel free to explore the data, rerun the analyses, or use the data for educational purposes.

In what follows, we will be using*10kTrees* for Primates, Version 2. Note, however, that you can use earlier versions. The details for the latest version are provided under "Dataset," but detailed data for the earlier versions is available under the "Archive" tab. We aim to update the trees every 9-15 months, depending on the availability of new data or methods. Currently, for example, we are running Version 3 (as of 01/03/2011).

## Downloading trees

Go to "Download Trees" on the "Primates" part of *10kTrees*. Here you can select whether to download a *consensus tree*, a *tree block* sampled from the posterior probability distribution, or a combination of these.

Briefly, a tree block is the sample of trees obtained from the Bayesian analysis. These trees are sampled in proportion to their posterior probability. Thus, nodal support can be represented simply as the proportion of trees in the tree block that have a particular node. The consensus tree represents a summary of the agreement of all trees from the tree block. One typical feature of consensus trees is that for each node, a value that represents the support of that particular node among all trees is displayed. Here, the consensus tree is a 50% majority rule consensus tree, which means that nodes are only present if 50% or more of all trees from the tree block show a particular node. You can see nodal support values by clicking on "50% majority rule consensus tree… (NEXUS format)" link under the "Dataset" tab, and then using exercise 2 to view the Nexus file that is provided.

We will download a tree block of 200 trees and a consensus tree of all 10,000 trees stored on the site. Type 200 in the box under, “How many trees do you want to download?” Notice that you can also select commonly used numbers using the pull-down list in this box.

The website also allows the user to switch between different taxonomies. For this exercise, let’s switch to the Wilson and Reeder taxonomy by clicking on the button next to that option.

You may have noticed that information on settings can be obtained by placing your cursor over the phrases next to yellow questions marks. Try this for the “Branch Lengths” option. You will see that the *phylogram* option gives branch lengths proportional to the number of substitutions, while the *chronogram* option gives branch lengths proportional to time. Select the chronogram option.

In the “Display and selection options” box, you can have the species appear in alphabetical order, or organized by major taxonomic groups. You can also load a previously selected set of species (e.g., if you save your settings on the screen that follows). For this exercise, let’s list the species alphabetically… it will make it easier to select them from the list below.

Next, you can select the species to be included on your tree; the program will prune the trees to just those species that you selected. All 213 species in the dataset that are recognized in Wilson and Reeder taxonomy are selected by default (note, however, that the Wilson and Reeder taxonomy includes more than the 213 recognized species for which data are available).

Scroll down, and notice that some species are “grayed” out. These are species that are not recognized in the selected taxonomy. They can be selected if you would like to override the taxonomic selection of species; simply click on the box and it will be included in the trees that are delivered to your computer.

For this exercise, we created an artificial dataset using a package in R that enabled us to simulate data on a tree with a known degree of correlated evolution (r=0.6), and under specific branch length transformations (internal branches were compressed, using a “lambda” transformation of 0.7; Freckleton et al. 2002). The details are not important to what follows, except to say that by simulating the data under a known evolutionary model, we can (in later exercises) determine whether the methods are actually extracting the correct information from the data.

Go ahead and select the species for which we have data in the table below. You will notice on the page that all species are selected by default. To start, you will find it helpful to click the “Select/Deselect” all species, so that all are unselected by default. The species that are not given on the selected taxonomy will still be grayed out.

Note in advance that two sub-species at the end of the table – *Gorilla beringei beringei* and *Propithecus verreauxi coquereli* – are not found on the *10kTrees* version that we are using. Just ignore them, and select the others. This is typical of comparative databases – not every species in the dataset will be found in the tree, and not every species on the trees will be found in your data. Exercise 3 will show how you can add these two species to a tree block.

| **Species Name** | **Trait Y** | **Trait X** |
|---|---|---|
| *Alouatta caraya* | -8.095824 | -1.13152509 |
| *Alouatta seniculus* | -16.496641 | -11.60648254 |
| *Aotus trivirgatus* | -10.8132139 | 4.22749401 |
| *Brachyteles arachnoides* | -2.6532112 | 4.1578414 |
| *Callithrix argentata* | -13.6225542 | 1.97378175 |
| *Callithrix jacchus* | -5.9467716 | -2.46740128 |
| *Callithrix pygmaea* | 2.0497898 | 2.85681177 |
| *Cebus apella* | -4.1927343 | 4.18742826 |
| *Eulemur fulvus* | 8.41412 | -3.09028271 |
| *Eulemur mongoz* | 1.9326126 | -1.72064871 |
| *Gorilla gorilla gorilla* | 1.2410109 | 0.89762521 |
| *Hylobates lar* | 8.4758482 | 5.07765763 |
| *Lophocebus albigena* | 2.2519099 | 13.52808359 |
| *Loris tardigradus* | 16.6220655 | 5.36742067 |
| *Macaca mulatta* | 4.2897586 | 9.35672515 |
| *Macaca nemestrina* | -2.8253884 | -5.04397374 |
| *Macaca radiata* | 0.4985151 | -0.04806715 |
| *Macaca silenus* | -0.4123201 | -9.95748869 |
| *Macaca sinica* | -4.0406364 | 7.37985841 |
| *Miopithecus talapoin* | 3.6057858 | 13.986117 |
| *Otolemur crassicaudatus* | -0.6024125 | -8.19305572 |
| *Otolemur garnettii* | 2.9631109 | -4.56602235 |
| *Pan paniscus* | 12.7008835 | 9.05545596 |
| *Papio anubis* | -0.926246 | 10.47746826 |
| *Papio hamadryas* | 5.4363688 | 8.32345764 |
| *Perodicticus potto* | -10.5194529 | -12.34076654 |
| *Propithecus diadema* | 11.5917927 | -1.54672153 |
| *Propithecus verreauxi* | 3.7558456 | -7.77196946 |
| *Saguinus oedipus* | -10.6000877 | 0.10799174 |
| *Semnopithecus entellus* | 5.3478877 | 4.66173289 |
| *Tarsius bancanus* | 5.9223065 | -5.65315634 |
| *Tarsius syrichta* | -6.6247245 | -10.11839168 |
| *Varecia variegata variegata* | 0.5078974 | -11.07540998 |
| *Gorilla gorilla beringei* | -1.6503404 | 2.19656967 |
| *Propithecus verreauxi coquereli* | -0.7450781 | -6.39965609 |

Glance back over the list of species and your settings to be sure they are correct. Notice that if you used the correct settings and excluded the two subspecies just mentioned, you should see that “33 species currently selected” at the top of the box that lists the species.

We are ready to download the set of 200 trees. Scroll down to the bottom and click “Download Trees.” It may take a minute or two to generate the file for a tree block. Please be patient.

On the next screen that comes up, we will first download the tree block. Find this file on your computer and rename it “treeblock_example.nex”.

Next, download the consensus tree and name it “treeconsensus_example.nex”.

Finally, notice that you can save the current selected species and settings in a separate file. Go ahead and do this, and we will see how it works in Exercise 2 (the file will be saved automatically in the location where you download files). This is actually very useful, for example in situations in which you might add or subtract a species at a later stage in the analysis, or change the other settings. It saves the trouble of re-selecting all the species (which can actually be quite time consuming when you are matching a list of species in your data to the list of species on the website).

You are now ready to view your trees. For that, proceed to Exercise 2!

**References**

Arnold, C., Matthews, L. J. & Nunn, C. L. 2010. The 10kTrees Website: A New Online Resource for Primate Phylogeny. Evolutionary Anthropology, 19, 114-118.

Carnes, L.M., C.L. Nunn, R.J. Lewis (in prep.). Effects of distribution of female primates on male reproductive monopolization.

Diaz-Uriarte, R., and T. Garland. 1998. Effects of branch length errors on the performance of phylogenetically independent contrasts. Systematic Biology 47:654-672.

Freckleton, R. P., P. H. Harvey, and M. Pagel. 2002. Phylogenetic analysis and comparative data: A test and review of evidence. American Naturalist 160:712-726.

Huelsenbeck, J. P., B. Rannala, and J. P. Masly. 2000. Accommodating phylogenetic uncertainty in evolutionary studies. Science 288:2349-2350.

Lutzoni, F., M. Pagel, and V. Reeb. 2001. Major fungal lineages are derived from lichen symbiotic ancestors. Nature 411:937-940.

Martins, E. P., and T. Garland. 1991. Phylogenetic analyses of the correlated evolution of continuous characters:  a simulation study. Evolution 45:534-557.

Pagel, M. & Lutzoni, F. 2002. Accounting for phylogenetic uncertainty in comparative studies of evolution and adaptation. In: Biological Evolution and Statistical Physics (Ed. by Lässig, M. & Valleriani, A.), pp. 148-161. Berlin: Springer-Verlag.