

10kTrees - Exercise #3

Adding Lineages to a Treeblock in R

It is common to have data on a species that is not included in a phylogeny. This is especially true when using with a phylogeny that incompletely samples all the species in a clade – which is the case for *10kTrees* – but it can happen in many cases when you have comparative research on different subspecies that are not recognized as separate species in widely used taxonomies. This incompleteness is not necessarily a fault of *10kTrees* itself; rather, species are often missing because insufficient genetic data of the appropriate quality are available.

When trying to add a missing species to a set of trees, one option is to find a closely related (ideally sister species) to the lineage of interest that is listed in *10kTrees*, but is not included in your dataset. To do this you will need another phylogeny that includes your species of interest, here called the “other” phylogeny. Based on the other phylogeny, the lineage you substitute from *10kTrees* should not insert your missing species in any position that would be incompatible with the placement of the species in the other phylogeny. If you can thereby substitute the species on *10kTrees* for the one in your dataset and achieve the same tree topology and branch lengths, by all means, do that!

In most cases, however, substitutions of that sort are not possible. Instead, you will need to add in a species. Again, you will need some additional phylogenetic information to decide where to place the new species, both in terms of topology and the time at which the species diverged from others in your tree. With a single tree, it is relatively easy to add a species to your tree; this can be done, for example, in Mesquite. But what should you do if you have 200 trees? Manually adding a lineage to each tree is a daunting task!

To help with this monumental task, we offer an R script that can add a lineage to an entire block of trees (created by Christian Arnold). Specifically, the script adds a lineage to each tree based on its sister species relationship to another lineage and the time at which those two species split (in units equivalent to those in the set of trees, which is 1 million year units for *10kTrees*). Thus, importantly, the added lineage will not represent phylogenetic uncertainty; it will be the same placement and split date for all trees in the treeblock. **Please note that a current limitation** of the script is that the new species can only be added to the trees if it is the sister species of a particular species that is already in the trees. That is, it is not possible with this script to add a species basal to a whole clade (two or more species).

We will demonstrate the script by adding the two lineages from Exercise 1 that were not found on the *10kTrees* website under the Wilson and Reeder taxonomy. These lineages are the subspecies: *Gorilla beringei beringei* and *Propithecus verreauxi coquereli*.

To begin, we need to identify the sister species for each of these, based on the set of taxa in our treeblock. We also need to identify the age at which these species split. For this exercise, the information is given in the following table:

Species to Add	Sister Lineage to Join	Age of Split
<i>Gorilla beringei beringei</i>	<i>Gorilla gorilla gorilla</i>	1.25 mya
<i>Propithecus verreauxi coquereli</i>	<i>Propithecus verreauxi</i>	1.00 mya

Next, we need to change the working directory to the folder where your tree is stored. For this, use the menu commands (under Misc on a Mac), or `setwd`. Assuming, for example, that my working directory is on my main documents folder, I'd type something like this: `setwd("/Users/Charlie/Documents")`

Now, assuming you have already installed ape, let's be sure ape is loaded and ready for action:

```
library("ape")
```

We are now ready to identify the treeblock file to alter (we will use the "treeblock_example.nex" file from Exercise 1:

```
fileName = "treeblock_example.nex"
```

And of course, we to pass along some crucial information related to the species to add (`speciesToBeInserted`), which lineage to add it to (`insertSpeciesNextTo`), and the date of the split (`branchLengthSplitAgo`). Note that branch lengths for the chrono trees on 10kTrees are in units of millions of years.

Let's start with *Gorilla gorilla beringei*.

```
speciesToBeInserted = "Gorilla_gorilla_beringei"
insertSpeciesNextTo = "Gorilla_gorilla_gorilla"
branchLengthSplitAgo = 1.25
```

We also need to name an output file, which will have the new species added to it:

```
filenameOutput = paste(fileName, ".ber.nex", sep="")
```

Next, paste the following code into R, and hit return:

```
#Create new phylogenetic tree, consisting of the species
that should be inserted
treeNewStr = paste("(", speciesToBeInserted, ":",
branchLengthSplitAgo, ");", sep="")
treeNew <- read.tree(text = treeNewStr)

# Read tree block
treeBlock = read.nexus(fileName)
treeBlockNew = list()

# For each tree in the tree block
for (i in 1:length(treeBlock)) {

  #Identify the position (i.e., integer number) where
insertNextT is located in the tree object
  pos = 0
```

```

    for (j in 1: length(treeBlock[[i]]$tip.label)) {
      if ( insertSpeciesNextTo ==
treeBlock[[i]]$tip.label[j]) {
        pos = j
      }
    }
    if (pos == 0) {
      stop ("Error: Could not find species
",insertSpeciesNextTo, " in tree ",i,"", exit\n")
    }

    #Merge both trees and write tree to new list of trees
    treeBlockNew[[i]] = bind.tree(treeBlock[[i]],
treeNew,where = pos, position = branchLengthSplitAgo)

  }

#Write new nexus file
write.nexus(treeBlockNew,file = filenameOutput)

```

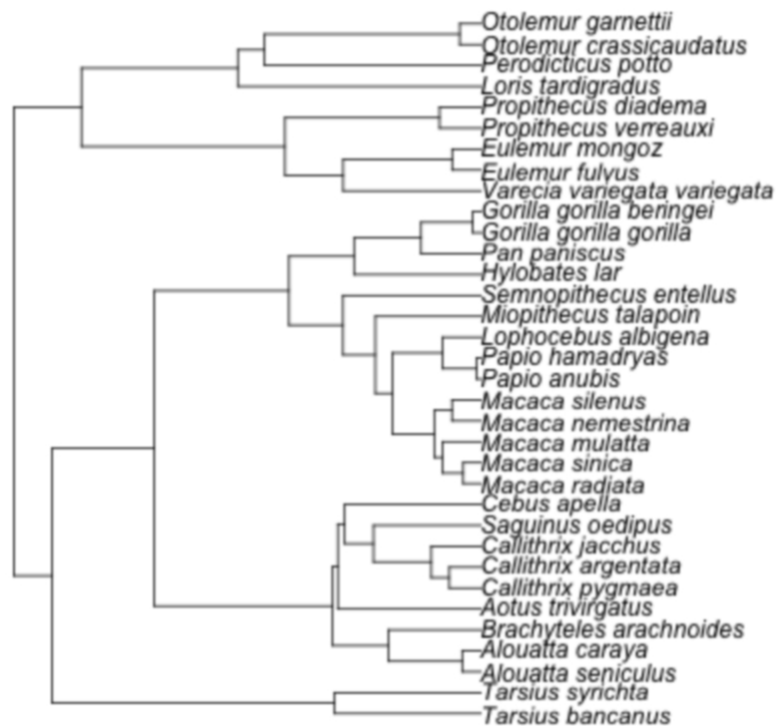
It may take a minute or more for the script to process; please be patient. When you see that R is finished with the script, check out the first tree in the newly created nexus file:

```

treeber=read.nexus("treeblock_example.nex.ber.nex")
plot(treeber[[1]])

```

You should see *Gorilla gorilla beringei* in your tree:



Now, to add *Propithecus verreauxi coquereli*, we repeat the procedure, but build on our altered tree, `treeblock_example.nex.ber.nex`, as follows. Note that the only changes are indicated by the “###” signs at the end of lines of code (which are not read by R):

```

fileName = "treeblock_example.nex.ber.nex"      ###

speciesToBeInserted = "Propithecus_verreauxi_coquereli" ###
insertSpeciesNextTo = "Propithecus_verreauxi"   ###
branchLengthSplitAgo = 1.0                       ###

filenameOutput = paste(fileName, ".coq.nex", sep="") ###

#Create new phylogenetic tree, consisting of the species
that should be inserted
treeNewStr = paste("(", speciesToBeInserted, ":",
branchLengthSplitAgo, ");", sep="")
treeNew <- read.tree(text = treeNewStr)

# Read tree block
treeBlock = read.nexus(fileName)
treeBlockNew = list()

# For each tree in the tree block
for (i in 1:length(treeBlock)) {

  #Identify the position (i.e., integer number) where
insertNextT is located in the tree object
  pos = 0
  for (j in 1: length(treeBlock[[i]]$tip.label)) {
    if ( insertSpeciesNextTo ==
treeBlock[[i]]$tip.label[j]) {
      pos = j
    }
  }

  if (pos == 0) {
    stop ("Error: Could not find species
",insertSpeciesNextTo, " in tree ",i," exit\n")
  }

  #Merge both trees and write tree to new list of trees
  treeBlockNew[[i]] = bind.tree(treeBlock[[i]],
treeNew,where = pos, position = branchLengthSplitAgo)

}

#Write new nexus file
write.nexus(treeBlockNew,file = filenameOutput)

```

To see your tree, type:

```

treebercoq=read.nexus("treeblock_example.nex.ber.nex.coq.
nex")

```

```
plot(treebercoq[[1]])
```

Now, we can use this tree (`treeblock_example.nex.ber.nex.coq.nex`) to include these two species in the analyses.