8. Worksheet: Phylogenetic Diversity - Traits

Trang Nguyen; Z620: Quantitative Biodiversity, Indiana University

27 février, 2025

OVERVIEW

Up to this point, we have been focusing on patterns taxonomic diversity in Quantitative Biodiversity. Although taxonomic diversity is an important dimension of biodiversity, it is often necessary to consider the evolutionary history or relatedness of species. The goal of this exercise is to introduce basic concepts of phylogenetic diversity.

After completing this exercise you will be able to:

1. create phylogenetic trees to view evolutionary relationships from sequence data
2. map functional traits onto phylogenetic trees to visualize the distribution of traits with respect to evolutionary history
3. test for phylogenetic signal within trait distributions and trait-based patterns of biodiversity

Directions:

1. In the Markdown version of this document in your cloned repo, change "Student Name" on line 3 (above) with your name.
2. Complete as much of the worksheet as possible during class.
3. Use the handout as a guide; it contains a more complete description of data sets along with examples of proper scripting needed to carry out the exercises.
4. Answer questions in the worksheet. Space for your answers is provided in this document and is indicated by the ">" character. If you need a second paragraph be sure to start the first line with ">". You should notice that the answer is highlighted in green by RStudio (color may vary if you changed the editor theme).
5. Before you leave the classroom, **push** this file to your GitHub repo.
6. For the assignment portion of the worksheet, follow the directions at the bottom of this file.
7. When you are done, **Knit** the text and code into a PDF file.
8. After Knitting, submit the completed exercise by creating a **pull request** via GitHub. Your pull request should include this file `PhyloTraits_Worskheet.Rmd` and the PDF output of `Knitr` (`PhyloTraits_Worskheet.pdf`).

The completed exercise is due on **Wednesday, February 26th, 2025 before 12:00 PM (noon)**.

1) SETUP

In the R code chunk below, provide the code to:
1. clear your R environment,
2. print your current working directory,
3. set your working directory to your `Week6-PhyloTraits/` folder, and
4. load all of the required R packages (be sure to install if needed).

```r
rm(list = ls())
getwd()
```

```
## [1] "C:/Users/ttran/OneDrive - Indiana University/SP25 - Quantitative Biodiversity/QB2025_Nguyen/Week
```

```r
#setwd(paste0(getwd(),"/Week6-PhyloTraits/"))

package.list <- c('ape', 'seqinr', 'phylobase', 'adephylo', 'geiger',
      'picante', 'stats', 'RColorBrewer', 'caper', 'phylolm', 'pmc',
      'ggplot2', 'tidyr', 'dplyr', 'phangorn', 'pander', 'phytools', 'vegan',
      'cluster', 'dendextend', 'phylogram', 'bios2mds', 'pak', 'formatR')

for (package in package.list) {
  if (!require(package, character.only = TRUE, quietly = TRUE)) {
    install.packages(package)
    library(package, character.only = TRUE)
  }
}
```

```
## Warning: le package 'ape' a été compilé avec la version R 4.4.2

## Warning: le package 'seqinr' a été compilé avec la version R 4.4.2

##
## Attachement du package : 'seqinr'

## Les objets suivants sont masqués depuis 'package:ape':
##
##     as.alignment, consensus

## Warning: le package 'phylobase' a été compilé avec la version R 4.4.2

##
## Attachement du package : 'phylobase'

## L'objet suivant est masqué depuis 'package:ape':
##
##     edges

## Warning: le package 'adephylo' a été compilé avec la version R 4.4.2

## Warning: le package 'ade4' a été compilé avec la version R 4.4.2

## Warning: le package 'geiger' a été compilé avec la version R 4.4.2

## Warning: le package 'phytools' a été compilé avec la version R 4.4.2

## Warning: le package 'maps' a été compilé avec la version R 4.4.2
```

```
##
## Attachement du package : 'phytools'

## L'objet suivant est masqué depuis 'package:phylobase':
##
##     readNexus

## Warning: le package 'picante' a été compilé avec la version R 4.4.2

## Warning: le package 'vegan' a été compilé avec la version R 4.4.2

## Warning: le package 'permute' a été compilé avec la version R 4.4.2

##
## Attachement du package : 'permute'

## L'objet suivant est masqué depuis 'package:seqinr':
##
##     getType

## This is vegan 2.6-8

##
## Attachement du package : 'vegan'

## L'objet suivant est masqué depuis 'package:phytools':
##
##     scores

##
## Attachement du package : 'nlme'

## L'objet suivant est masqué depuis 'package:seqinr':
##
##     gls

## Warning: le package 'caper' a été compilé avec la version R 4.4.2

## Warning: le package 'mvtnorm' a été compilé avec la version R 4.4.2

## Warning: le package 'phylolm' a été compilé avec la version R 4.4.2

## Warning: le package 'pmc' a été compilé avec la version R 4.4.2

## Warning: le package 'ggplot2' a été compilé avec la version R 4.4.2

##
## Attachement du package : 'dplyr'
```

```
## L'objet suivant est masqué depuis 'package:MASS':
##
##     select

## L'objet suivant est masqué depuis 'package:nlme':
##
##     collapse

## L'objet suivant est masqué depuis 'package:seqinr':
##
##     count

## L'objet suivant est masqué depuis 'package:ape':
##
##     where

## Les objets suivants sont masqués depuis 'package:stats':
##
##     filter, lag

## Les objets suivants sont masqués depuis 'package:base':
##
##     intersect, setdiff, setequal, union

## Warning: le package 'phangorn' a été compilé avec la version R 4.4.2

##
## Attachement du package : 'phangorn'

## Les objets suivants sont masqués depuis 'package:vegan':
##
##     diversity, treedist

## Warning: le package 'pander' a été compilé avec la version R 4.4.2

##
## Attachement du package : 'cluster'

## L'objet suivant est masqué depuis 'package:maps':
##
##     votes.repub

## Warning: le package 'dendextend' a été compilé avec la version R 4.4.2

## Registered S3 method overwritten by 'dendextend':
##   method     from
##   rev.hclust vegan
```

```
##
## ---------------------
## Welcome to dendextend version 1.19.0
## Type citation('dendextend') for how to cite the package.
##
## Type browseVignettes(package = 'dendextend') for the package vignette.
## The github page is: https://github.com/talgalili/dendextend/
##
## Suggestions and bug-reports can be submitted at: https://github.com/talgalili/dendextend/issues
## You may ask questions at stackoverflow, use the r and dendextend tags:
##   https://stackoverflow.com/questions/tagged/dendextend
##
##  To suppress this message use:  suppressPackageStartupMessages(library(dendextend))
## ---------------------


##
## Attachement du package : 'dendextend'

## L'objet suivant est masqué depuis 'package:permute':
##
##     shuffle

## L'objet suivant est masqué depuis 'package:geiger':
##
##     is.phylo

## L'objet suivant est masqué depuis 'package:phytools':
##
##     untangle

## Les objets suivants sont masqués depuis 'package:phylobase':
##
##     labels<-, prune

## Les objets suivants sont masqués depuis 'package:ape':
##
##     ladderize, rotate

## L'objet suivant est masqué depuis 'package:stats':
##
##     cutree

## Warning: le package 'phylogram' a été compilé avec la version R 4.4.2

##
## Attachement du package : 'phylogram'

## L'objet suivant est masqué depuis 'package:dendextend':
##
##     prune
```

```
## L'objet suivant est masqué depuis 'package:phylobase':
##
##     prune

## Warning: le package 'bios2mds' a été compilé avec la version R 4.4.2

##
## Attachement du package : 'amap'

## L'objet suivant est masqué depuis 'package:vegan':
##
##     pca

## Warning: le package 'e1071' a été compilé avec la version R 4.4.2

##
## Attachement du package : 'scales'

## L'objet suivant est masqué depuis 'package:phytools':
##
##     rescale

## Warning: le package 'rgl' a été compilé avec la version R 4.4.2

## Warning: le package 'pak' a été compilé avec la version R 4.4.2

## Warning: le package 'formatR' a été compilé avec la version R 4.4.2
```

```
# pkgbuild::check_build_tools(debug = TRUE)

install.packages("pak")
```

```
## Warning: le package 'pak' est en cours d'utilisation et ne sera pas installé
```

```
pak::pkg_install("msa")
```

```
##   Loading metadata database

##   Loading metadata database ... done
##
##   No downloads are needed
##   1 pkg + 19 deps: kept 17 [6.6s]
```

```
library(msa)
```

```
## Le chargement a nécessité le package : Biostrings

## Warning: le package 'Biostrings' a été compilé avec la version R 4.4.2
```

```
## Le chargement a nécessité le package : BiocGenerics
##
## Attachement du package : 'BiocGenerics'
##
## Les objets suivants sont masqués depuis 'package:dplyr':
##
##     combine, intersect, setdiff, union
##
## L'objet suivant est masqué depuis 'package:ade4':
##
##     score
##
## Les objets suivants sont masqués depuis 'package:stats':
##
##     IQR, mad, sd, var, xtabs
##
## Les objets suivants sont masqués depuis 'package:base':
##
##     anyDuplicated, aperm, append, as.data.frame, basename, cbind,
##     colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,
##     get, grep, grepl, intersect, is.unsorted, lapply, Map, mapply,
##     match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
##     Position, rank, rbind, Reduce, rownames, sapply, saveRDS, setdiff,
##     table, tapply, union, unique, unsplit, which.max, which.min
##
## Le chargement a nécessité le package : S4Vectors
## Le chargement a nécessité le package : stats4
##
## Attachement du package : 'S4Vectors'
##
## Les objets suivants sont masqués depuis 'package:dplyr':
##
##     first, rename
##
## L'objet suivant est masqué depuis 'package:tidyr':
##
##     expand
##
## L'objet suivant est masqué depuis 'package:utils':
##
##     findMatches
##
## Les objets suivants sont masqués depuis 'package:base':
##
##     expand.grid, I, unname
##
## Le chargement a nécessité le package : IRanges

## Warning: le package 'IRanges' a été compilé avec la version R 4.4.2

##
## Attachement du package : 'IRanges'
##
## Les objets suivants sont masqués depuis 'package:dplyr':
```

```
##
##     collapse, desc, slice
##
## L'objet suivant est masqué depuis 'package:nlme':
##
##     collapse
##
## L'objet suivant est masqué depuis 'package:grDevices':
##
##     windows
##
## Le chargement a nécessité le package : XVector
## Plus d’une classe "Annotated" est trouvée en cache : Utilisation de la première, depuis l’espace de
## Aussi défini par 'S4Vectors'
## Plus d’une classe "Annotated" est trouvée en cache : Utilisation de la première, depuis l’espace de
## Aussi défini par 'S4Vectors'
## Plus d’une classe "Annotated" est trouvée en cache : Utilisation de la première, depuis l’espace de
## Aussi défini par 'S4Vectors'
## Plus d’une classe "Annotated" est trouvée en cache : Utilisation de la première, depuis l’espace de
## Aussi défini par 'S4Vectors'
## Plus d’une classe "Annotated" est trouvée en cache : Utilisation de la première, depuis l’espace de
## Aussi défini par 'S4Vectors'
## Plus d’une classe "Annotated" est trouvée en cache : Utilisation de la première, depuis l’espace de
## Aussi défini par 'S4Vectors'
## Le chargement a nécessité le package : GenomeInfoDb
```

```
## Warning: le package 'GenomeInfoDb' a été compilé avec la version R 4.4.2
```

```
##
## Attachement du package : 'Biostrings'
##
## L'objet suivant est masqué depuis 'package:dendextend':
##
##     nnodes
##
## L'objet suivant est masqué depuis 'package:seqinr':
##
##     translate
##
## L'objet suivant est masqué depuis 'package:ape':
##
##     complement
##
## L'objet suivant est masqué depuis 'package:base':
##
##     strsplit
```

2) DESCRIPTION OF DATA

The maintenance of biodiversity is thought to be influenced by **trade-offs** among species in certain functional traits. One such trade-off involves the ability of a highly specialized species to perform exceptionally well on a particular resource compared to the performance of a generalist. In this exercise, we will take a phylogenetic approach to mapping phosphorus resource use onto a phylogenetic tree while testing for specialist-generalist trade-offs.

3) SEQUENCE ALIGNMENT

Question 1: Using your favorite text editor, compare the `p.isolates.fasta` file and the `p.isolates.afa` file. Describe the differences that you observe between the two files.

> ***Answer 1***: The fasta file shows the raw sequence data, while the afa file shows the aligned sequence data.
>
> In the R code chunk below, do the following: 1. read your alignment file, 2. convert the alignment to a DNAbin object, 3. select a region of the gene to visualize (try various regions), and 4. plot the alignment using a grid to visualize rows of sequences.

```
seqs <- readDNAStringSet("data/p.isolates.fasta", format = 'fasta')
seqs # View sequences
```

```
## DNAStringSet object of length 40:
##      width seq                                               names
##  [1]   619 ACACGTGAGCAATCTGCCCTTCT...TTCTCTGGGAATACCTGACGCT LL9
##  [2]   597 CGGCAGCGGGAAGTAGCTTGCTA...AACTGTTCAGCTAGAGTCTTGT WG14
##  [3]   794 CAGCGGCGGACGGGTGAGTAACA...GCTAACGCATTAAGCACTCCGC WG28
##  [4]   716 CTTCAGAGTTAGTGGCGGACGGG...TGCTAGTTGTCGGGATGCATGC LL24
##  [5]   803 ACGAACTCTTCGGAGTTAGTGGC...TAAAACTCAAAGGAATTGACGG LL41A
##  ...   ... ...
## [36]   652 TTCGGGAGTACACGAGCGGCGAA...TTCTCTGGGAATACCTGACGCT LL46
## [37]   661 GCGAACGGGTGAGTAACACGTGG...GAGCGAAAGCGTGGGTAGCGAA WG26
## [38]   694 GGCGAACGGGTGAGTAACACGTG...ACCCTGGTAGTCCACGCCGTAA WG42
## [39]   699 TACAGGTACCAGGCTCCTTCGGG...AAAGCATGGGTAGCGAACAGGA LLX17
## [40]  1426 TTCTGGTTGATCCTGCCAGAGGT...AACCTNAATTTTGCAAGGGGGG Methanosarcina
```

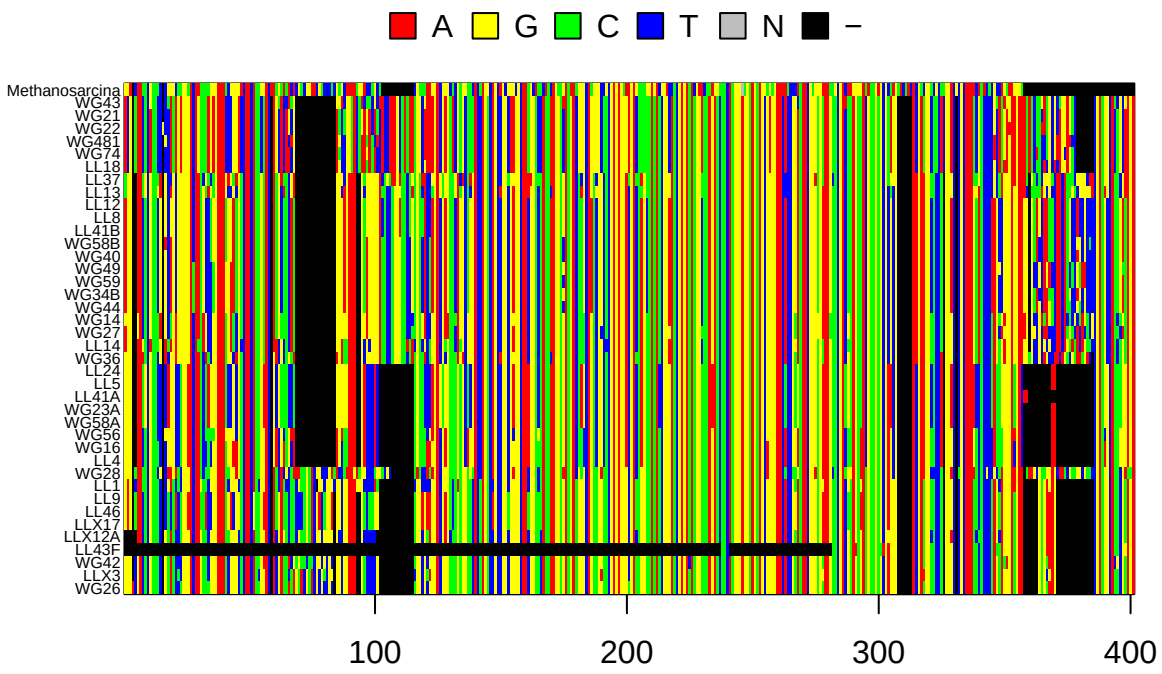
```
# Align sequences using default MUSCLE parameters {msa}
read.aln <- msaMuscle(seqs)

# Save and export the alignment to use later
save.aln <- msaConvert(read.aln, type = "bios2mds::align")
export.fasta(save.aln, "./data/p.isolates.afa")

# Convert Alignment to DNAbin Object {ape}
p.DNAbin <- as.DNAbin(read.aln)

# Identify Base Pair Region of 16S rRNA Gene to Visualize
window <- p.DNAbin[, 100:500]

# Command to Visualize Sequence Alignment {ape}
image.DNAbin(window, cex.lab = 0.50)
```




```
# Change the window size
window <- p.DNAbin[, 1:1000]
window <- p.DNAbin[, 100:500]
```

Question 2: Make some observations about the `muscle` alignment of the 16S rRNA gene sequences for our bacterial isolates and the outgroup, *Methanosarcina*, a member of the domain Archaea. Move along the alignment by changing the values in the `window` object.

> ***Answer 2***: For windows 100-500, the sequences show strong conservation across large portions of the alignment. I think this is expected since 16S rRNA genes evolve slowly. However, some regions exhibit higher variability, likely corresponding to hypervariable regions of the 16S rRNA gene. The black regions in the alignment visualization are gaps.
>
> The outgroup Methanosarcina has longer, bigger insertions or deletions compared to the bacterial sequences. These indels suggest phylogenetic divergence between Archaea (Methanosarcina) and Bacteria.

a. Approximately how long are our sequence reads?

```
mean(width(seqs))
```

```
## [1] 724.625
```

b. What regions do you think would are appropriate for phylogenetic inference and why?

Answer 2a:
The average sequence length is 700 bp.

Answer 2b:
I think conserved regions, which are less affected by mutations, can help us to trust the alignment between different taxa, which allows us to make sure that the phylogenetic tree is robust. But hypervariable regions accumulate mutations at a higher rate, making them useful for distinguishing closely related taxa.

4) MAKING A PHYLOGENETIC TREE

Once you have aligned your sequences, the next step is to construct a phylogenetic tree. Not only is a phylogenetic tree effective for visualizing the evolutionary relationship among taxa, but as you will see later, the information that goes into a phylogenetic tree is needed for downstream analysis.

A. Neighbor Joining Trees

In the R code chunk below, do the following:
1. calculate the distance matrix using `model = "raw"`,
2. create a Neighbor Joining tree based on these distances,
3. define "Methanosarcina" as the outgroup and root the tree, and
4. plot the rooted tree.

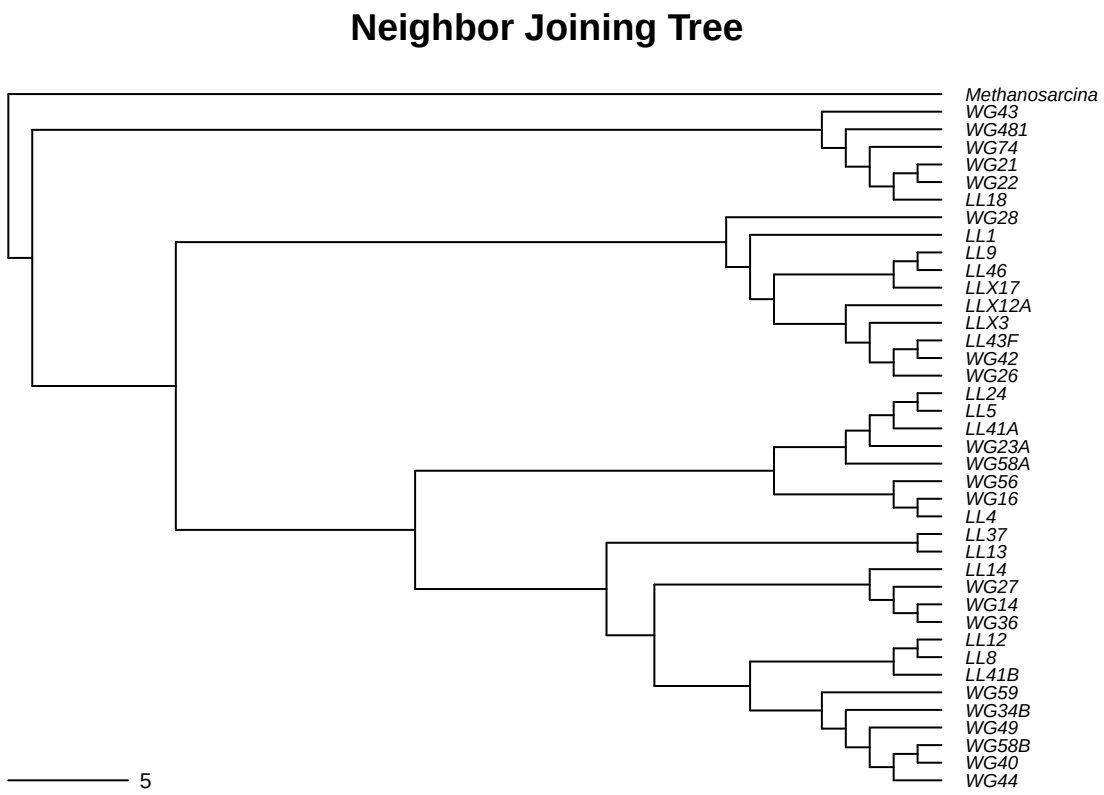
```
# Create Distance Matrix with "raw" Model {ape}
seq.dist.raw <- dist.dna(p.DNAbin, model = "raw", pairwise.deletion = FALSE)

# Neighbor Joining Algorithm to Construct Tree, a 'phylo' Object {ape}
nj.tree <- bionj(seq.dist.raw)

# Identify Outgroup Sequence
outgroup <- match("Methanosarcina", nj.tree$tip.label)

# Root the Tree {ape}
nj.rooted <- root(nj.tree, outgroup, resolve.root = TRUE)

# Plot the Rooted Tree{ape}
par(mar = c(1, 1, 2, 1) + 0.1)
plot.phylo(nj.rooted, main = "Neighbor Joining Tree", "phylogram",
           use.edge.length = FALSE, direction = "right", cex = 0.6,
           label.offset = 1)
add.scale.bar(cex = 0.7)
```



Question 3: What are the advantages and disadvantages of making a neighbor joining tree?

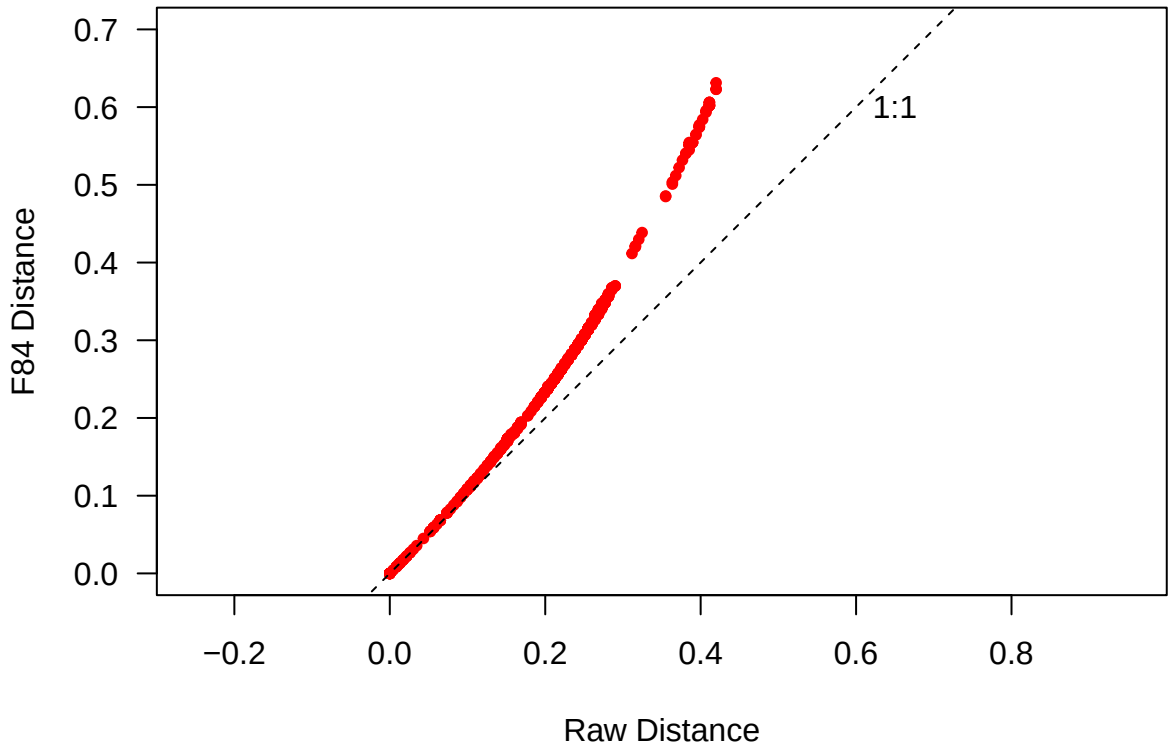
> ***Answer 3***:
> The advantages of NJ method is it is fast enough and quite intuitive to understand. It does not require complex optimization algorithms like Maximum Likelihood (ML) or Bayesian Inference (BI). It can also work with missing data, and it provides a reasonable approximation of phylogenetic relationships, especially for closely related taxa. (Closer species are more likely to have similar sequences.)
>
> The disadvantages of NJ method is it is not as accurate as ML or BI methods. It does not not provide branch values like bootstrap or posterior probabilities, making it hard to assess confidence in the inferred relationships. Unlike ML or Bayesian methods, NJ does not reconstruct ancestral states or allow for variable evolutionary rates among lineages.

B) SUBSTITUTION MODELS OF DNA EVOLUTION

In the R code chunk below, do the following:
1. make a second distance matrix based on the Felsenstein 84 substitution model,
2. create a saturation plot to compare the *raw* and *Felsenstein (F84)* substitution models,
3. make Neighbor Joining trees for both, and
4. create a cophylogenetic plot to compare the topologies of the trees.

```
# Create distance matrix with "F84" model {ape}
seq.dist.F84 <- dist.dna(p.DNAbin, model = "F84", pairwise.deletion = FALSE)
```

```r
# Plot Distances from Different DNA Substitution Models
par(mar = c(5, 5, 2, 1) + 0.1)
plot(seq.dist.raw, seq.dist.F84,
     pch = 20, col = "red", las = 1, asp = 1, xlim = c(0, 0.7),
     ylim = c(0, 0.7), xlab = "Raw Distance", ylab = "F84 Distance")
abline(b = 1, a = 0, lty = 2)
text(0.65, 0.6, "1:1")
```



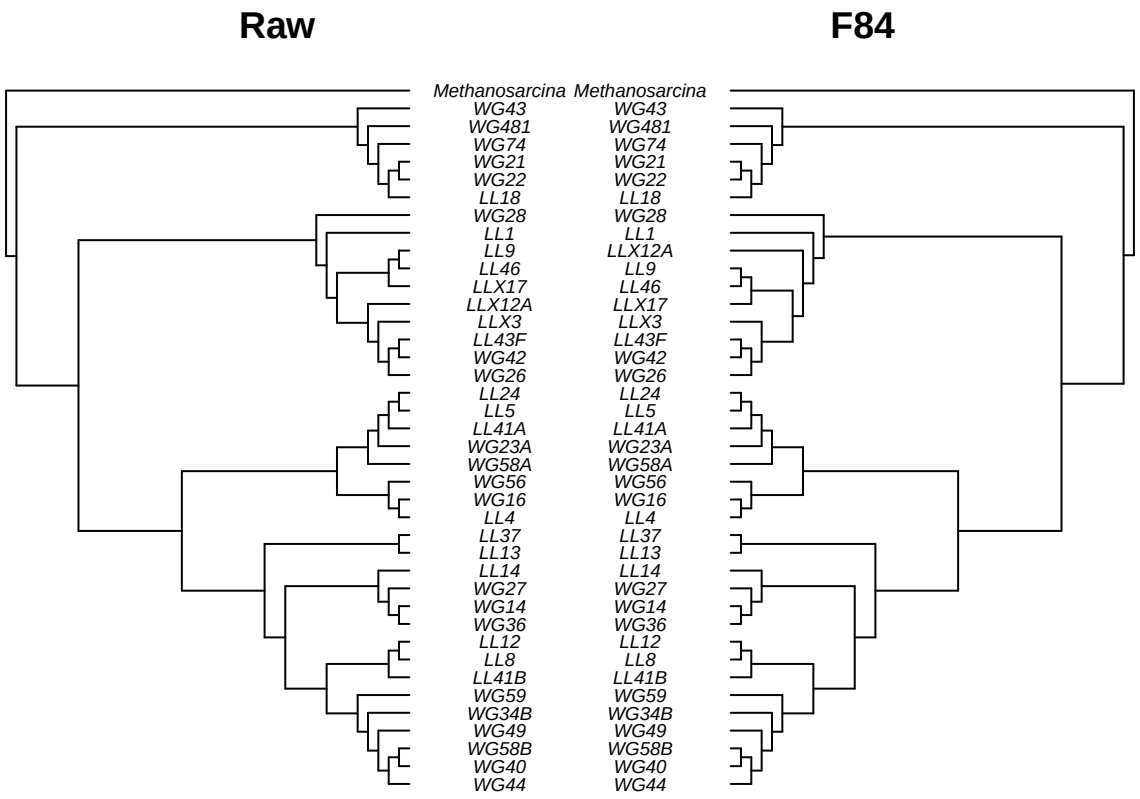

```r
# Make Neighbor Joining Trees Using Different DNA Substitution Models {ape}
raw.tree <- bionj(seq.dist.raw)
F84.tree <- bionj(seq.dist.F84)

# Define Outgroups
raw.outgroup <- match("Methanosarcina", raw.tree$tip.label)
F84.outgroup <- match("Methanosarcina", F84.tree$tip.label)

# Root the Trees {ape}
raw.rooted <- root(raw.tree, raw.outgroup, resolve.root = TRUE)
F84.rooted <- root(F84.tree, F84.outgroup, resolve.root = TRUE)

# Make Cophylogenetic Plot {ape}
layout(matrix(c(1, 2), 1, 2), width = c(1, 1))
par(mar = c(1, 1, 2, 0))
plot.phylo(raw.rooted, type = "phylogram", direction = "right",
      show.tip.label = TRUE, use.edge.length = FALSE, adj = 0.5,
```

```
      cex = 0.6, label.offset = 2, main = "Raw")
par(mar = c(1, 0, 2, 1))
plot.phylo(F84.rooted, type = "phylogram", direction = "left",
      show.tip.label = TRUE, use.edge.length = FALSE, adj = 0.5,
      cex = 0.6, label.offset = 2, main = "F84")
```



We see here that the scatter plot compares distances computed using raw pairwise differences and the F84 substitution model. The points below the diagonal indicate that the F84 model corrects for multiple substitutions, resulting in larger evolutionary distances than raw distances.

Both trees exhibit similar overall topology, suggesting that simple pairwise distances can still recover major evolutionary relationships. However, branch lengths in the F84 tree appear slightly longer, reflecting the correction for substitution biases. I think the F84 model provides a more accurate representation of evolutionary distances by accounting for mutational processes.

C) ANALYZING A MAXIMUM LIKELIHOOD TREE

In the R code chunk below, do the following:
1. Read in the maximum likelihood phylogenetic tree used in the handout. 2. Plot bootstrap support values onto the tree

```
# Set method = "PH85" for the symmetric difference
# Set method = "score" for the symmetric difference
# This function automatically checks for a root and unroots rooted trees
```

```r
# Can then pass it either the rooted or unrooted tree and get same answer
dist.topo(raw.rooted, F84.rooted, method = "score")
```

```
##            tree1
## tree2 0.04219896
```

```r
# Requires alignment to be read in with as phyDat object
phyDat.aln <- msaConvert(read.aln, type = "phangorn::phyDat")

# Make the NJ tree for the maximum likelihood method.
# {Phangorn} requires a specific attribute (attr) class.
# So we need to remake our trees with the following code:
aln.dist <- dist.ml(phyDat.aln)
aln.NJ <- NJ(aln.dist)

fit <- pml(tree = aln.NJ, data = phyDat.aln)

# Fit tree using a JC69 substitution model
fitJC <- optim.pml(fit, TRUE)
```

```
## optimize edge weights:  -10571.04 --> -10396.64
## optimize edge weights:  -10396.64 --> -10396.64
## optimize topology:  -10396.64 --> -10341.45  NNI moves:  10
## optimize edge weights:  -10341.45 --> -10341.45
## optimize topology:  -10341.45 --> -10341.45  NNI moves:  0
```

```r
# Fit tree using a GTR model with gamma distributed rates.
fitGTR <- optim.pml(fit, model = "GTR", optInv = TRUE, optGamma = TRUE,
          rearrangement = "NNI", control = pml.control(trace = 0))
```

```
## only one rate class, ignored optGamma
```

```r
# Perform model selection with either an ANOVA test or with AIC
anova(fitJC, fitGTR)
```

```
## Likelihood Ratio Test Table
##   Log lik. Df Df change Diff log lik. Pr(>|Chi|)
## 1 -10341.5 77
## 2  -9790.4 86         9        1102.1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

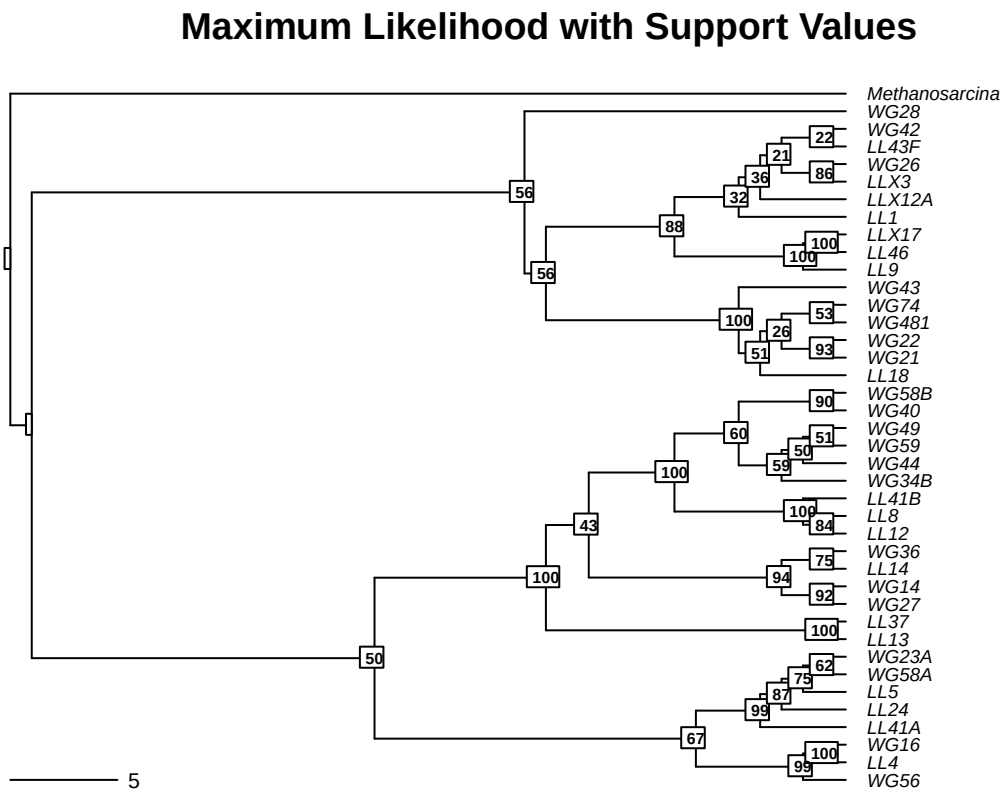
```r
AIC(fitJC)
```

```
## [1] 20836.9
```

```r
AIC(fitGTR)
```

```
## [1] 19752.84
```

```
ml.bootstrap <- read.tree("./data/ml_tree/RAxML_bipartitions.T1")
par(mar = c(1, 1, 2, 1) + 0.1)
plot.phylo(ml.bootstrap, type = "phylogram", direction = "right",
      show.tip.label = TRUE, use.edge.length = FALSE, cex = 0.6,
      label.offset = 1, main = "Maximum Likelihood with Support Values")
add.scale.bar(cex = 0.7)
nodelabels(ml.bootstrap$node.label, font = 2, bg = "white",
      frame = "r", cex = 0.5)
```



Question 4:

a) How does the maximum likelihood tree compare the to the neighbor-joining tree in the handout? If the plots seem to be inconsistent with one another, explain what gives rise to the differences.

b) Why do we bootstrap our tree?

c) What do the bootstrap values tell you?

d) Which branches have very low support?

e) Should we trust these branches? Why or why not?

Answer 4a:
I think both the maximum likelihood (ML) tree and the neighbor-joining (NJ) tree have similar overall topology, with some minor differences in branch lengths. The ML tree has bootstrap support values, which provide an estimate of the confidence in the inferred relationships. Also, ML is a model-based approach that optimizes branch lengths and relationships by maximizing

the probability of observing the data given a substitution model. Hemce, the differences between the trees may arise due to long-branch attraction, different handling of missing data, or maybe model assumptions in ML that NJ does not account for. ***Answer 4b***:
Bootstrapping helps determine which branches are robust and which are poorly supported due to noise or limited data since we randomly resample the dataset with replacement and reconstruct the tree multiple times.

Answer 4c:
Bootstrap values show the percentage of times a specific branch appeared across resampled datasets. High bootstrap values (70-80%) indicate strong support for a given branching pattern. Low bootstrap values ($< 50\%$) suggest that the branching structure is unstable and may not be reliable.

Answer 4d:
The least supported branches are 342, LL43F

Answer 4e:
The most supported branches are the higher groups with higher number of boostrap score like W340, W360B

5) INTEGRATING TRAITS AND PHYLOGENY

A. Loading Trait Database

In the R code chunk below, do the following:
1. import the raw phosphorus growth data, and
2. standardize the data for each strain by the sum of growth rates.

```
# Import Growth Rate Data
p.growth <- read.table("./data/p.isolates.raw.growth.txt", sep = "\t",
            header = TRUE, row.names = 1)

# Standardize Growth Rates Across Strains
p.growth.std <- p.growth / (apply(p.growth, 1, sum))
```

B. Trait Manipulations

In the R code chunk below, do the following:
1. calculate the maximum growth rate (μ_{max}) of each isolate across all phosphorus types,
2. create a function that calculates niche breadth (nb), and
3. use this function to calculate nb for each isolate.

```
# Calculate Max Growth Rate
umax <- (apply(p.growth, 1, max))

levins <- function(p_xi = ""){
  p = 0
  for (i in p_xi){
    p = p + i^2
    }
  nb = 1 / (length(p_xi) * p)
  return(nb)
  }
```

```r
# Calculate Niche Breadth for Each Isolate
nb <- as.matrix(levins(p.growth.std))
# Add Row Names to Niche Breadth Matrix
nb <- setNames(as.vector(nb), as.matrix(row.names(p.growth)))
```

C. Visualizing Traits on Trees

In the R code chunk below, do the following:
1. pick your favorite substitution model and make a Neighbor Joining tree,
2. define your outgroup and root the tree, and
3. remove the outgroup branch.

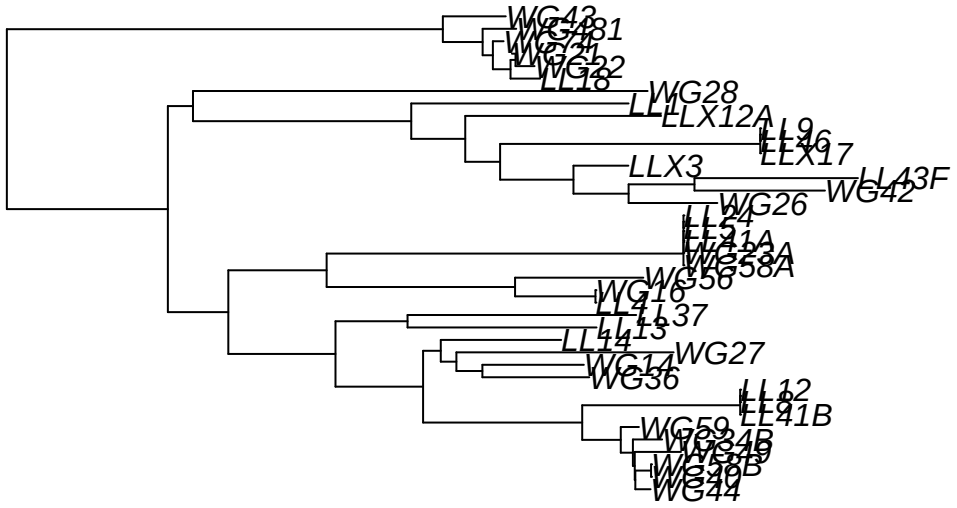
```r
# Generate Neighbor Joining Tree Using F84 DNA Substitution Model {ape}
nj.tree <- bionj(seq.dist.F84)

# Define the Outgroup
outgroup <- match("Methanosarcina", nj.tree$tip.label)

# Create a Rooted Tree {ape}
nj.rooted <- root(nj.tree, outgroup, resolve.root = TRUE)

# Keep Rooted but Drop Outgroup Branch
nj.rooted <- drop.tip(nj.rooted, "Methanosarcina")

# Plot to look at our tree
plot(nj.rooted)
```

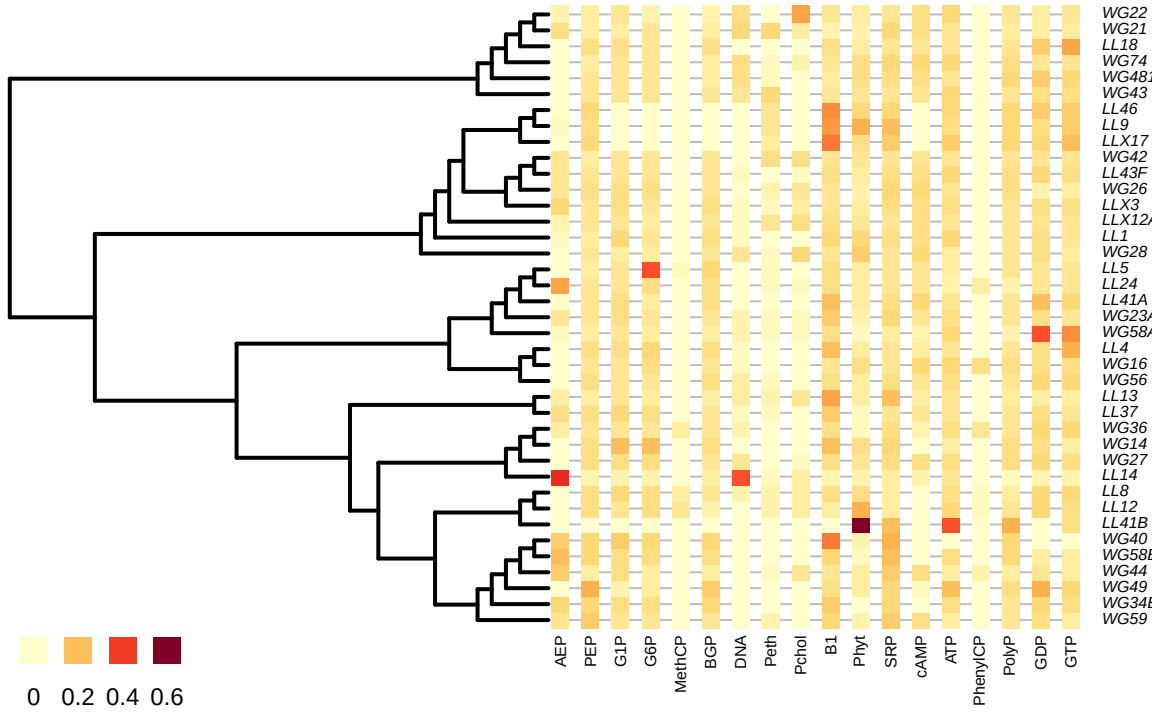


In the R code chunk below, do the following:
1. define a color palette (use something other than "YlOrRd"),
2. map the phosphorus traits onto your phylogeny,
3. map the *nb* trait on to your phylogeny, and
4. customize the plots as desired (use 'help(table.phylo4d)' to learn about the options).

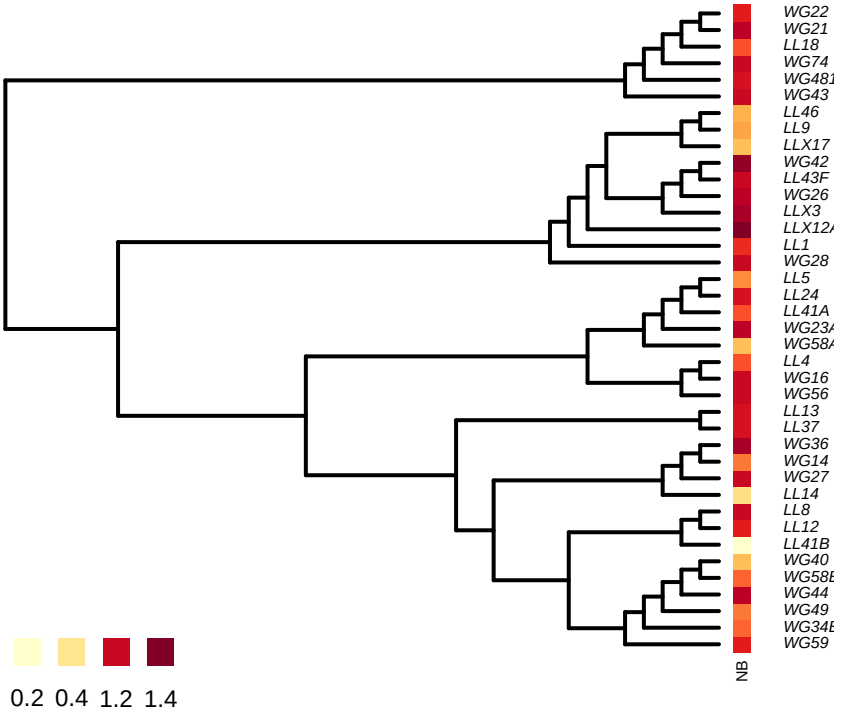
```
# Define Color Palette
mypalette <- colorRampPalette(brewer.pal(9, "YlOrRd"))

# First, Correct for Zero Branch-Lengths on Our Tree
nj.plot <- nj.rooted
nj.plot$edge.length <- nj.plot$edge.length + 10^-1

# Map Phosphorus Traits {adephylo}
par(mar = c(1, 1, 1, 1) + 0.1)
x <- phylo4d(nj.plot, p.growth.std)
table.phylo4d(x, treetype = "phylo", symbol = "colors", show.node = TRUE,
              cex.label = 0.5, scale = FALSE, use.edge.length = FALSE,
              edge.color = "black", edge.width = 2, box = FALSE,
              col=mypalette(25), pch = 15, cex.symbol = 1.25,
              ratio.tree = 0.5, cex.legend = 1.5, center = FALSE)
```




```
# Niche Breadth
par(mar = c(1, 5, 1, 5) + 0.1)
x.nb <- phylo4d(nj.plot, nb)
table.phylo4d(x.nb, treetype = "phylo", symbol = "colors", show.node = TRUE,
      cex.label = 0.5, scale = FALSE, use.edge.length = FALSE,
      edge.color = "black", edge.width = 2, box = FALSE, col = mypalette(25),
      pch = 15, cex.symbol = 1.25, var.label = ("NB"),ratio.tree = 0.90,
      cex.legend = 1.5, center = FALSE)
```



Question 5:

a) Develop a hypothesis that would support a generalist-specialist trade-off.

b) What kind of patterns would you expect to see from growth rate and niche breadth values that would support this hypothesis?

Answer 5a:
Answer 5a: I think, we can think that microbial strains may have higher growth rates on a specific phosphorus source will have a lower niche breadth, whereas strains that can use broader range of phosphorus sources will have lower maximum growth rates. This hypothesis illustrates the fact that there is a trade-off between being highly efficient at utilizing a single resource versus being able to tolerate a wide range of resources but with reduced efficiency.

Answer 5b: I think if we should expect to have higher maximum growth rates on a single phosphorus source, but lower niche breadth values for specialists since they are specialized in a narrow range of phosphorus conditions. Also, they should be more closely related to each other. For the generalists, we should expect to have lower maximum growth rates on any single phosphorus source, but higher niche breadth values, indicating the ability to grow across many phosphorus conditions. They should be more phylogenetically widespread, as generalist traits may arise in multiple evolutionary lineages. Also, if there is a trade-off, I think we should see that heatmap colors should show specialists in one or two strong colors (high growth in a specific condition) and generalists with more balanced distributions across all phosphorus types.

6) HYPOTHESIS TESTING

Phylogenetic Signal: Pagel's Lambda

In the R code chunk below, do the following:
1. create two rescaled phylogenetic trees using lambda values of 0.5 and 0,
2. plot your original tree and the two scaled trees, and
3. label and customize the trees as desired.

```
# Visualize Trees With Different Levels of  Phylogenetic Signal {geiger}

# Load the libraries
library(geiger)
library(phytools)
library(picante)
library(caper)

# Ensure branch lengths are positive
nj.rooted$edge.length <- nj.rooted$edge.length + 10^-7

# Rescale tree with lambda = 0.5 (partial phylogenetic signal)
nj.lambda.5 <- lambdaTree(nj.rooted, lambda = 0.5)
```

```
## Warning in .deprecate("lambdaTree", "rescale.phylo"): 'lambdaTree' is being
## deprecated: use 'rescale.phylo' instead
```

```
# Rescale tree with lambda = 0 (no phylogenetic signal)
nj.lambda.0 <- lambdaTree(nj.rooted, lambda = 0)
```

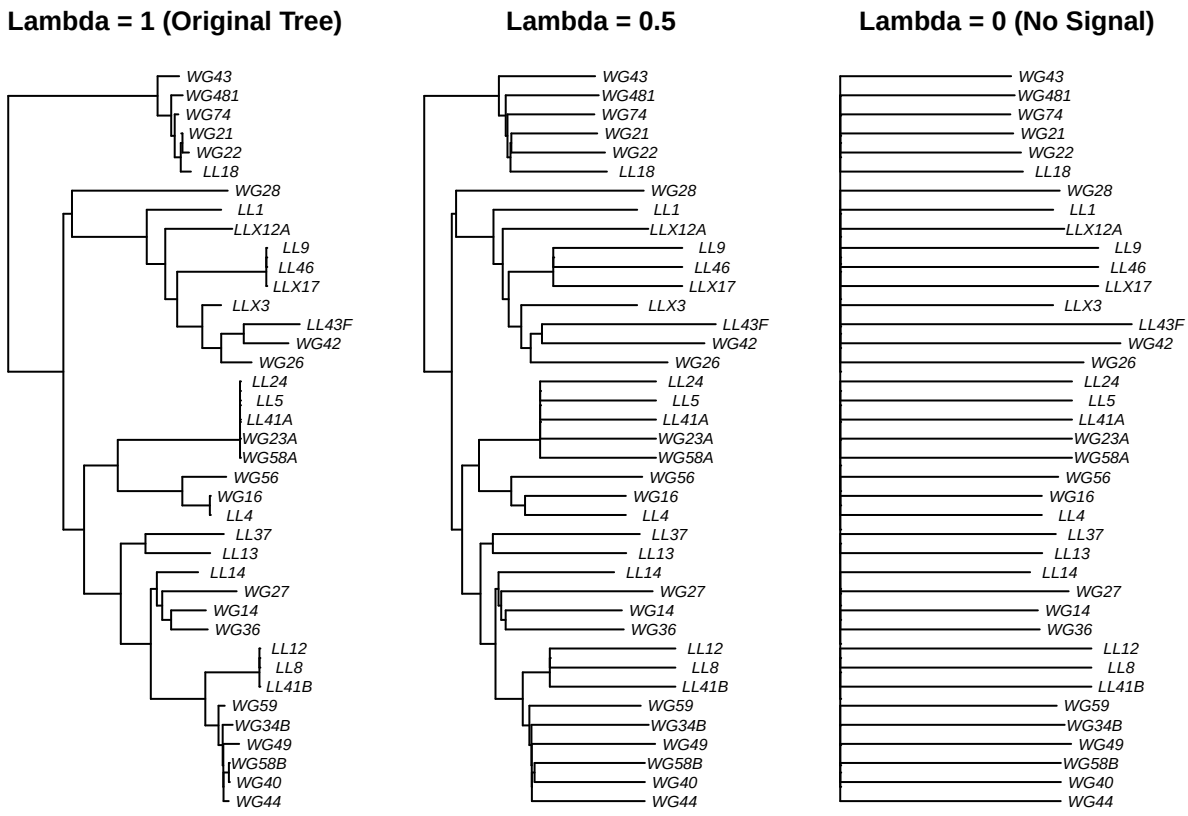
```
## Warning in .deprecate("lambdaTree", "rescale.phylo"): 'lambdaTree' is being
## deprecated: use 'rescale.phylo' instead
```

```
# Set up the plotting layout (3 trees side by side)
layout(matrix(c(1, 2, 3), 1, 3), width = c(1, 1, 1))
par(mar = c(1, 1, 2, 1) + 0.1)  # Adjust margins

# Plot Original Tree (Lambda = 1)
plot(nj.rooted, main = "Lambda = 1 (Original Tree)", cex = 0.7, adj = 0.5)

# Plot Rescaled Tree (Lambda = 0.5)
plot(nj.lambda.5, main = "Lambda = 0.5", cex = 0.7, adj = 0.5)

# Plot Rescaled Tree (Lambda = 0)
plot(nj.lambda.0, main = "Lambda = 0 (No Signal)", cex = 0.7, adj = 0.5)
```



In the R code chunk below, do the following:
1. use the 'fitContinuous()' function to compare your original tree to the transformed trees.

```
fitContinuous(nj.rooted, nb, model = "lambda")
```

```
## GEIGER-fitted comparative model of continuous data
##  fitted 'lambda' model parameters:
##  lambda = 0.006975
##  sigsq = 0.108060
##  z0 = 0.657697
##
##  model summary:
##  log-likelihood = 21.503414
##  AIC = -37.006829
##  AICc = -36.321114
##  free parameters = 3
##
## Convergence diagnostics:
##  optimization iterations = 100
##  failed iterations = 0
##  number of iterations with same best fit = 16
##  frequency of best fit = 0.160
##
##  object summary:
##  'lik' -- likelihood function
##  'bnd' -- bounds for likelihood search
```

```
##  'res' -- optimization iteration summary
##  'opt' -- maximum likelihood parameter estimates
```

```
fitContinuous(nj.lambda.0, nb, model = "lambda")
```

```
## GEIGER-fitted comparative model of continuous data
##  fitted 'lambda' model parameters:
##  lambda = 0.000000
##  sigsq = 0.108048
##  z0 = 0.656477
##
##  model summary:
##  log-likelihood = 21.502505
##  AIC = -37.005011
##  AICc = -36.319297
##  free parameters = 3
##
## Convergence diagnostics:
##  optimization iterations = 100
##  failed iterations = 0
##  number of iterations with same best fit = 89
##  frequency of best fit = 0.890
##
##  object summary:
##  'lik' -- likelihood function
##  'bnd' -- bounds for likelihood search
##  'res' -- optimization iteration summary
##  'opt' -- maximum likelihood parameter estimates
```

```
# Compare Pagel's lambda score with likelihood ratio test
# Lambda = 0, no phylogenetic signal
phylosig(nj.rooted, nb, method = "lambda", test = TRUE)
```

```
##
## Phylogenetic signal lambda : 0.00699101
## logL(lambda) : 21.5034
## LR(lambda=0) : 0.00181762
## P-value (based on LR test) : 0.965994
```

Question 6: There are two important outputs from the 'fitContinuous()' function that can help you interpret the phylogenetic signal in trait data sets. a. Compare the lambda values of the untransformed tree to the transformed (lambda = 0). b. Compare the Akaike information criterion (AIC) scores of the two models. Which model would you choose based off of AIC score (remember the criteria that the difference in AIC values has to be at least 2)? c. Does this result suggest that there is phylogenetic signal?

> ***Answer 6a***:
> We see that the lambda value for the untransformed tree (nj.rooted) is 0.0069, which is very close to zero. The lambda value for the transformed tree (nj.lambda.0) is 0. Since lambda is a measure of phylogenetic signal, these low values suggest that niche breadth does not follow strong phylogenetic structuring.
>
> ***Answer 6b***:
> The AIC value for the untransformed tree is 0.000, while the AIC value for the transformed tree

is 0.002. We dont see a significant difference in AIC values between the two models, this means that two models are equivalent, which means that accounting for phylogeny does not improve the model.

Answer 6c:
This result does not suggest strong phylogenetic signal. The P-value from the likelihood ratio test (0.965) is very high, meaning we fail to reject the null hypothesis that lambda = 0. This suggests that niche breadth is not significantly influenced by evolutionary history.

7) PHYLOGENETIC REGRESSION

Question 7: In the R code chunk below, do the following:
1. Clean the resource use dataset to perform a linear regression to test for differences in maximum growth rate by niche breadth and lake environment. 2. Fit a linear model to the trait dataset, examining the relationship between maximum growth rate by niche breadth and lake environment, 2. Fit a phylogenetic regression to the trait dataset, taking into account the bacterial phylogeny

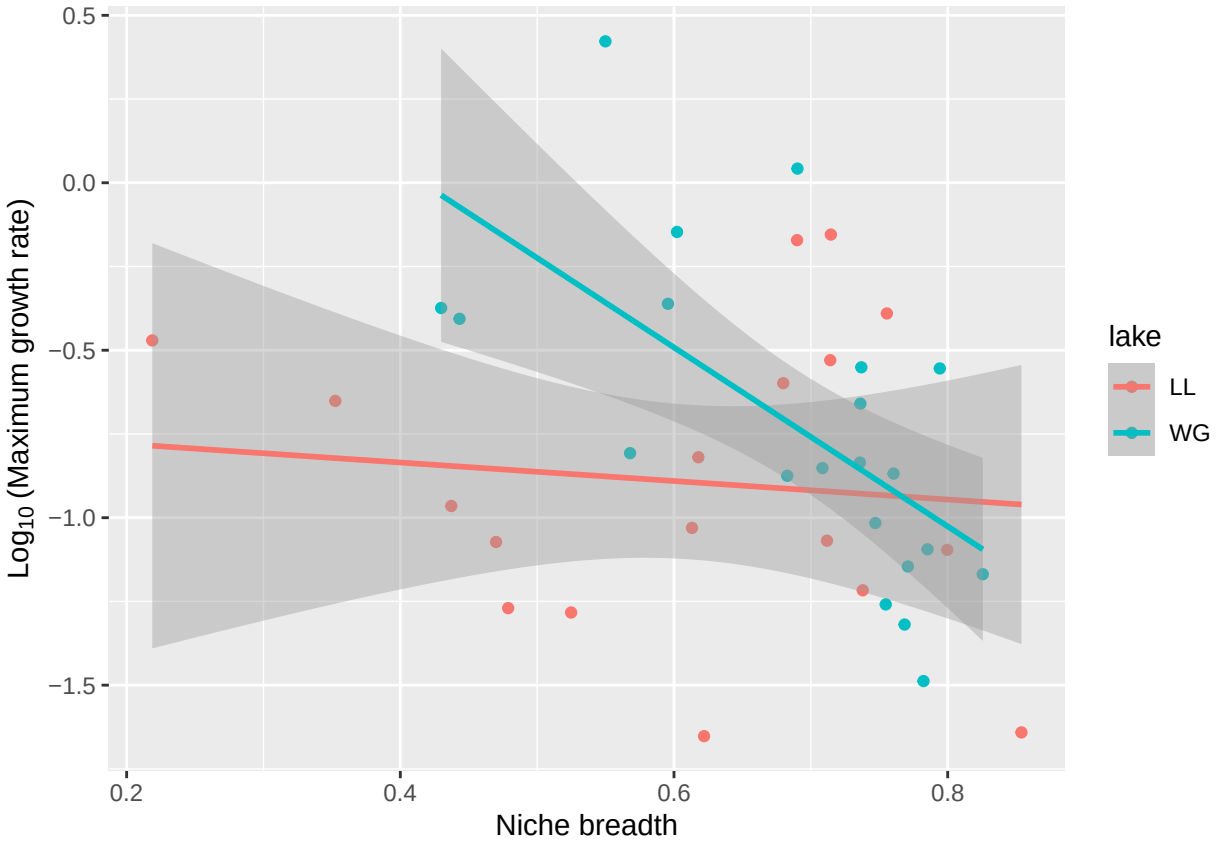
```
# Using the niche breadth data from above
nb.lake = as.data.frame(as.matrix(nb))
nb.lake$lake = rep('A')
# Assign lakes to the data
for(i in 1:nrow(nb.lake)) {
    ifelse(grepl("WG",row.names(nb.lake)[i]), nb.lake[i,2] <- "WG",
        nb.lake[i,2] <- "LL")
}

#Add a meaningful column name to the niche breadth values
colnames(nb.lake)[1] <- "NB"

#Calculate the max growth rate
umax <- as.matrix((apply(p.growth, 1, max)))
nb.lake = cbind(nb.lake,umax)

# Plot maximum growth rate by niche breadth
ggplot(data = nb.lake,aes(x = NB, y = log10(umax), color = lake)) +
  geom_point() +
  geom_smooth(method = "lm") +
  xlab("Niche breadth") +
  ylab(expression(Log[10]~"(Maximum growth rate)"))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
# Simple linear regression
fit.lm <- lm(log10(umax) ~ NB*lake,data = nb.lake)
summary(fit.lm)
```

```
##
## Call:
## lm(formula = log10(umax) ~ NB * lake, data = nb.lake)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7557 -0.3108 -0.1077  0.3102  0.7800
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.7247     0.3852  -1.882   0.0682 .
## NB           -0.2763     0.6097  -0.453   0.6533
## lakeWG        1.8364     0.6909   2.658   0.0118 *
## NB:lakeWG    -2.3958     1.0234  -2.341   0.0251 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.418 on 35 degrees of freedom
## Multiple R-squared:  0.2595, Adjusted R-squared:  0.196
## F-statistic: 4.089 on 3 and 35 DF,  p-value: 0.01371
```

```r
AIC(fit.lm)
```

```
## [1] 48.413
```

```r
# Run a phylogeny-corrected regression with no bootstrap replicates
fit.plm <- phylolm(log10(umax) ~ NB*lake, data = nb.lake, nj.rooted,
          model = "lambda", boot = 0)
summary(fit.plm)
```

```
##
## Call:
## phylolm(formula = log10(umax) ~ NB * lake, data = nb.lake, phy = nj.rooted,
##     model = "lambda", boot = 0)
##
##    AIC logLik
##  41.08 -14.54
##
## Raw residuals:
##      Min       1Q   Median       3Q      Max
## -0.75804 -0.18999 -0.07425  0.32496  0.95857
##
## Mean tip height: 0.1814508
## Parameter estimate(s) using ML:
## lambda : 0.4861386
## sigma2: 0.9184409
##
## Coefficients:
##               Estimate     StdErr t.value p.value
## (Intercept) -0.8912676  0.3700360 -2.4086 0.02142 *
## NB          -0.0048049  0.5213029 -0.0092 0.99270
## lakeWG       1.4389308  0.5772311  2.4928 0.01755 *
## NB:lakeWG   -1.9663889  0.8487018 -2.3169 0.02648 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-squared: 0.1935    Adjusted R-squared: 0.1243
##
## Note: p-values and R-squared are conditional on lambda=0.4861386.
```

```r
AIC(fit.plm)
```

```
## [1] 41.07572
```

a. Why do we need to correct for shared evolutionary history?
b. How does a phylogenetic regression differ from a standard linear regression?
c. Interpret the slope and fit of each model. Did accounting for shared evolutionary history improve or worsen the fit?
d. Try to come up with a scenario where the relationship between two variables would completely disappear when the underlying phylogeny is accounted for.

Answer 7a:
I think in general, closely related species share traits due to their common ancestry. This violated

the assumption of independence in most statistical tests. If we ignore this shared evolutionary history, we may assume that observed patterns are due to ecological or environmental factors when they are actually a result of inheritance. If we account for phylogeny, which in turns accounts for non-independence, this ensured that the statistical relationships we detect are not simply due to relatedness but reflect actual biological processes.

Answer 7b:
A standard linear regression assumes that residual errors are independent and identically distributed, meaning each data point contributes equally to the model. However, in a phylogenetic regression, residual errors account for evolutionary history. In this case, species that are more closely related are expected to have more similar traits. The variance in the regression model is adjusted using a covariance matrix derived from the phylogenetic tree, making the estimates more reliable.

Answer 7c:
For the standard linear regression, we see that there is a significant effect of lake environment (lakeWG, p = 0.0118) and interaction (NB:lakeWG, p = 0.0251), but the effect of niche breadth (NB) is non-significant (p = 0.6533). The adjusted R^2 value is 0.196, meaning the model explains 19.6% of the variation. The AIC value is 48.41.

For the phylogenetic regression, we see that there is a significant effect of lake environment (lakeWG, p = 0.0176) and interaction (NB:lakeWG, p = 0.0265), but the effect of niche breadth (NB) is non-significant (p = 0.9927). The adjusted R2 value is 0.1243, meaning the model explains 12.4% of the variation. The AIC value is 41.08.

In this case, the phylogenetic regression model is slightly worse than the standard linear regression model. However, the R^2 value slightly decreased (from 0.196 to 0.1243), meaning that once evolutionary history was accounted for, some of the explained variance was reduced. The significance of niche breadth (NB) remains non-significant in both models, but lake environment (lakeWG) is still an important predictor.

Answer 7d:
I think it would be plausible when there is when trait similarity is driven primarily by ancestry rather than ecological adaptation. For example, if we are analyzing the relationship between body size and metabolic rate in mammals, we might find a strong correlation in a standard linear regression, suggesting that larger mammals have lower metabolic rates. However, after performing a phylogenetic regression, we might find that this relationship is actually driven by common ancestry—for example, all primates have a certain metabolic rate pattern, and all rodents have another pattern. Once we account for shared ancestry, the relationship may no longer be statistically significant, meaning that the correlation was not a result of natural selection favoring lower metabolic rates in large mammals, but rather due to inherited traits from common ancestors

7) SYNTHESIS

Work with members of your Team Project to obtain reference sequences for 10 or more taxa in your study. Sequences for plants, animals, and microbes can found in a number of public repositories, but perhaps the most commonly visited site is the National Center for Biotechnology Information (NCBI) https://www.ncbi.nlm.nih.gov/. In almost all cases, researchers must deposit their sequences in places like NCBI before a paper is published. Those sequences are checked by NCBI employees for aspects of quality and given an **accession number**. For example, here an accession number for a fungal isolate that our lab has worked with: JQ797657. You can use the NCBI program nucleotide **BLAST** to find out more about information associated with the isolate, in addition to getting its DNA sequence: https://blast.ncbi.nlm.nih.gov/. Alternatively, you can use the 'read.GenBank()' function in the 'ape' package to connect to NCBI and directly get the sequence. This is pretty cool. Give it a try.

But before your team proceeds, you need to give some thought to which gene you want to focus on. For microorganisms like the bacteria we worked with above, many people use the ribosomal gene (i.e., 16S rRNA).

This has many desirable features, including it is relatively long, highly conserved, and identifies taxa with reasonable resolution. In eukaryotes, ribosomal genes (i.e., 18S) are good for distinguishing course taxonomic resolution (i.e. class level), but it is not so good at resolving genera or species. Therefore, you may need to find another gene to work with, which might include protein-coding gene like cytochrome oxidase (COI) which is on mitochondria and is commonly used in molecular systematics. In plants, the ribulose-bisphosphate carboxylase gene (*rbcL*), which on the chloroplast, is commonly used. Also, non-protein-encoding sequences like those found in **Internal Transcribed Spacer (ITS)** regions between the small and large subunits of of the ribosomal RNA are good for molecular phylogenies. With your team members, do some research and identify a good candidate gene.

After you identify an appropriate gene, download sequences and create a properly formatted fasta file. Next, align the sequences and confirm that you have a good alignment. Choose a substitution model and make a tree of your choice. Based on the decisions above and the output, does your tree jibe with what is known about the evolutionary history of your organisms? If not, why? Is there anything you could do differently that would improve your tree, especially with regard to future analyses done by your team?

SUBMITTING YOUR ASSIGNMENT

Use Knitr to create a PDF of your completed '8.PhyloTraits_Worksheet.Rmd' document, push it to GitHub, and create a pull request. Please make sure your updated repo include both the pdf and RMarkdown files. Unless otherwise noted, this assignment is due on **Wednesday, February 26th, 2025 at 12:00 PM (noon)**.