8. Worksheet: Phylogenetic Diversity - Traits

Jocelyn Huang; Z620: Quantitative Biodiversity, Indiana University

27 February, 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).

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

```
## [1] "/cloud/project/QB2025_Huang/Week6-PhyloTraits"
```

```
setwd("/cloud/project/QB2025_Huang/Week6-PhyloTraits")

package.list <-c("ape","seqinr","phylobase","adephylo","geiger","picante","stats","RColorBrewer","caper

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

```
##
## Attaching package: 'seqinr'

## The following objects are masked from 'package:ape':
##
##     as.alignment, consensus

##
## Attaching package: 'phylobase'

## The following object is masked from 'package:ape':
##
##     edges

##
## Attaching package: 'phytools'

## The following object is masked from 'package:phylobase':
##
##     readNexus

##
## Attaching package: 'permute'

## The following object is masked from 'package:seqinr':
##
##     getType

##
## Attaching package: 'vegan'

## The following object is masked from 'package:phytools':
##
##     scores

##
## Attaching package: 'nlme'

## The following object is masked from 'package:seqinr':
##
##     gls

##
## Attaching package: 'dplyr'

## The following object is masked from 'package:MASS':
##
##     select
```

```
## The following object is masked from 'package:nlme':
##
##     collapse

## The following object is masked from 'package:seqinr':
##
##     count

## The following object is masked from 'package:ape':
##
##     where

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

##
## Attaching package: 'phangorn'

## The following objects are masked from 'package:vegan':
##
##     diversity, treedist

##
## Attaching package: 'cluster'

## The following object is masked from 'package:maps':
##
##     votes.repub

## 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))
## ---------------------

##
## Attaching package: 'dendextend'

## The following object is masked from 'package:permute':
##
##     shuffle
```

```
## The following object is masked from 'package:geiger':
##
##     is.phylo

## The following object is masked from 'package:phytools':
##
##     untangle

## The following objects are masked from 'package:phylobase':
##
##     labels<-, prune

## The following objects are masked from 'package:ape':
##
##     ladderize, rotate

## The following object is masked from 'package:stats':
##
##     cutree

##
## Attaching package: 'phylogram'

## The following object is masked from 'package:dendextend':
##
##     prune

## The following object is masked from 'package:phylobase':
##
##     prune

##
## Attaching package: 'amap'

## The following object is masked from 'package:vegan':
##
##     pca

##
## Attaching package: 'scales'

## The following object is masked from 'package:phytools':
##
##     rescale

## Warning in rgl.init(initValue, onlyNULL): RGL: unable to open X11 display

## Warning: 'rgl.init' failed, will use the null device.
## See '?rgl.useNULL' for ways to avoid this warning.
```

```r
# pak::pkg_install("msa")
library(msa)
```

```
## Loading required package: Biostrings

## Loading required package: BiocGenerics

##
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:dplyr':
##
##     combine, intersect, setdiff, union
```

```
## The following object is masked from 'package:ade4':
##
##     score

## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs

## The following objects are masked from '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

## Loading required package: S4Vectors

## Loading required package: stats4

##
## Attaching package: 'S4Vectors'

## The following objects are masked from 'package:dplyr':
##
##     first, rename

## The following object is masked from 'package:tidyr':
##
##     expand

## The following object is masked from 'package:utils':
##
##     findMatches

## The following objects are masked from 'package:base':
##
##     expand.grid, I, unname

## Loading required package: IRanges

##
## Attaching package: 'IRanges'

## The following objects are masked from 'package:dplyr':
##
##     collapse, desc, slice

## The following object is masked from 'package:nlme':
##
##     collapse

## Loading required package: XVector

## Found more than one class "Annotated" in cache; using the first, from namespace 'RNeXML'

## Also defined by 'S4Vectors'

## Found more than one class "Annotated" in cache; using the first, from namespace 'RNeXML'

## Also defined by 'S4Vectors'
```

```
## Found more than one class "Annotated" in cache; using the first, from namespace 'RNeXML'
## Also defined by 'S4Vectors'
## Found more than one class "Annotated" in cache; using the first, from namespace 'RNeXML'
## Also defined by 'S4Vectors'
## Found more than one class "Annotated" in cache; using the first, from namespace 'RNeXML'
## Also defined by 'S4Vectors'
## Found more than one class "Annotated" in cache; using the first, from namespace 'RNeXML'
## Also defined by 'S4Vectors'
## Loading required package: GenomeInfoDb
##
## Attaching package: 'Biostrings'
## The following object is masked from 'package:dendextend':
##
##     nnodes
## The following object is masked from 'package:seqinr':
##
##     translate
## The following object is masked from 'package:ape':
##
##     complement
## The following object is masked from '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

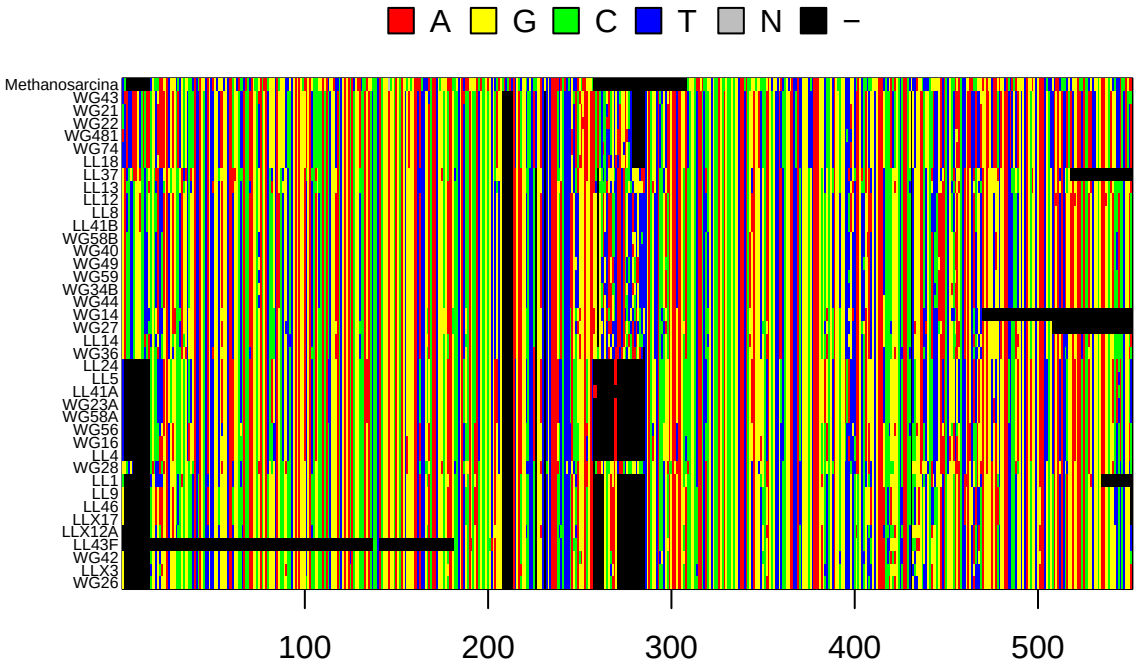
Question 1: Using your favorite text editor, compare the `p.isolates.fasta` file and the `p.isolates.afa` file. Describe the differences that you observe between the two files.

> ***Answer 1***: While both file types seperate sequences into samples of the two sites, sequences in the `.fasta` file does not conatin dashes that represent gaps, and `.afa` does.

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.

```
#if (!require("BiocManager", quietly = TRUE)){
#  install.packages("BiocManager")
#}
#BiocManager::install("Biostrings")
```

```
library(Biostrings)

# read the FASTA file:
seqs <- readDNAStringSet("data/p.isolates.fasta", format = "fasta")
# align the sequence using default MUSCLE parameter:
read.aln <- msaMuscle(seqs)

# convert to a DNAbin object:
p.DNAbin <- as.DNAbin(read.aln)
window <- p.DNAbin[, 200:750]
image.DNAbin(window, cex.lab = 0.5)
```



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. a. Approximately how long are our sequence reads? b. What regions do you think would are appropriate for phylogenetic inference and why?

> ***Answer 2a***: It's 1492 units long.
> ***Answer 2b***:
> I think region 250-450 and 500-750 would be appropriate because they seem to have a high degree of similarity.

4) MAKING A PHYLOGENETIC TREE

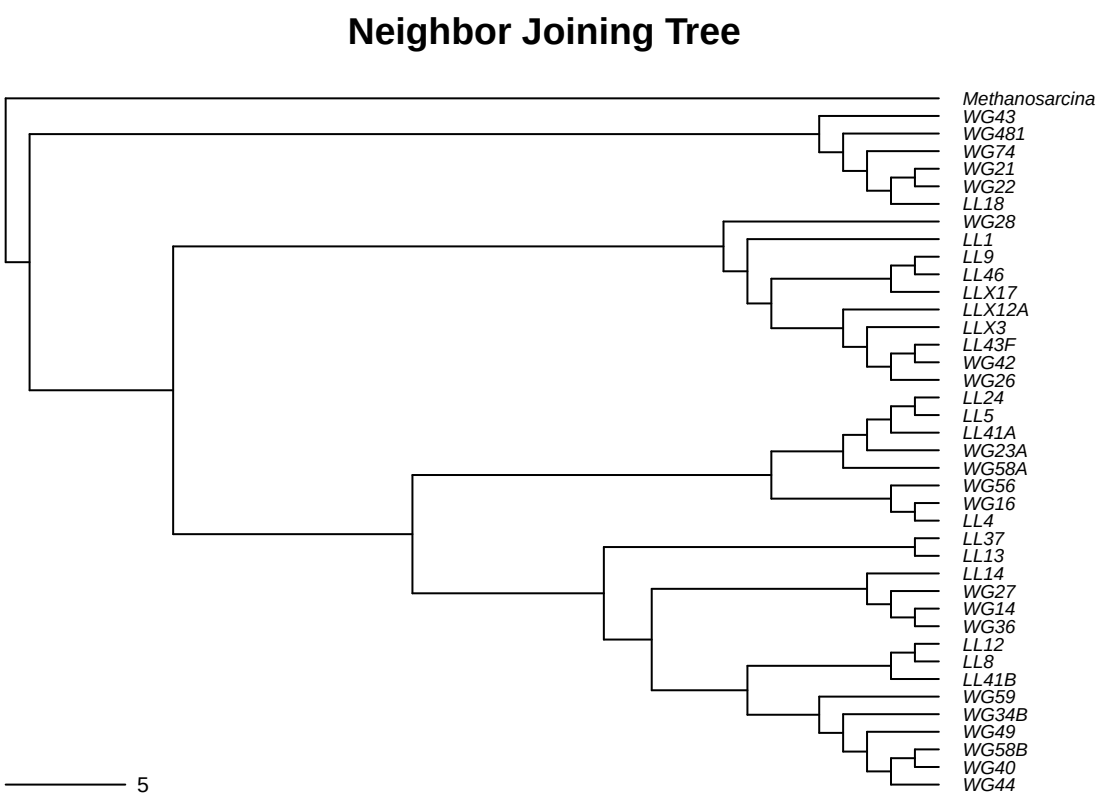
Once you have aligned your sequences, the next step is to construct a phylogenetic tree. Not only is a phylogenetic tree effective for visualizing the evolutionary relationship among taxa, but as you will see later, the information that goes into a phylogenetic tree is needed for downstream analysis.

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.

```
seq.dist.raw <- dist.dna(p.DNAbin, model = "raw", pairwise.deletion = FALSE) #distance matrix using raw

nj.tree <- bionj(seq.dist.raw) #create a Neighbor Joining tree
outgroup <- match("Methanosarcina", nj.tree$tip.label) #identify outgroup
nj.rooted <- root(nj.tree, outgroup, resolve.root = TRUE)

# Plot the rooted tree:
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?

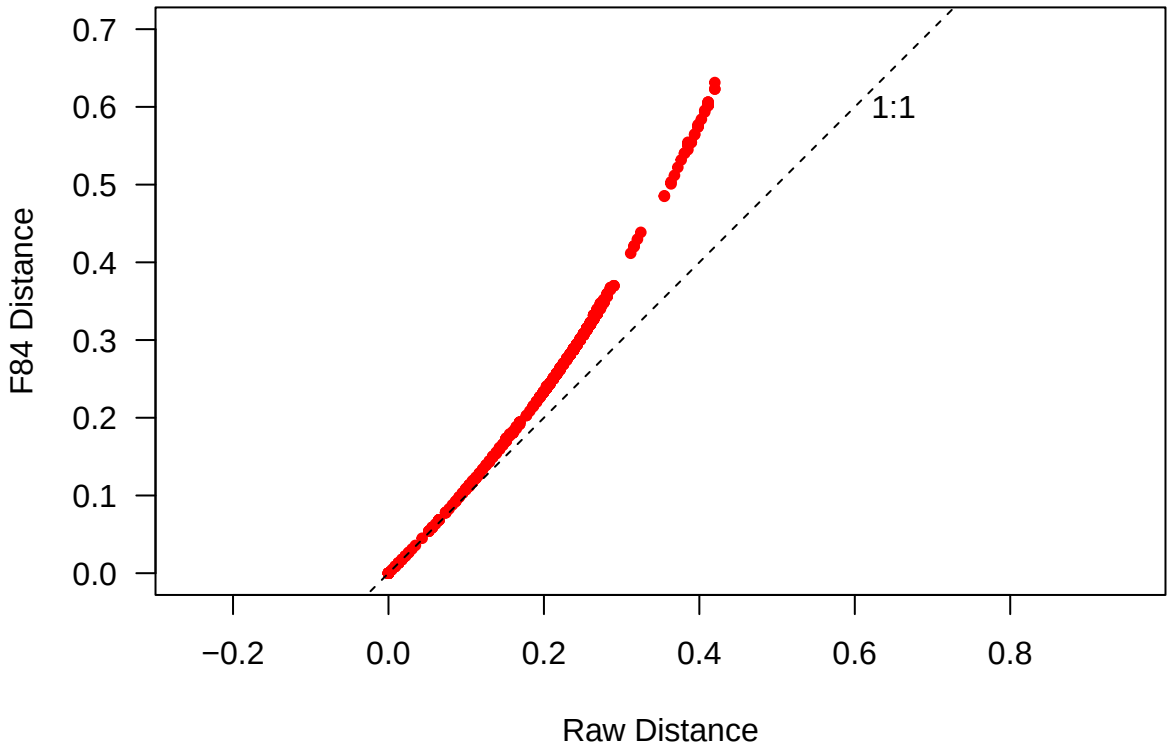
> ***Answer 3***:
> Neighbour joining tree is very simple and efficient to make since all one needs is a pairwise distance matrix that calculate the sequential difference between each taxa, and know which taxa is the outgroup. However, the tree that one can make by this method is only a "guide tree" and is relatively rough because it only takes into account the simplist substitution, while substitution can happen multiple times at one sites and many biases exist. It also does not take into account the specific nucleotide states.

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,

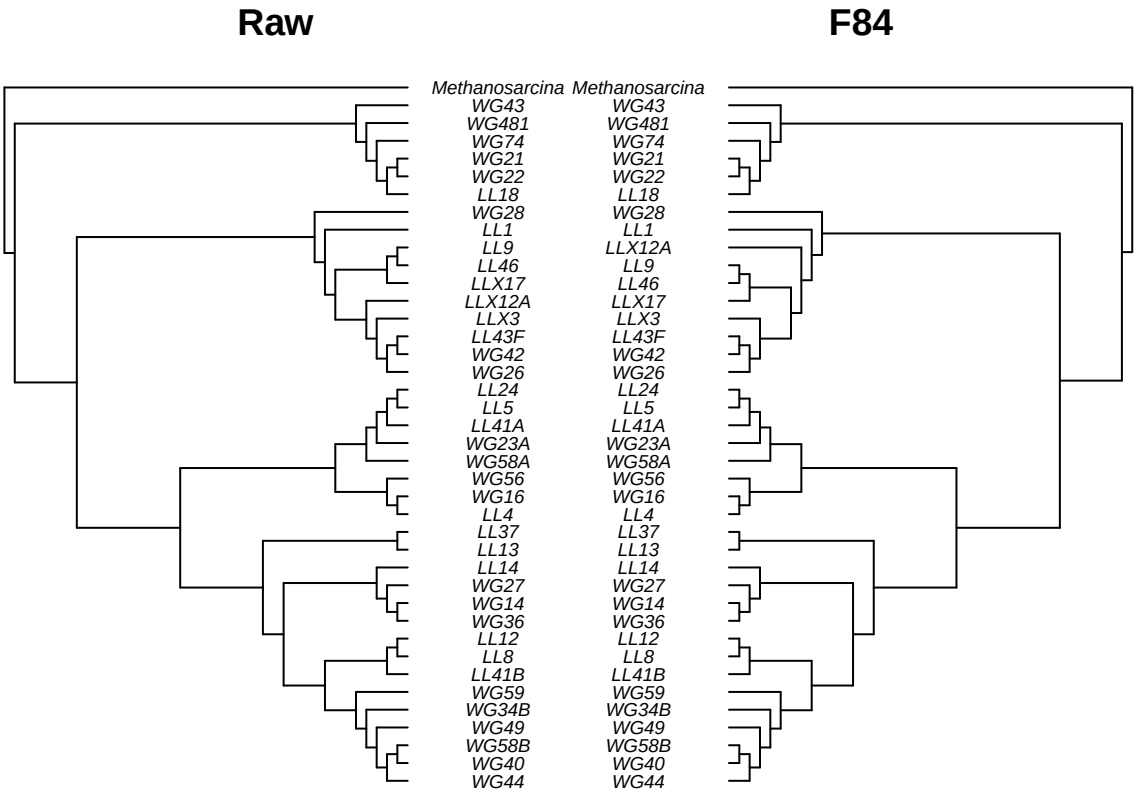
3. make Neighbor Joining trees for both, and
4. create a cophylogenetic plot to compare the topologies of the trees.

```
seq.dist.F84 <- dist.dna(p.DNAbin, model = "F84", pairwise.deletion = F) #distance matrix based on Fels

# Saturation plot
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")
```




```
# Make neighbour joining tree for both:
raw.tree <- bionj(seq.dist.raw) #NJ for raw
F84.tree <- bionj(seq.dist.F84) #NJ for F84

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

# Root the Trees:
raw.rooted <- root(raw.tree, raw.outgroup, resolve.root = T)
F84.rooted <- root(F84.tree, F84.outgroup, resolve.root = T)

# Cophylogenetic Plot:
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 = T, use.edge.length = F, 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 = T, use.edge.length = F, adj = 0.5,
           cex = 0.6, label.offset = 2, main = "F84")
```



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

```
# Read alignment file as phyDat:
phyDat.aln <- msaConvert(read.aln, type = "phangorn::phyDat")

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, T)
```

```
## 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
```

```
# Fit tree using a GTR model with gamma distributed rates:
fitGTR <- optim.pml(fit, model = "GTR", optInv = T, optGamma = T,
```

```
                    rearrangement = "NNI",
                    control = pml.control(trace = 0))
```

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

```
# Perform model selection with either an ANOVA 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
```

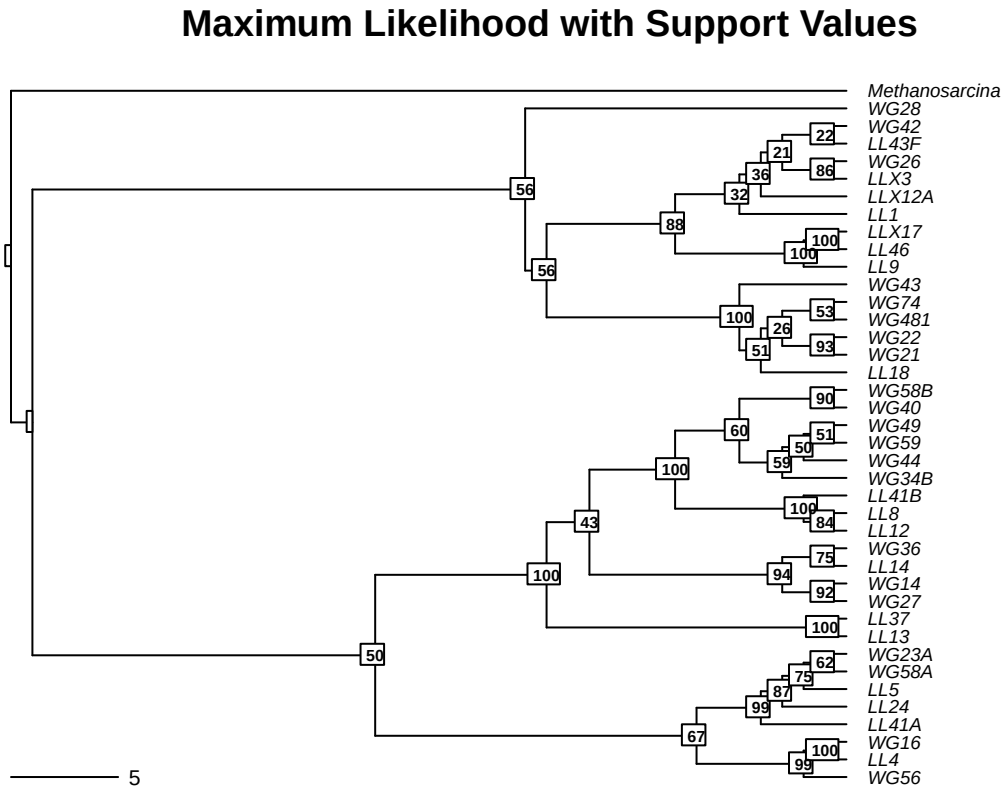
```
AIC(fitJC)
```

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

```
AIC(fitGTR)
```

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

```
# Bootstrapping support value:
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 = T, use.edge.length = F, 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:
The maximum likelihood tree is different from the neighbour joining tree since the ML tree split into two large polyphyletic groups shortly after the divergence from Methanosarcina, while the NJ tree divergence are less complicated. The difference might results from that the NJ tree is made only based on pairwise difference between taxa, while the ML tree is calculated from highest probability in terms of which and when taxa diverges from each other based on the alignment sequences given.

Answer 4b:
Phylogentic trees are built based on statistical methods, bootstrapping is to use statistical simulations (with large number of re-sampling trails) to calculate how confidence we are in the correctness of the tree we made.

Answer 4c:
Bootstrap values are the percentage of confidence or support we have in the tree taxa to be operationally correct; or in another way, how correct the existing tree branches are.

Answer 4d:
The lowest support occure in the divergence between the samples of WG42 and LL43F, and that of WG26 and LLX3. However, any branch that has bootstrap value less than 50% are considered insufficiently supported

Answer 4e:
We shouldn't because these branches mean they have low statistical confidence, which means these might not be the correct evolutionary relation among the taxa for these branches (that are less than 50%).

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.

```
p.growth <- read.table("./data/p.isolates.raw.growth.txt", sep = "\t",
                       header = T, row.names = 1) #load raw P growth data
p.growth.std <- p.growth/(apply(p.growth, 1, sum)) #standardize
```

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.

```
umax <- (apply(p.growth, 1, max)) #calculate max growth rate
levins <- function(p_xi = ""){
  p = 0
  for(i in p_xi){
    p = p + i^2
  }
  nb = 1 / (length(p_xi)*p)
  return(nb)
}

nb <- as.matrix(levins(p.growth.std)) #niche breadth for each isolates
nb <- setNames(as.vector(nb), as.matrix(row.names(p.growth)))
```

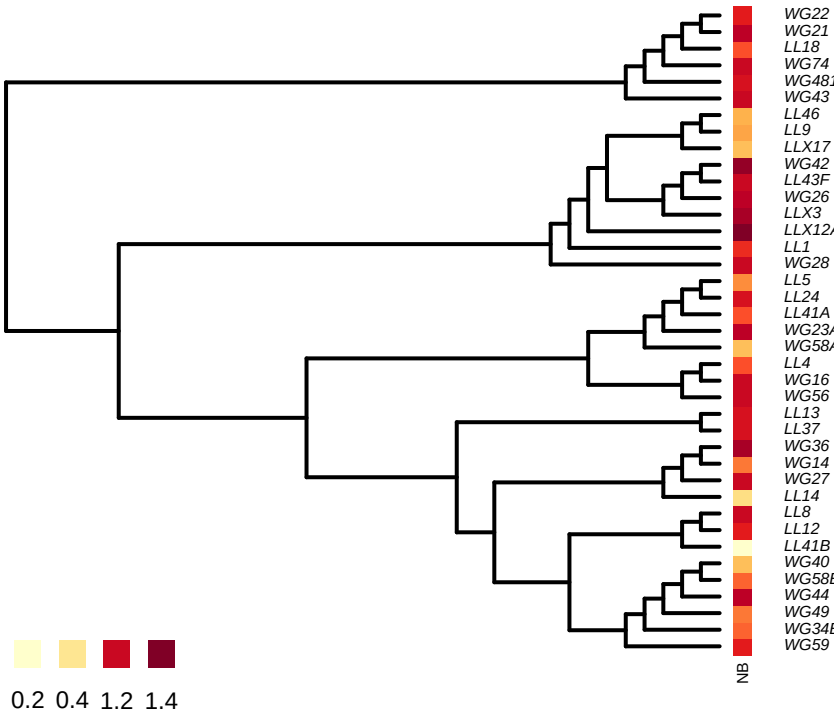
C. Visualizing Traits on Trees

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.

```
nj.F84.tree <- bionj(seq.dist.F84)
outgroup.F84 <- match("Methanosarcina", nj.F84.tree$tip.label)
nj.F84.rooted <- root(nj.F84.tree, outgroup.F84, resolve.root = T)
nj.F84.rooted <- drop.tip(nj.F84.rooted, "Methanosarcina")
```

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).

```
mypalette <- colorRampPalette(brewer.pal(9, "YlOrRd"))
nj.F84.plot <- nj.F84.rooted
nj.F84.plot$edge.length <- nj.F84.plot$edge.length + 10^(-1)

# Visualize the growth rate with niche breadth:
par(mar = c(1,5,1,5) +0.1)
x.nb <- phylo4d(nj.F84.plot, nb)
table.phylo4d(x.nb, treetype = "phylo", symbol = "colors",
              show.node = T, cex.label = 0.5, scale = F,
              use.edge.length = F, edge.color = "black",
              edge.width = 2, box = F, col = mypalette(25),
              pch = 15, cex.symbol = 1.25, var.label = ("NB"),
              ratio.tree = 0.9, cex.legend = 1.5, center = F)
```



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:
Organisms in resource-limited environment (LL) are more likely to be specialists, while generalists are more likely to thrive in resource-rich environment (WG).

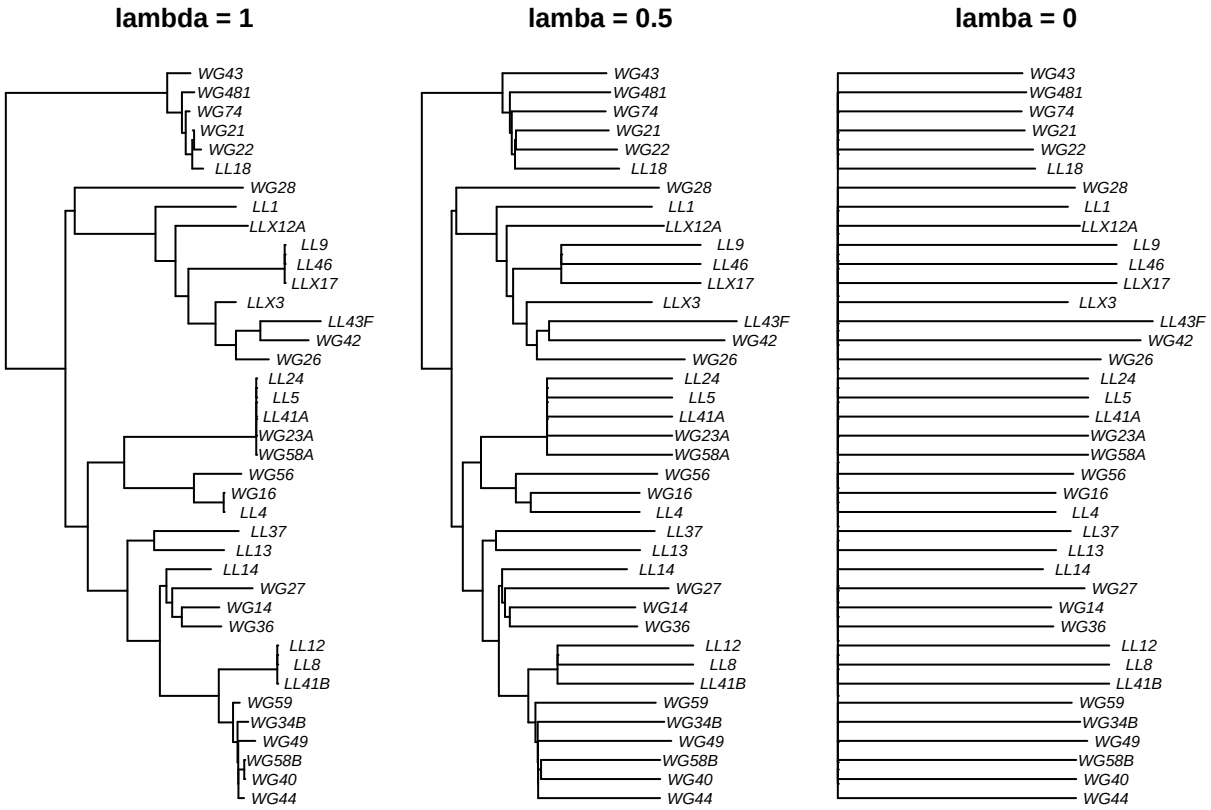
Answer 5b:
Growth rate is not affected by niche breadth.

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.

```
library(geiger)
nj.lambda.5 <- geiger:::rescale.phylo(nj.F84.rooted, model = "lambda", lambda = 0.5)
nj.lambda.0 <-geiger:::rescale.phylo(nj.F84.rooted, model = "lambda", lambda = 0)
layout(matrix(c(1, 2, 3), 1, 3), width = c(1, 1, 1))
par(mar=c(1, 0.5, 2, 0.5) + 0.1)
plot(nj.F84.rooted, main = "lambda = 1", cex = 0.7, adj = 0.5) #lambda = 1
plot(nj.lambda.5, main = "lamba = 0.5", cex = 0.7, adj = 0.5)
plot(nj.lambda.0, main = "lamba = 0",  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.F84.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 = -37.006827
##  AICc = -36.321113
##  free parameters = 3
##
## Convergence diagnostics:
##  optimization iterations = 100
##  failed iterations = 58
##  number of iterations with same best fit = NA
##  frequency of best fit = NA
##
##  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 = -37.005010
##  AICc = -36.319295
##  free parameters = 3
##
## Convergence diagnostics:
##  optimization iterations = 100
##  failed iterations = 0
##  number of iterations with same best fit = 85
##  frequency of best fit = 0.850
##
##  object summary:
##  'lik' -- likelihood function
##  'bnd' -- bounds for likelihood search
##  'res' -- optimization iteration summary
##  'opt' -- maximum likelihood parameter estimates
```

```
phylosig(nj.F84.rooted, nb, method = "lambda", test = TRUE)
```

```
##
## Phylogenetic signal lambda : 0.00699105
## logL(lambda) : 21.5034
## LR(lambda=0) : 0.00181763
## 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's phylogenetic signal?

> ***Answer 6a***:
> The lambda value of the untransformed tree is 0.006, which is larger than the transformed tree. This means the shared traits are less affected by the evolutionary relatedness but are more influenced by the ecological or environmental conditions in the two locations.
>
> ***Answer 6b***:
> Since the AIC scores for the untransformed tree is -37.006823, and the one for the transformed tree is -37.005010, the two trees are considered equivalent in terms of their lambda values.
>
> ***Answer 6c***:
> Since the p-value of 0.966 suggests that the untransformed tree has a lambda value that is not significantly different from zero, the tree doesn't really have phylogentic signals.

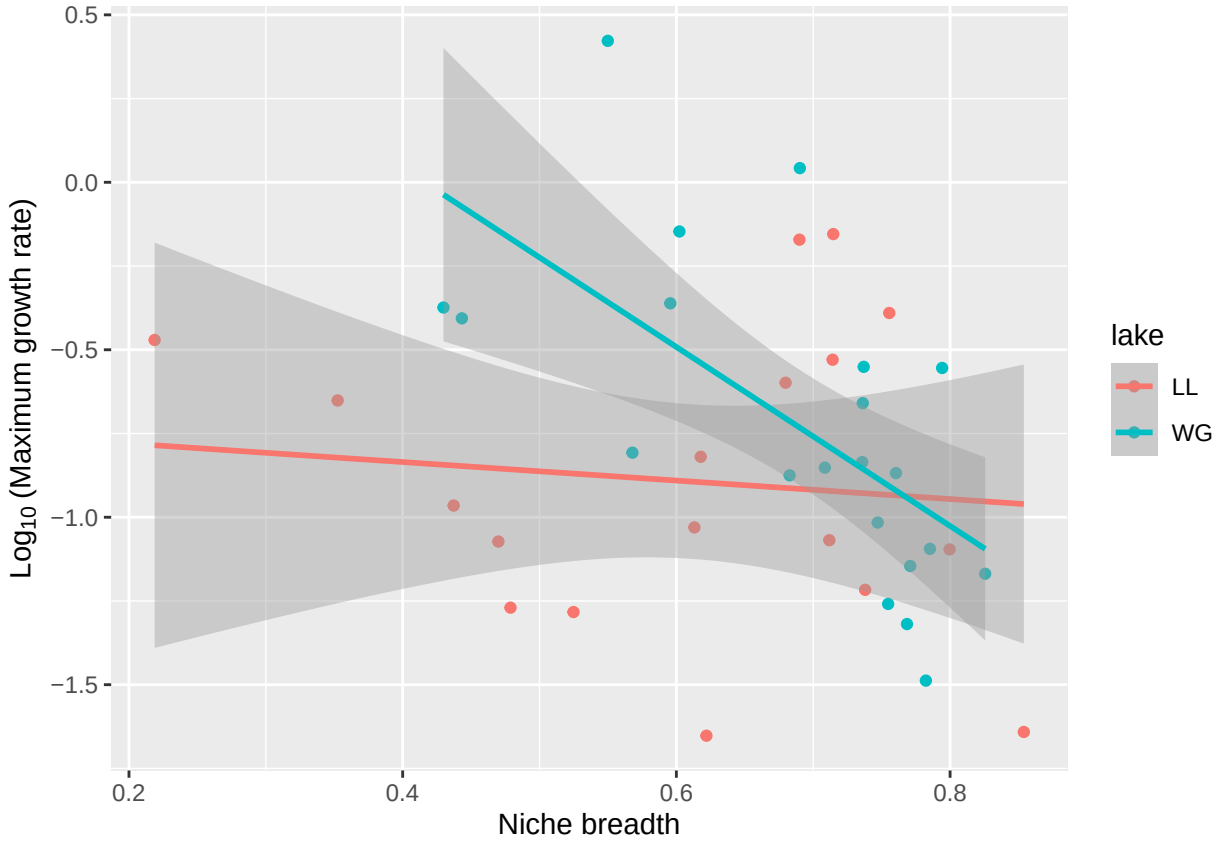
7) PHYLOGENETIC REGRESSION

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

```
nb.lake <- as.data.frame(as.matrix(nb))
nb.lake$lake <- rep('A')

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")
}

colnames(nb.lake)[1] <- "NB"

umax <- as.matrix(apply(p.growth, 1, max))
nb.lake <- cbind(nb.lake, umax)

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 regression model
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
```

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

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

```
# Correcting for phylogeny:
fit.plm <- phylolm(log10(umax) ~ NB*lake, data = nb.lake, nj.rooted,
                   model = "lambda", boot = 0)
```

```
## Warning in phylolm(log10(umax) ~ NB * lake, data = nb.lake, nj.rooted, model =
## "lambda", : will drop from the tree 1 taxa with missing data
```

```
summary(fit.plm)
```

```
##
## Call:
## phylolm(formula = log10(umax) ~ NB * lake, data = nb.lake, phy = nj.rooted,
##     model = "lambda", boot = 0)
##
##    AIC logLik
##  41.12 -14.56
##
## Raw residuals:
##      Min       1Q   Median       3Q      Max
## -0.75573 -0.18983 -0.07978  0.32375  0.95388
##
## Mean tip height: 0.1411147
## Parameter estimate(s) using ML:
## lambda : 0.4838753
## sigma2: 1.152639
##
## Coefficients:
##              Estimate    StdErr t.value p.value
## (Intercept) -0.908378  0.367115 -2.4744 0.01834 *
## NB           0.018987  0.523770  0.0363 0.97129
## lakeWG       1.464616  0.576672  2.5398 0.01569 *
```

```
## NB:lakeWG   -1.997261  0.846830 -2.3585 0.02406 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-squared: 0.1968    Adjusted R-squared: 0.1279
##
## Note: p-values and R-squared are conditional on lambda=0.4838753.
```

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

```
## [1] 41.12295
```

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:
Because of the shared evolutionary history among the taxon exists, the samples are not independent from each other, which is a vialation of regression analysis assumption. Correcting the evolutionary history, or the phylogeny, can enable us to analyze the relationships between independent variables such as environmental or ecological traits.

Answer 7b:
variances of residual errors of phylogenetic regression are a covariance matrix while that of simple linear regression is assumed to be independent.

Answer 7c:
Both models report significance relationship between maximum growth rate and lake environment as well as the interaction between the niche breadth and lake environment. This means that lake resource environmental condition does affect whether the taxa is a generalist or a specialist, and that lake environment also significanly affect the maximum growth rate in these taxa. However, after correcting for the shared evolutionary history, the coefficient between the above variables decreases, meaning that the shared evolutionary history play a role in controlling how strong the environmental condition influences the maximum growth rate, and that there is an evolutionary relatedness of being a generalist or a specialist.

Answer 7d:
Ecological invasion.

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?

```
## I have slacked Emma for help in this section, because even though I have found
# the "accession ID" in the paper, I wasn't able to locate the exact
# phylogenetic sequences the study submitted.
# Our group has also decided to use 16S RNA sequence
# because this is what the original study group used for their bacteria data.
```

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)**.