Phylogenetic Diversity - Traits

Savannah Bennett; Z620: Quantitative Biodiversity, Indiana University

21 February, 2017

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. Change "Student Name" on line 3 (above) with your name.
2. Complete as much of the exercise as possible during class; what you do not complete in class will need to be done on your own outside of class.
3. Use the handout as a guide; it contains a more complete description of data sets along with the proper scripting needed to carry out the exercise.
4. Be sure to **answer the questions** in this exercise document; they also correspond to the handout. Space for your answer is provided in this document and indicated by the ">" character. If you need a second paragraph be sure to start the first line with ">".
5. Before you leave the classroom, **push** this file to your GitHub repo.
6. For homework, 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, please submit the completed exercise by creating a **pull request** via GitHub. Your pull request should include this file *PhyloTraits_exercise.Rmd* and the PDF output of `Knitr` (*PhyloTraits_exercise.pdf*).

1) SETUP

Typically, the first thing you will do in either an R script or an RMarkdown file is setup your environment. This includes things such as setting the working directory and loading any packages that you will need.

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/Savannah/GitHub/QB2017_Bennett/Week6-PhyloTraits"
```

```
setwd("C:/Users/Savannah/GitHub/QB2017_Bennett/Week6-PhyloTraits")
```

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 less or your favorite text editor, compare the `p.isolates.fasta` file and the `p.isolates.afa` file. Describe the differences that you observe between the files.

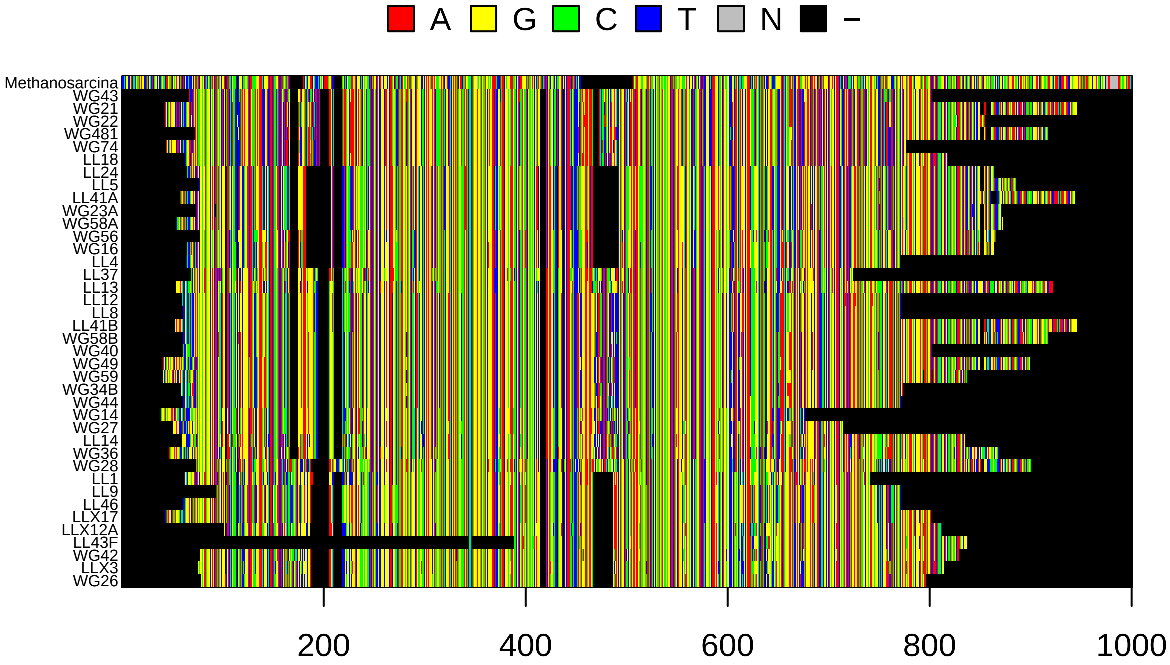
> ***Answer 1***: The p.isolates.afa file is aligned, while the p.isolates.fasta file is not. They both have the same information/data, but they are formatted slightly different.

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.

```
package.list <- c("ape", "seqinr", "phylobase", "adephylo", "geiger",
"picante", "stats", "RColorBrewer", "caper", "phylolm", "pmc",
"ggplot2", "tidyr", "dplyr", "phangorn", "pander")
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: 'adephylo'

## The following object is masked from 'package:ade4':
##
##     orthogram

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

## This is vegan 2.4-2

##
## Attaching package: 'vegan'

## The following object is masked from 'package:ade4':
##
##     cca

##
## 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 objects are masked from 'package:seqinr':
##
##     count, query

## 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
```

```r
#Read Alignment File
read.aln <- read.alignment(file= "./data/p.isolates.afa", format = "fasta")

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

#Identify Base Pair Region of 16S rRNA Gene to Visualize
window <- p.DNAbin[, 0:1000]

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



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 reads?

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

Answer 2a: Our reads are approximately 700 nucleotides.
Answer 2b: Sequences where species share most/many of the same nucleotides could be conserved sequences, which would be useful in making phylogenetic inferences. This would enable one to see which species are closely related to one another, as they would share many of the same sequences.

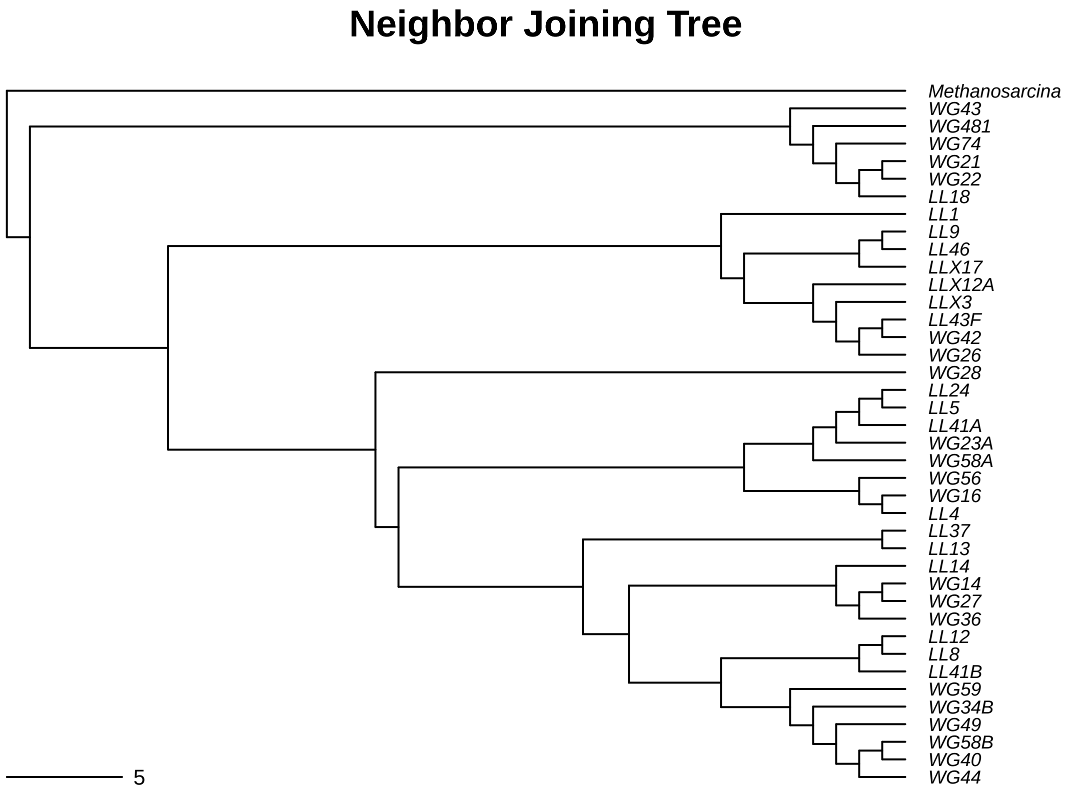
4) MAKING A PHYLOGENETIC TREE

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.

```r
#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?

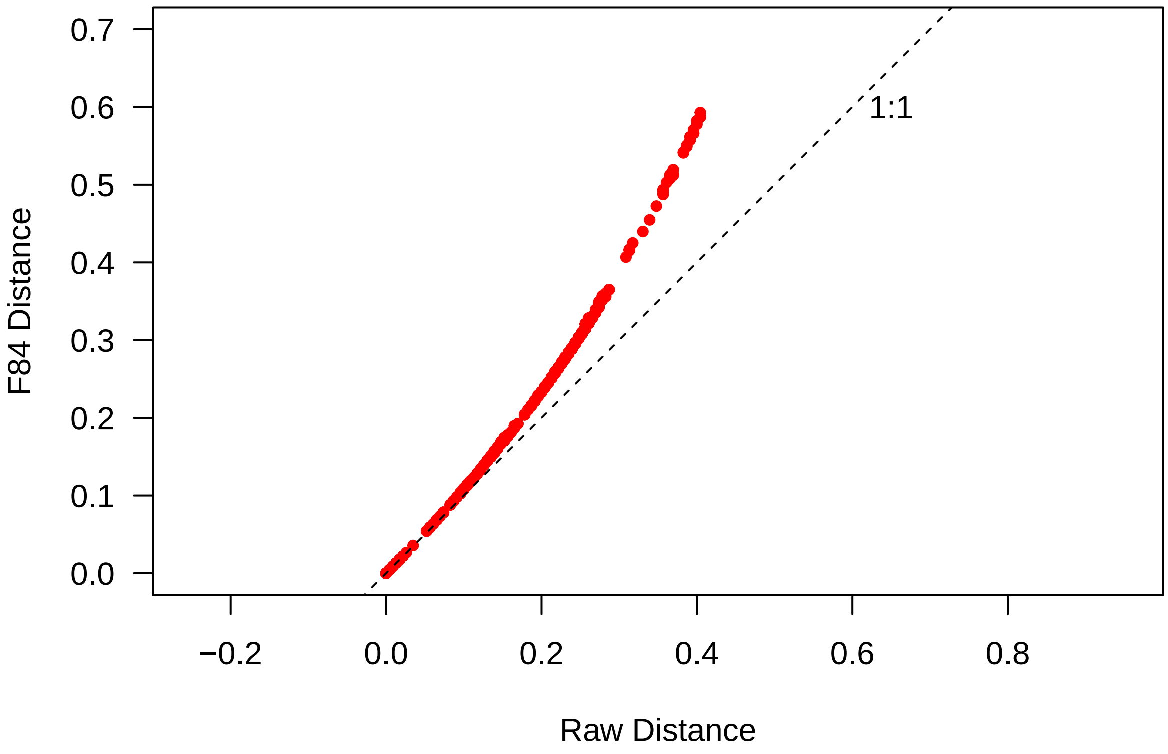
> ***Answer 3***: Neighbor joining trees are useful in that they can be used to incorporate more sophisticated/complex models. They can be used to analyze data sets with many taxa, and is faster than some of the other methods. However, it doesn't take specific nucleotide states into account because it uses a distance matrix, it only provides one tree, and it can be sensitive to the underlying substitution model. Neighbor joining trees also assign negative values as branch

lengths.

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.

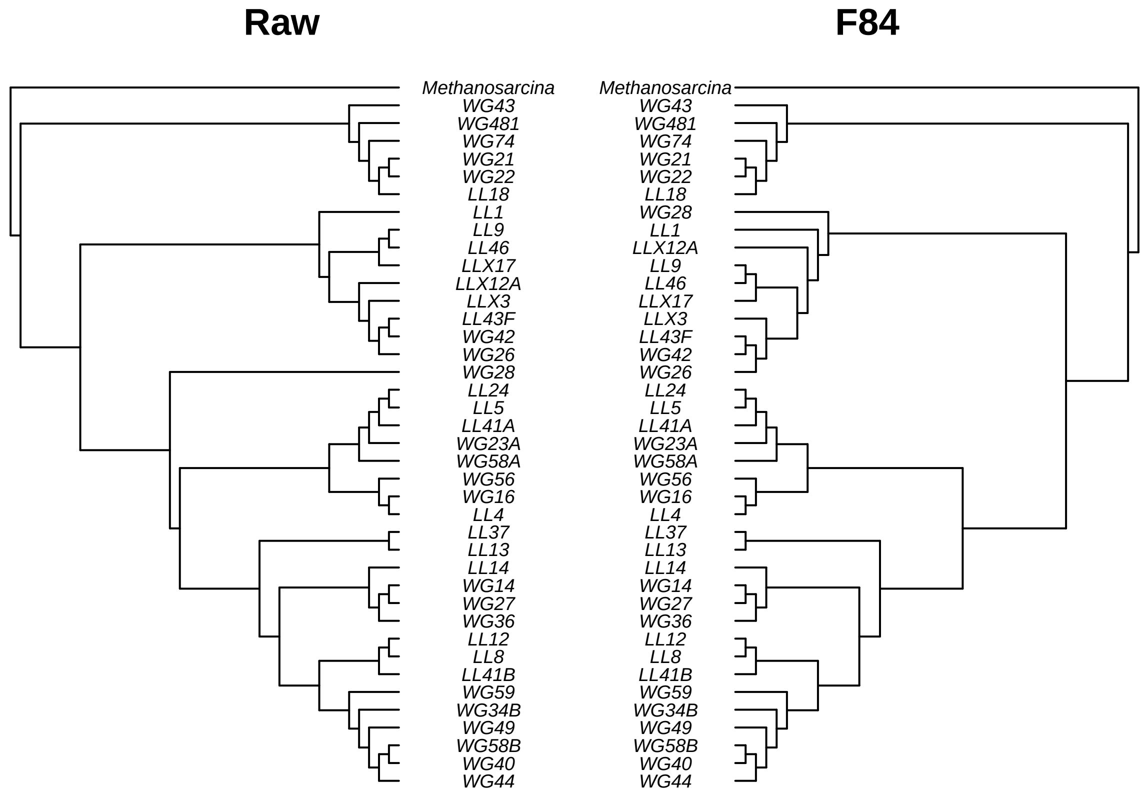
```r
#Create distance matrix with "F84" model {ape}
seq.dist.F84 <- dist.dna(p.DNAbin, model = "F84", pairwise.deletion = FALSE)

#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
```

```r
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.ofset = 2, main = "F84")
```

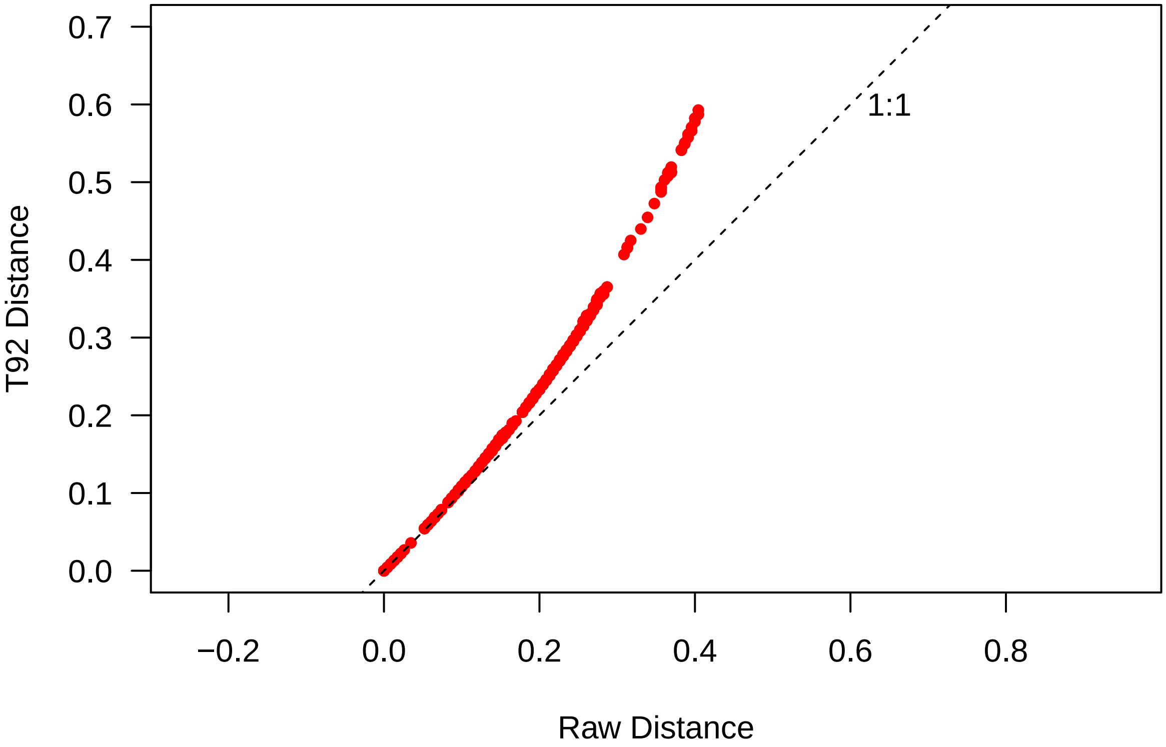
```
## Warning in plot.window(...): "label.ofset" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "label.ofset" is not a graphical
## parameter

## Warning in title(...): "label.ofset" is not a graphical parameter
```



In the R code chunk below, do the following:
1. pick another substitution model,
2. create and distance matrix and tree for this model,

3. make a saturation plot that compares that model to the *Felsenstein (F84)* model,
4. make a cophylogenetic plot that compares the topologies of both models, and
5. be sure to format, add appropriate labels, and customize each plot.

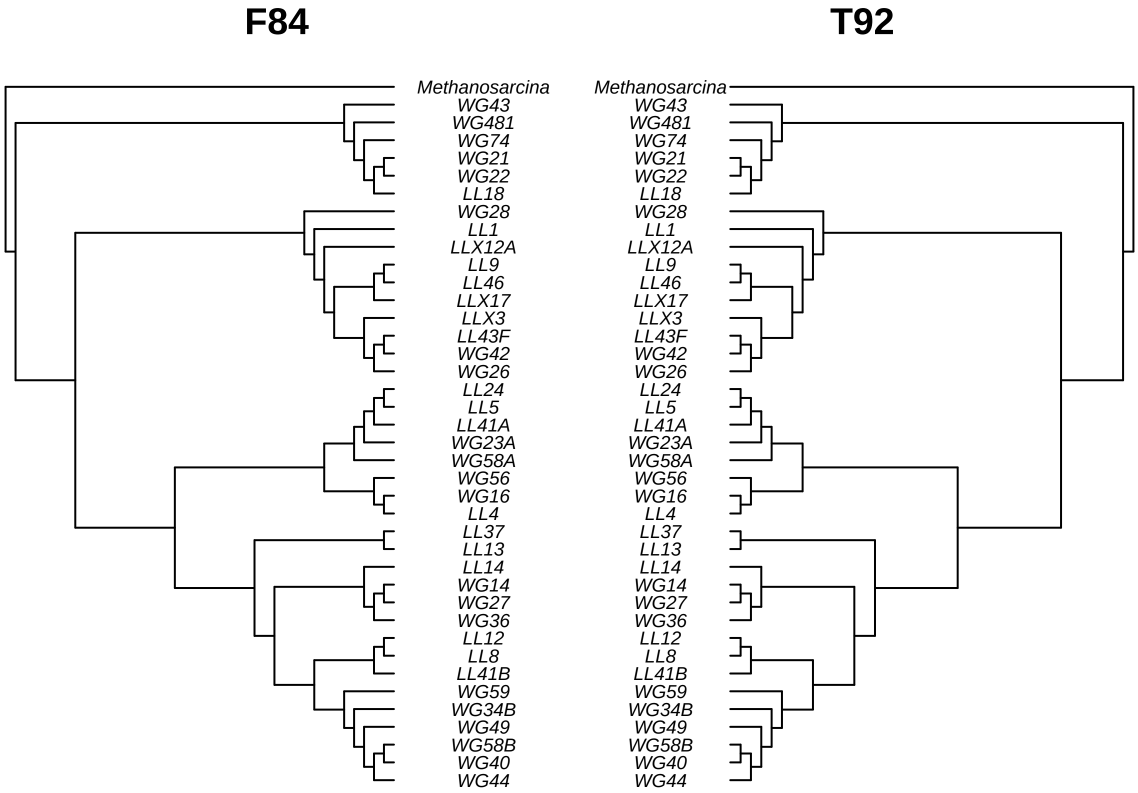
```r
#Create Distance Matrix for T92
seq.dist.T92 <- dist.dna(p.DNAbin, model = "F84", pairwise.deletion = FALSE)

#Plot Distances from Different DNA Substitution Models
par(mar = c(5,5,2,1) + 0.1)
plot(seq.dist.raw, seq.dist.T92,
     pch = 20, col = "red", las = 1, asp = 1, xlim = c(0, 0.7), ylim = c(0, 0.7),
     xlab = "Raw Distance", ylab = "T92 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}
F84.tree <- bionj(seq.dist.F84)
T92.tree <- bionj(seq.dist.T92)

#Define Outgroups
F84.outgroup <- match("Methanosarcina", F84.tree$tip.label)
T92.outgroup <- match("Methanosarcina", T92.tree$tip.label)

#Root the Trees {ape}
F84.rooted <- root(F84.tree, F84.outgroup, resolve.root= TRUE)
T92.rooted <- root(T92.tree, T92.outgroup, resolve.root= TRUE)
```

```r
#Make Cophylogenetic Plot {ape}
layout(matrix(c(1,2), 1, 2), width = c(1,1))
par(mar = c(1, 1, 2, 0))
plot.phylo(F84.rooted, type = "phylogram", direction = "right", show.tip.label=TRUE,
           use.edge.length = FALSE, adj = 0.5, cex = 0.6, label.offset = 2, main = "F84")

par(mar = c(1, 0, 2, 1))
plot.phylo(T92.rooted, type="phylogram", direction = "left", show.tip.label=TRUE,
           use.edge.length = FALSE, adj = 0.5, cex = 0.6, label.ofset = 2, main = "T92")
```

```
## Warning in plot.window(...): "label.ofset" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "label.ofset" is not a graphical
## parameter

## Warning in title(...): "label.ofset" is not a graphical parameter
```



Question 4:

a. Describe the substitution model that you chose. What assumptions does it make and how does it compare to the F84 model?
b. Using the saturation plot and cophylogenetic plots from above, describe how your choice of substitution model affects your phylogenetic reconstruction. If the plots are inconsistent with one another, explain why.
c. How does your model compare to the *F84* model and what does this tell you about the substitution rates of nucleotide transitions?

Answer 4a: I chose the Tamura model. This model assumes there are equal frequencies of nucleotides and recognizes that the probability of transition mutations is higher than that of transversion mutations. The Tamura model accounts for G + C content as well.

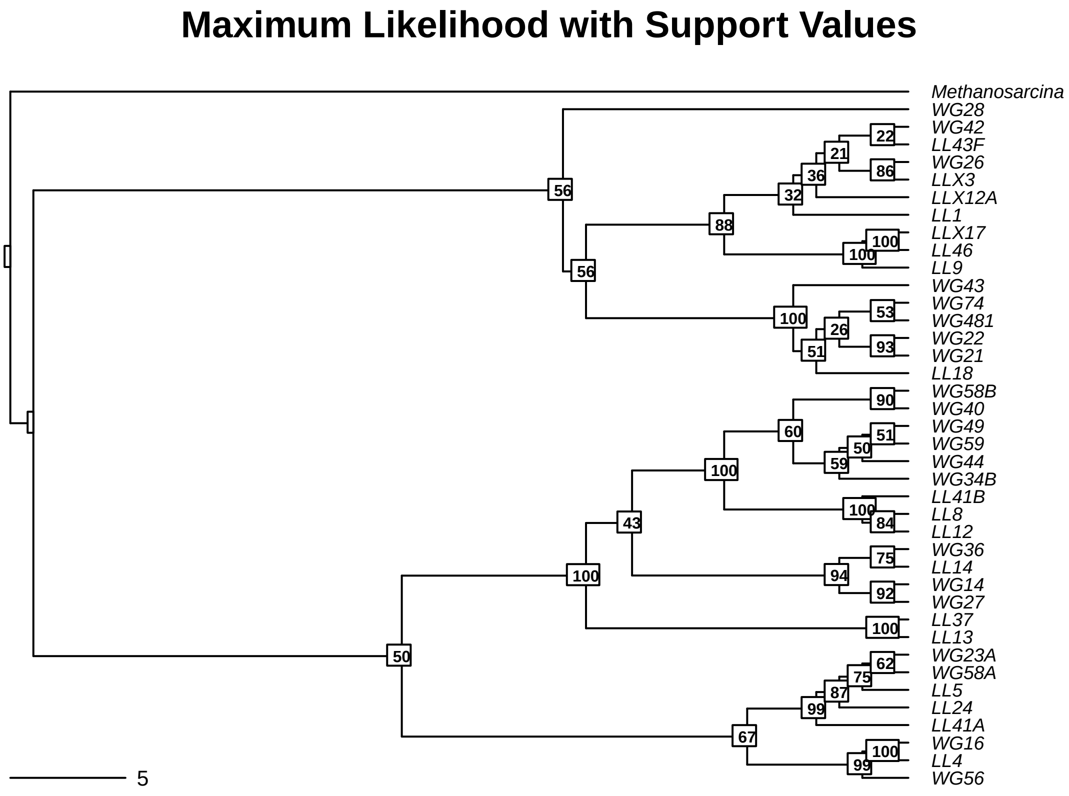
Answer 4b: The saturation plot compares the DNA substitution model to a distance matrix that does not have the substitution model, which allows you to determine whether there is correction for multiple substitutions. The cophylogenetic plot, on the other hand, allows you to compare phylogenetic trees made from two models. Both plots therefore have different functions in a phylogenetic reconstruction.

Answer 4c: The models provided the same output, which suggests that the substitution rates are similar.

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

```
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 5:

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?

Answer 5a: The maximum likelihood tree is built upon maximum likelihood, while the neighbor-joining tree is not. It is affected by sampling error to a lesser extent than the neighbor-joining tree. The plots generated in this assignment seem fairly similar, with a few differences in the branching patterns for some of the taxa. These differences exist because the maximum likelihood method produces a tree that gives the data the highest probability, while the neighbor-joining tree does not.

Answer 5b: We bootstrap our tree to tell us how confident we are in the placement of each branch. It involves resampling the data to tell us how reliable the tree is.

Answer 5c: Bootstrap values tell you what how many times (in a percentage) a branch was observed when the data was resampled and the tree was reconstructed.

Answer 5d: The branches corresponding to WG42 and LL43F seem to have fairly low support (support values = 21, 22).

Answer 5e: We should not trust these branches because they do not have bootstrap values 95% or higher.

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)

# Standadize 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)
}

# Calculate Niche Breadth for Each Isolate
nb <- as.matrix(levins(p.growth.std))

# Add Row & Column Names to Niche Breadth Matrix
rownames(nb) <- row.names(p.growth)
colnames(nb) <- c("NB")
```

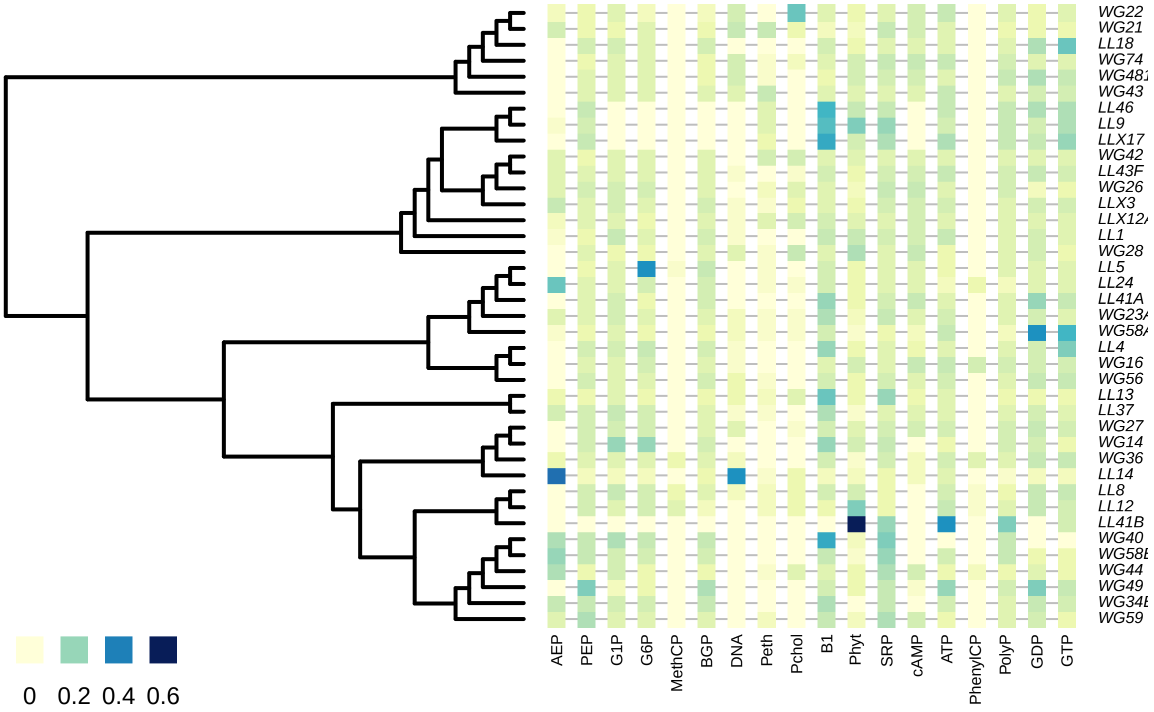
C. Visualizing Traits on Trees

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.

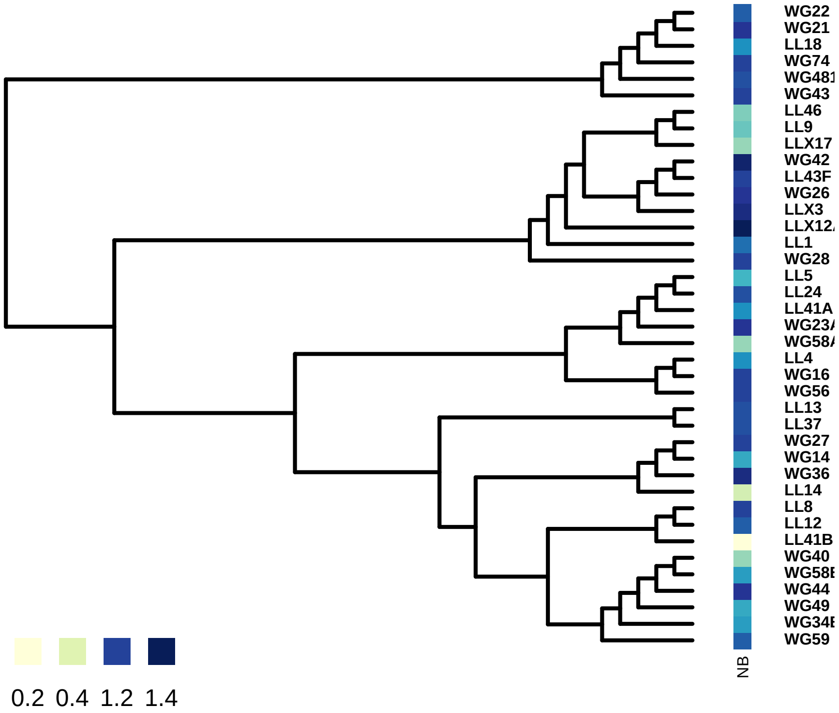
```
# Generate Neighbor Joining Tree Using F84 DNA Substitution Model {ape}
nj.tree <- bionj(seq.dist.F84)

# 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")
```

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, "YlGnBu"))

# Map Phosphorus Traits {adephylo}
par(mar=c(1,1,1,1) + 0.1)
x <- phylo4d(nj.rooted, 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.rooted, 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, font = 2)
```



Question 6:

a) Make 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 6a: Specialists can thrive on one or a few forms of phosphorus. However, they can only grow well on that one form, or a small variety of resource types (phosphorus). Generalists, on the other hand, can grow with a greater variety of resource types, but only grow moderately well in the presence of these resources.

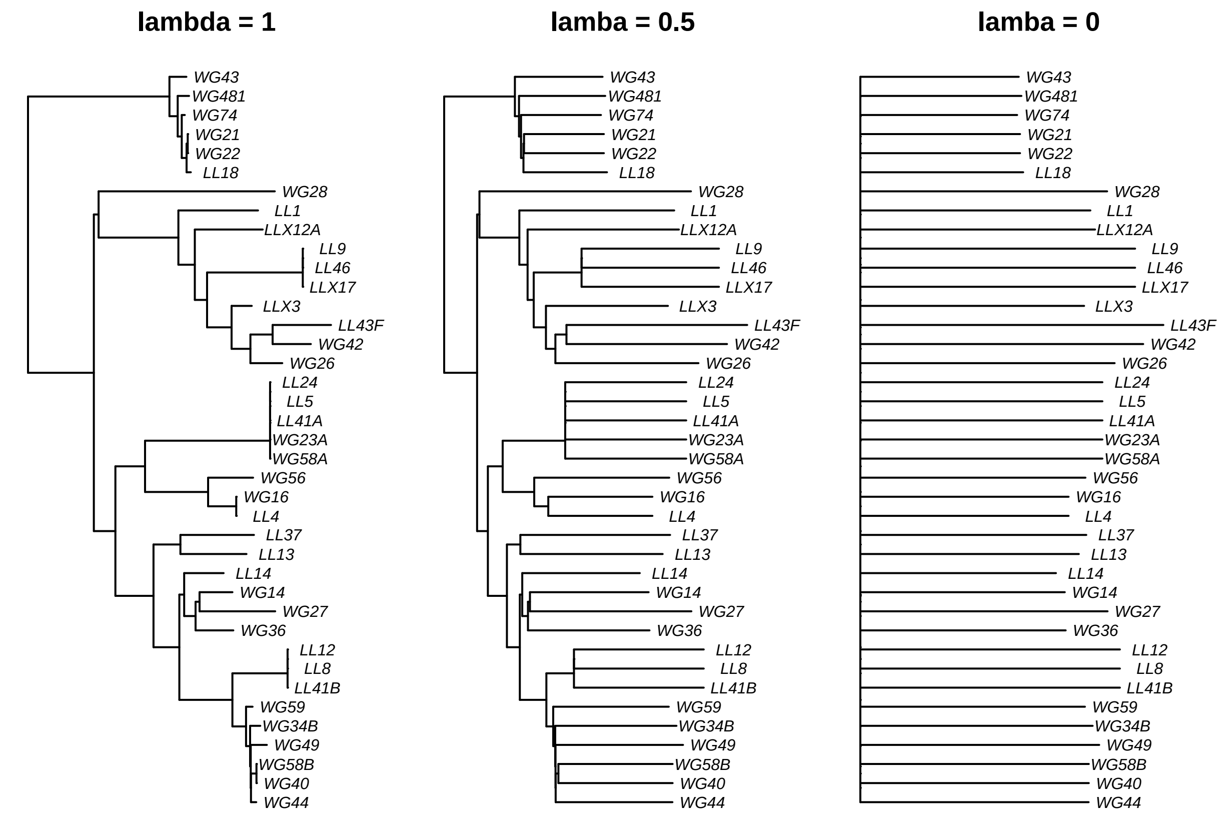
Answer 6b: Generalists should have lower maximum growth rates than specialists as a cost for being able to use many forms of phosphorus. Generalists would also have larger niche breaddth values than specialists. Specialists should have higher maximum growth rates, but only on the form of phosphorus they are specialized in utilizing. They would have lower niche breadth values than generalists because they utilize fewer forms of phosphorus.

6) HYPOTHESIS TESTING

A) 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.

```r
# Visualize Trees With Different Levels of Phylogenetic Signal {geiger}
nj.lambda.5 <- rescale(nj.rooted, "lambda", 0.5)
nj.lambda.0 <- rescale(nj.rooted, "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.rooted, main = "lambda = 1", cex = 0.7, adj = 0.5)
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.

```r
# Generate Test Statistics for Comparing Phylogenetic Signal {geiger}
fitContinuous(nj.rooted, nb, model = "lambda")
```

```
## GEIGER-fitted comparative model of continuous data
##  fitted 'lambda' model parameters:
##  lambda = 0.000343
##  sigsq = 0.106392
##  z0 = 0.657843
##
##  model summary:
##  log-likelihood = 21.652588
##  AIC = -37.305177
##  AICc = -36.619462
##  free parameters = 3
```

```
##
## Convergence diagnostics:
##  optimization iterations = 100
##  failed iterations = 52
##  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 = 0.106395
##  z0 = 0.657777
##
##  model summary:
##  log-likelihood = 21.652293
##  AIC = -37.304587
##  AICc = -36.618872
##  free parameters = 3
##
## Convergence diagnostics:
##  optimization iterations = 100
##  failed iterations = 0
##  frequency of best fit = 0.86
##
##  object summary:
##  'lik' -- likelihood function
##  'bnd' -- bounds for likelihood search
##  'res' -- optimization iteration summary
##  'opt' -- maximum likelihood parameter estimates
```

Question 7: 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 7a***: The lambda value of the untransformed tree is 0.02, while the lambda value of the transformed tree is 0.
>
> ***Answer 7b***: The AIC score when lambda = 1 is -37.32, while the AIC score when lambda = 0 is -37.30. Based on this information, the models are considered equivalent because the difference between the AIC scores is less than two.
>
> ***Answer 7c***: This suggests that there is not a phylogenetic signal because the models are equivalent when the signal is present or removed.

B) Phylogenetic Signal: Blomberg's K

In the R code chunk below, do the following:
1. correct tree branch-lengths to fix any zeros,
2. calculate Blomberg's K for each phosphorus resource using the `phylosignal()` function,
3. use the Benjamini-Hochberg method to correct for false discovery rate, and
4. calculate Blomberg's K for niche breadth using the `phylosignal()` function.

```
# First, Correct for Zero Branch-Lengths on Our Tree
nj.rooted$edge.length <- nj.rooted$edge.length + 10^-7

# Calculate Phylogenetic Signal for Growth on All Phosphorus Resources
# First, Create a Blank Output Matrix
p.phylosignal <- matrix(NA, 6, 18)
colnames(p.phylosignal) <- colnames(p.growth.std)
rownames(p.phylosignal) <- c("K", "PIC.var.obs", "PIC.var.mean",
  "PIC.var.P", "PIC.var.z", "PIC.P.BH")

# Use a For Loop to Calculate Blomberg's K for Each Resource
for (i in 1:18){
  x <- as.matrix(p.growth.std[ ,i, drop = FALSE])
  out <- phylosignal(x, nj.rooted)
  p.phylosignal[1:5, i] <- round(t(out), 3)
}

# Use the BH Correction on P-values:
p.phylosignal[6, ] <- round(p.adjust(p.phylosignal[4, ], method = "BH"), 3)
p.phylosignal[6, ]
```

```
##      AEP      PEP      G1P      G6P   MethCP      BGP      DNA     Peth
##    0.594    0.310    0.364    0.808    0.635    0.198    0.018    0.635
##    Pchol       B1     Phyt      SRP     cAMP      ATP PhenylCP    PolyP
##    0.655    0.594    0.714    0.635    0.018    0.735    0.819    0.714
##      GDP      GTP
##    0.735    0.687
```

```
# Calcualate Phylogenetic Signal for Niche Breadth
signal.nb <- phylosignal(nb, nj.rooted)
signal.nb
```

```
##              K PIC.variance.obs PIC.variance.rnd.mean PIC.variance.P
## 1 3.427719e-06         49966.78              50029.49          0.537
##   PIC.variance.Z
## 1   -0.003069224
```

Question 8: Using the K-values and associated p-values (i.e., "PIC.var.P"") from the `phylosignal` output, answer the following questions:

a. Is there significant phylogenetic signal for niche breadth or standardized growth on any of the phosphorus resources?

b. If there is significant phylogenetic signal, are the results suggestive of clustering or overdispersion?

   Answer 8a: Niche breadth does not have a significant phylogenetic signal (p=0.544), as phylogenetic signal value (3.42 x 10^-6) for niche breadth is relatively low. There is a significant phylogenetic signal for standardized growth for BGP (p=0.03), DNA (p=0.001), and CAMP (p=0.005).

Answer 8b: The K-values are less than one, which suggests that traits are overdispersed.

C. Calculate Dispersion of a Trait

In the R code chunk below, do the following:
1. turn the continuous growth data into categorical data,
2. add a column to the data with the isolate name,
3. combine the tree and trait data using the `comparative.data()` function in `caper`, and
4. use `phylo.d()` to calculate *D* on at least three phosphorus traits.

```
# Turn Continuous Data into Categorical Data
p.growth.pa <- as.data.frame((p.growth > 0.01) * 1)

# Look at Phosphorus Use for Each Resource
apply(p.growth.pa, 2, sum)
```

```
##      AEP      PEP      G1P      G6P   MethCP      BGP      DNA     Peth
##       20       38       35       34        3       35       19       21
##    Pchol       B1     Phyt      SRP     cAMP      ATP PhenylCP    PolyP
##       18       38       36       39       29       38        6       39
##      GDP      GTP
##       37       38
```

```
# Add Names Column to Data
p.growth.pa$name <- rownames(p.growth.pa)

# Merge Trait and Phylogenetic Data; Run `phylo.d`
p.traits <- comparative.data(nj.rooted, p.growth.pa, "name")
phylo.d(p.traits, binvar = AEP)
```

```
##
## Calculation of D statistic for the phylogenetic structure of a binary variable
##
##   Data :  p.growth.pa
##   Binary variable :  AEP
##   Counts of states:  0 = 19
##                      1 = 20
##   Phylogeny :  nj.rooted
##   Number of permutations :  1000
##
## Estimated D :  0.4644254
## Probability of E(D) resulting from no (random) phylogenetic structure :  0.003
## Probability of E(D) resulting from Brownian phylogenetic structure    :  0.018
```

```
phylo.d(p.traits, binvar = PhenylCP)
```

```
##
## Calculation of D statistic for the phylogenetic structure of a binary variable
##
##   Data :  p.growth.pa
##   Binary variable :  PhenylCP
##   Counts of states:  0 = 33
##                      1 = 6
##   Phylogeny :  nj.rooted
##   Number of permutations :  1000
##
```

```
## Estimated D :  0.8620865
## Probability of E(D) resulting from no (random) phylogenetic structure :  0.267
## Probability of E(D) resulting from Brownian phylogenetic structure    :  0.022
```

```
phylo.d(p.traits, binvar = DNA)
```

```
##
## Calculation of D statistic for the phylogenetic structure of a binary variable
##
##   Data :  p.growth.pa
##   Binary variable :  DNA
##   Counts of states:  0 = 20
##                      1 = 19
##   Phylogeny :  nj.rooted
##   Number of permutations :  1000
##
## Estimated D :  0.6052314
## Probability of E(D) resulting from no (random) phylogenetic structure :  0.032
## Probability of E(D) resulting from Brownian phylogenetic structure    :  0.004
```

```
phylo.d(p.traits, binvar = cAMP)
```

```
##
## Calculation of D statistic for the phylogenetic structure of a binary variable
##
##   Data :  p.growth.pa
##   Binary variable :  cAMP
##   Counts of states:  0 = 10
##                      1 = 29
##   Phylogeny :  nj.rooted
##   Number of permutations :  1000
##
## Estimated D :  0.1475362
## Probability of E(D) resulting from no (random) phylogenetic structure :  0.001
## Probability of E(D) resulting from Brownian phylogenetic structure    :  0.308
```

Question 9: Using the estimates for D and the probabilities of each phylogenetic model, answer the following questions:

a. Choose three phosphorus growth traits and test whether they are significantly clustered or overdispersed?

b. How do these results compare the results from the Blomberg's K analysis?

c. Discuss what factors might give rise to differences between the metrics.

   Answer 9a: The estimated D value for AEP is 0.468, and the p-value for the probability of D resulting from no phylogenetic structure is 0.005. This suggests that the ability of the bacterial isolate to grow on this form of phosphorus is significantly overdispersed. The estimated D value for PhenylCP is 0.869, and the p-value for the probability of D resulting from no phylogenetic structure is 0.27, which indicates that it is not significantly overdispersed. This trait is significantly clustered (p= 0.015). The D value for CAMP is 0.14, and the p-value (0.001) indicates the trait is significantly overdispersed. The D value being closer to 0 than 1 suggests that the traits are randomly clumped.

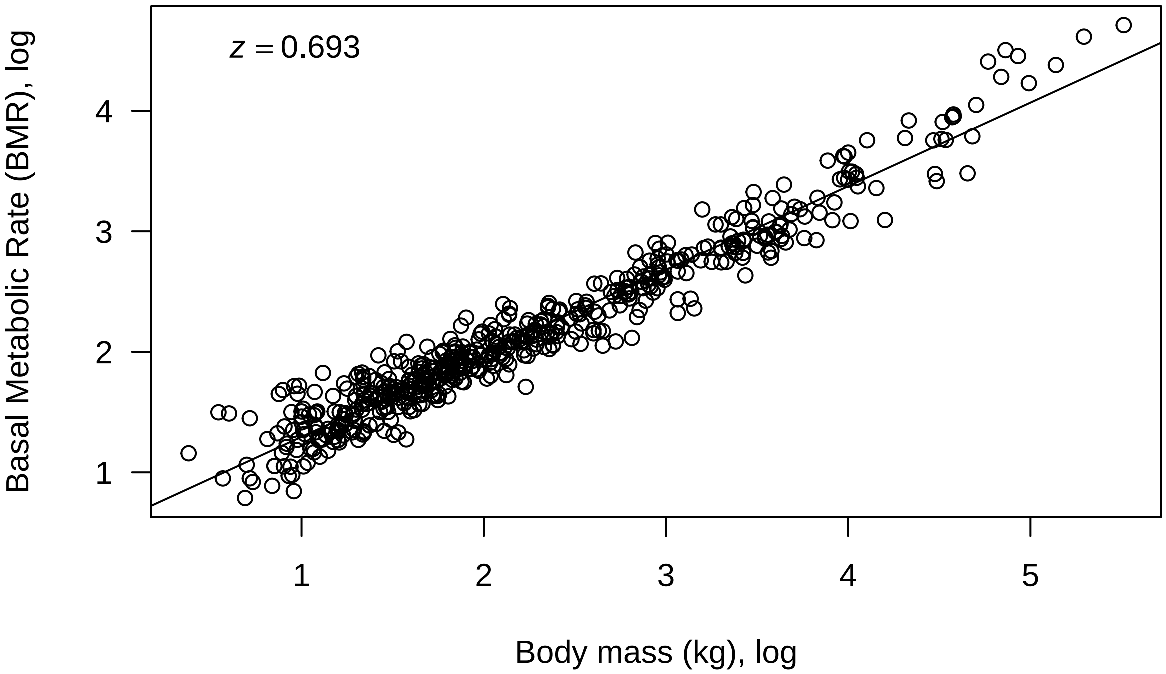
   Answer 9b: Overall, the results are fairly similar to the Blomberg's K analyses because most of the traits (except PhenylCP) are overdispersed.

   Answer 9c: D and Blomberg's K values are calculated differently, so analyses with the same data may yield different D and K values.

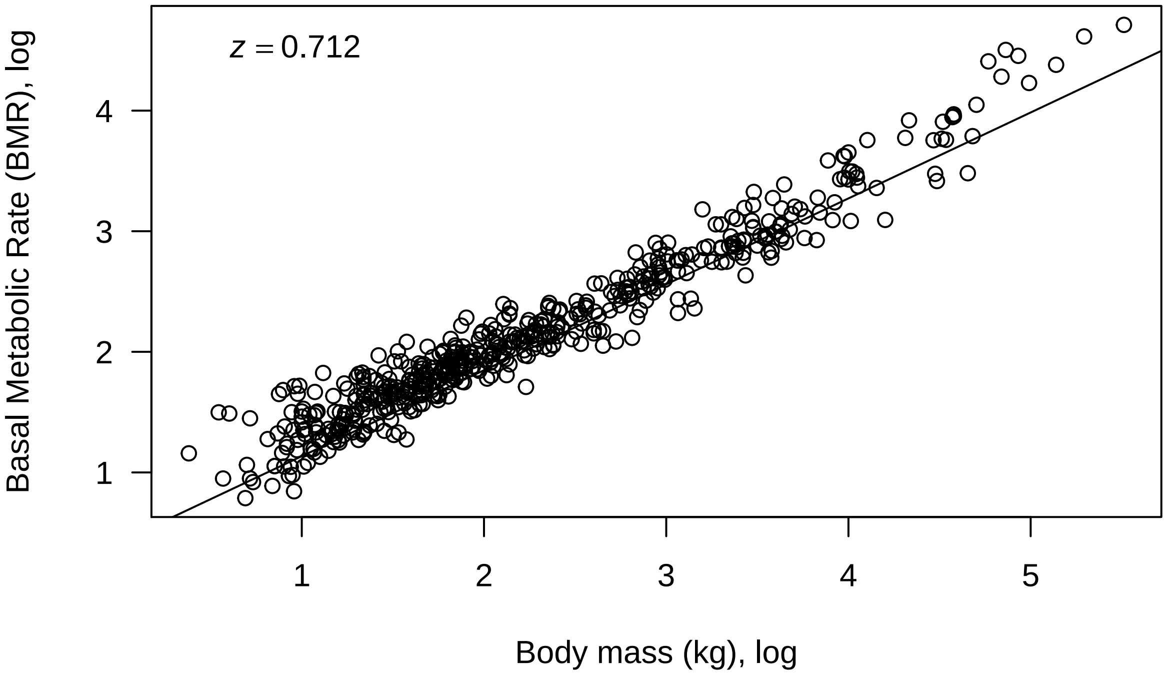
7) PHYLOGENETIC REGRESSION

In the R code chunk below, do the following:
1. Load and clean the mammal phylogeny and trait dataset, 2. Fit a linear model to the trait dataset, examining the relationship between mass and BMR, 2. Fit a phylogenetic regression to the trait dataset, taking into account the mammal supertree

```r
# Input the tree and dataset
mammal.Tree <- read.tree("./data/mammal_best_super_tree_fritz2009.tre")
mammal.data <- read.table("./data/mammal_BMR.txt", sep = "\t",
header = TRUE)

# Select the variables we want to analyze
mammal.data <- mammal.data[, c("Species", "BMR_.mlO2.hour.",
"Body_mass_for_BMR_.gr.")]
mammal.species <- array(mammal.data$Species)

# Select the tips in the mammal tree that are also in the
# dataset
pruned.mammal.tree <- drop.tip(mammal.Tree, mammal.Tree$tip.label[-na.omit(match(mammal.species,
mammal.Tree$tip.label))])

# Select the species from the dataset that are in our prunned
# tree
pruned.mammal.data <- mammal.data[mammal.data$Species %in% pruned.mammal.tree$tip.label,
]

# Turn column of Species names into rownames
rownames(pruned.mammal.data) <- pruned.mammal.data$Species

# Run a simple linear regression
fit <- lm(log10(BMR_.mlO2.hour.) ~ log10(Body_mass_for_BMR_.gr.),
          data = pruned.mammal.data)
plot(log10(pruned.mammal.data$Body_mass_for_BMR_.gr.),
     log10(pruned.mammal.data$BMR_.mlO2.hour.),
las = 1, xlab = "Body mass (kg), log", ylab = "Basal Metabolic Rate (BMR), log")
abline(a = fit$coefficients[1], b = fit$coefficients[2])
b1 <- round(fit$coefficients[2], 3)
eqn <- bquote(italic(z) == .(b1))

# plot the slope
text(0.5, 4.5, eqn, pos = 4)
```



```r
# Run a phylogeny-corrected regression with no bootstrap
# replicates
fit.phy <- phylolm(log10(BMR_.mlO2.hour.) ~ log10(Body_mass_for_BMR_.gr.),
                   data = pruned.mammal.data, pruned.mammal.tree, model = "lambda",
                   boot = 0)
plot(log10(pruned.mammal.data$Body_mass_for_BMR_.gr.), log10(pruned.mammal.data$BMR_.mlO2.hour.),
     las = 1, xlab = "Body mass (kg), log", ylab = "Basal Metabolic Rate (BMR), log")
abline(a = fit.phy$coefficients[1], b = fit.phy$coefficients[2])
b1.phy <- round(fit.phy$coefficients[2], 3)
eqn <- bquote(italic(z) == .(b1.phy))

#Plot the slope
text(0.5, 4.5, eqn, pos = 4)
```



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 worsten 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 10a: We need to correct for shared evolutionary history because our samples are not independent. In other words, it violates the assumption of independence.

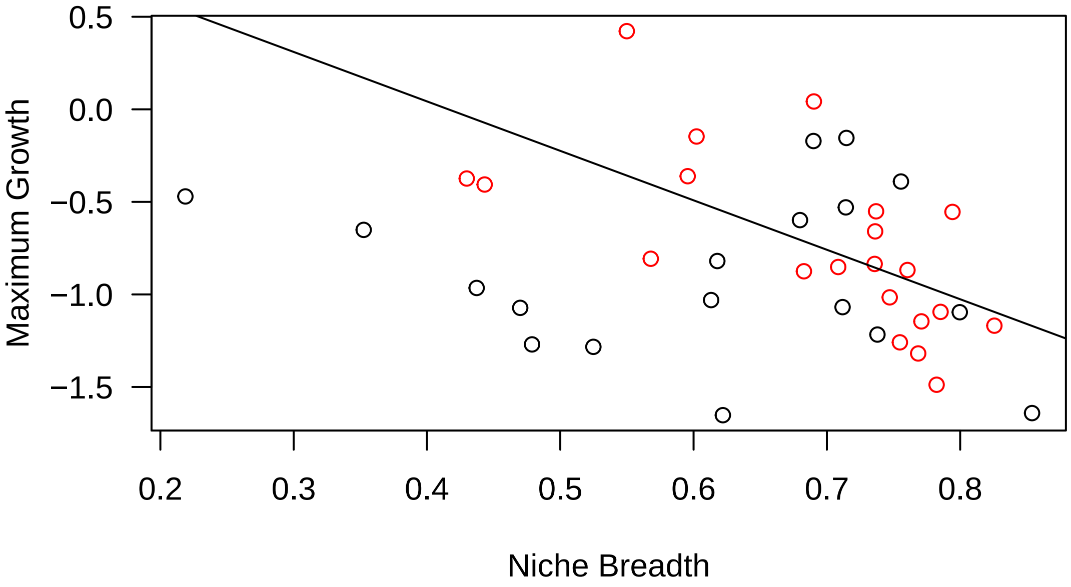
Answer 10b: In a linear regression, it is assumed that the residual errors are independent, whereas in a phylogenetic regression, branch lengths of the underlying phylogeny are taken into account with a covariance matrix.

Answer 10c: Both models have a positive slope, indicating that basal metabolic rate increases with body size. Accounting for shared evolutionary history improved the fit of the regression and the slope.

Answer 10d: Traits that are highly conserved, and are found more so due to phylogeny as opposed to environmental factors would cause the relationship to disappear when the underlying phylogeny is accounted for. An example of this could perhaps be maternal body size and number of eggs/offspring produced in different fish species. There should a strong positive relationship between maternal body size and number of eggs produced/female, and this relationship should be due to phylogeny.

7) SYNTHESIS

Below is the output of a multiple regression model depicting the relationship between the maximum growth rate (μ_{max}) of each bacterial isolate and the niche breadth of that isolate on the 18 different sources of phosphorus. One feature of the study which we did not take into account in the handout is that the isolates came from two different lakes. One of the lakes is an very oligotrophic (i.e., low phosphorus) ecosystem named Little Long (LL) Lake. The other lake is an extremely eutrophic (i.e., high phosphorus) ecosystem named Wintergreen (WG) Lake. We included a "dummy variable" (D) in the multiple regression model (0 = WG, 1 = LL) to account for the environment from which the bacteria were obtained. For the last part of the assignment, plot nich breadth vs. μ_{max} and the slope of the regression for each lake. Be sure to color the data from each lake differently.



Question 11: Based on your knowledge of the traits and their phylogenetic distributions, what conclusions would you draw about our data and the evidence for a generalist-specialist tradeoff?

> ***Answer 11***: There is a negative relationship between maximum growth and niche breadth. In other words, maximum growth decreases with increasing niche breadth. This supports the hypothesis that there is a generalist-specialist tradeoff. Specialists would have a lower niche breadth, and would therefore perform/survive best in one or a few environments. Generalists would have a wider niche breadth, but would have lower maximum growth rates. However, this relationship is only significant in the Wintergreen lake where phosphorus levels are higher.