11. Worksheet: Phylogenetic Diversity - Traits

Brianna Crawley; Z620: Quantitative Biodiversity, Indiana University

20 February, 2019

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 today, it is *imperative* that you **push** this file to your GitHub repo, at whatever stage you are. This will enable you to pull your work onto your own computer.
6. When you have completed the worksheet, **Knit** the text and code into a single PDF file by pressing the `Knit` button in the RStudio scripting panel. This will save the PDF output in your '8.BetaDiversity' folder.
7. After Knitting, please submit the worksheet by making a **push** to your GitHub repo and then create a **pull request** via GitHub. Your pull request should include this file (**11.PhyloTraits_Worksheet.Rmd**) with all code blocks filled out and questions answered) and the PDF output of `Knitr` (**11.PhyloTraits_Worksheet.pd**

The completed exercise is due on **Wednesday, February 20th, 2019 before 12:00 PM (noon)**.

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 "*/11.PhyloTraits*" folder, and
4. load all of the required R packages (be sure to install if needed).

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

```
## [1] "C:/Users/wolve/GitHub/QB2019_Crawley/2.Worksheets/11.PhyloTraits"
```

```
setwd("/Users/wolve/GitHub/QB2019_Crawley/2.Worksheets/11.PhyloTraits")
package.list <- c('ape', 'seqinr', 'phylobase', 'adephylo', 'geiger', 'picante', 'stats', 'RColorBrewer
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: 'permute'

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

## This is vegan 2.5-3

##
## 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 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
```

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.

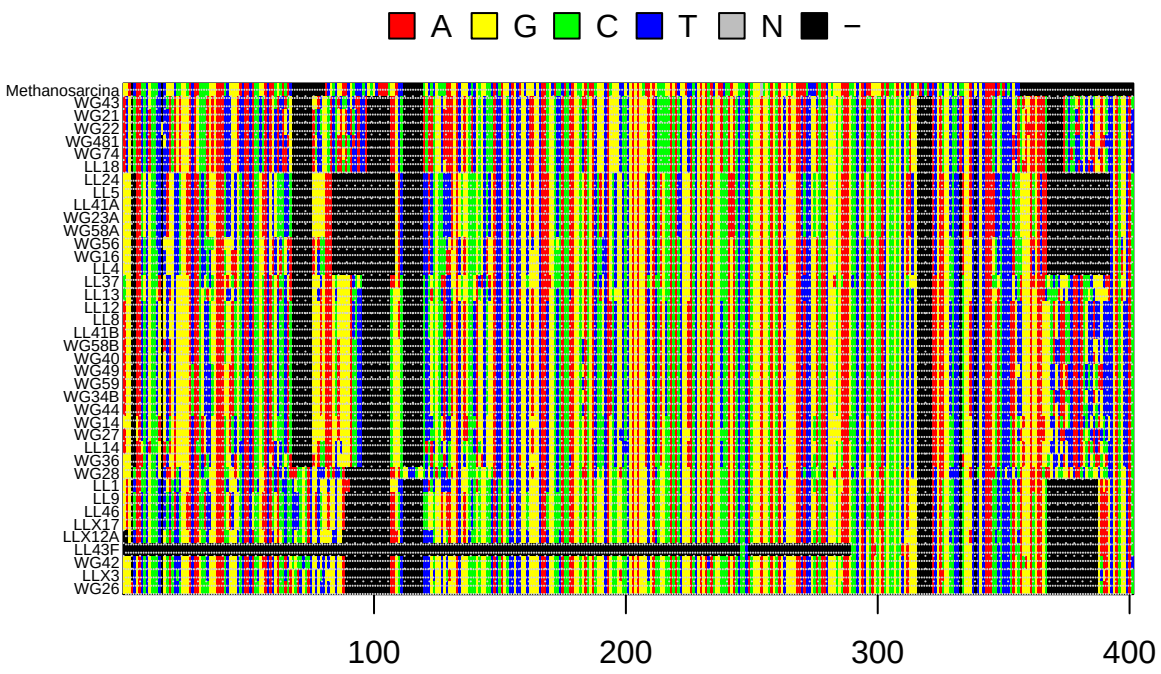
3) SEQUENCE ALIGNMENT

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.

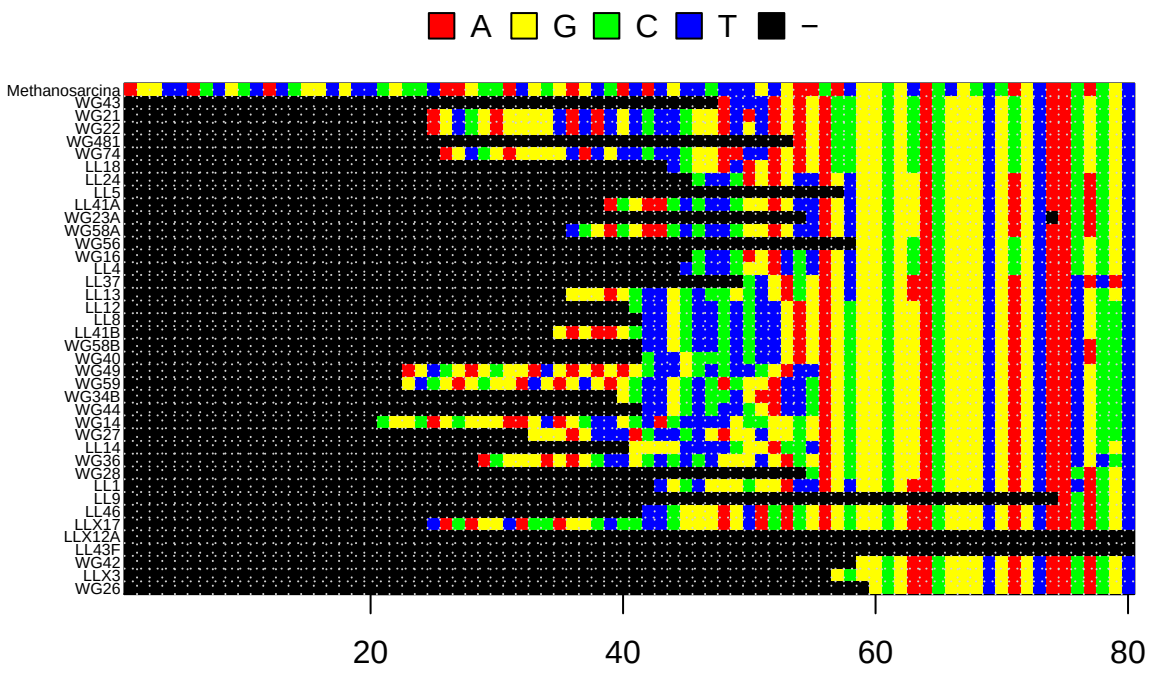
> ***Answer 1***: Characters in the .fasta file are lowercase, while characters in the .afa file are capitalized. Spaces in the .fasta file are replaced by dashes in the .afa file; the .afa file also contains dashes in some samples that are not present in the .fasta file. Samples are ordered differently in the .afa file.

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.

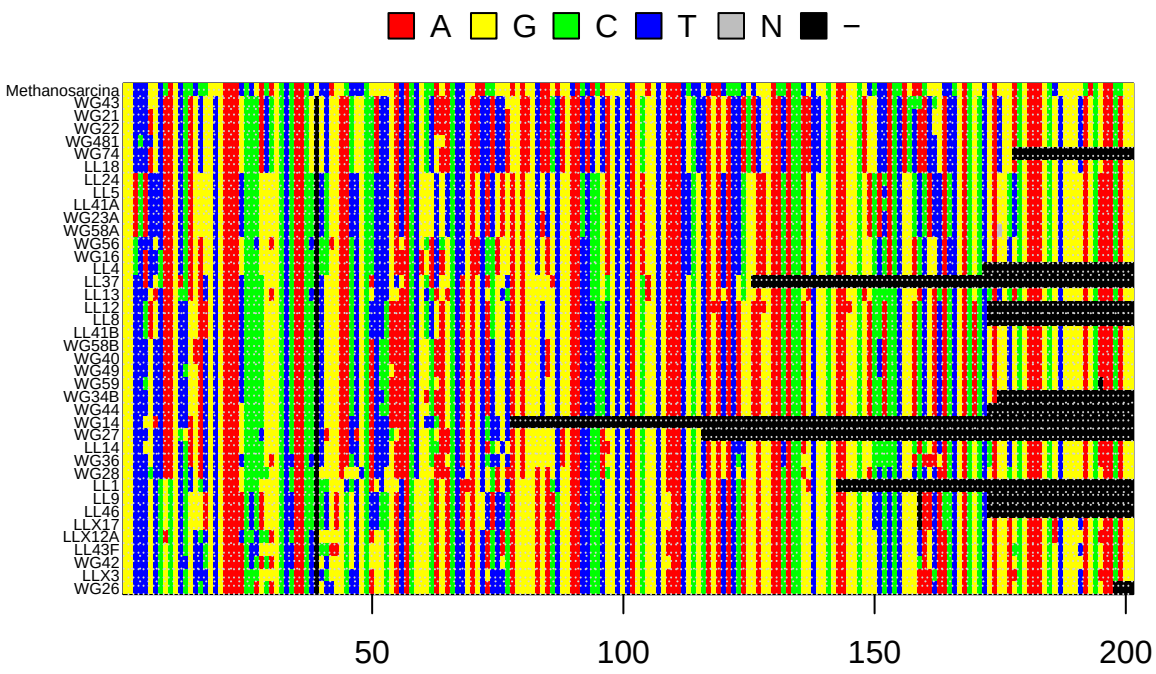
```
read.aln <- read.alignment(file="./data/p.isolates.afa", format = "fasta")
p.DNAbin <- as.DNAbin(read.aln)

#region of 16S rRNA gene from handout
window <- p.DNAbin[, 100:500]
image.DNAbin(window, cex.lab = 0.50)
grid(ncol(window), nrow(window), col = "lightgrey")
```



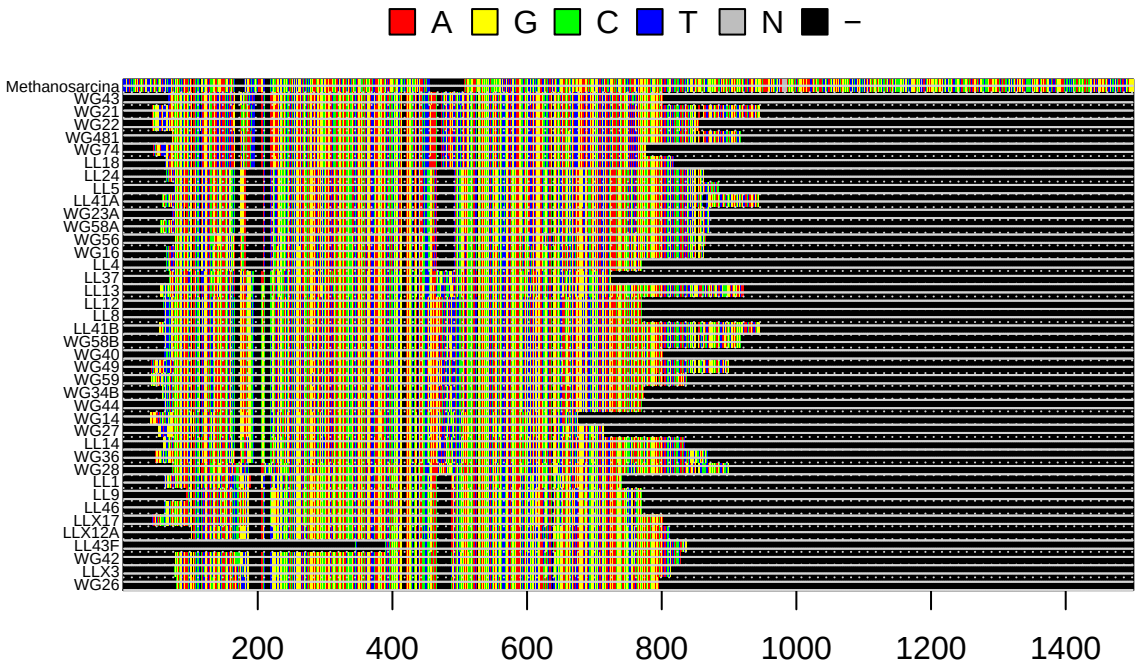

```
#other regions
window <- p.DNAbin[, 20:99]
image.DNAbin(window, cex.lab = 0.50)
grid(ncol(window), nrow(window), col = "lightgrey")
```




```
window <- p.DNAbin[, 600:800]
image.DNAbin(window, cex.lab = 0.50)
grid(ncol(window), nrow(window), col = "lightgrey")
```




```
window <- p.DNAbin[, 1:1500]
image.DNAbin(window, cex.lab = 0.50)
grid(ncol(window), nrow(window), col = "lightgrey")
```



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 be appropriate for phylogenetic inference and why?

Answer 2a: 1500 base pairs ***Answer 2b***: Regions that are mostly lacking in alignment gaps; differences can be seen between basepairs that are present.

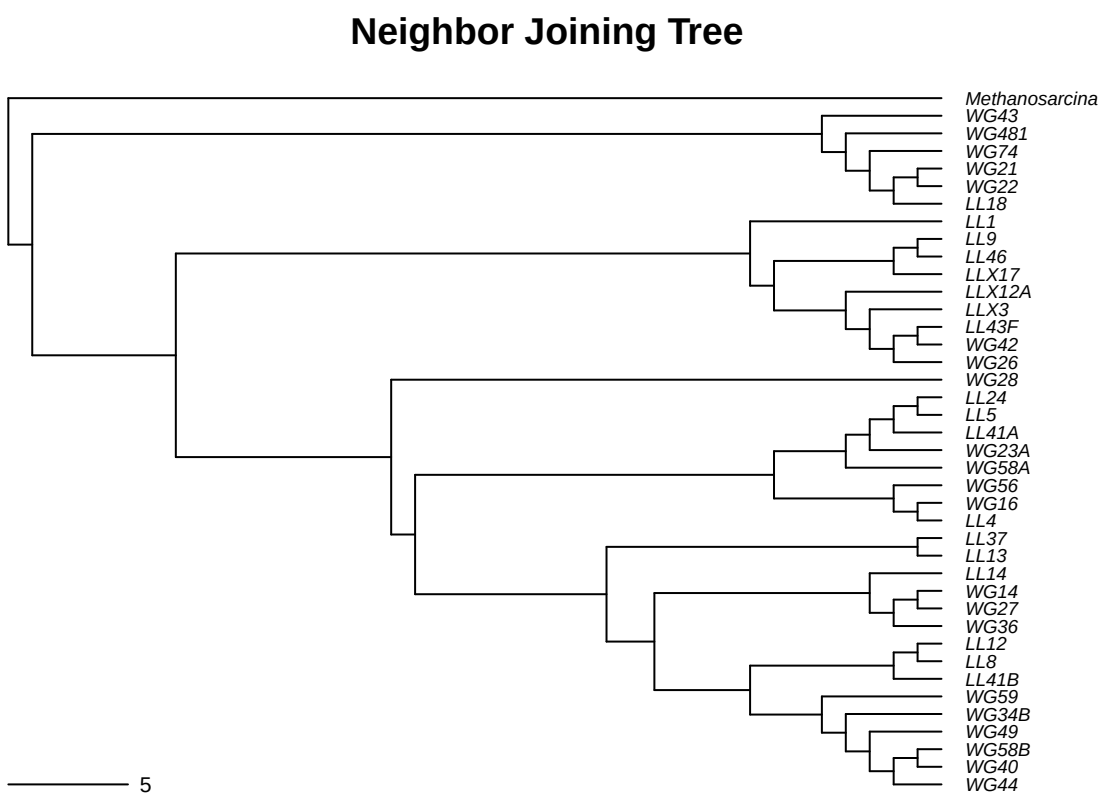
4) MAKING A PHYLOGENETIC TREE

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
seq.dist.raw <- dist.dna(p.DNAbin, model = "raw", pairwise.deletion = FALSE)

nj.tree <- bionj(seq.dist.raw)

outgroup <- match("Methanosarcina", nj.tree$tip.label)
nj.rooted <- root(nj.tree, outgroup, resolve.root = TRUE)

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)
```



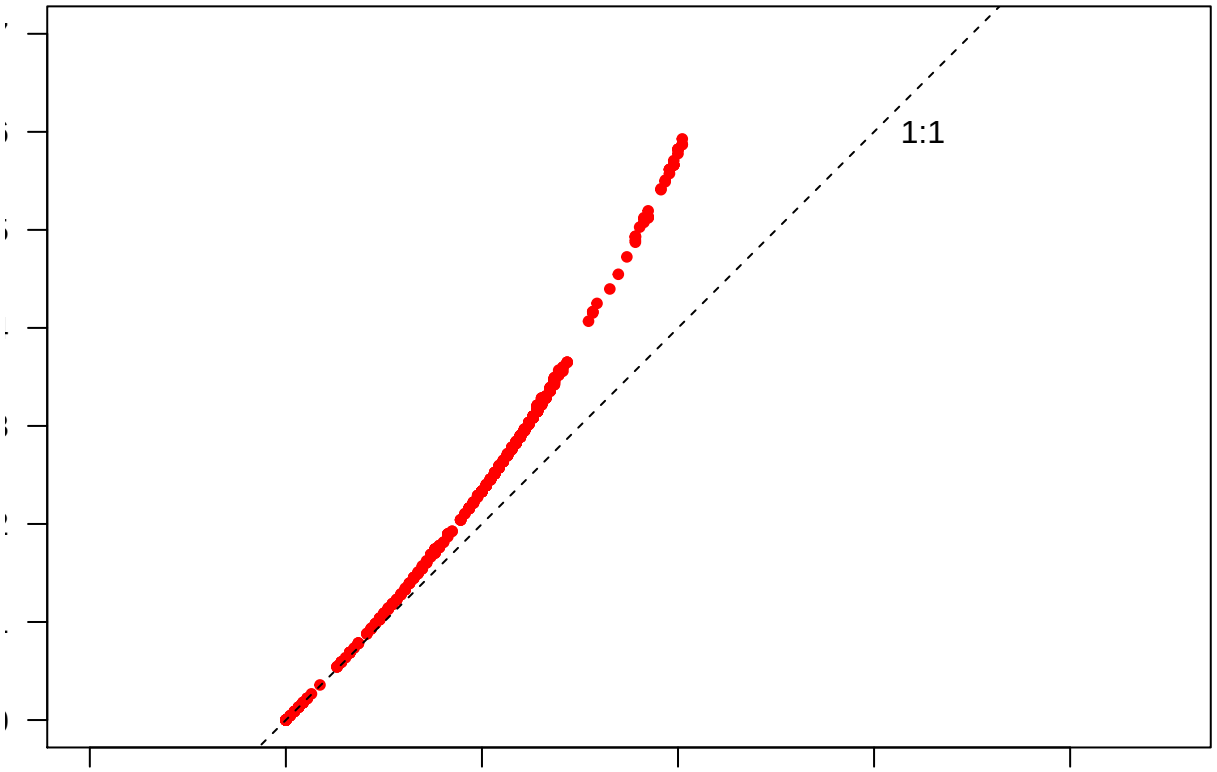
Question 3: What are the advantages and disadvantages of making a neighbor joining tree?

> ***Answer 3***: Neighbor joining is fast, and easily used for analyzing large datasets. However, it uses raw estimates of phylogenetic distance. A Neighbor joining tree may not account for multiple substitutions that may have occured at the same site or substitution biases.

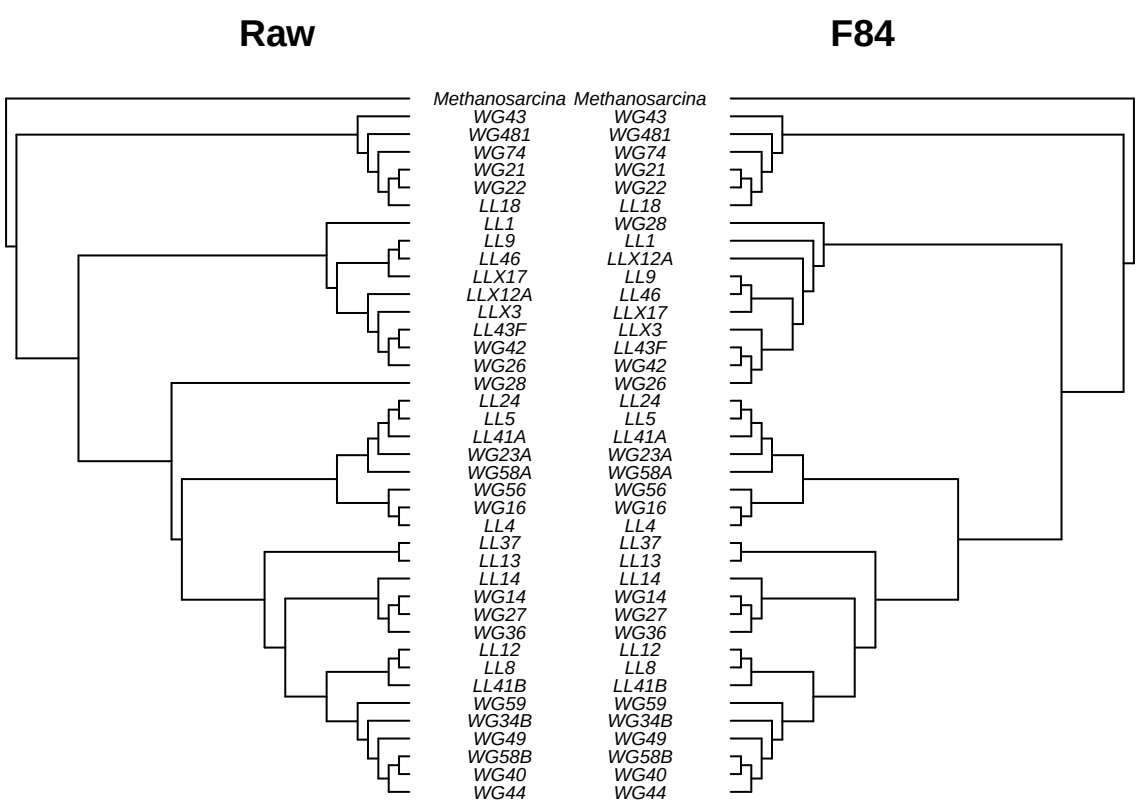
B) SUBSTITUTION MODELS OF DNA EVOLUTION

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 = FALSE)

par(mar = c(1, 1, 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")
```




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

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")
```

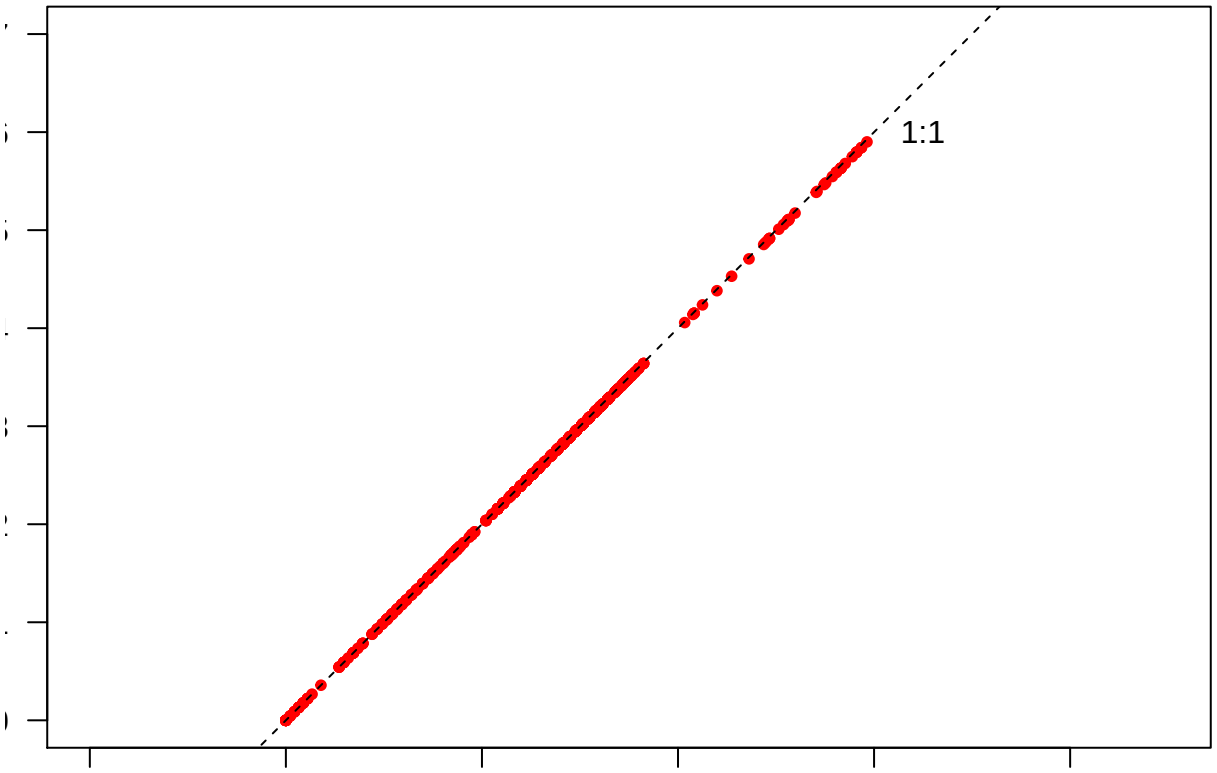



```
dist.topo(raw.rooted, F84.rooted, method = "score")
```

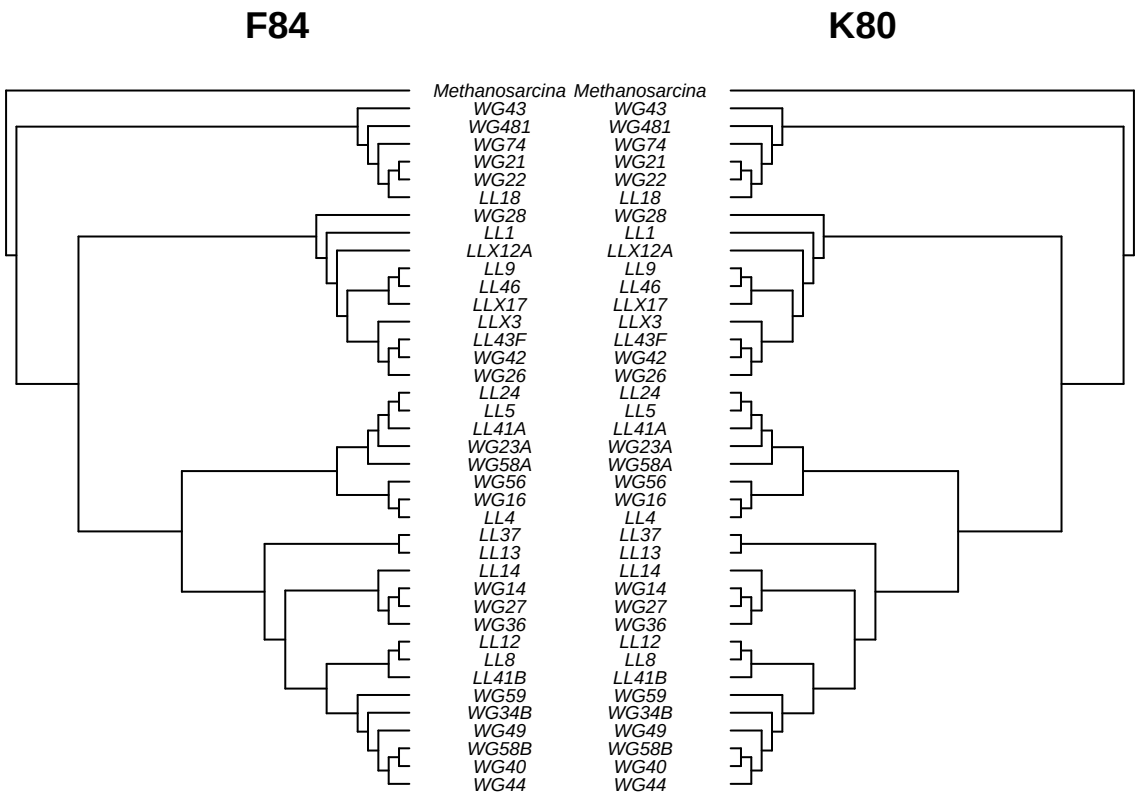
```
##            tree1
## tree2 0.04387426
```

In the R code chunk below, do the following:
1. pick another substitution model,
2. create a 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.

```
seq.dist.K80 <- dist.dna(p.DNAbin, model = "K80", pairwise.deletion = FALSE)
K80.tree <- bionj(seq.dist.K80)
K80.outgroup <- match("Methanosarcina", K80.tree$tip.label)
K80.rooted <- root(K80.tree, K80.outgroup, resolve.root = TRUE)

par(mar = c(1, 1, 2, 1) + 0.1)
plot(seq.dist.F84, seq.dist.K80,
     pch = 20, col = "red", las = 1, asp = 1, xlim = c(0, 0.7), ylim = c(0, 0.7),
     xlab = "F84 Distance", ylab = "K80 Distance")
abline(b = 1, a = 0, lty = 2)
text(0.65, 0.6, "1:1")
```




```
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(K80.rooted, type = "phylogram", direction = "left", show.tip.label = TRUE,
           use.edge.length = FALSE, adj = 0.5, cex = 0.6, label.offset = 2, main = "K80")
```



Question 4:

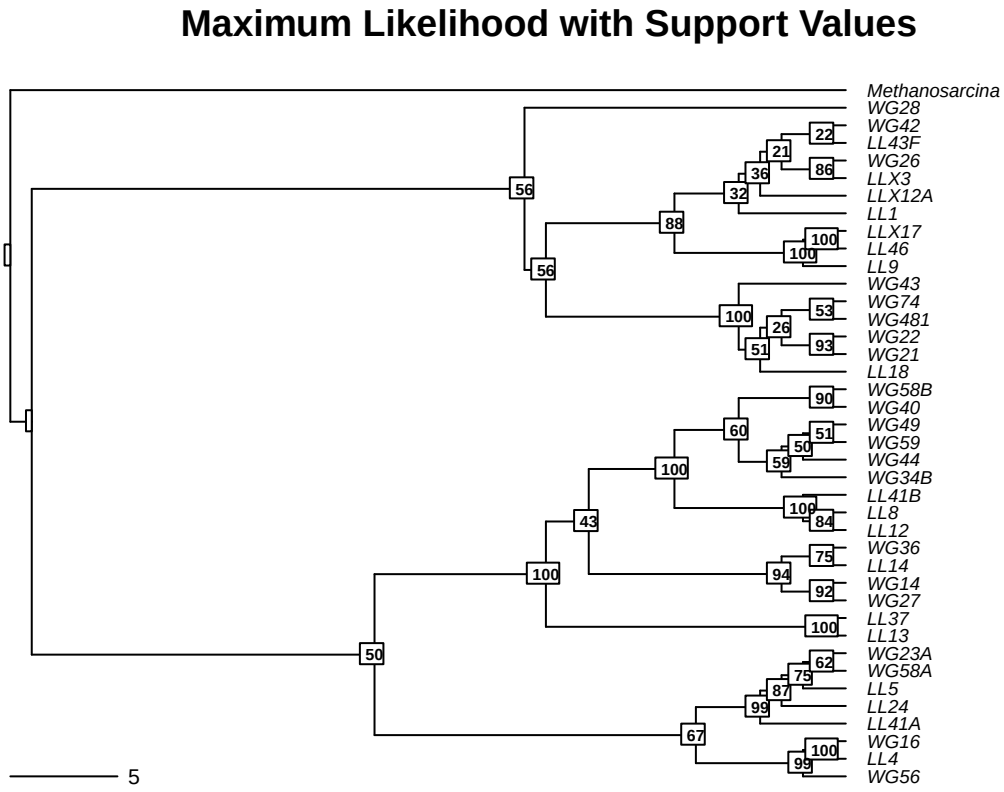
a. Describe the substitution model that you chose. What assumptions does it make and how does it compare to the F84 model?
b. Using the saturation plot and cophylogenetic plots from above, describe how your choice of substitution model affects your phylogenetic reconstruction. If the plots are inconsistent 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 used the Kimura model, which assumes equal frequencies of nucleotides, but recognizes that transitionn mutations occur with higher probability than transversion mutations. ***Answer 4b***: With regards to the F84 and K80 models, the choice of substitution model has no effect on the resulting phylogeny. This is not always the case, as is seen when comparing the neighbor-joining tree with the F84 model. Although the differences between the neighbour joining tree and F84 tree are slight, they are still important. ***Answer 4c***: The K80 model yielded the same results as the F84 model; substition rates are slow.

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 Suppor
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: Many of the lower-tier clades contain the same samples (ex: the ML group containing WG58B through LL13). Major differences between the two trees appear to be related to changes in early nodes. ***Answer 5b***: Bootstrapping a tree gives you an estimate of the reliability of the tree in its current form. ***Answer 5c***: Bootstrap values tell you the percentage of bootstrapped replicates in which the node appears; values above 95 indicate a well-supported node.

Answer 5d: The branches with the least support are found in the WG42-LL1 clade (with the exception of WG26-LLX3 pair). ***Answer 5e***: Since most of the given values are well under 95, the branches are not reliable.

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 = TRUE, row.names = 1)
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.

```
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)
}
nb <- as.matrix(levins(p.growth.std))
rownames(nb) <- row.names(p.growth)
colnames(nb) <- c("NB")
nb
```

```
##               NB
## LL1    0.6798191
## LL12   0.6899362
## LL13   0.7146458
## LL14   0.3525101
## LL18   0.6178110
## LL24   0.7117767
## LL37   0.7141804
## LL4    0.6131567
## LL41A  0.6219701
## LL41B  0.2187649
## LL43F  0.7379376
## LL46   0.4699429
## LL5    0.5248238
## LL8    0.7555647
## LL9    0.4788159
```

```
## LLX12A 0.8539080
## LLX17  0.4372624
## LLX3   0.7996862
## WG14   0.5678840
## WG16   0.7358387
## WG21   0.7852797
## WG22   0.6827565
## WG23A  0.7709106
## WG26   0.7823286
## WG27   0.7362067
## WG28   0.7547562
## WG34B  0.6022315
## WG36   0.7942277
## WG40   0.4298220
## WG42   0.8256545
## WG43   0.7604551
## WG44   0.7685069
## WG481  0.7085050
## WG49   0.5498899
## WG56   0.7368923
## WG58A  0.4432747
## WG58B  0.5955820
## WG59   0.6902266
## WG74   0.7471288
```

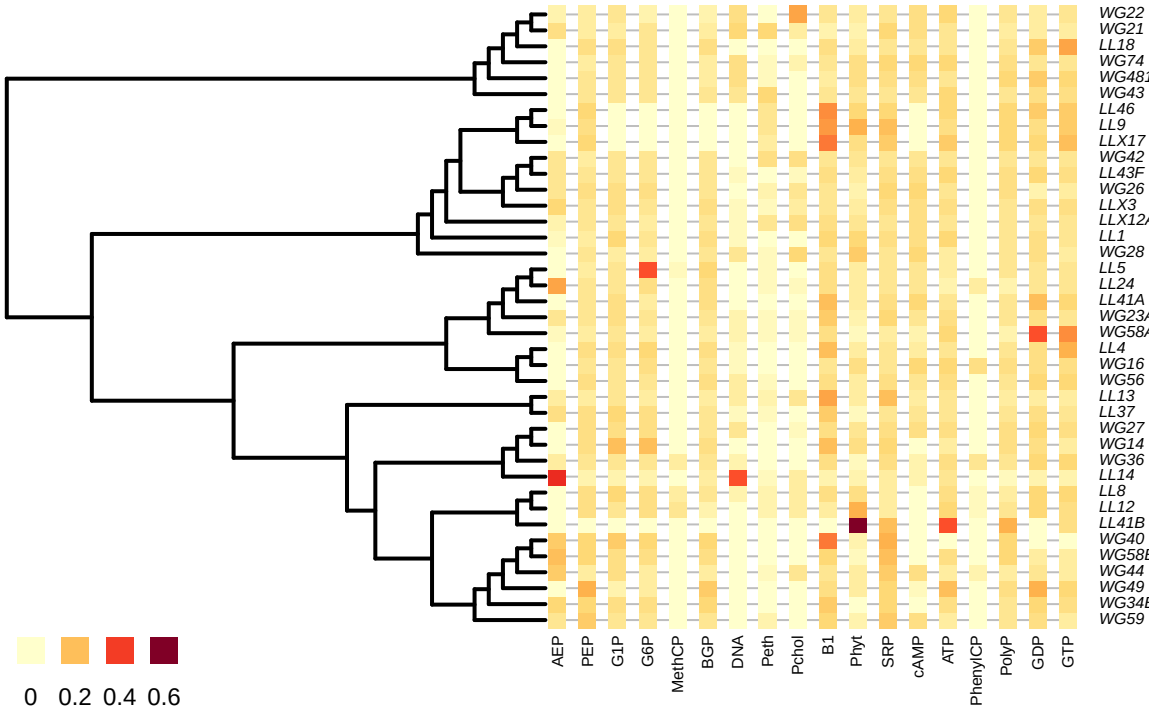
C. Visualizing Traits on Trees

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.

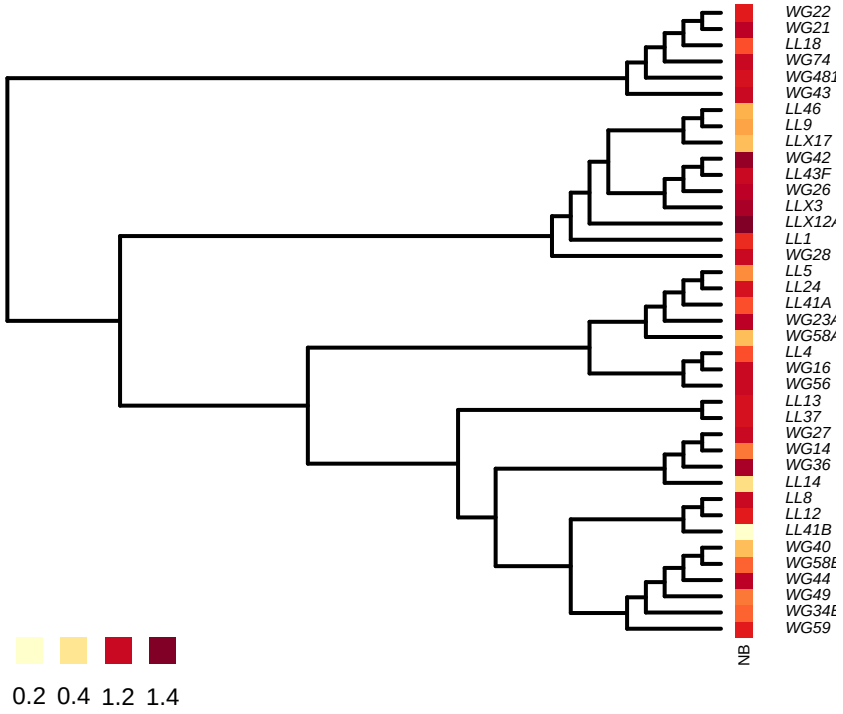
```
nj.tree <- bionj(seq.dist.F84)
outgroup <- match("Methanosarcina", nj.tree$tip.label)
nj.rooted <- root(nj.tree, outgroup, resolve.root = TRUE)
nj.rooted <- drop.tip(nj.rooted, "Methanosarcina")
```

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).

```
mypalette <- colorRampPalette(brewer.pal(9, "YlOrRd"))
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)
```




```
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)
```



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?

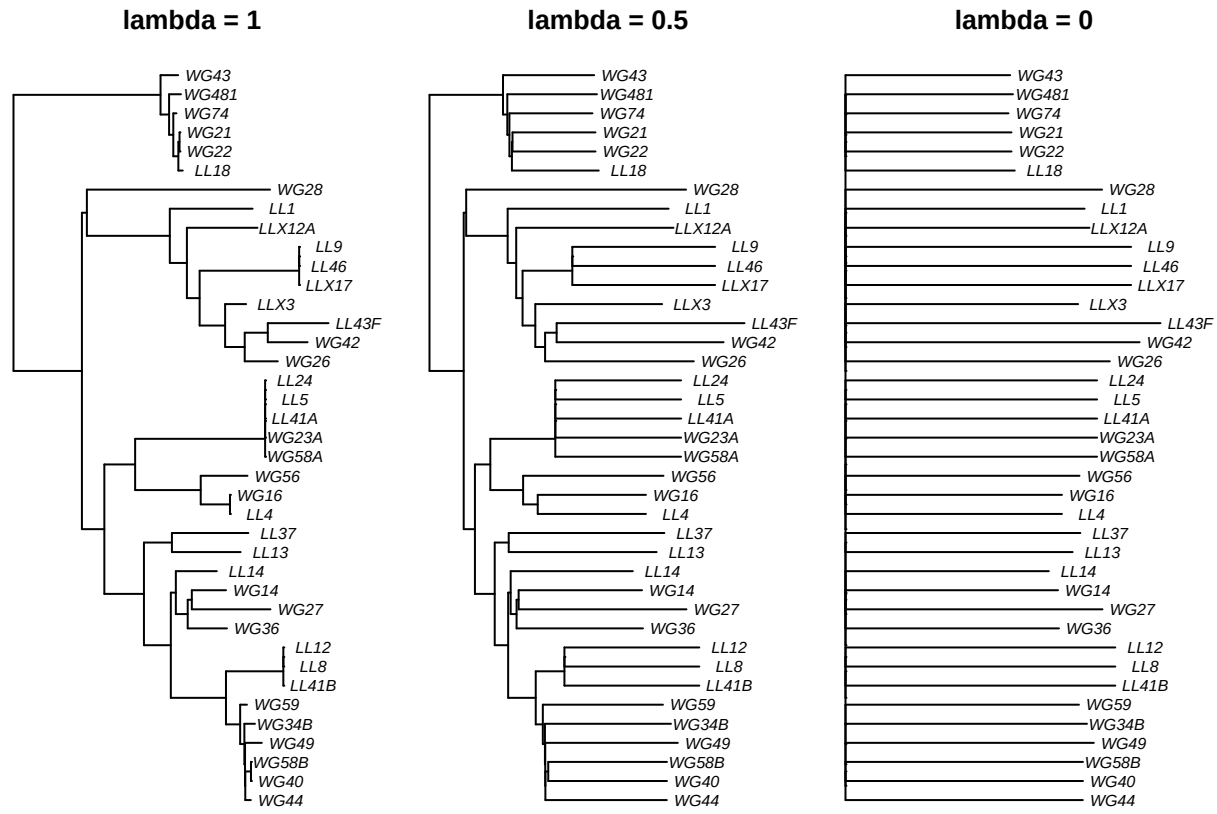
Answer 6a: In a nutrient-limited environment, bacteria that are able to utilize whatever forms of phosphorous are available will survive better than bacteria that specialize on a few sources. Bacteria isolated from the oligotrophic lake will have lower growth rates than bacteria isolated from the eutrophic lake.
Answer 6b: Bacteria isolates from the same lake will show similar patterns with regards to growth rate and niche breadth. The generalists will show fairly moderate and uniform growth rates across all forms of phosphorous; in addition, niche breadth values will be fairly high. Specialists will exhibit very high growth rates on one or two forms of phosphorous and very low growth rates on all others; niche breadth values will be fairly low.

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.

```
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 = "lambda = 0.5", cex = 0.7, adj = 0.5)
plot(nj.lambda.0, main = "lambda = 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.rooted, nb, model = "lambda")
```

```
## GEIGER-fitted comparative model of continuous data
##  fitted 'lambda' model parameters:
##  lambda = 0.020848
##  sigsq = 0.106492
##  z0 = 0.661368
##
##  model summary:
##  log-likelihood = 21.661104
##  AIC = -37.322208
##  AICc = -36.636494
##  free parameters = 3
##
```

```
## Convergence diagnostics:
##  optimization iterations = 100
##  failed iterations = 46
##  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 = 0.80
##
##  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 values of the untransformed and transformed trees are 0.020848 and 0.000000, respectively. ***Answer 7b***: The AIC score of the untransformed tree is -37.322208, while the AIC score of the transformed tree is -37.304587. Since the difference between the two scores is less than two, they are functionally equivalent. ***Answer 7c***: The results suggest there is no phylogenetic signal.

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.

```
nj.rooted$edge.length <- nj.rooted$edge.length + 10^-7

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")
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)
}

p.phylosignal[6, ] <- round(p.adjust(p.phylosignal[4, ], method = "BH"), 3)
p.phylosignal
```

```
##                    AEP      PEP      G1P      G6P  MethCP      BGP      DNA
## K                0.000    0.000    0.000    0.000   0.000    0.000    0.000
## PIC.var.obs   4373.157  664.095  948.941 5924.730 350.894  536.104  259.084
## PIC.var.mean 8337.672 1565.605 1892.723 3651.795 484.955 1758.214 4878.198
## PIC.var.P       0.221    0.057    0.119    0.764   0.382    0.025    0.003
## PIC.var.z      -0.868   -1.340   -1.175    0.912  -0.396   -1.741   -1.256
## PIC.P.BH        0.604    0.256    0.428    0.809   0.628    0.150    0.027
##                   Peth    Pchol       B1     Phyt      SRP     cAMP
## K                0.000    0.000    0.000    0.000    0.000    0.000
## PIC.var.obs   1446.463 2368.391 3517.018 9240.368 1307.025  690.723
## PIC.var.mean 1884.248 3351.130 5233.201 9334.981 1629.586 3048.836
## PIC.var.P       0.323    0.384    0.235    0.558    0.332    0.003
## PIC.var.z      -0.519   -0.561   -0.773   -0.012   -0.540   -2.441
## PIC.P.BH        0.628    0.628    0.604    0.717    0.628    0.027
##                    ATP PhenylCP    PolyP      GDP      GTP
## K                0.000    0.000    0.000    0.000    0.000
## PIC.var.obs   4040.137 1224.017 1126.345 4473.878 2721.766
## PIC.var.mean 3070.573  782.936 1236.365 3505.331 2914.380
## PIC.var.P       0.615    0.818    0.487    0.678    0.470
## PIC.var.z       0.429    0.947   -0.197    0.471   -0.142
## PIC.P.BH        0.738    0.818    0.674    0.763    0.674
```

```
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              50400.96          0.528
##   PIC.variance.Z
## 1     -0.0210621
```

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: There is no significant phylogenetic signal for niche breadth. Only two phosphorous resources yield significant phylogenetic signal (DNA and cAMP). ***Answer 8b***: The results suggest overdispersion for the two significant results.

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.

```
p.growth.pa <- as.data.frame((p.growth > 0.1) * 1)

p.growth.pa$name <- rownames(p.growth.pa)

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 = 31
##                      1 = 8
##   Phylogeny :  nj.rooted
##   Number of permutations :  1000
##
## Estimated D :  0.1192861
## Probability of E(D) resulting from no (random) phylogenetic structure :  0.004
## Probability of E(D) resulting from Brownian phylogenetic structure    :  0.369
```

```
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 = 37
##                      1 = 2
##   Phylogeny :  nj.rooted
##   Number of permutations :  1000
##
## Estimated D :  1.358269
## Probability of E(D) resulting from no (random) phylogenetic structure :  0.683
## Probability of E(D) resulting from Brownian phylogenetic structure    :  0.079
```

```
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 = 33
##                      1 = 6
##   Phylogeny :  nj.rooted
##   Number of permutations :  1000
##
## Estimated D :  0.7966953
## Probability of E(D) resulting from no (random) phylogenetic structure :  0.179
## Probability of E(D) resulting from Brownian phylogenetic structure    :  0.019
```

```
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 = 30
##                      1 = 9
##   Phylogeny :  nj.rooted
##   Number of permutations :  1000
##
## Estimated D :  0.6432848
## Probability of E(D) resulting from no (random) phylogenetic structure :  0.06
## Probability of E(D) resulting from Brownian phylogenetic structure    :  0.018
```

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?

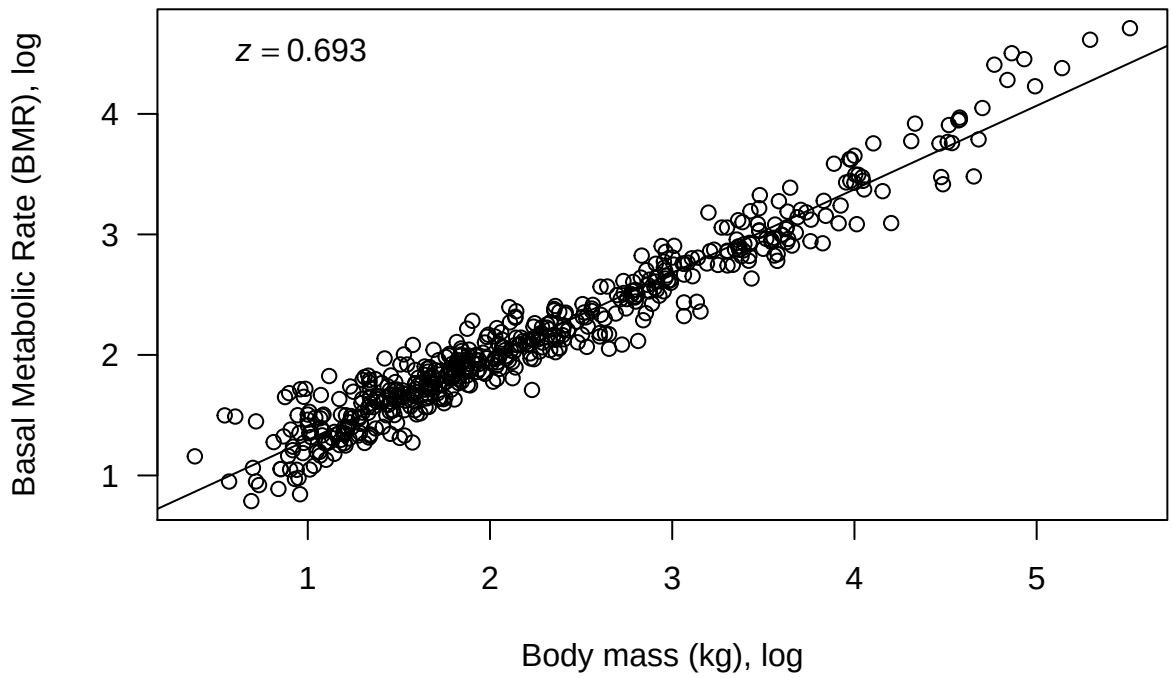
c. Discuss what factors might give rise to differences between the metrics.

Answer 9a: DNA and cAMP appear to be overdispersed, while AEP is randomly clumped.
Answer 9b: The results for DNA and cAMP align with the results of Blomberg's K analysis.
Answer 9c: Distance between species on a phylogenetic tree.

7) PHYLOGENETIC REGRESSION

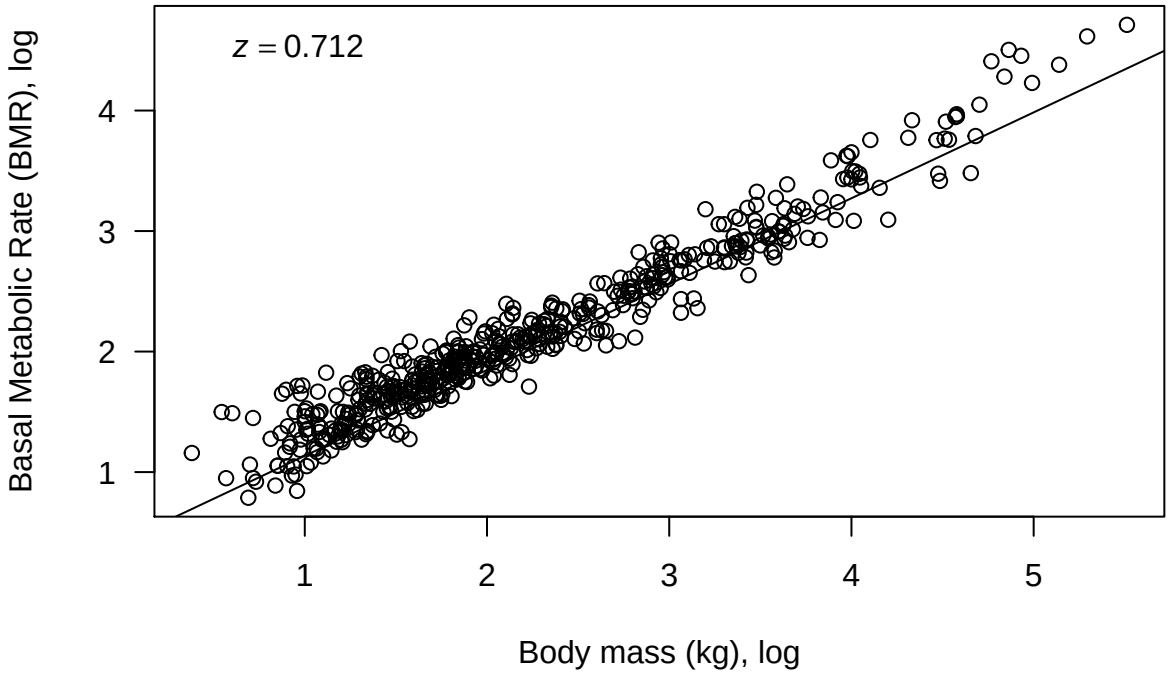
In the R code chunk below, do the following:
1. Load and clean the mammal phylogeny and trait dataset, 2. Fit a linear model to the trait dataset, examining the relationship between mass and BMR, 2. Fit a phylogenetic regression to the trait dataset, taking into account the mammal supertree

```
mammal.Tree <- read.tree("./data/mammal_best_super_tree_fritz2009.tre")
mammal.data <- read.table("./data/mammal_BMR.txt", sep = "\t", header = TRUE)
mammal.data <- mammal.data[,c("Species", "BMR_.mlO2.hour.", "Body_mass_for_BMR_.gr.")]
mammal.species <- array(mammal.data$Species)
pruned.mammal.tree <- drop.tip(mammal.Tree, mammal.Tree$tip.label[-na.omit(match(mammal.species, mammal
pruned.mammal.data <- mammal.data[mammal.data$Species %in% pruned.mammal.tree$tip.label,]
rownames(pruned.mammal.data) <- pruned.mammal.data$Species

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 =
     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))
text(0.5, 4.5, eqn, pos = 4)
```




```
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 =
     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))
```

```
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 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 10a: The variables are otherwise not independent. ***Answer 10b***: In a phylogenetic, the variance of the residual errors are described by a covariance matrix that takes into account the branch lengths of the underlying phylogeny. ***Answer 10c***: Slope increased when accounting for shared evolutionary history, but fit appears to worsen. ***Answer 10d***: If one of the variables is associated with closely related species; i.e. charting specific leaf area with disease susceptibility.

7) SYNTHESIS

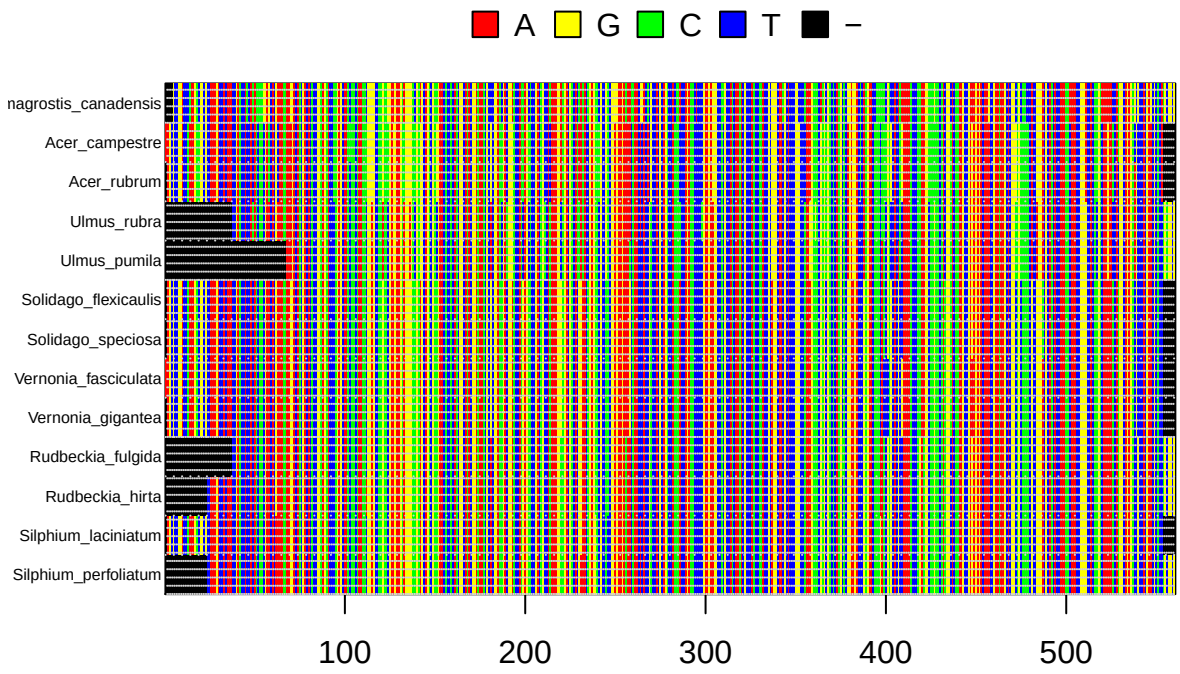
Work with members of your Team Project to obtain reference sequences for 10 or more taxa in your study. Sequences for plants, animals, and microbes can found in a number of public repositories, but perhaps the most commonly visited site is the National Center for Biotechnology Information (NCBI) https://www.ncbi.nlm.nih.gov/. 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.

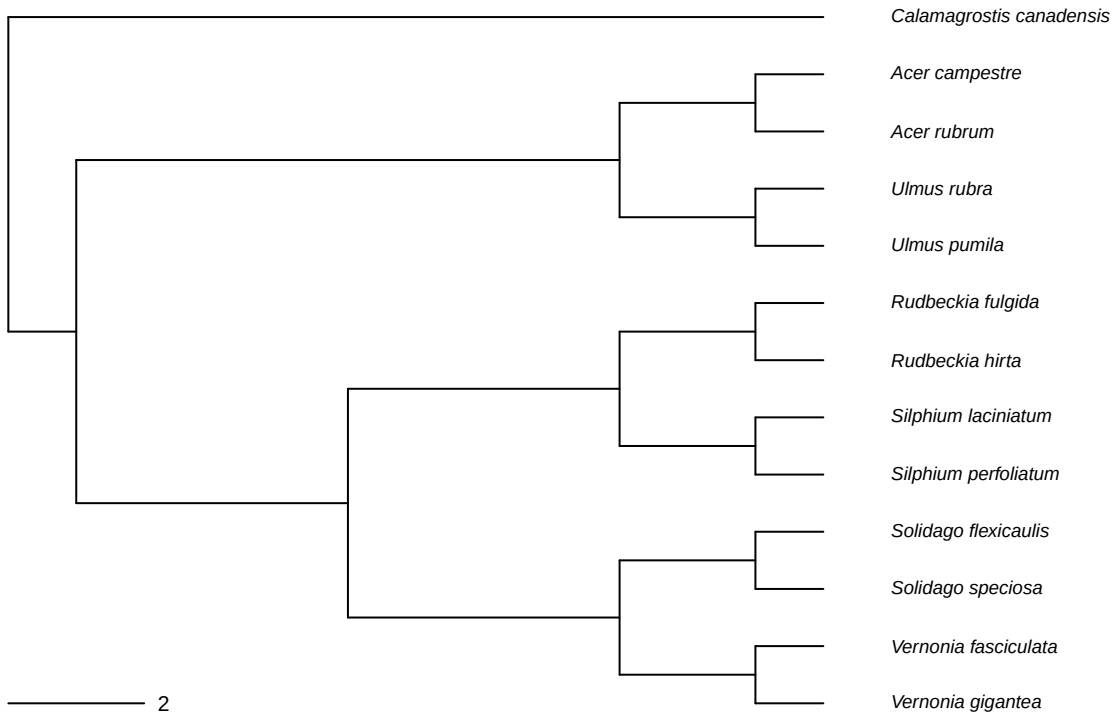
But before your team proceeds, you need to give some thought to which gene you want to focus on. For microorganisms like the bacteria we worked with above, many people use the ribosomal gene (i.e., 16S rRNA). This has many desirable features, including it is relatively long, highly conserved, and identifies 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?

```
read.plant <- read.alignment(file="./plant_seq.afa", format = "fasta")
plant.DNAbin <- as.DNAbin(read.plant)

window <- plant.DNAbin[, 1:560]
image.DNAbin(window, cex.lab = 0.50)
grid(ncol(window), nrow(window), col = "lightgrey")
```




```
plant.dist.F84 <- dist.dna(plant.DNAbin, model = "F84", pairwise.deletion = FALSE)
plant.tree <- bionj(plant.dist.F84)
plant.outgroup <- match("Calamagrostis_canadensis", plant.tree$tip.label)
plant.rooted <- root(plant.tree, plant.outgroup, resolve.root = TRUE)

par(mar = c(1, 1, 2, 1) + 0.1)
plot.phylo(plant.rooted, main = "F84 Tree", "phylogram", use.edge.length = FALSE,
           direction = "right", cex = 0.6, label.offset = 1)
add.scale.bar(cex = 0.7)
```



Yes; all congeners are grouped together, as well as all asters. Unsure how this could be changed with regards to our team's analyses; we could possibly make a tree looking at the relatedness of all planted/volunteer species?