Phylogenetic Diversity - Traits

Erik Parker; Z620: Quantitative Biodiversity, Indiana University

22 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).

```
## [1] "/media/Datas/GitHub/QB2017_Parker/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.

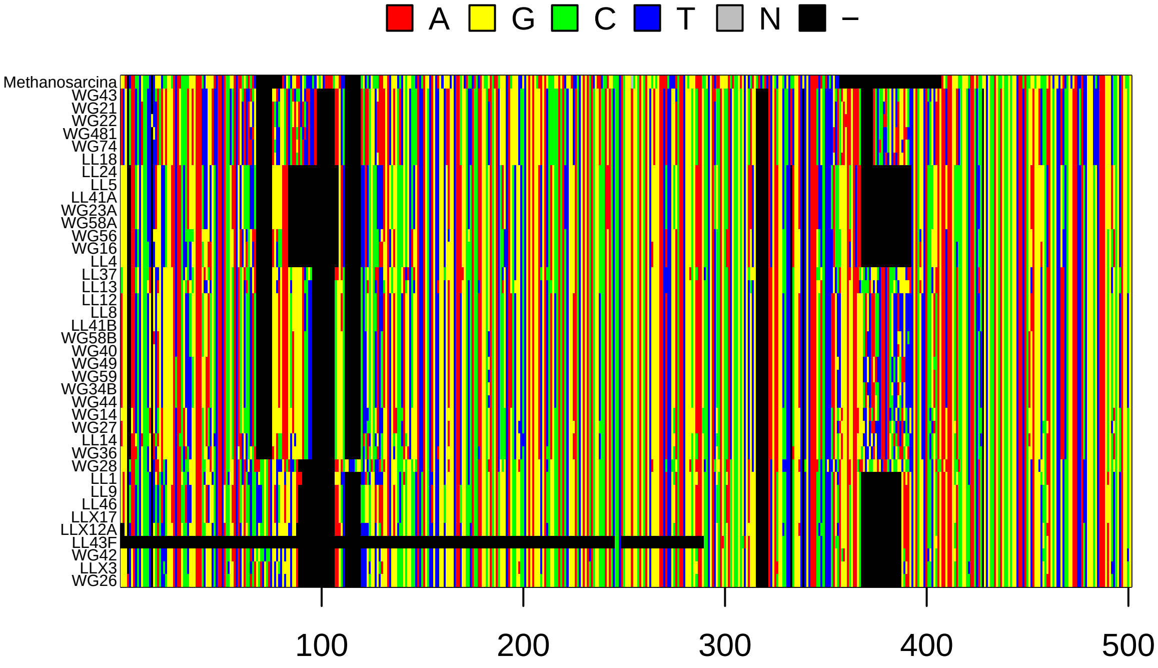
3) SEQUENCE ALIGNMENT

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.

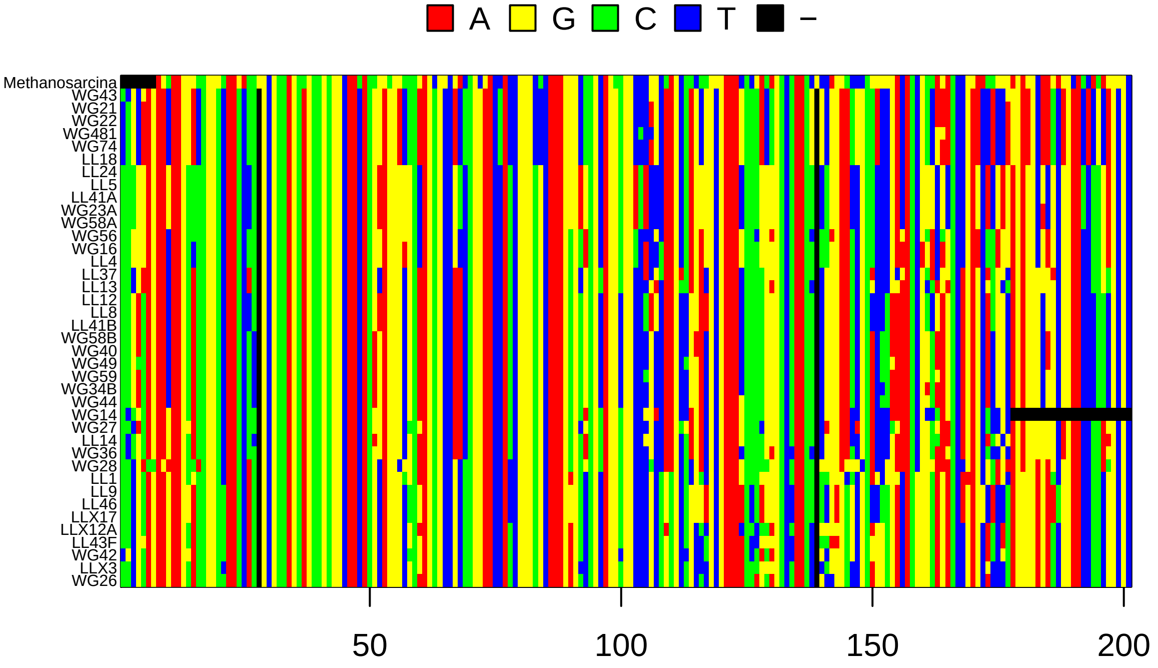
> ***Answer 1***: The fasta file is a list of sequences for a bunch of bacterial genes, while the afa file seems to be some sort of alignment? The fasta file is just showing nucleic acids, while the afa file also has dashes which seem like they represent gaps in some alignment of the reported genes to a reference.

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.

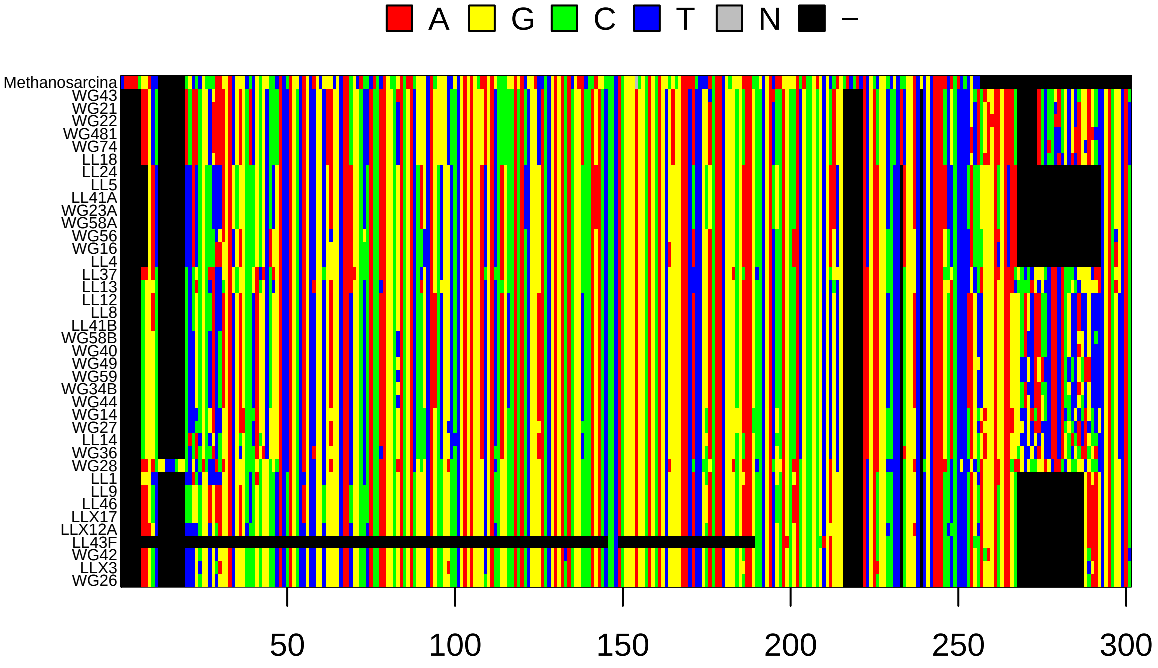
```r
read.aln <- read.alignment(file = "./data/p.isolates.afa", format = "fasta")

p.DNAbin <- as.DNAbin(read.aln)

window1 <- p.DNAbin[, 100:600]
window2 <- p.DNAbin[,500:700]
window3 <- p.DNAbin[,200:500]
window4 <- p.DNAbin[,1:1000]


#visualize windows

image.DNAbin(window1, cex.lab = .5)
```



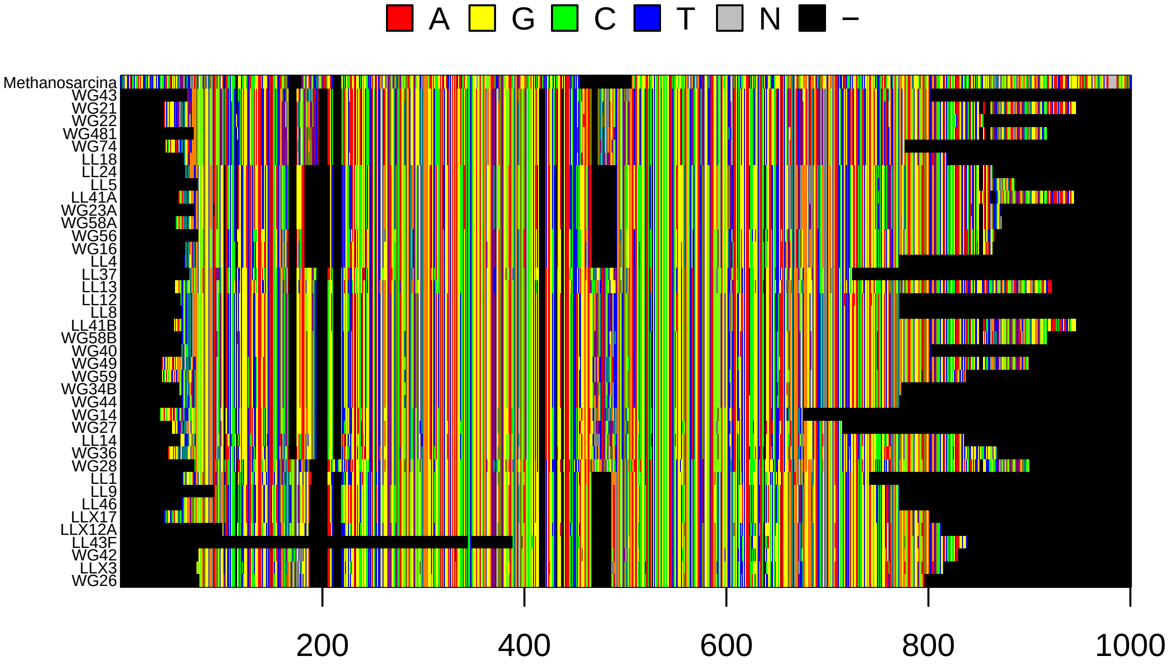
```r
image.DNAbin(window2, cex.lab = .5)
```



```r
image.DNAbin(window3, cex.lab = .5)
```



```r
image.DNAbin(window4, cex.lab = .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 reads?

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

Answer 2a:
The reads are all shorter than 1000bp, most seem to be between 600bp at the longest and about 100 at the shortest.

Answer 2b: It seems like the regions between ~200 and ~500 bp, and 500 and ~700 bp are the most appropriate for phylogenetic analysis as they are present in almost all of the samples. 500 to 700bp in particular seems the best.

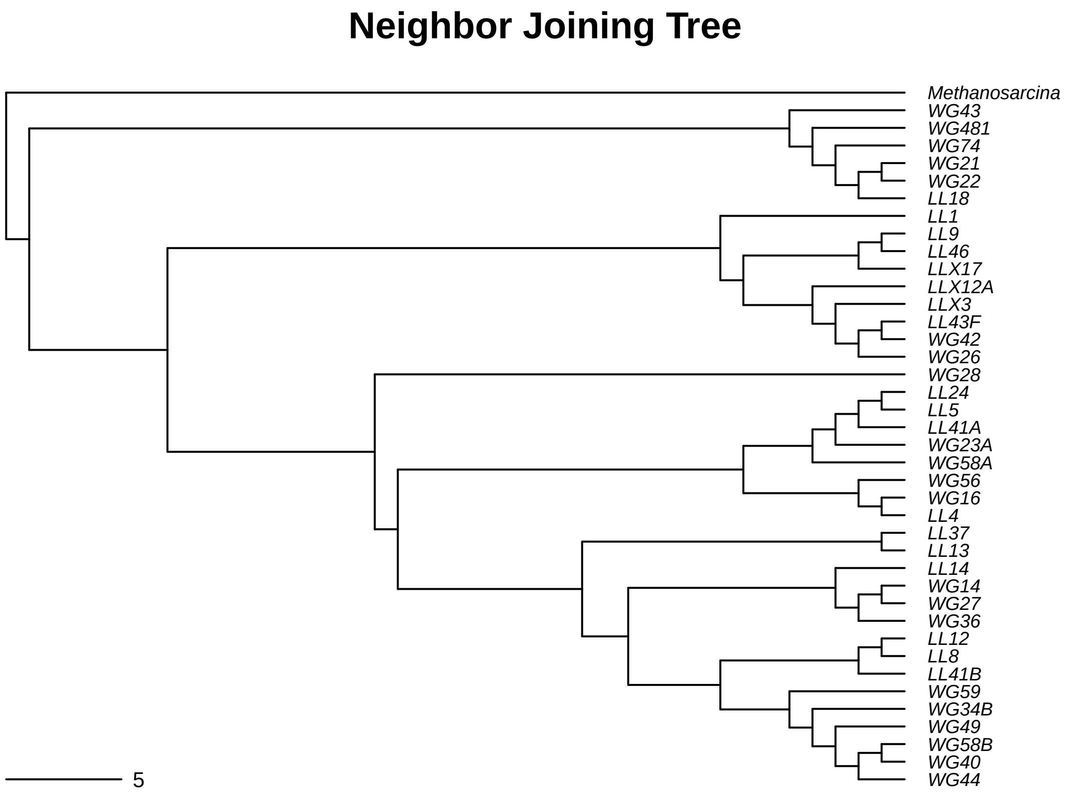
4) MAKING A PHYLOGENETIC TREE

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)

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)
```



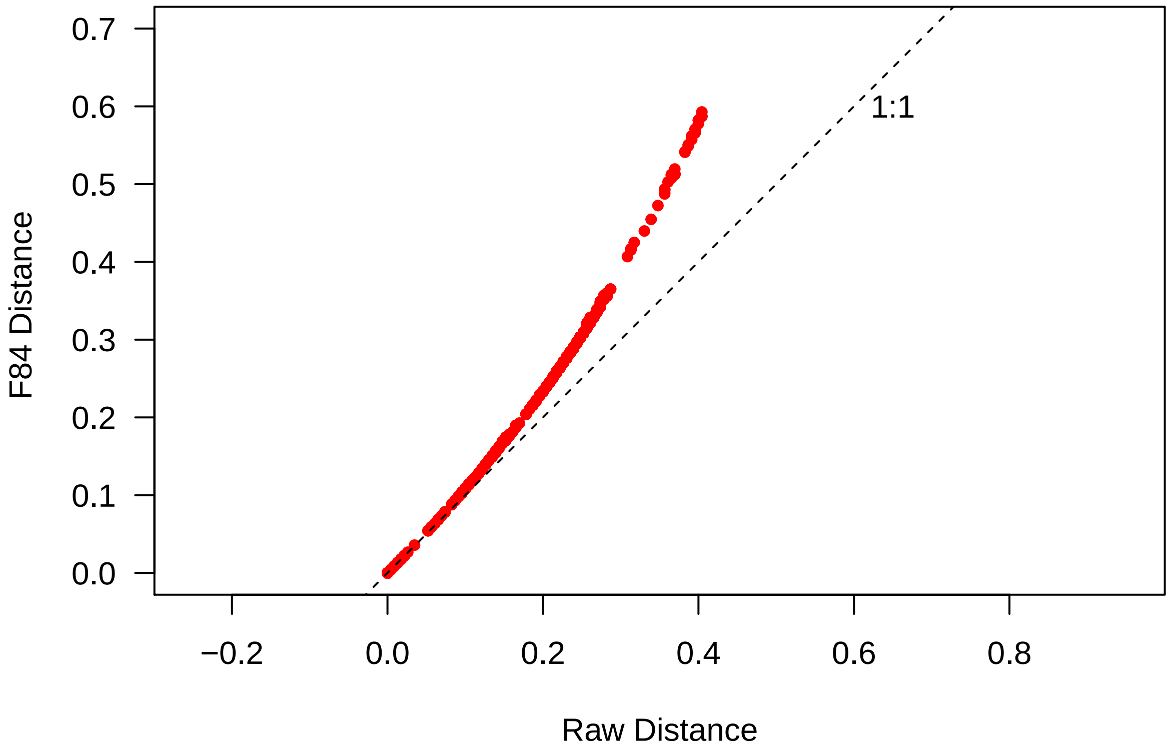
Question 3: What are the advantages and disadvantages of making a neighbor joining tree?

> ***Answer 3***: The main advantage of neighbor joining trees seems to be that it is relatively fast due to the lack of complex assumptions. The main disadvantage seems to be that it is very basic and not very reflective of evolutonary reality in many cases where relationships can't be conclusively determined just from data from a few sequences.

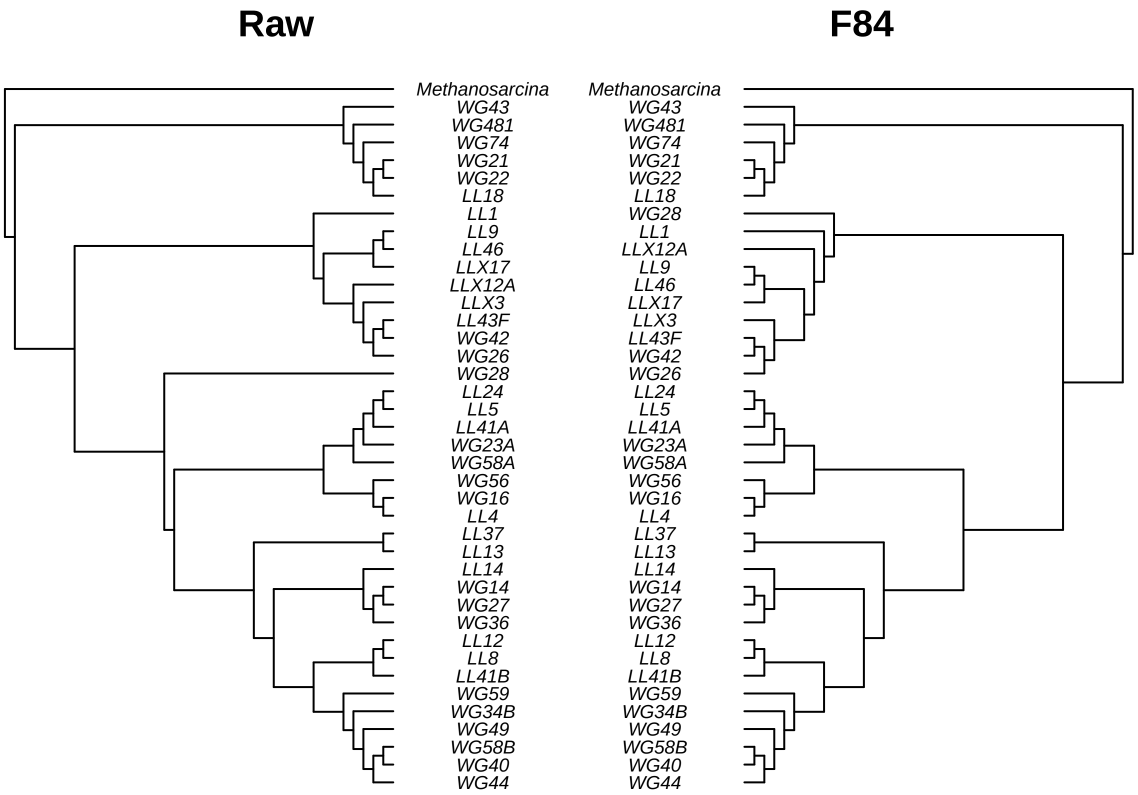
B) SUBSTITUTION MODELS OF DNA EVOLUTION

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
seq.dist.F84 <- dist.dna(p.DNAbin, model = "F84", pairwise.deletion = FALSE)

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")
```

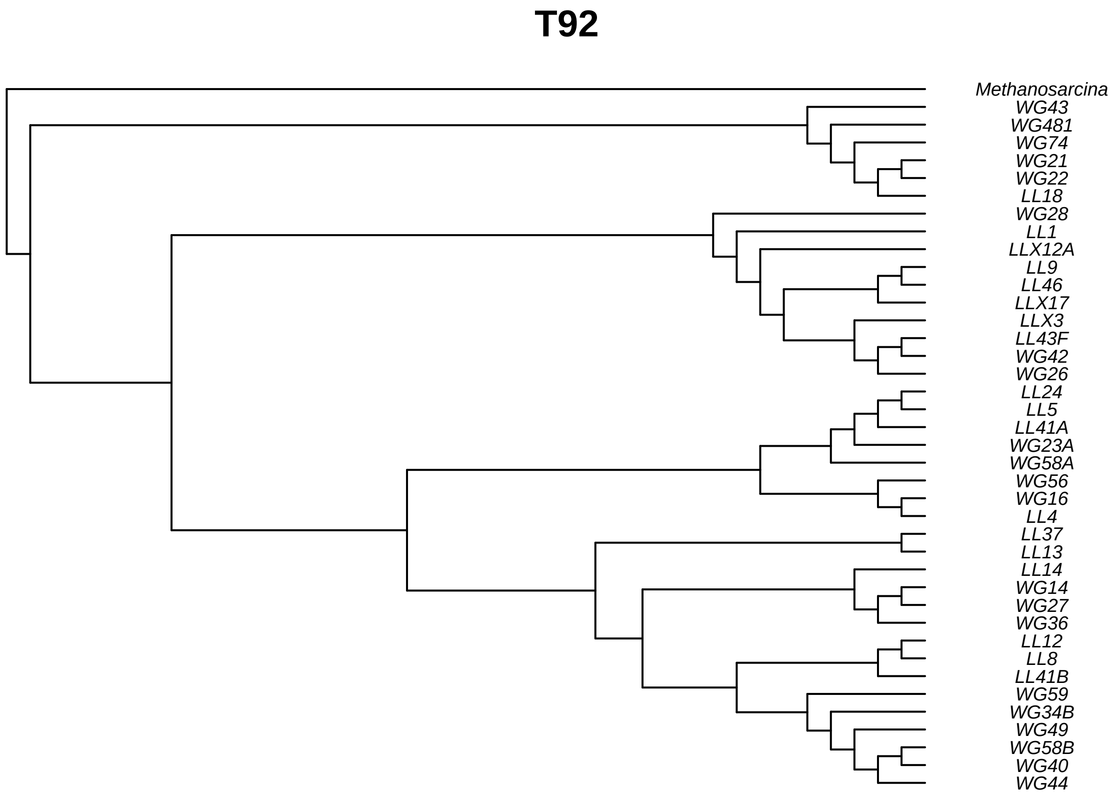


```
#Figure shows that the F84 distance is greater than the raw distance, apparently means that it is corre

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 =
par(mar = c(1,0,2,1))
plot.phylo(F84.rooted, type = "phylogram", direction = "left", show.tip.label = TRUE, use.edge.length =
```

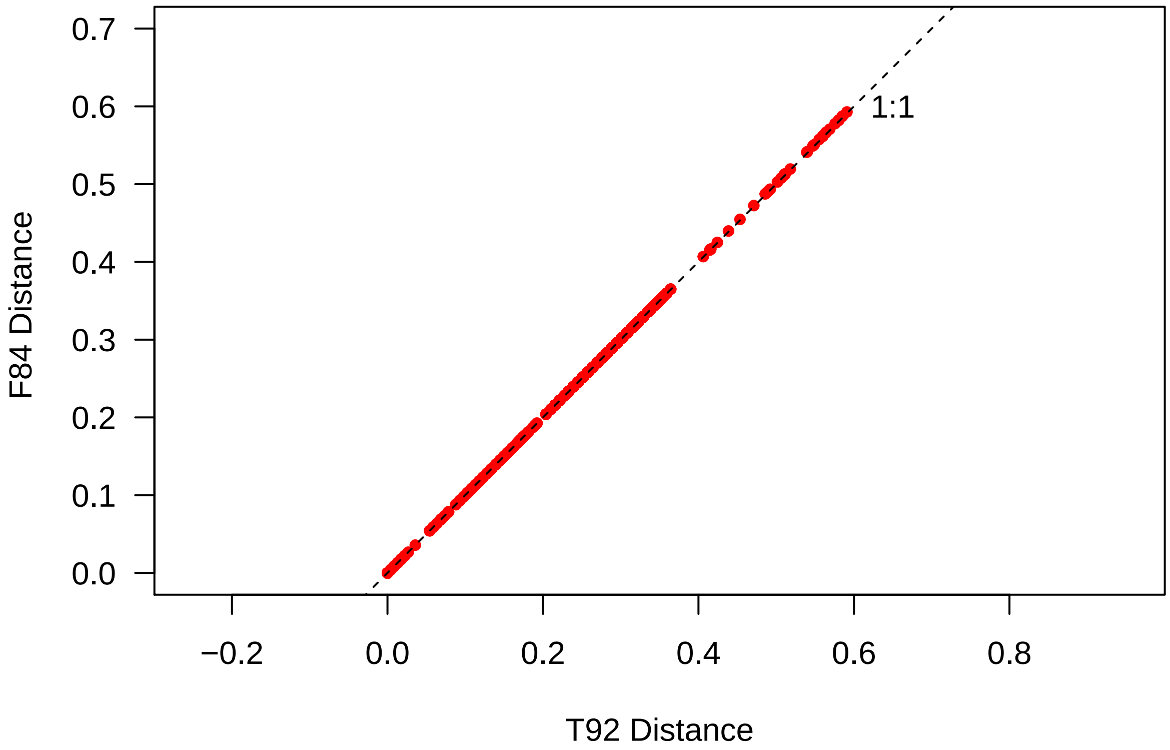


In the R code chunk below, do the following:
1. pick another substitution model,
2. create and distance matrix and tree for this model,
3. make a saturation plot that compares that model to the *Felsenstein (F84)* model,
4. make a cophylogenetic plot that compares the topologies of both models, and
5. be sure to format, add appropriate labels, and customize each plot.

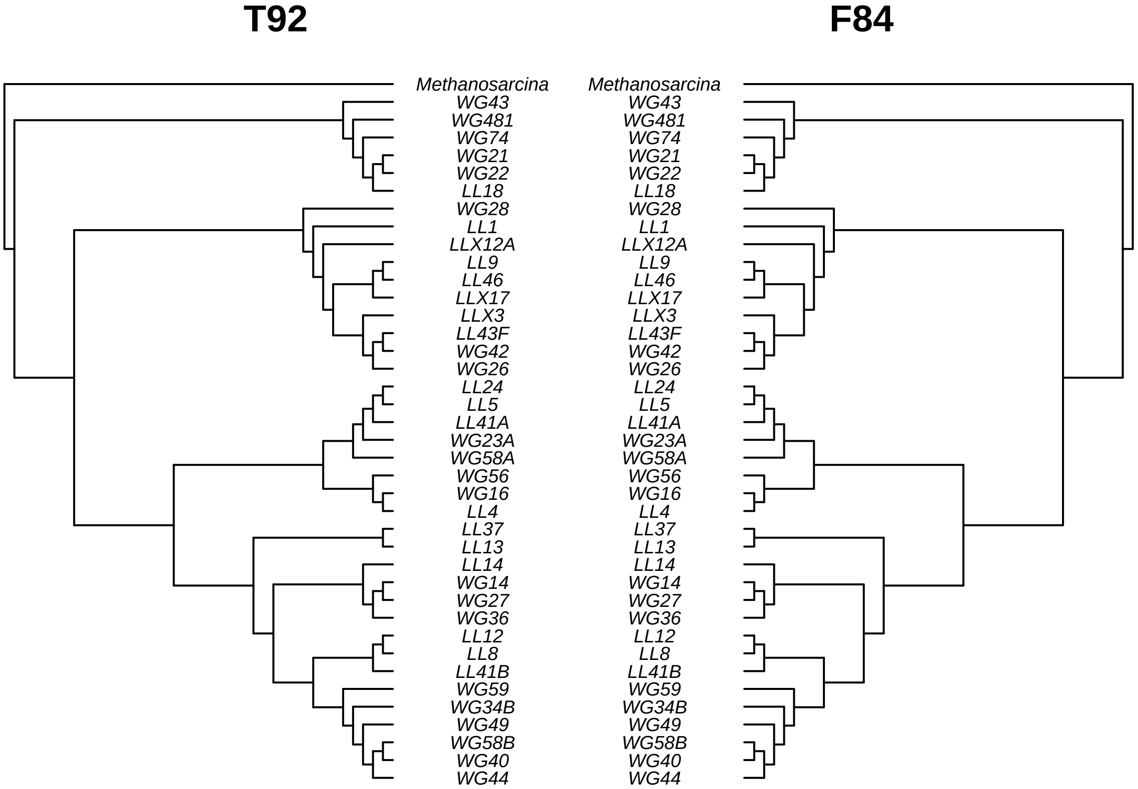
```r
seq.dist.T92 <- dist.dna(p.DNAbin, model = "T92", pairwise.deletion = FALSE)

T92.tree <- bionj(seq.dist.T92)

T92.outgroup <- match("Methanosarcina", T92.tree$tip.label)

T92.rooted <- root(T92.tree, T92.outgroup, resolve.root = TRUE)

par(mar = c(1, 1, 2, 0))
plot.phylo(T92.rooted, type = "phylogram", direction = "right", show.tip.label = TRUE, use.edge.length =
```



```
par(mar = c(5,5,2,1) + 0.1)
plot(seq.dist.T92, seq.dist.F84,
     pch = 20, col = "red", las = 1, asp = 1, xlim = c(0,0.7), ylim = c(0, 0.7),
     xlab = "T92 Distance", ylab = "F84 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(T92.rooted, type = "phylogram", direction = "right", show.tip.label = TRUE, use.edge.length =
par(mar = c(1,0,2,1))
plot.phylo(F84.rooted, type = "phylogram", direction = "left", show.tip.label = TRUE, use.edge.length =
```



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:
The T92 model recognizes that transition mutations (A<->G of C<->T) occur more frequently and with a higher probability than do transversion mutations (purine to pyrimidine or vice versa) while also accounting for the levels of G and C content in the reads.

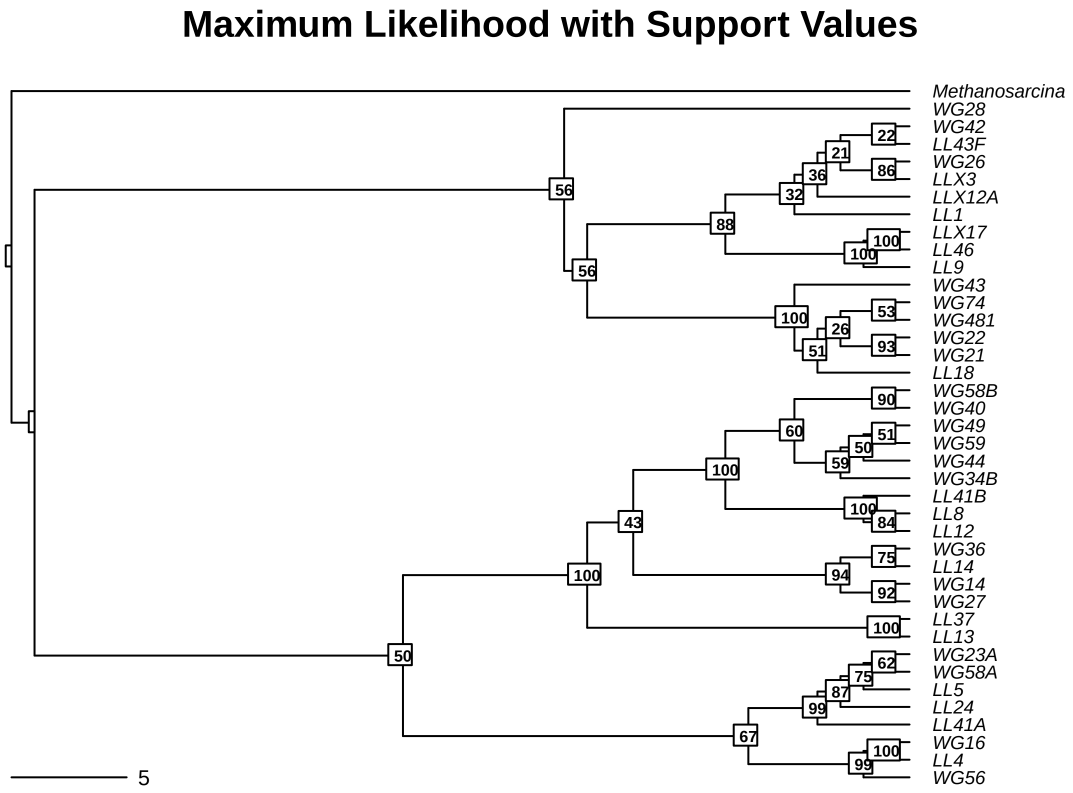
Answer 4b: Apparently in this case, my choice of substitution model did nothing to affect a difference in the end phylogenetic reconstruction. Both models (I assume) were different from the raw distance, but were identical to oneanother in both the saturation and cophylogenetic plots.

Answer 4c: The trees constructed using the T92 and F84 models are identical. This tells me that, for these samples, accounting for G and C content is not important - it gets the same end results as the model assuming that the nueclotides are at equal frequencies.

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 and neighbor-joining trees appear to be extremely different, both in terms of overall topology, but also the positions and relatedness of different species on the tree.

Answer 5b:
Bootstrapping is done to provide an idea of the accuracy of our estimated tree. Was that tree

arrived at because of a wealth of strong evidence? Or was the program building the tree less sure of its final product?

Answer 5c: Bootstrap values tell us how sure we can be that a particular node is assigned correctly. These values are reached through repeated random resampling of the data followed by assembly of a new tree using only those randomly selected points. These new trees are then all compared to the original estimated tree, with congruent results increasing the bootstrap value.

Answer 5d:
Unfortunately quite a few of the branches have low support, and a good number also show very low support. The lowest of these values are at branches near the top of the tree containing WG42, LL43F, WG26, LLX3, LLX12A, and LL1.

Answer 5e:
Those named branches above all show such low bootstrap support along much of their length that I don't think it would be wise to put much trust in their arrangement. By our tree, the only thing that seems to be relatively likely is that those named groups are seperated from the nearest sister cluster quite convincingly - but within that disputed cluster it's mostly anyone's guess what the reality is.

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")
```

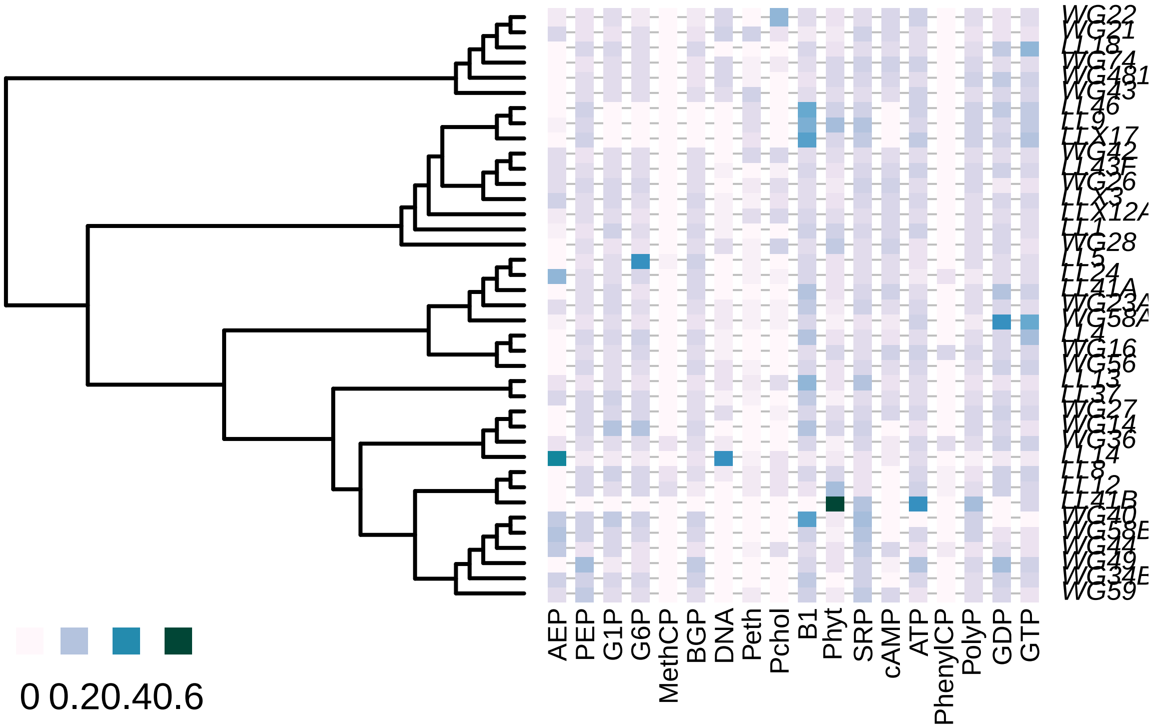
C. Visualizing Traits on Trees

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.

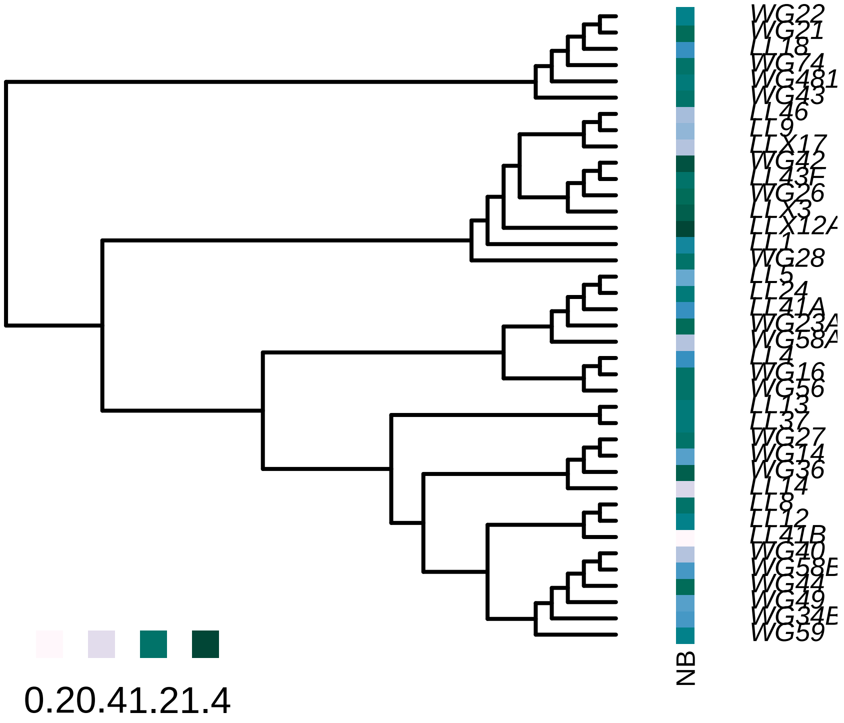
```
nj.tree <- bionj(seq.dist.T92)

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, "PuBuGn"))


#Tree mapping phosphorus traits
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 = .8, 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)
```



```
#tree mapping 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 = 0.8, 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 = 0.8, 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 order to outcompete other microbes for a particular phosphorous source, species must devote a great number of resources to the creation of enzymes for their chosen media at the expense of enzymes which break down other phosphorous sources.

Answer 6b:
Species with high growth rates on specific sources of phosphorous should show small niche breadths (specialists), while species with interediate growth rates on a variety of sources should show larger niche breadth (generalists).

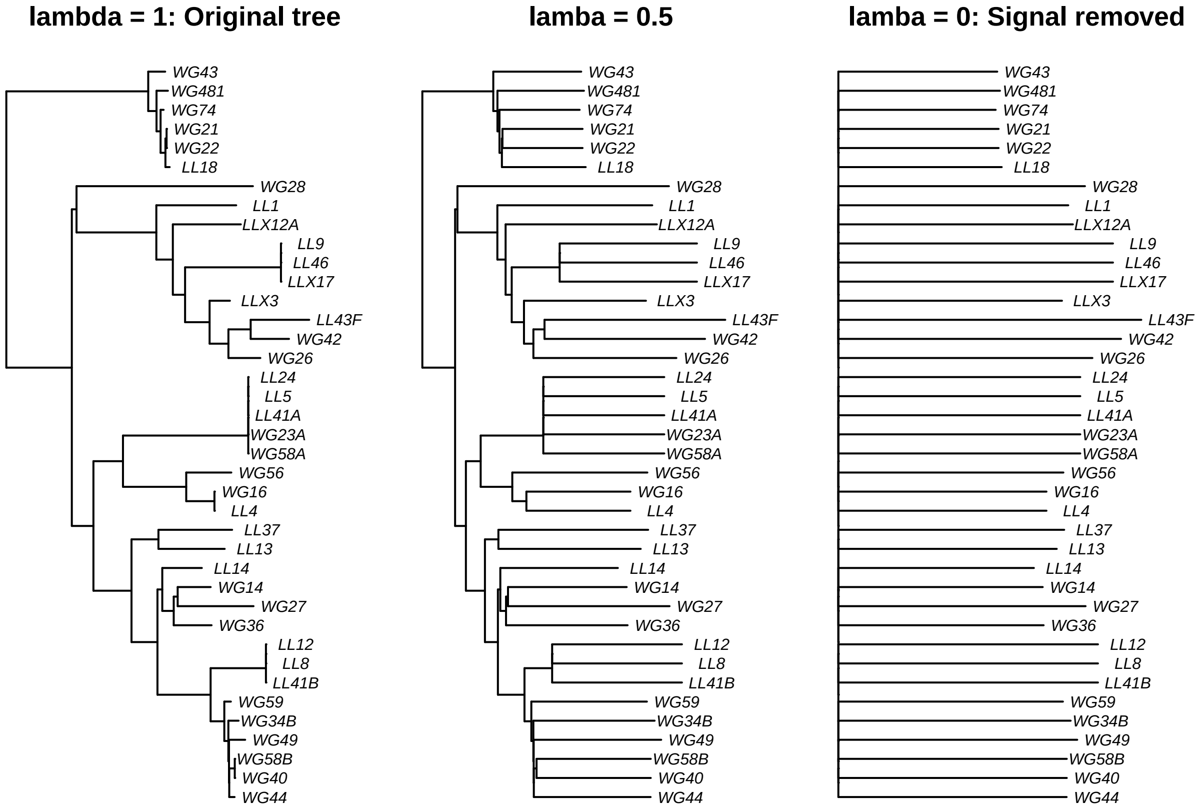
6) HYPOTHESIS TESTING

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: Original tree", cex = 0.7, adj = 0.5)
```

```r
plot(nj.lambda.5, main = "lamba = 0.5", cex = 0.7, adj = 0.5)
plot(nj.lambda.0, main = "lamba = 0: Signal removed", 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
# Untransformed
fitContinuous(nj.rooted, nb, model = "lambda")
```

```
## GEIGER-fitted comparative model of continuous data
##  fitted 'lambda' model parameters:
##  lambda = 0.020685
##  sigsq = 0.106809
##  z0 = 0.661298
##
##  model summary:
##  log-likelihood = 21.656475
##  AIC = -37.312951
##  AICc = -36.627236
##  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
```

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

```
## GEIGER-fitted comparative model of continuous data
##  fitted 'lambda' model parameters:
##  lambda = 0.000000
##  sigsq = 0.106713
##  z0 = 0.657740
##
##  model summary:
##  log-likelihood = 21.647816
##  AIC = -37.295632
##  AICc = -36.609918
##  free parameters = 3
##
## Convergence diagnostics:
##  optimization iterations = 100
##  failed iterations = 0
##  frequency of best fit = 0.85
##
##  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 almost identical. Running fitContinuous on the untransformed tree returns a lambda value of 0.0207, while the transformed tree shows a lambda of 0.
>
> ***Answer 7b***:
> Again, the AIC scores for the two trees are nearly identical. We see a difference of ~0.02 between them - far less than the difference of 2 required to actually be meaningful. Given this, I wouldn't choose one model over the other based just on AIC score as they seem to be equivalent.
>
> ***Answer 7c***:
> I'm not completely sure, but my interpretation of these results is that there is no significant phylogenetic signal in our data, as the two trees were found to be functionally equivalent based on their AIC scores.

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.

```
# Correcting branch lengths to get rid of any lengths of 0.
nj.rooted$edge.length <- nj.rooted$edge.length + 10^-7

# Calculating phylogenetic signal for each phosphorus source
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)
}

# Correcting for false discovery rate (repeated testing) using Benjamini-Hochberg method

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 8299.808 1478.823 1899.795 3715.868  509.360 1707.309 5239.255
## PIC.var.P       0.239    0.078    0.109    0.754    0.331    0.029    0.005
## PIC.var.z      -0.845   -1.268   -1.223    0.878   -0.483   -1.709   -1.296
## PIC.P.BH        0.615    0.351    0.392    0.798    0.623    0.174    0.045
##                  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 1872.161 3304.413 5369.171 9469.557 1617.946 3042.578
## PIC.var.P       0.346    0.394    0.215    0.550    0.323    0.003
## PIC.var.z      -0.504   -0.537   -0.836   -0.029   -0.549   -2.600
## PIC.P.BH        0.623    0.645    0.615    0.707    0.623    0.045
##                   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 3065.132  763.523 1242.621 3693.049 2958.098
## PIC.var.P       0.604    0.808    0.496    0.631    0.474
## PIC.var.z       0.438    0.949   -0.203    0.361   -0.177
## PIC.P.BH        0.710    0.808    0.687    0.710    0.687
```

```
# Calculating Bloomberg's K for niche breadth as well

signal.nb <- phylosignal(nb, nj.rooted)
signal.nb
```

```
##              K PIC.variance.obs PIC.variance.rnd.mean PIC.variance.P
## 1 3.435768e-06         49966.78              49595.39          0.547
##   PIC.variance.Z
## 1     0.01836982
```

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: Based on the p values calculated for niche breadth and individual phosphorus resources and a significance cutoff of 0.05, the variables showing a significant phylogenetic signal are cAMP, BGP, and DNA based growth media. None of the other phosphorus sources, nor the niche breadth showed a significant phylogenetic signal, though PEP and G1P were both close to significance based on our chosen cutoff meaning that there likely is something going on with those media types that we could elucidate with further testing.

   Answer 8b: As $K = 0$ for the three significant media types (and all media types, actually) this might mean one of two things that I can think of. 1) I did it wrong. 2) $K < 1$ so the traits being investigated are overdispersed and closely related species are less similar than expected by chance.

   In the end I am going to assume that I did the calculations right and that there is overdispersion of traits on this tree.

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.

```r
p.growth.pa <- as.data.frame((p.growth > 0.01) * 1)

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
```

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

p.traits <- comparative.data(nj.rooted, p.growth.pa, "name")
phylo.d(p.traits, binvar = BGP)
```

```
##
## Calculation of D statistic for the phylogenetic structure of a binary variable
##
##   Data :  p.growth.pa
##   Binary variable :  BGP
##   Counts of states:  0 = 4
##                      1 = 35
##   Phylogeny :  nj.rooted
##   Number of permutations :  1000
##
## Estimated D :  -0.3824044
## Probability of E(D) resulting from no (random) phylogenetic structure :  0.001
## Probability of E(D) resulting from Brownian phylogenetic structure    :  0.665
```

```r
#phylo.d(p.traits, binvar = PEP)
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 :  0.6078468
## Probability of E(D) resulting from no (random) phylogenetic structure :  0.033
## Probability of E(D) resulting from Brownian phylogenetic structure    :  0.004
```

```r
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 :  0.1447216
## Probability of E(D) resulting from no (random) phylogenetic structure :  0
## Probability of E(D) resulting from Brownian phylogenetic structure    :  0.323
```

```r
#phylo.d(p.traits, binvar = G1P)
```

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:
   Using the same three phosphorus growth traits identified as significant in Blomberg's K analysis (BGP, DNA, cAMP), we see that the relationships of all three have a low probability of resulting from no phylogenetic structure to the data. Both BGP and cAMP, while showing different potential outcomes (clustered vs overdispersed), have good probabilities of resulting from a Brownian phylogenetic structure. However, the DNA trait shows no strong signal of random or Brownian structure, while also strongly suggesting overdispersal.

   Answer 9b:
   These results are widly different from the Blomberg's K analysis performed above where all traits returned the same K value of 0. Here, the three traits analyzed show large differences in their

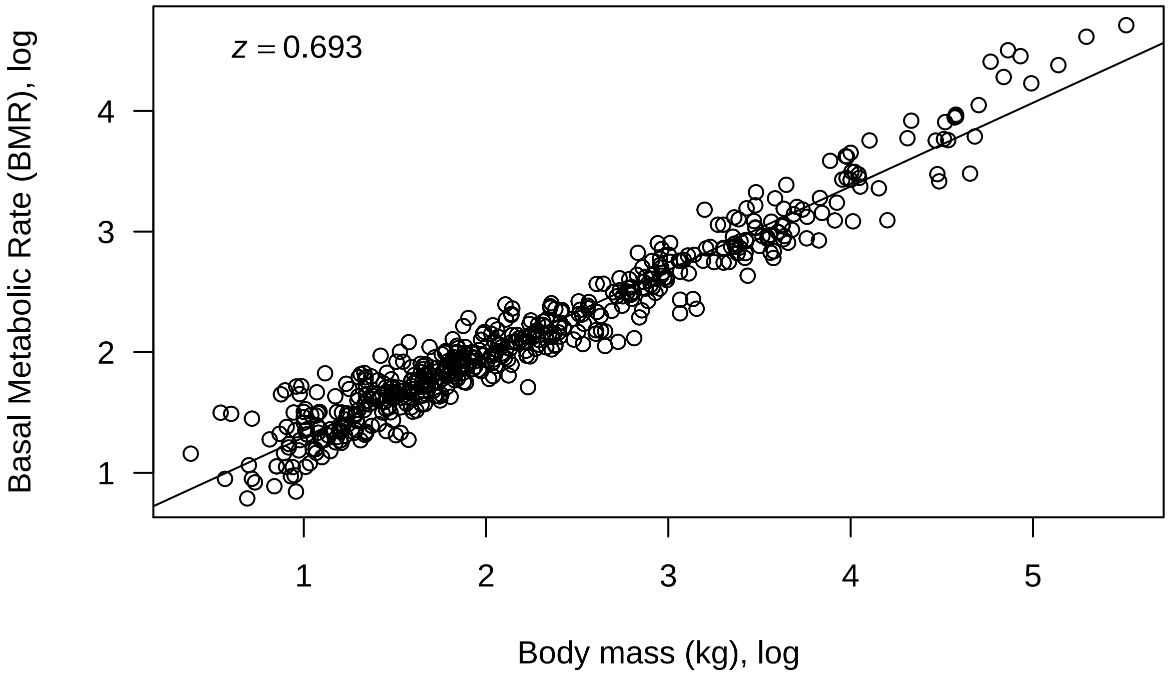
estimated dispersion values and these differences make me more inclined to believe that this test is more reflective of reality.

Answer 9c:
From what I can tell, the main difference between K and D is that they are calculated from different forms of data. Blomberg's K works with continuous data, while trait dispersion (D) uses categorical, binary data. By reducing the dataset to binary differences, it seems like there is less room for ambiguity in interpretation and finding differences in the data becomes easier to do.

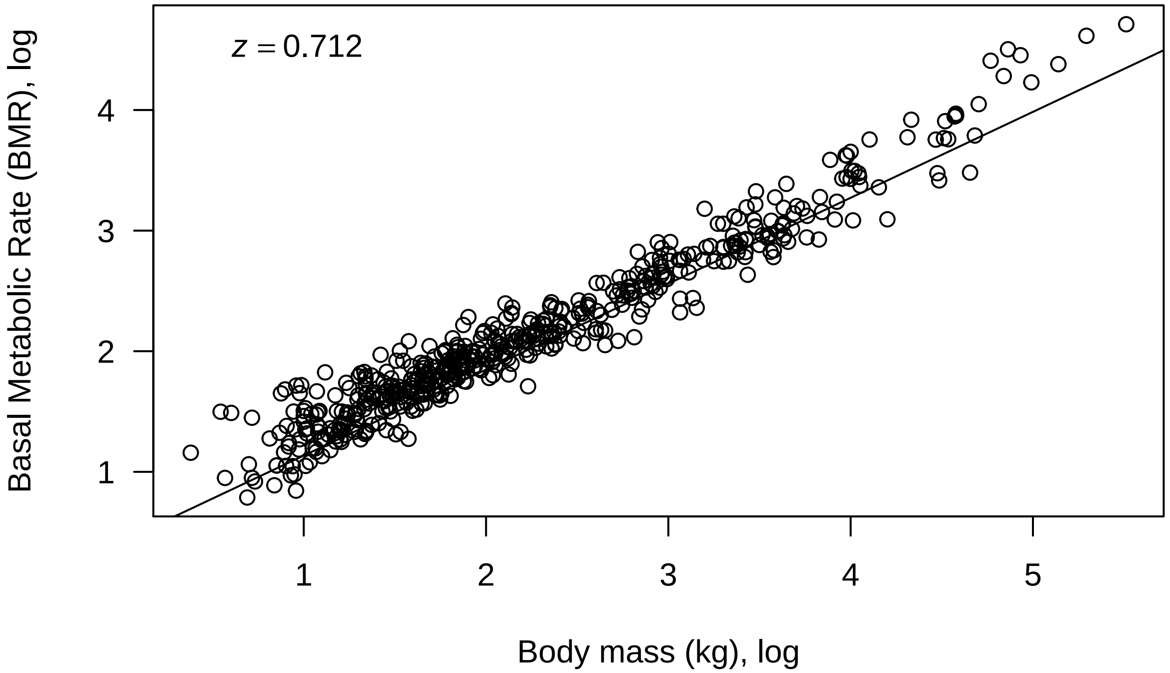
7) PHYLOGENETIC REGRESSION

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
mammal.Tree <- read.tree("./data/mammal_best_super_tree_fritz2009.tre")
mammal.data <- read.table("./data/mammal_BMR.txt", sep = "\t",
header = TRUE)

# Pull out variables of interest
mammal.data <- mammal.data[, c("Species", "BMR_.mlO2.hour.",
"Body_mass_for_BMR_.gr.")]
mammal.species <- array(mammal.data$Species)

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

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

rownames(pruned.mammal.data) <- pruned.mammal.data$Species


# Run a 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))
text(0.5, 4.5, eqn, pos = 4)
```



```
# Phylogeny-corrected regression with no bootstrapping
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))
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:
Kind of like autocorrelation, might see that there are a lot of shared similarities between species not due to any important pattern, just that they are closely related and closely related species are more similar to eachother.

Answer 10b:
Phylogenetic regression, unlike standard regression, takes into account phylogenetic relatedness when calculating the residuals of the model. Normally the residual errors are assumed to be independent, but phylogenetic regression methods take into account branch lengths of the phylogeny of the species being regressed.

Answer 10c:
The slope and fit of both models, standard and phylogeny adjusted, are quite good and compelling. In both we see that as mammal body mass increases, so too does BMR in a linear log-log fashion. This means that the untransformed relationship between mass and BMR is best represented by a power law function where BMR is equal to the mass of the animal to the slope of the regression (0.712) times some constant.

The second plot accounting for evolutionary history resulted in a slightly better model fit, meaning that accounting for the underlying phylogeny of these data gives us a stronger relationship between mass and BMR than we would expect under the assumption of no underlying evolutionary relationship to the data.

Answer 10d:
Imagnie a scenario where we have two sister clades of lizards. All lizards in clade A are large in size, and have a minimum of 20 bumps on their heads, with some variance around a mean of 23 bumps per head sampled. Lizards in clade B are all significantly smaller, and never have bumps on their heads. There are some outliers from A and B which show intermediate body sizes, but the majority of lizards sampled clump into two distinct groups, and clade B still never has bumps.

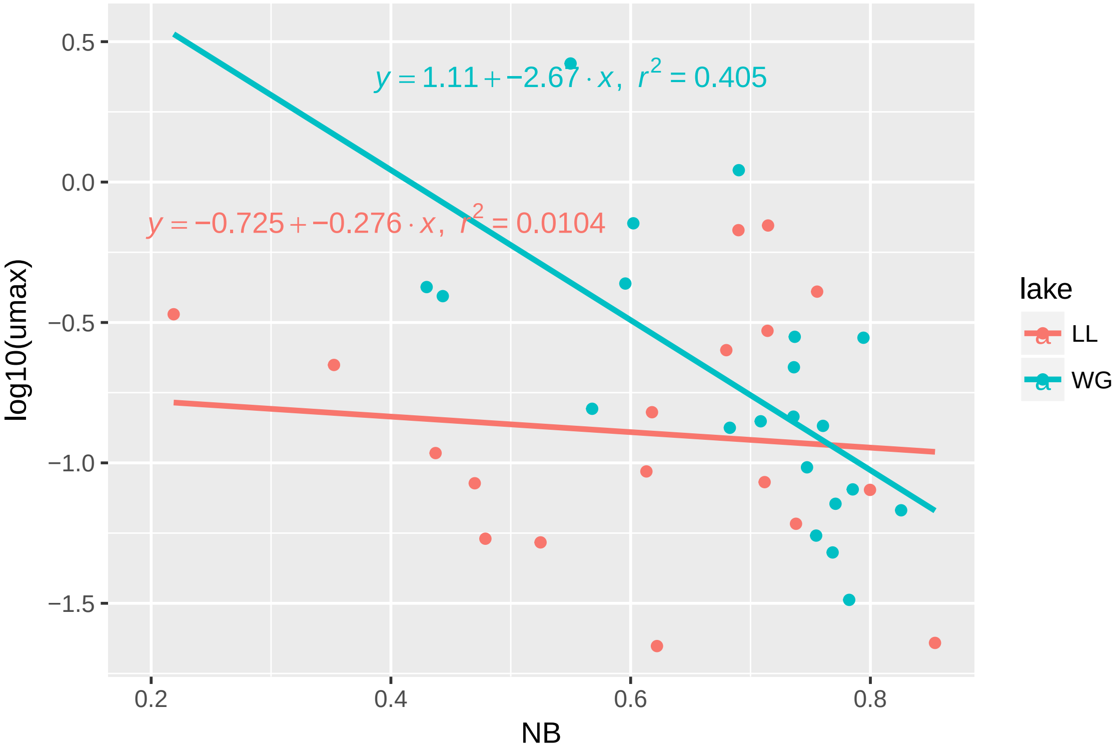
A normal regression on these variables would show a strong positive relationship between body size and bump number. But this is just due to the shared evolutionary history of lizards within the clades, and accounting for this history should lead to the disappearance of the relationship.

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.

```
##
## Attaching package: 'devtools'

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

```
##
##     check
## Sourcing https://gist.githubusercontent.com/kdauria/524eade46135f6348140/raw/676acaca9a0a144ef320ae2
## SHA-1 hash of file is c0b163b9fd2d7fe7bd5541e3266d8d36ff3b895d
```



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***: Based on the regression results above, and the earlier phylogenetic results, there seems to be relatively strong support for a generalist-specialist tradeoff in these data. This new piece of information (the regression plot) pretty clearly shows that there is a reasonably strong inverse relationship between the maximum growth rate of bacterial strains, and the niche breadth of that strain when found in a phosphorus rich environment. This fits well with my earlier prediction that specialist species (low NB) should thrive in environments rich in their chosen phosphorus source (eutrophic lakes). The relationship in the low phosphorus lake (LL) is less clear, seems to show not much difference in growth rate of the bacterial strains regardless of their niche breadth overall.
>
> In the end though, the eutrophic lake should be expected to show the largest difference between specialist and generalist strains if there is a tradeoff, and it does - thus providing good support for the existence of said tradeoff.