Phylogenetic Diversity - Traits

Roy Moger-Reischer; 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).

```
#rm(list=ls())
#setwd("~/GitHub/QB2017_Moger-Reischer/Week6-PhyloTraits")
setwd("C:\\Users\\rmoge\\GitHub\\QB2017_Moger-Reischer\\Week6-PhyloTraits")
package.list <- c('vegan', 'tidyr', 'dplyr', 'codyn', 'ggplot2',
'cowplot', 'MullerPlot', 'RColorBrewer', 'reshape2', 'lubridate',
```

```r
'TTR', 'xtable', 'multcomp', 'pander', 'png', 'grid', 'tseries', 'nlme', 'forecast', 'lsmeans', 'ape',
for (package in package.list) {
if (!require(package, character.only = TRUE, quietly = TRUE)) {
install.packages(package, repos='http://cran.us.r-project.org')
library(package, character.only = TRUE) }
}
```

```
## This is vegan 2.4-2
```

```
##
## Attaching package: 'dplyr'
```

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

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

```
##
## Attaching package: 'cowplot'
```

```
## The following object is masked from 'package:ggplot2':
##
##     ggsave
```

```
##
## Attaching package: 'reshape2'
```

```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
##
## Attaching package: 'MASS'
```

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

```
##
## Attaching package: 'TH.data'
```

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

```
##
## Attaching package: 'nlme'
```

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

```
##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

##
## Attaching package: 'timeDate'

## The following object is masked from 'package:xtable':
##
##     align

## This is forecast 7.3

##
## Attaching package: 'forecast'

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

##
## Attaching package: 'seqinr'

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

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

## The following objects are masked from 'package:dplyr':
##
##     count, query

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

##
## Attaching package: 'phylobase'

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

##
## Attaching package: 'ade4'

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

##
## Attaching package: 'adephylo'

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

```
##
## Attaching package: 'phangorn'

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

##
## Attaching package: 'devtools'

## The following object is masked from 'package:lsmeans':
##
##     test

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

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.

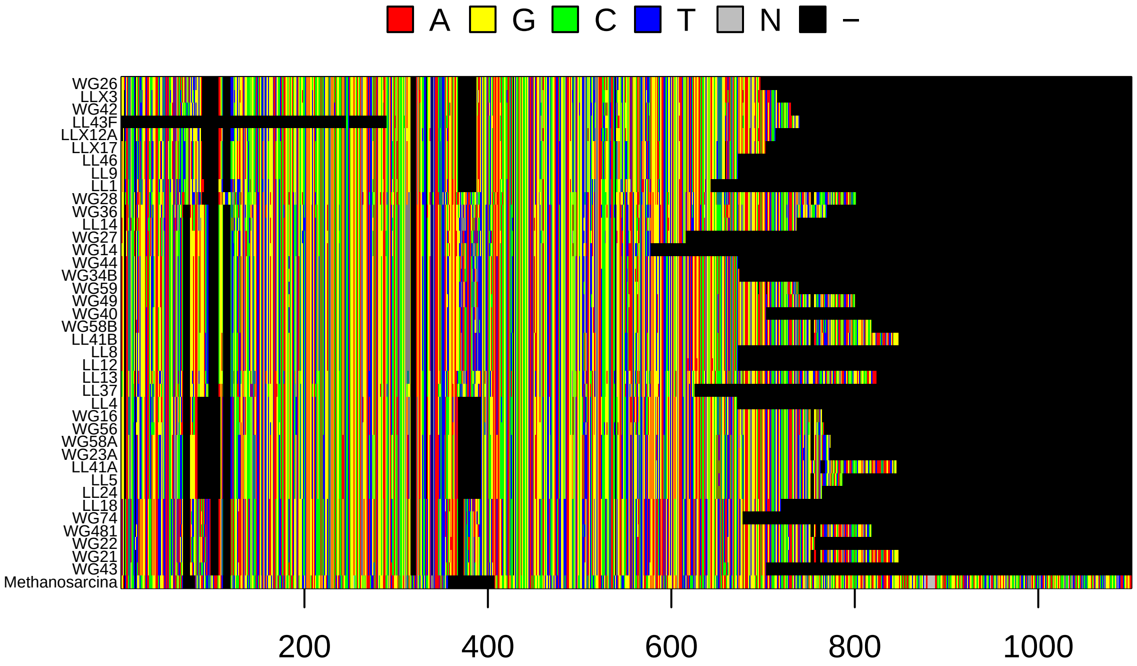
3) SEQUENCE ALIGNMENT

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***:
>
> It looks like the .afa file has been aligned. It included gaps. The .fasta file has not been aligned.

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 Alignment File {seqinr}
read.aln <- read.alignment(file = "./data/p.isolates.afa", format = "fasta")
#read.aln<-WINDOWS
p.DNAbin <- as.DNAbin(read.aln)#ape object

window <- p.DNAbin[, 100:1200]
# 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:

I am unable to discern the length of the reads from the information given. It looks like the length of the sequence of interest is ~1500 in the ancestor. So the read length would not be more than that (which would be ludicrously long, an12yway). However, depending on the sequencing technology that was used, the read lengths could vary.

Answer 2b:

Variable regions can aid phylogenetic inference. If a site is invariant we cannot say anything about which taxa are more recently diverged.

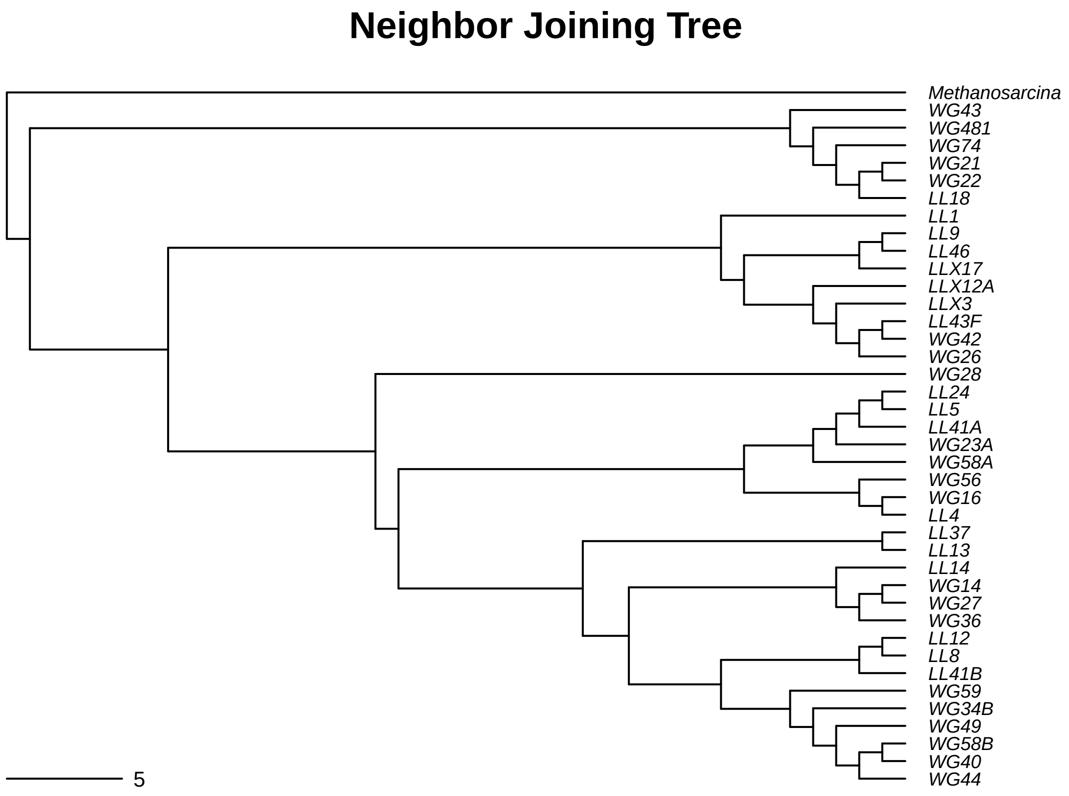
4) MAKING A PHYLOGENETIC TREE

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.

```
#1
seq.dist.raw <- dist.dna(p.DNAbin, model = "raw", pairwise.deletion = FALSE)
#2
nj.tree <- bionj(seq.dist.raw)
#3 Identify OG
outgroup <- match("Methanosarcina", nj.tree$tip.label)
# Root
nj.rooted <- root(nj.tree, outgroup, resolve.root = TRUE)
#4 Plot rooted
par(mar = c(1, 1, 2, 1) + 0.1)
plot.phylo(nj.rooted, main = "Neighbor Joining Tree", "phylogram",    use.edge.length = FALSE, directio
add.scale.bar(cex = 0.7)
```



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

> ***Answer 3***:
>
> NJ is a useful starting tree because it is simple to understand, straightforward, and fast to generate. Disadvantages could include that it must be ultrametric; doesn’t consider Bayesian probabilities; I think that it assumes constant rate of evolution on all branches (but I could be wrong); it must

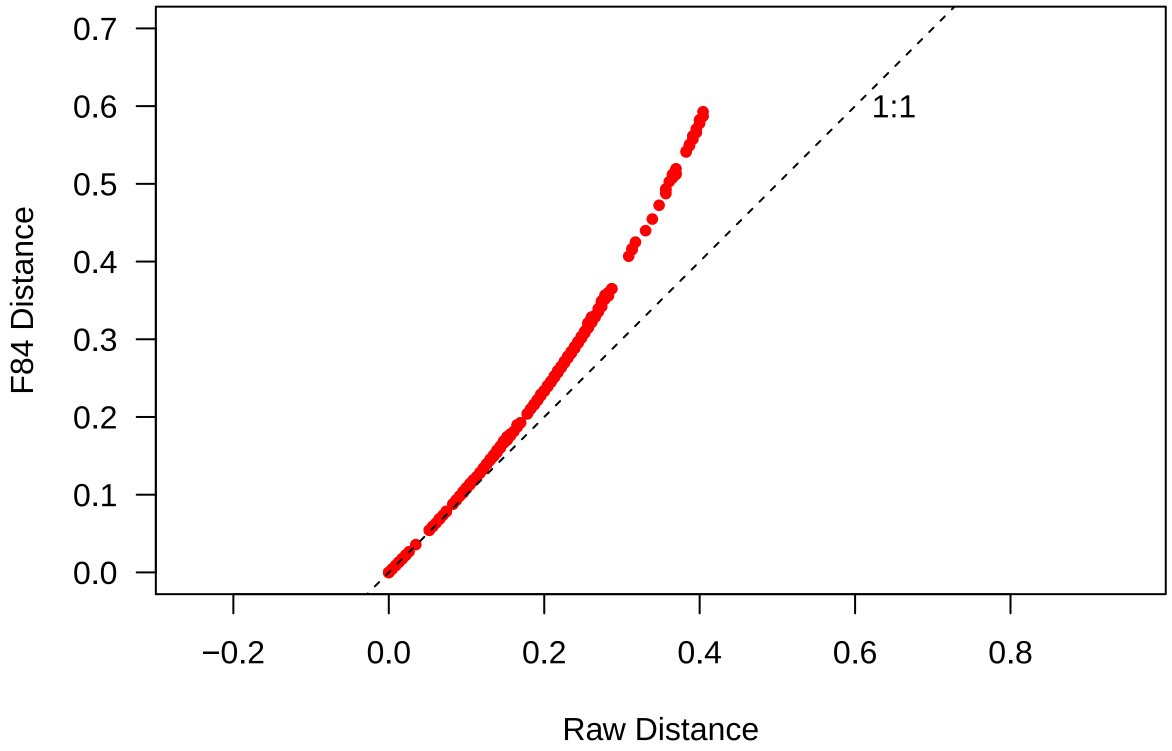
assume that the distance matrix is true; it only produces one tree, so we can't give it a P-value.

Within medium-sized clades, bacteria taxa from the same lake tend to cluster more closely related into clades near the tips. However, the more-ancient (larger) medium-to-large-sized clades do readily include taxa from lakes.

B) SUBSTITUTION MODELS OF DNA EVOLUTION

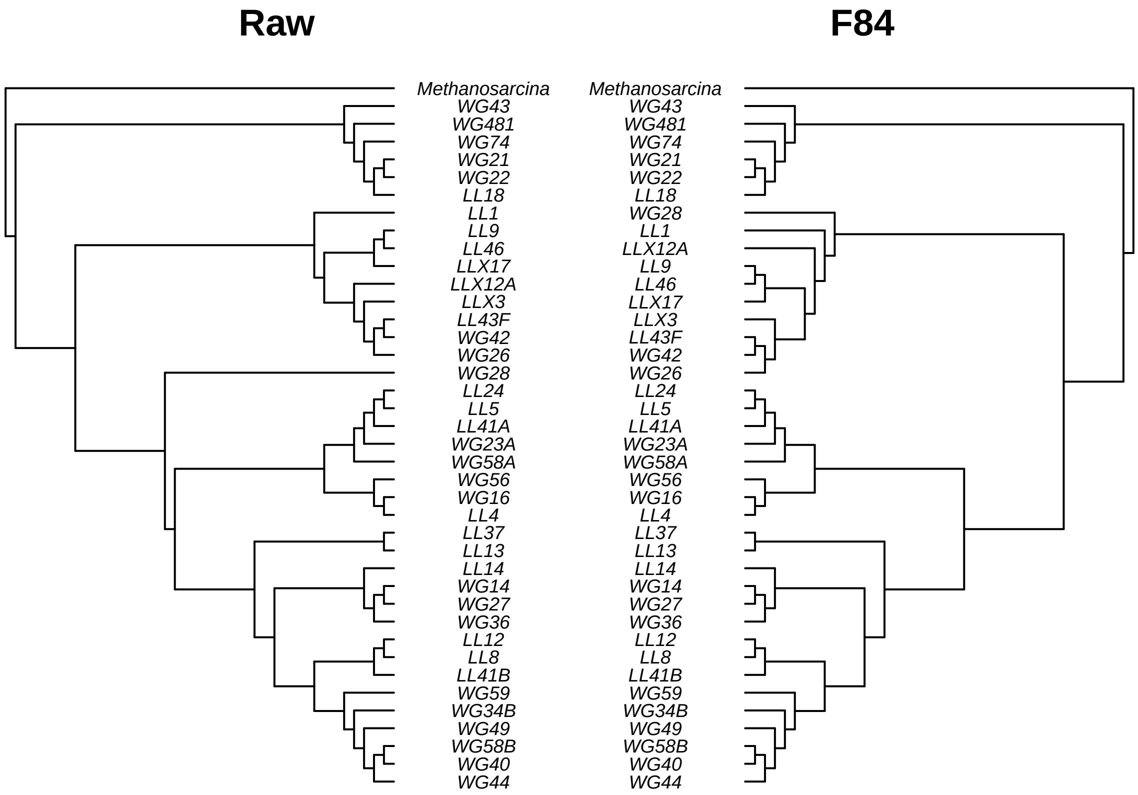
In the R code chunk below, do the following:
1. make a second distance matrix based on the Felsenstein 84 substitution model,
2. create a saturation plot to compare the *raw* and *Felsenstein (F84)* substitution models,
3. make Neighbor Joining trees for both, and
4. create a cophylogenetic plot to compare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

```
#1
seq.dist.F84 <- dist.dna(p.DNAbin, model = "F84", pairwise.deletion = FALSE)
#2
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,
abline(b = 1, a = 0, lty = 2)
text(0.65, 0.6, "1:1")
```



```
#3
raw.tree <- bionj(seq.dist.raw)
F84.tree <- bionj(seq.dist.F84)

# Define OGs
```

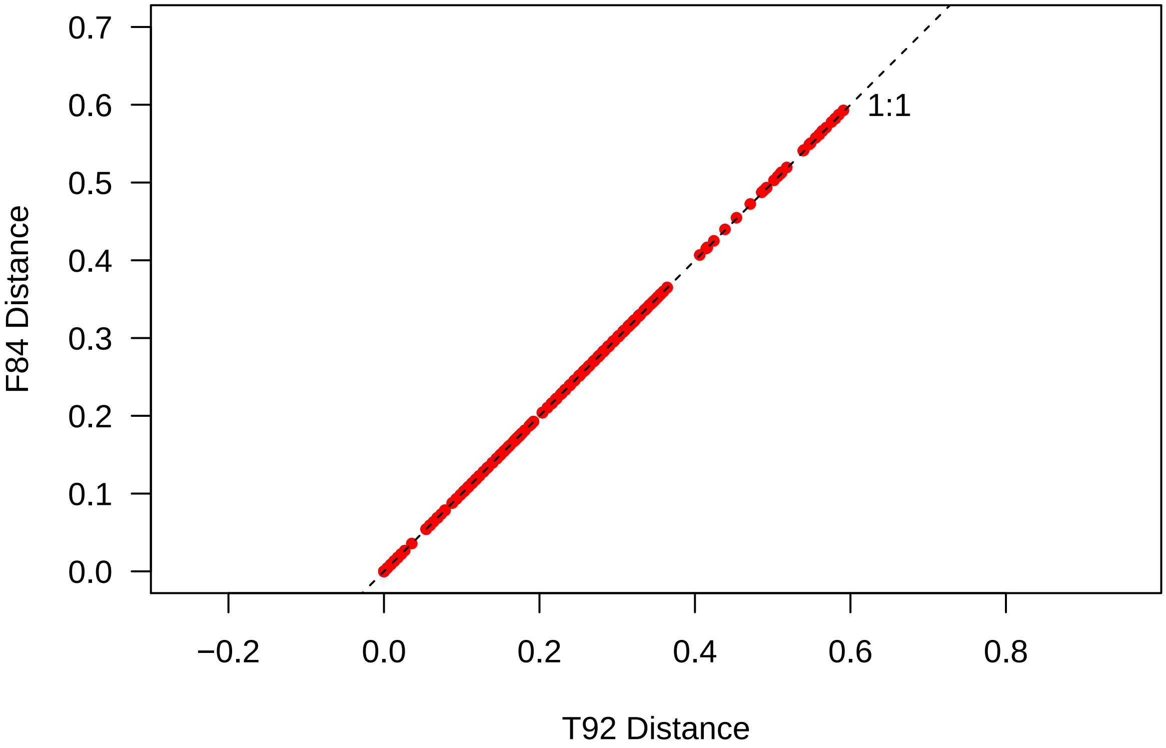
```
raw.outgroup <- match("Methanosarcina", raw.tree$tip.label)
F84.outgroup <- match("Methanosarcina", F84.tree$tip.label)
# Root
raw.rooted <- root(raw.tree, raw.outgroup, resolve.root=TRUE)
F84.rooted <- root(F84.tree, F84.outgroup, resolve.root=TRUE)
#4 Make Cophylogenetic
layout(matrix(c(1,2), 1, 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 = F

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



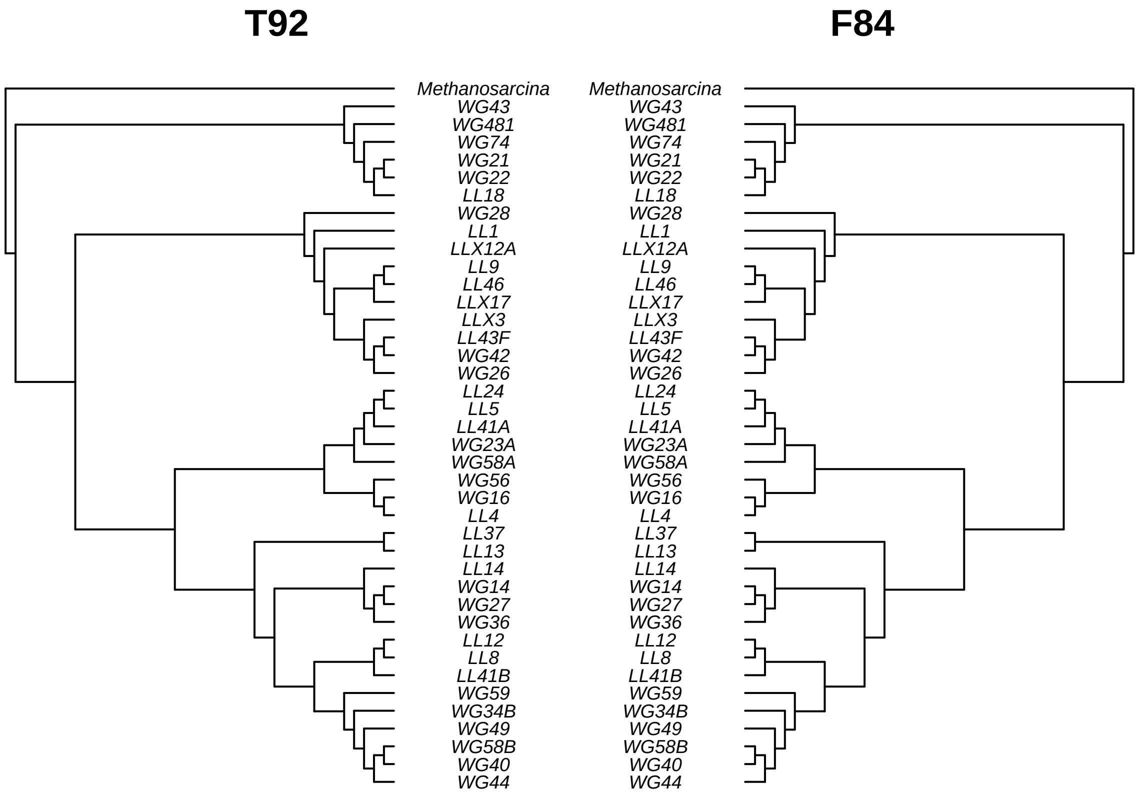
In the R code chunk below, do the following:
1. pick another substitution model,
2. create and distance matrix and tree for this model,
3. make a saturation plot that compares that model to the *Felsenstein (F84)* model,
4. make a cophylogenetic plot that compares the topologies of both models, and
5.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

```
#1
seq.dist.T92 <- dist.dna(p.DNAbin, model = "T92", pairwise.deletion = FALSE)


#2
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,
abline(b = 1, a = 0, lty = 2)
text(0.65, 0.6, "1:1")
```



```
#3
T92.tree <- bionj(seq.dist.T92)
F84.tree <- bionj(seq.dist.F84)

# Define OGs
T92.outgroup <- match("Methanosarcina", T92.tree$tip.label)
F84.outgroup <- match("Methanosarcina", F84.tree$tip.label)
# Root
T92.rooted <- root(T92.tree, T92.outgroup, resolve.root=TRUE)
F84.rooted <- root(F84.tree, F84.outgroup, resolve.root=TRUE)
#4 Make Cophylogenetic
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 = F

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



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 T92 model. It is similar to the F84 model in that it both allows for differing probabilities of particular types of basepair substitutions, and in that allows for unequal total base composition.

Answer 4b:

In this particular case, the actual shape of the phylogeny is not affected by the choice of a substitution model (T92 vs. F84). However, the saturation plot indicates that the JC model, which does not account for the possibility of revertant mutations, underestimates evolutionary distance in comparison to some other models.

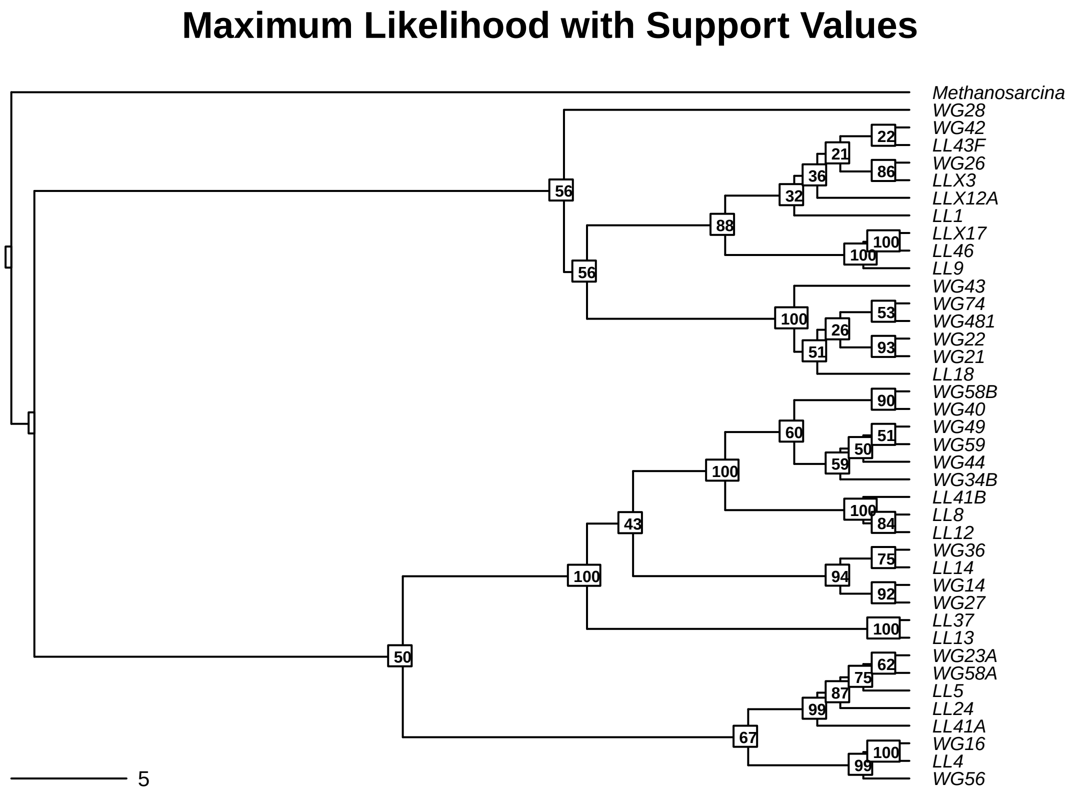
Answer 4c:

See 4b.

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.lengt
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 ML tree is generated by resampling the data and asking how likely, given a particular tree, it is that we would obtain the data that we did. The ML tree allows us to quantify the strength of our certainty about the clustering in each clade.

Answer 5b:
We bootstrap so that we can know how many times out of 100 a certain clade was included in the ML tree in all of our resamplings. The we can know how confident we can be that a certain clade really does represent a monophyletic group.

Answer 5c:
The bootstrap values tell me how many times out of 100 a particular clade was indeed generated as part of the ML tree.

Answer 5d:

It looks like the clades for which WG28 is the sister outgroup have lower support than do the clades clustering such that LL4, WG56, WG16, LL41A, LL24, LL5, WG58A, WG23A are in the outgroup sister clade.

Answer 5e:

Typically we should trust branches with >95% support.

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.

```r
p.growth <- read.table("./data/p.isolates.raw.growth.txt", sep = "\t", header = TRUE, row.names = 1)
#stdz GR
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.

```r
umax <- (apply(p.growth, 1, max))
#2
levins <- function(p_xi = ""){
  p = 0
  for (i in p_xi){
    p = p + i^2
    }
  nb = 1 / (length(p_xi) * p)
  return(nb)
}
#3
nb <- as.matrix(levins(p.growth.std))
# Add rownames, colnames
rownames(nb) <- row.names(p.growth)
colnames(nb) <- c("NB")
```

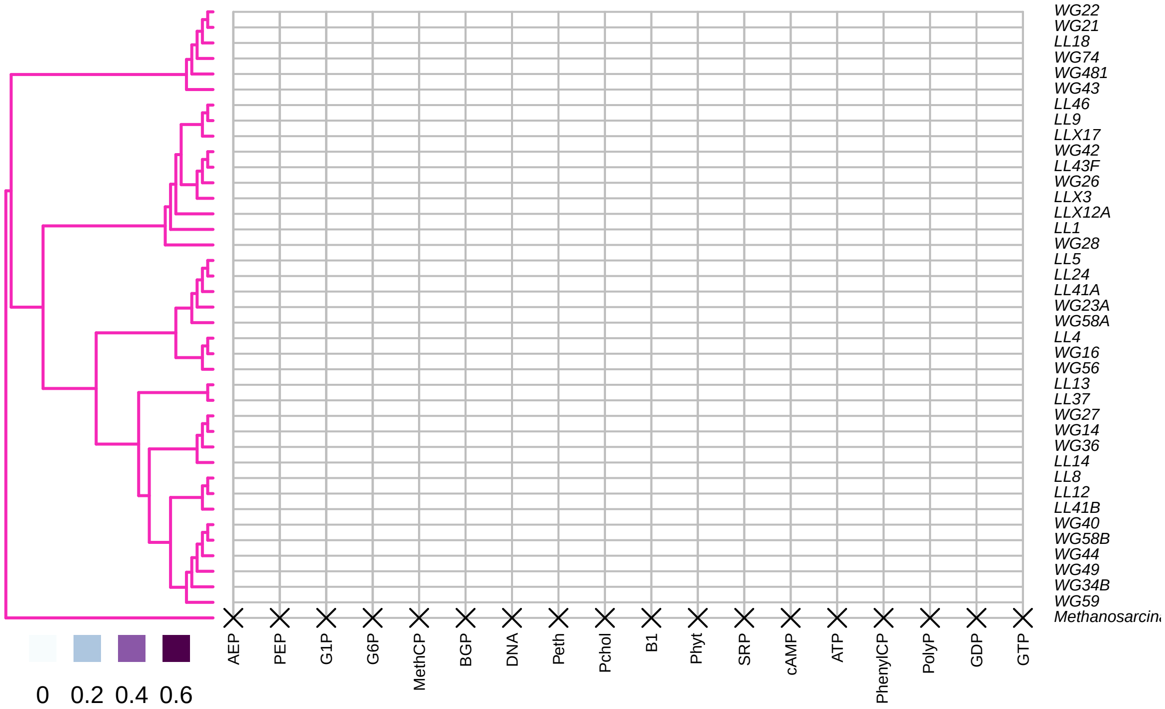
C. Visualizing Traits on Trees

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.

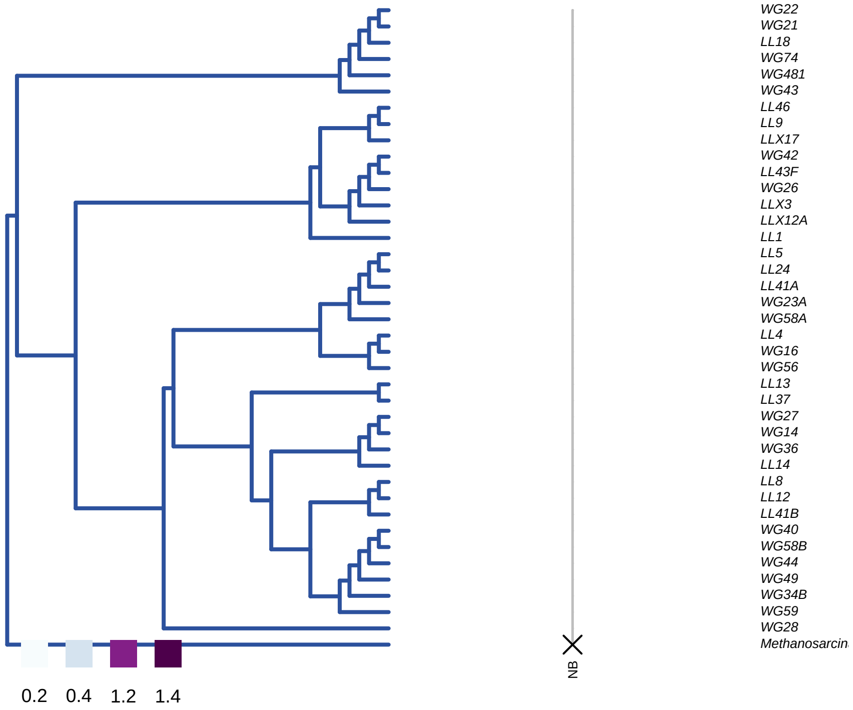
```
nj.tree <- bionj(seq.dist.T92)
#2
outgroup <- match("Methanosarcina", nj.tree$tip.label)
nj.rooted2 <- root(nj.tree, outgroup, resolve.root = TRUE)
#3
#nj.rooted <- drop.tip(nj.rooted, "Methanosarcina")
#I need to include Methanosarcina in the phylogeny, else I get an error on a particular line below.
```

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, "BuPu"))
#2
par(mar=c(1,1,1,1) + 0.1)
x <- phylo4d(nj.rooted2, p.growth.std, missing.data="OK")
table.phylo4d(x, treetype = "phyl", symbol = "colors", show.node = TRUE,cex.label = 0.5, scale = FALSE,
```



```r
#3
par(mar=c(1,5,1,5) + 0.1)
x.nb <- phylo4d(nj.rooted, nb, missing.data="OK")
table.phylo4d(x.nb, treetype = "phyl", symbol = "colors", show.node = TRUE, cex.label = 0.4, scale = FA
```



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:

Consider anammox bacteria, which fix nitrogen using a highly unstable molecule, hydrazine (N2 H4). These bacteria must invest in an anammoxosome, a special subcellular compartment, to store its volatile supply of hydrazine. However, although specialist bacteria must invest additional resources to be able to carry their method of N fixation, they have the advantage of being able to grow in certain environments which would be inhospitable to species that could not handle hydrazine and would not be able to obtain N.

Answer 6b:

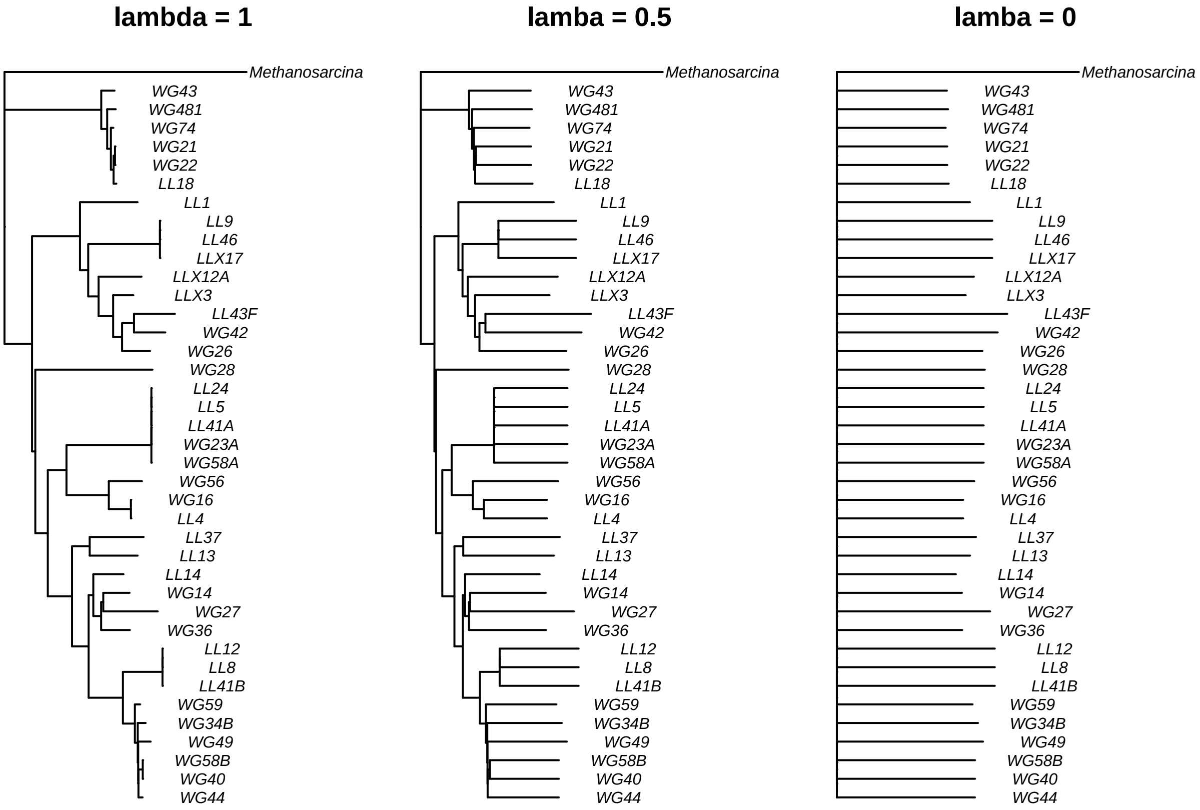
If there truly is a tradeoff, then we expect species with a very large uMax to only be able to grow well on a limited array of P sources; contrariwise, generalist species should be able to grow on a wide array of P sources, but the uMax should not be particular large on any of them.

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


lambda = 1

lamba = 0.5

lamba = 0

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")
```

```
## Warning in treedata(phy, dat): The following tips were not found in 'data' and were dropped from 'ph
##  Methanosarcina
```

```
## GEIGER-fitted comparative model of continuous data
##  fitted 'lambda' model parameters:
##  lambda = 0.061980
##  sigsq = 0.140020
##  z0 = 0.664039
##
##  model summary:
##  log-likelihood = 21.456738
##  AIC = -36.913476
##  AICc = -36.227761
##  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.5, nb, model = "lambda")


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

```
## Warning in treedata(phy, dat): The following tips were not found in 'data' and were dropped from 'phy
##  Methanosarcina
```

```
## GEIGER-fitted comparative model of continuous data
##  fitted 'lambda' model parameters:
##  lambda = 0.000000
##  sigsq = 0.139249
##  z0 = 0.656203
##
##  model summary:
##  log-likelihood = 21.399126
##  AIC = -36.798252
##  AICc = -36.112537
##  free parameters = 3
##
## Convergence diagnostics:
##  optimization iterations = 100
##  failed iterations = 0
##  frequency of best fit = 0.87
##
##  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:

Untransformed lambda is 0.020682; transformed lambda is 0

Answer 7b:

Untransformed tree is given AIC = -37.312951 Transformed tree is given AIC = -37.295632. The transformed tree is given lower AIC. The difference in AIC is <2.

Answer 7c:

My interpretation is that if the AIC for untransformd tree were >2 lower than that for the transformed tree, there would be significant (important) phylogenetic signal. AIC was lower for the lambda = 0 model. I would therefore say that no, we cannot conclude there is phylogenetic signal in the data.

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.

```
#1
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")
#2
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)
  }
```

```
## [1] "Dropping tips from the tree because they are not present in the data:"
## [1] "Methanosarcina"
## [1] "Dropping tips from the tree because they are not present in the data:"
## [1] "Methanosarcina"
## [1] "Dropping tips from the tree because they are not present in the data:"
## [1] "Methanosarcina"
## [1] "Dropping tips from the tree because they are not present in the data:"
## [1] "Methanosarcina"
## [1] "Dropping tips from the tree because they are not present in the data:"
## [1] "Methanosarcina"
## [1] "Dropping tips from the tree because they are not present in the data:"
## [1] "Methanosarcina"
## [1] "Dropping tips from the tree because they are not present in the data:"
## [1] "Methanosarcina"
## [1] "Dropping tips from the tree because they are not present in the data:"
## [1] "Methanosarcina"
## [1] "Dropping tips from the tree because they are not present in the data:"
## [1] "Methanosarcina"
## [1] "Dropping tips from the tree because they are not present in the data:"
## [1] "Methanosarcina"
## [1] "Dropping tips from the tree because they are not present in the data:"
## [1] "Methanosarcina"
## [1] "Dropping tips from the tree because they are not present in the data:"
## [1] "Methanosarcina"
## [1] "Dropping tips from the tree because they are not present in the data:"
## [1] "Methanosarcina"
## [1] "Dropping tips from the tree because they are not present in the data:"
## [1] "Methanosarcina"
## [1] "Dropping tips from the tree because they are not present in the data:"
## [1] "Methanosarcina"
## [1] "Dropping tips from the tree because they are not present in the data:"
## [1] "Methanosarcina"
## [1] "Dropping tips from the tree because they are not present in the data:"
## [1] "Methanosarcina"
## [1] "Dropping tips from the tree because they are not present in the data:"
## [1] "Methanosarcina"
```

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

#4
signal.nb <- phylosignal(nb, nj.rooted)
```

```
## [1] "Dropping tips from the tree because they are not present in the data:"
## [1] "Methanosarcina"
```

```
print(signal.nb)
```

```
##              K PIC.variance.obs PIC.variance.rnd.mean PIC.variance.P
## 1 4.187392e-06         49966.79              50199.32           0.53
##   PIC.variance.Z
## 1     -0.0113577
```

```
signal.nb
```

```
##              K PIC.variance.obs PIC.variance.rnd.mean PIC.variance.P
## 1 4.187392e-06         49966.79              50199.32           0.53
##   PIC.variance.Z
## 1     -0.0113577
```

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:

   Marginally significant phylogenetic signal on PEP, G1P. Significant phylogenetic signal for growth on DNA, cyclic AMP. No significant phylogenetic signal for growth on other P sources or for niche breadth.

   Answer 8b:

   In all cases the K-values are indicative of overdispersion.

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.

```
#1
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
```

```
#2
p.growth.pa$name <- rownames(p.growth.pa)
#3
p.traits <- comparative.data(nj.rooted, p.growth.pa, "name")
#4
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 :  0.425306
## Probability of E(D) resulting from no (random) phylogenetic structure :  0.003
## Probability of E(D) resulting from Brownian phylogenetic structure    :  0.036
```

```
phylo.d(p.traits, binvar = PhenylCP)#\neq phencyclidine
```

```
##
## 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 :  0.8958007
## Probability of E(D) resulting from no (random) phylogenetic structure :  0.308
## Probability of E(D) resulting from Brownian phylogenetic structure    :  0.013
```

```r
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 :  0.6044173
## Probability of E(D) resulting from no (random) phylogenetic structure :  0.022
## 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 :  0.109393
## Probability of E(D) resulting from no (random) phylogenetic structure :  0.001
## Probability of E(D) resulting from Brownian phylogenetic structure    :  0.352
```

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:

For cAMP, DNA, and AEP the results are all the same. Estimated D is always positive, meaning that the trait is overdispersed. For all three P resources for neither Brownian phylogenetic structure nor random/ no phylogenetic structure is there a probability >95% that the estimated value for D results from that true underlying phylogenetic structure.

Answer 9b:

Unlike E(D), Blomberg's K indicated significant phylogenetic overdispersion of bacterial ability to grow on DNA and cAMP.

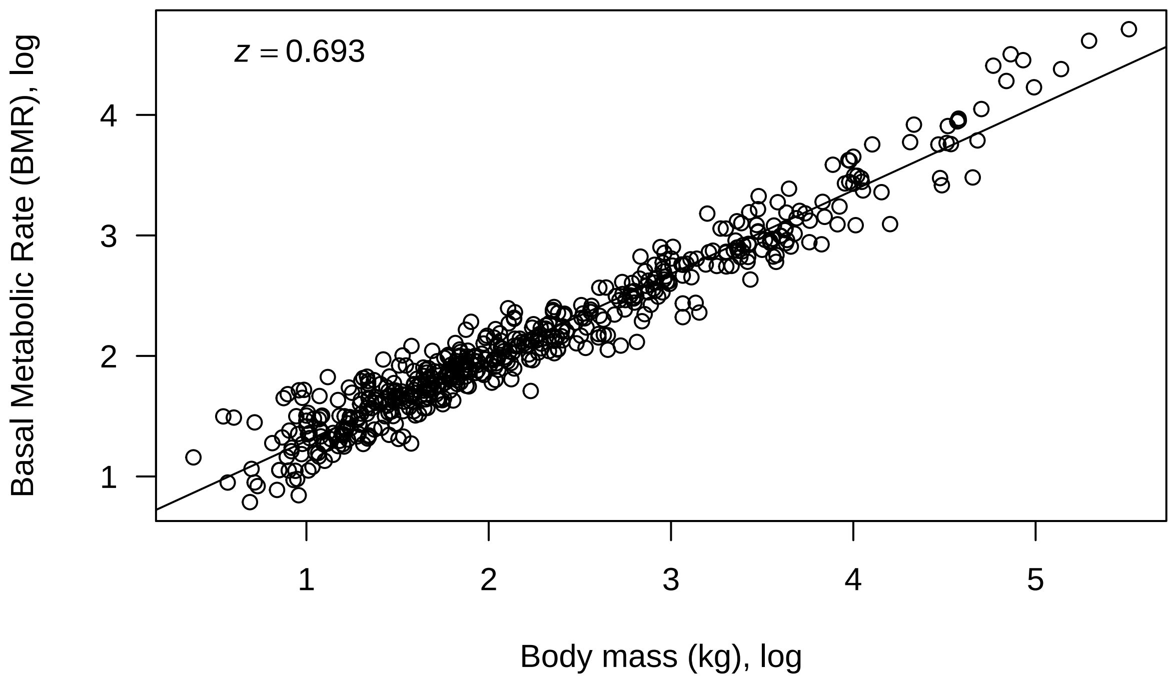
Answer 9c:

Fritz and Purvis' D, originaly designed to predict extinction risk for particular taxa, uses binary traits. This could mean that some of the information in our data were sacrificed when they were altered to be binary.

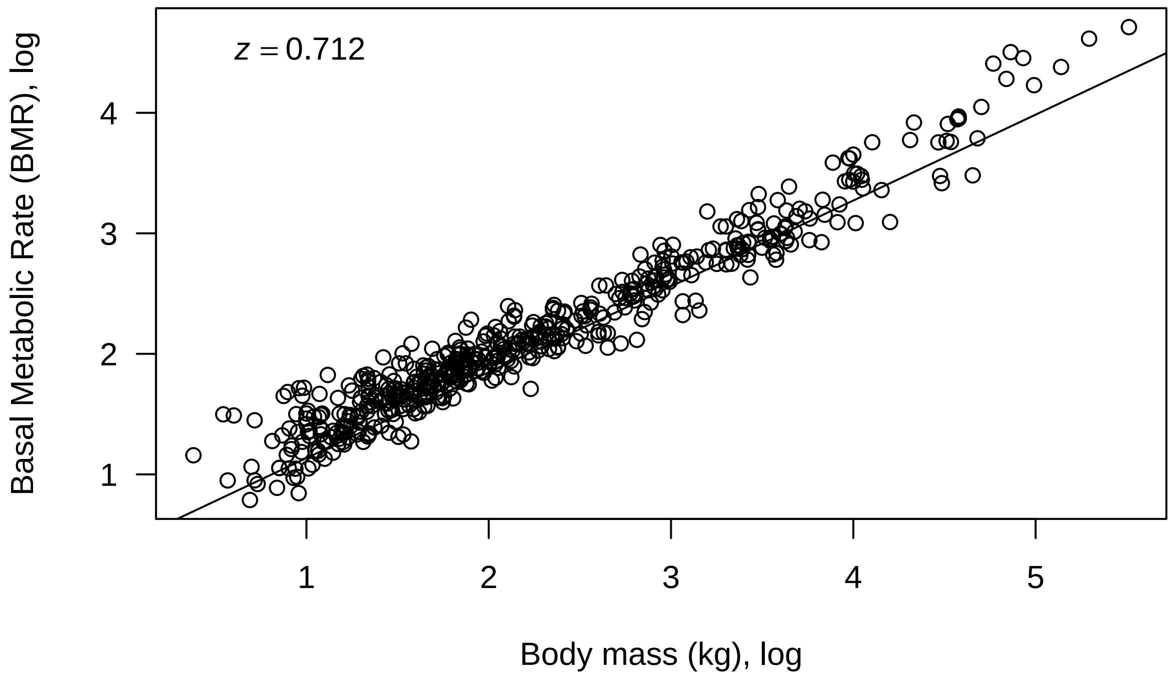
7) PHYLOGENETIC REGRESSION

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, 3. 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

#2
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 =
abline(a = fit$coefficients[1], b = fit$coefficients[2])
b1 <- round(fit$coefficients[2], 3)
eqn <- bquote(italic(z) == .(b1))
#slope
text(0.5, 4.5, eqn, pos = 4)
```



```
#3
fit.phy <- phylolm(log10(BMR_.mlO2.hour.) ~ log10(Body_mass_for_BMR_.gr.),data = pruned.mammal.data, pr
plot(log10(pruned.mammal.data$Body_mass_for_BMR_.gr.), log10(pruned.mammal.data$BMR_.mlO2.hour.), las =
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:

To use one variable's value to predict the value of another variable, we would like to use linear regression. However, regression assume that all of our datapoints are independent of one another. This is not true if some of the organisms are more closely phlogenetically related to one another than they are to certain other taxa. Therefore it is wise to data which been corrected to be phylogenetically independent.

Answer 10b:

In a simple linear regression, we need residuals with mean = 0 and a simple variance (whatever its magnitude). For a phylogenetic regression we need residuals with the same mean but the variance now has a variance-covariance matrix term. This helps account for phylogenetic signal and phylogenetic relatedness.

Answer 10c:

For the uncorrected model, the amount of variation explained by the model is indicated by Rˆ2 = 0.9438. For the phylogenetic regression, the goodness of the model is given by a AIC value (-646.9). I do not know how to compare Rˆ2 to AIC. For both types of regression the P-value for the regression relationship is P < 0.001.

Answer 10d:

Imagine a population of shrimp living in Yaponskoye more. Because anthropogenic climate change is perceived as a Chinese fabrication, it proceeds unabated and the level of the sea rises. The shrimp are free to swim back and forth from the sea to Lake Khanka. After several centuries, humanity has departed, fleeing Earth for reaches of space unknown at 1000c via collapsars. Meanwhile, Earth enters a moderate ice age, and sea levels decrease. This causes a vicariance event for the shrimp population. Part of the population is stranded in Lake Khanka, and part is relegated back into the Yaponskoye more. Over evolutionary time and after many speciation events, the two vicariantly-separated populations now represent several species each. The lake species all are more phylogenetically related and all are adapted to live in fresh water. The sea species are all more phylogenitically related to one another, too, and all are adapted to live in seawater. In addition, the legacy of ocean acidification has endured: The members of the saltwater shrimp clade also evolve increased tolerance for high [H+].

Many kiloyears later, the ice age ends, sea levels rise, the shrimp species are free to intermingle again.

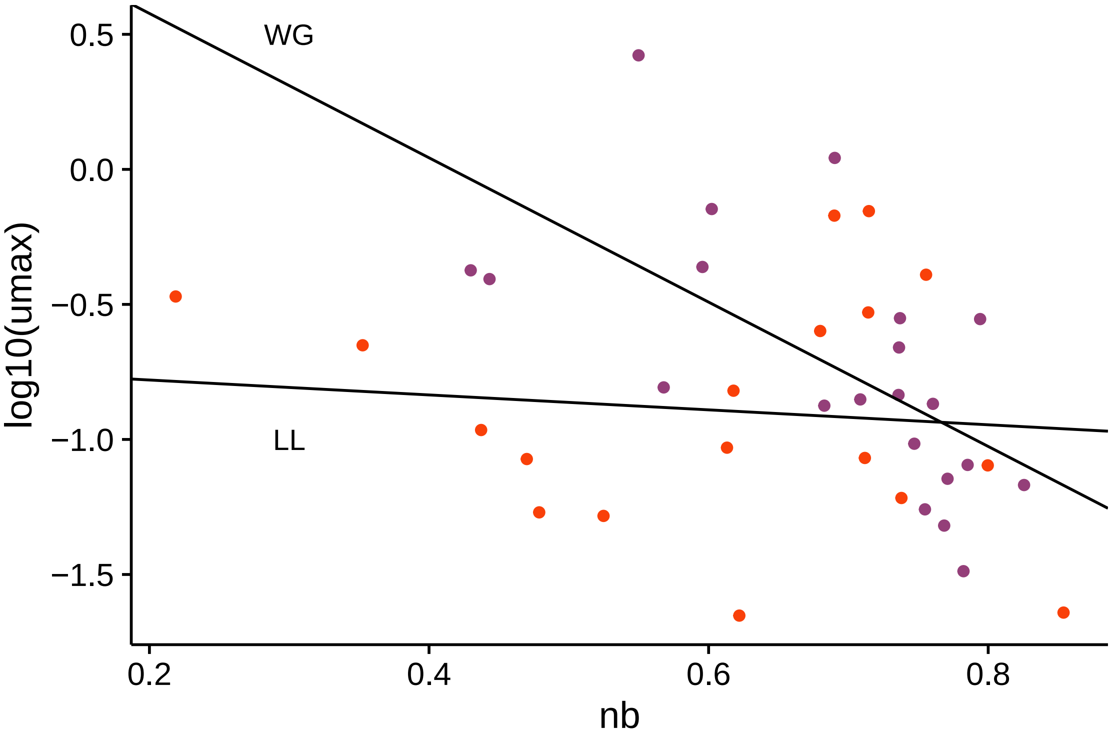
More importantly, if you were to regress acid tolerance against salinity tolerance, you would see a positive predictive relationship. However, if you corrected for phylogeny, the relationship would disappear because it is completely dependent on relatedness.

7) SYNTHESIS

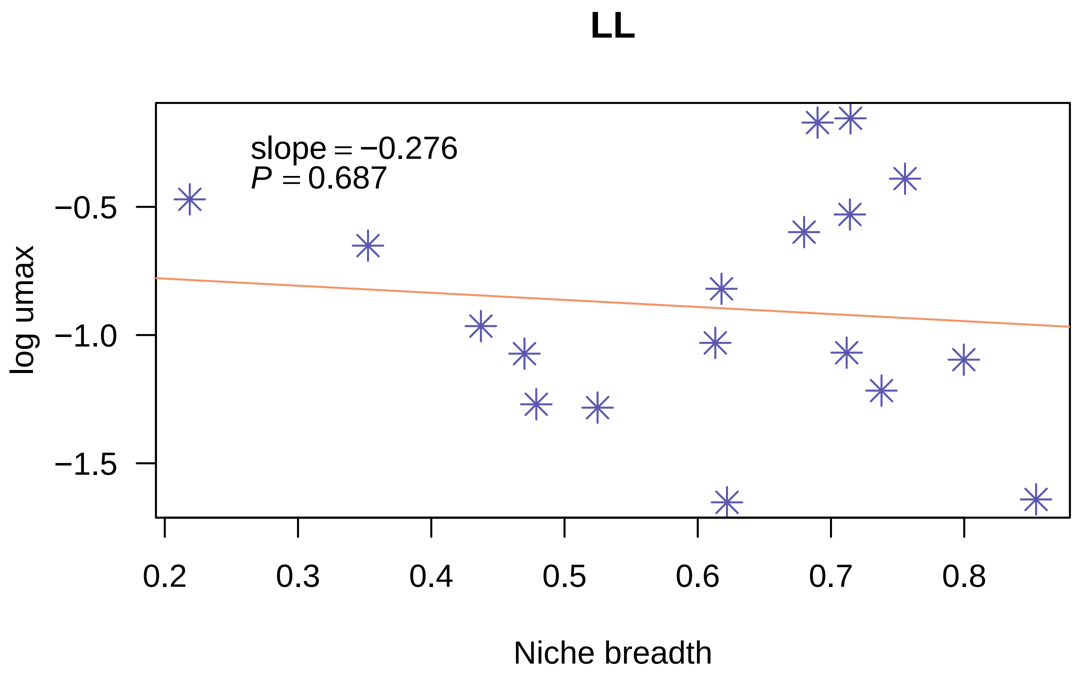
Below is the output of a multiple regression model depicting the relationship between the maximum growth rate (μ_{max}) of each bacterial isolate and the niche breadth of that isolate on the 18 different sources of phosphorus. One feature of the study which we did not take into account in the handout 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.

```
p.growth <- read.table("./data/p.isolates.raw.growth.txt", sep = "\t", header = TRUE, row.names = 1)
umax <- (apply(p.growth, 1, max)) # calculate max growth
lake <- ifelse(grepl("WG",row.names(p.growth)),'#943f79', '#f94009') # make an empty vector for lake id
tradeoff <- data.frame(nb,umax,lake) # make new data frame

###943f79=WG; LL=#f94009
##
D <- (lake == "LL") * 1
fit<-lm(log10(umax) ~ nb + D + nb * D)

LLTRADE<-filter(tradeoff, lake=="#f94009")
WGTRADE<-filter(tradeoff, lake=="#943f79")
#fitL<-lm(log10(umax)~nb,data=LLTRADE)
```

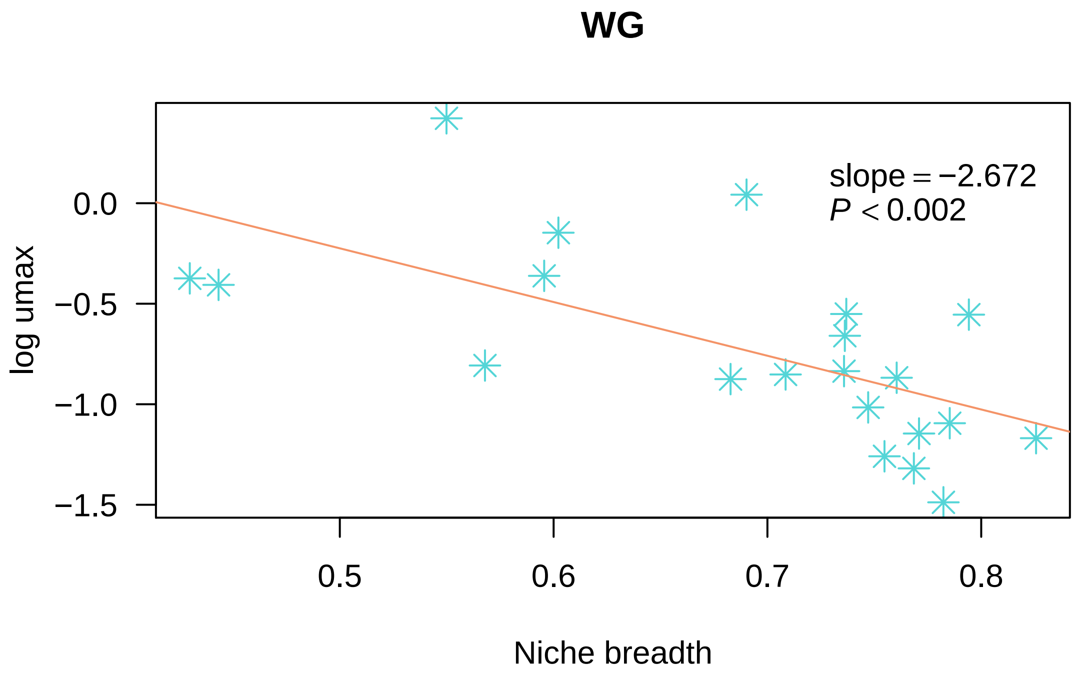
```
#fitW<-lm(log10(umax)~nb,data=WGTRADE)

#scatterLL <- qplot(x=nb, y=umax, data=LLTRADE)
#scatterWG<- qplot(x=nb, y=umax, data=WGTRADE)
#plot.new()
#scatterplot
#plot(abline(fitL))
ggplot(tradeoff,aes(nb,log10(umax)) + geom_point(aes(nb,log10(umax)),color=lake) + geom_abline(slope=-
```



```
#ggplot(tradeoff,aes(nb,log10(umax))+ stat_smooth(aes(nb,log10(umax))) + geom_point(aes(nb,log10(umax)
#ggplot(tradeoff,aes(nb,log10(umax)),color=lake)+ geom_quantile(aes(nb,log10(umax))) + geom_point(aes(n
#ggplot(tradeoff,aes(nb,umax))
#+ geom_quantile(x=nb,y=umax,color=lake)
# + geom_point(x=nb,y=umax,color=lake)


######################
#Also plot each lake separately

fitLL <- lm(log10(umax) ~ NB,data = LLTRADE)
plot(LLTRADE$NB, log10(LLTRADE$umax), las = 1, xlab = "Niche breadth", ylab = "log umax",col="#5b57af",
abline(a = fitLL$coefficients[1], b = fitLL$coefficients[2],col="#f49367")
b1L <- round(fitLL$coefficients[2], 3)
eqn <- bquote("slope" == .(b1L))
#slope
text(.25, -0.3, eqn, pos = 4)
text(.25,-.4, bquote(italic("P") == 0.687),pos=4)
```



```
fitWG <- lm(log10(umax) ~ NB,data = WGTRADE)
plot(WGTRADE$NB, log10(WGTRADE$umax), las = 1, xlab = "Niche breadth", ylab = "log umax",col="#54d5d7",
abline(a = fitWG$coefficients[1], b = fitWG$coefficients[2],col="#f49367")
b1L <- round(fitWG$coefficients[2], 3)
eqn <- bquote("slope" == .(b1L))
#slope
text(.72, .1, eqn, pos = 4)
text(.72,-.05, bquote(italic("P") < "0.002"),pos=4)
```



```
#I do not understand: At the top, you need to use the column name "nb" for R to run the chunk. In the p
```

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***:
>
> Within medium-to-larger-sized clades, there are bacterial taxa from both lakes. Within those clades, the clades near the tips tend to segregate by lake. One could, therfore, imagine that it is possible that regression relationships would disappear after accounting for phylogeny.
>
> In the analysis I ran in the chunk above, we saw that for both individual lakes, there was a negative predictive relationship of niche breadth for maximum growth rate. This is indicative of a generalist-specialist tradeoff: An individual taxon is not likely to both have a high maximum growth rate and to have a wide niche breadth. The linear model was significant only for the WG lake, however.