# 9.Phylogenetic Diversity - Communities

Elaine Hoffman; Z620: Quantitative Biodiversity, Indiana University

06 March, 2025

## OVERVIEW

Complementing taxonomic measures of $\alpha$- and $\beta$-diversity with evolutionary information yields insight into a broad range of biodiversity issues including conservation, biogeography, and community assembly. In this worksheet, you will be introduced to some commonly used methods in phylogenetic community ecology.

After completing this assignment you will know how to:

1. incorporate an evolutionary perspective into your understanding of community ecology
2. quantify and interpret phylogenetic $\alpha$- and $\beta$-diversity
3. evaluate the contribution of phylogeny to spatial 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 '9.PhyloCom' 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 *9.PhyloCom_Worksheet.Rmd* and the PDF output of `Knitr` (*9.PhyloCom_Worksheet.pdf*).

The completed exercise is due on **Wednesday, March 5$^{\text{th}}$, 2025 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 `Week7-PhyloCom/` folder,
4. load all of the required R packages (be sure to install if needed), and
5. load the required R source file.

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

```
## [1] "/cloud/project/Week7-PhyloCom"
```

```r
setwd("/cloud/project/Week7-PhyloCom/")

package.list <- c('picante', 'ape', 'seqinr', 'vegan', 'fossil',
                  'reshape', 'devtools', 'BiocManager', 'ineq',
                  'labdsv', 'matrixStats', 'pROC')

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 2.6-8
```

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

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

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

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

```
##
## Attaching package: 'shapefiles'
```

```
## The following objects are masked from 'package:foreign':
##
##     read.dbf, write.dbf
```

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

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

```
##
## Attaching package: 'BiocManager'
```

```
## The following object is masked from 'package:devtools':
##
##     install
```

```
## This is mgcv 1.9-1. For overview type 'help("mgcv-package")'.
```

```
## Registered S3 method overwritten by 'labdsv':
##   method       from
##   summary.dist ade4

## This is labdsv 2.1-0
## convert existing ordinations with as.dsvord()

##
## Attaching package: 'labdsv'

## The following objects are masked from 'package:vegan':
##
##     calibrate, pca, pco, scores

## The following objects are masked from 'package:stats':
##
##     density, loadings

##
## Attaching package: 'matrixStats'

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

## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```r
source("./bin/MothurTools.R")
```

## 2) DESCRIPTION OF DATA

**need to discuss data set from spatial ecology!**

We sampled >50 forested ponds in Brown County State Park, Yellowood State Park, and Hoosier National Forest in southern Indiana. In addition to measuring a suite of geographic and environmental variables, we characterized the diversity of bacteria in the ponds using molecular-based approaches. Specifically, we amplified the 16S rRNA gene (i.e., the DNA sequence) and 16S rRNA transcripts (i.e., the RNA transcript of the gene) of bacteria. We used a program called `mothur` to quality-trim our data set and assign sequences to operational taxonomic units (OTUs), which resulted in a site-by-OTU matrix.
In this module we will focus on taxa that were present (i.e., DNA), but there will be a few steps where we need to parse out the transcript (i.e., RNA) samples. See the handout for a further description of this week's dataset.

## 3) LOAD THE DATA

In the R code chunk below, do the following:
1. load the environmental data for the Brown County ponds (*20130801_PondDataMod.csv*),
2. load the site-by-species matrix using the `read.otu()` function,
3. subset the data to include only DNA-based identifications of bacteria,
4. rename the sites by removing extra characters,
5. remove unnecessary OTUs in the site-by-species, and
6. load the taxonomic data using the `read.tax()` function from the source-code file.

```r
env <- read.table("data/20130801_PondDataMod.csv", sep = ",", header = TRUE)
env <- na.omit(env)
```

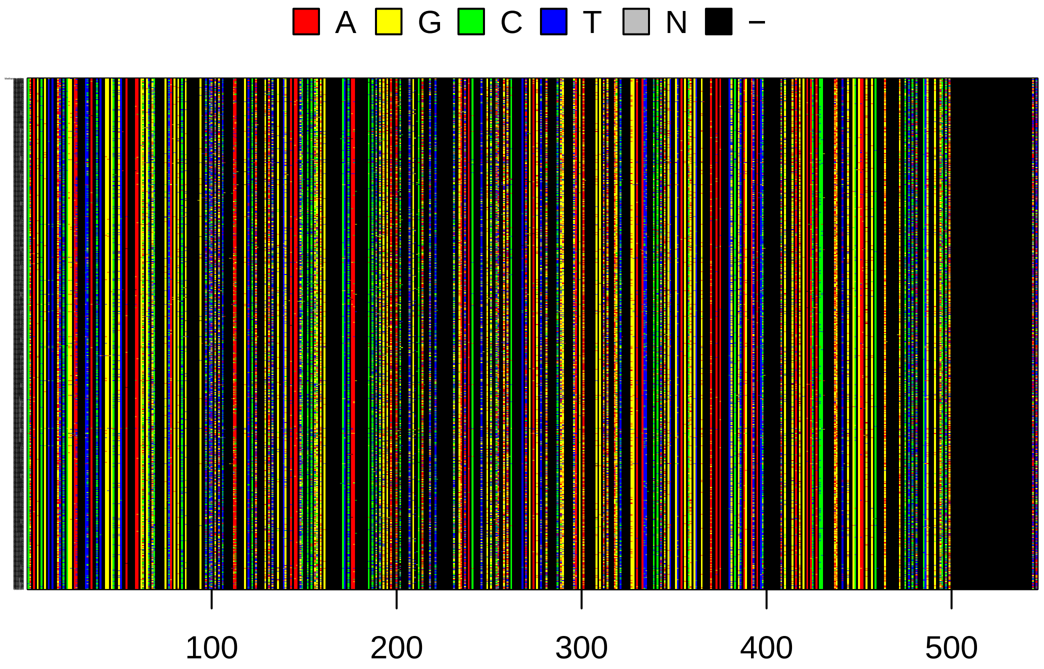
```r
#Load site-by-species matrix
comm <- read.otu(shared = "./data/INPonds.final.rdp.shared", cutoff = "1")

#Select DNA data using 'grep()'
comm <- comm[grep("*-DNA", rownames(comm)), ]

#Perform replacement of all matches with 'gsub()'
rownames(comm) <- gsub("\\-DNA", "", rownames(comm))
rownames(comm) <- gsub("\\_", "", rownames(comm))

#Remove sites not in the environmental data set
comm <- comm[rownames(comm) %in% env$Sample_ID, ]

#Remove zero-abundance OTUs from data set
comm <- comm[ , colSums(comm) > 0]

tax <- read.tax(taxonomy = "./data/INPonds.final.rdp.1.cons.taxonomy")
```

```
## Warning in type.convert.default(as.character(x)): 'as.is' should be specified
## by the caller; using TRUE
## Warning in type.convert.default(as.character(x)): 'as.is' should be specified
## by the caller; using TRUE
## Warning in type.convert.default(as.character(x)): 'as.is' should be specified
## by the caller; using TRUE
## Warning in type.convert.default(as.character(x)): 'as.is' should be specified
## by the caller; using TRUE
## Warning in type.convert.default(as.character(x)): 'as.is' should be specified
## by the caller; using TRUE
## Warning in type.convert.default(as.character(x)): 'as.is' should be specified
## by the caller; using TRUE
```

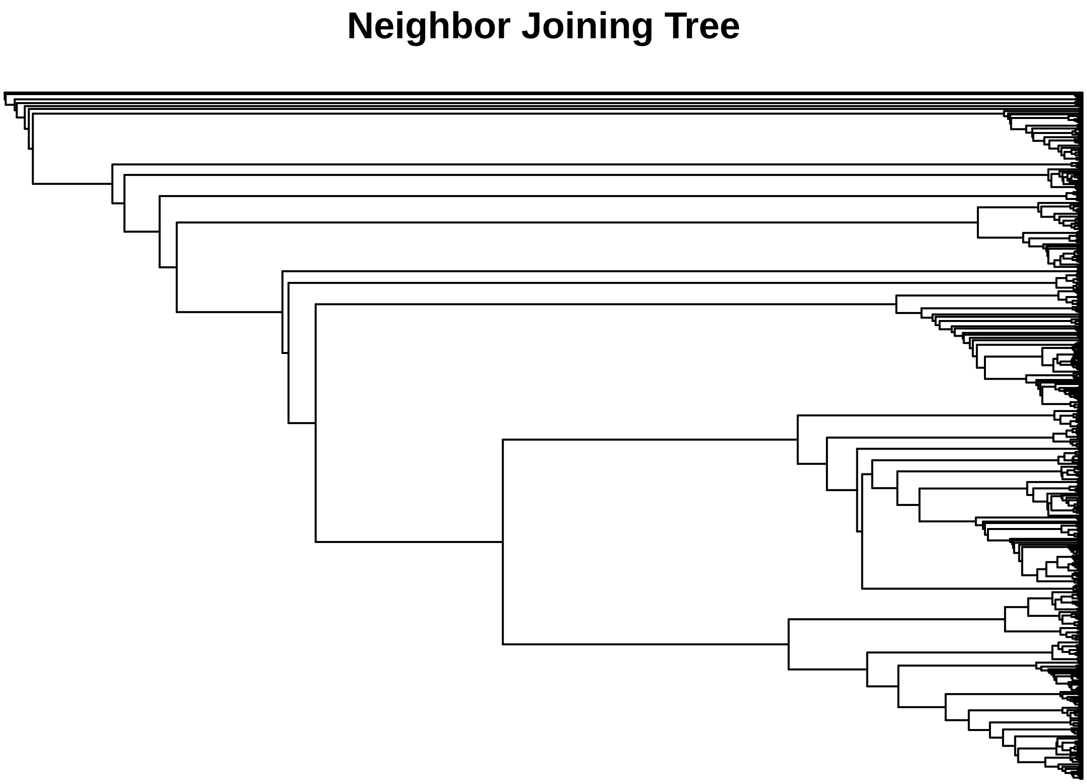
Next, in the R code chunk below, do the following:
1. load the FASTA alignment for the bacterial operational taxonomic units (OTUs),
2. rename the OTUs by removing everything before the tab (\t) and after the bar (|),
3. import the *Methanosarcina* outgroup FASTA file,
4. convert both FASTA files into the DNAbin format and combine using `rbind()`,
5. visualize the sequence alignment,
6. using the alignment (with outgroup), pick a DNA substitution model, and create a phylogenetic distance matrix,
7. using the distance matrix above, make a neighbor joining tree,
8. remove any tips (OTUs) that are not in the community data set,
9. plot the rooted tree.

```r
#Import the alignment file ('seqinr')
ponds.cons <- read.alignment(file = "./data/INPonds.final.rdp.1.rep.fasta",
                             format = "fasta")

#Rename OTUs in the FASTA file
ponds.cons$nam <- gsub(".*\t", "", ponds.cons$nam)
ponds.cons$nam <- gsub("\\|.*", "", ponds.cons$nam)

#Import outgroup sequence
outgroup <- read.alignment(file = "./data/methanosarcina.fasta", format = "fasta")
```

```
#Convert alignment file to DNAbin
DNAbin <- rbind(as.DNAbin(outgroup), as.DNAbin(ponds.cons))

#Visualize alignment
image.DNAbin(DNAbin, show.labels = T, cex.lab = 0.05, las = 1)
```



```
#Make distance matrix ('ape')
seq.dist.jc <- dist.dna(DNAbin, model = "JC", pairwise.deletion = FALSE)

#Make a neighbor-joining tree file ('ape')
phy.all <- bionj(seq.dist.jc)

#Drop tops of zero-occurance OTUs ('ape')
phy <- drop.tip(phy.all, phy.all$tip.label[!phy.all$tip.label %in%
                                              c(colnames(comm), "Methanosarcina")])

#Identify outgroup sequence
outgroup <- match("Methanosarcina", phy$tip.label)

#Root the tree {ape}
phy <- root(phy, outgroup, resolve.root = TRUE)

#Plot the rooted tree {ape}
par(mar = c(1, 1, 2, 1) + 0.1)
plot.phylo(phy, main = "Neighbor Joining Tree", "phylogram",
           show.tip.label = FALSE, use.edge.length = FALSE,
           direction = "right", cex = 0.6, label.offset = 1)
```



## 4) PHYLOGENETIC ALPHA DIVERSITY

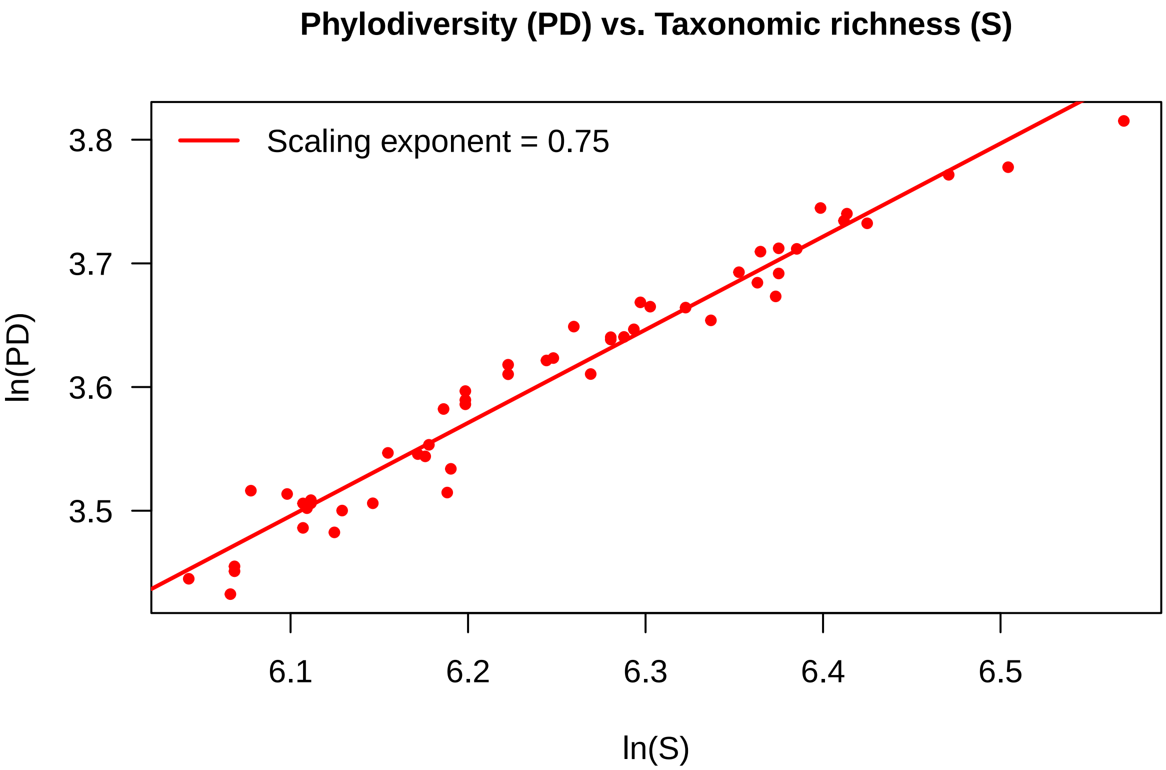
### A. Faith's Phylogenetic Diversity (PD)

In the R code chunk below, do the following:
1. calculate Faith's D using the `pd()` function.

```
#Calculate PD and S {picante}
pd <- pd(comm, phy, include.root = FALSE)
```

In the R code chunk below, do the following:
1. plot species richness (S) versus phylogenetic diversity (PD),
2. add the trend line, and
3. calculate the scaling exponent.

```
#Biplot of S and PD
par(mar = c(5, 5, 4, 1) + 0.1)

plot(log(pd$S), log(pd$PD),
     pch = 20, col = "red", las = 1,
     xlab = "ln(S)", ylab = "ln(PD)", cex.main = 1,
     main = "Phylodiversity (PD) vs. Taxonomic richness (S)")

#Test of power-law relationship
fit <- lm('log(pd$PD) ~  log(pd$S)')
abline(fit, col = "red", lw = 2)
exponent <- round(coefficients(fit) [2], 2)
legend("topleft", legend = paste("Scaling exponent = ", exponent, sep = ""),
       bty = "n", lw = 2, col = "red")
```



***Question 1***: Answer the following questions about the PD-S pattern.
a. Based on how PD is calculated, how and why should this metric be related to taxonmic richness? b. When would you expect these two estimates of diversity to deviate from one another? c. Interpret the significance of the scaling PD-S scaling exponent.

> ***Answer 1a***: PD and S are inherently related because PD is calculated by adding up total branch length and as more species are added (S) that will automatically increase the total branch length but by how much is not a constant factor. ***Answer 1b***: If species are either very closely related or very distantly related this will result in a small PD vs a large PD with the same S which would cause the metrics to deviate from each other significantly. ***Answer 1c***: The scaling exponent is how much PD increases for each addition to S. In the graph above, the scaling exponent is 0.75 which means PD increases by that factor for every species added. This suggests that the species in this system are somewhat closely related because PD is increasing at a slower rate than S.

### i. Randomizations and Null Models

In the R code chunk below, do the following:
1. estimate the standardized effect size of PD using the `richness` randomization method.

```
#Estimate standardized effect size of PD via randomization ('picante')
ses.pd.r <- ses.pd(comm[1:2,], phy, null.model = "richness", runs = 25,
                   include.root = FALSE)

#help("ses.pd")

#Estimate standardized effect size of PD via frequency ('picante')
ses.pd.f <- ses.pd(comm[1:2,], phy, null.model = "frequency", runs = 25,
                   include.root = FALSE)

#Estimate standardized effect size of PD via phylogeny.pool ('picante')
ses.pd.p <- ses.pd(comm[1:2,], phy, null.model = "phylogeny.pool", runs = 25,
                   include.root = FALSE)
```

***Question 2***: Using `help()` and the table above, run the `ses.pd()` function using two other null models and answer the following questions:

a. What are the null and alternative hypotheses you are testing via randomization when calculating `ses.pd`?
b. How did your choice of null model influence your observed ses.pd values? Explain why this choice affected or did not affect the output.

> ***Answer 2a***: The null hypothesis is that the PD observed is not significantly different from a PD generated solely from a randomized pool with no ecological processes acting on it. The alternative hypothesis is that the PD observed is significantly different from a PD generated solely from a randomized pool with no ecological processes acting on it, indicating that there are non-random effects influencing the PD. ***Answer 2b***: The observed values for each site did not change depending on the null model because they are based on the actual samples and not the null model. The different models did yield different p-values and z-scores though which could change how the results are interpreted. However, none of the p-values ended up being less than 0.05 even though there was a great deal of variation between the different null models, so none of the results were statistically significant. Nevertheless this still illustrates the importance of considering null model selection when interpreting results.

### B. Phylogenetic Dispersion Within a Sample

Another way to assess phylogenetic $\alpha$-diversity is to look at dispersion within a sample.

#### i. Phylogenetic Resemblance Matrix

In the R code chunk below, do the following:
1. calculate the phylogenetic resemblance matrix for taxa in the Indiana ponds data set.

```
#Create a phylogenetic distance matrix ('picante')
phydist <- cophenetic.phylo(phy)
```

#### ii. Net Relatedness Index (NRI)

In the R code chunk below, do the following:
1. Calculate the NRI for each site in the Indiana ponds data set.

```
#Estimate standardization effect size of NRI via randomization ('picante')
ses.mpd <- ses.mpd(comm, phydist, null.model = "taxa.labels",
                   abundance.weighted = FALSE, runs = 25)

#Calculate NRI
NRI <- as.matrix(-1 * ((ses.mpd[,2] - ses.mpd[,3]) / ses.mpd[,4]))
rownames(NRI) <- row.names(ses.mpd)
colnames(NRI) <- "NRI"


ses.mpd <- ses.mpd(comm, phydist, null.model = "taxa.labels",
                   abundance.weighted = TRUE, runs = 25)

NRI.A <- as.matrix(-1 * ((ses.mpd[,2] - ses.mpd[,3]) / ses.mpd[,4]))
rownames(NRI) <- row.names(ses.mpd)
colnames(NRI) <- "NRI"
```

#### iii. Nearest Taxon Index (NTI)

In the R code chunk below, do the following: 1. Calculate the NTI for each site in the Indiana ponds data set.

```
#Estimate standardized effect size of NRI via randomization {picante}
ses.mntd <- ses.mntd(comm, phydist, null.model = "taxa.labels",
                     abundance.weighted = FALSE, runs = 25)

#Calculate NTI
NTI <- as.matrix(-1 * ((ses.mntd[,2] - ses.mntd[,3]) / ses.mntd[,4]))
rownames(NTI) <- row.names(ses.mntd)
colnames(NTI) <- "NTI"


ses.mntd <- ses.mntd(comm, phydist, null.model = "taxa.labels",
                     abundance.weighted = TRUE, runs = 25)

NTI.A <- as.matrix(-1 * ((ses.mntd[,2] - ses.mntd[,3]) / ses.mntd[,4]))
rownames(NTI) <- row.names(ses.mntd)
colnames(NTI) <- "NTI"
```

***Question 3***:

a. In your own words describe what you are doing when you calculate the NRI.
b. In your own words describe what you are doing when you calculate the NTI.
c. Interpret the NRI and NTI values you observed for this dataset.
d. In the NRI and NTI examples above, the arguments "abundance.weighted = FALSE" means that the indices were calculated using presence-absence data. Modify and rerun the code so that NRI and NTI are calculated using abundance data. How does this affect the interpretation of NRI and NTI?

> ***Answer 3a***: You are first calculating mean phylogenetic distance (MPD) and then subtracting a randomized MPD generated from a null model and then dividing that by the standard deviation from the randomized null model. Negative outputs then indicate that the taxa are not closely related, whereas positive outputs indicate that the taxa are closely related. A value of zero would indicate that there is no difference between the observed MPD and the randomnly generated MPD.
> ***Answer 3b***: NTI is calculated the same way as NRI, but it uses mean nearest phylogenetic neighbor distance (MNND) in place of MPD. This method only considers the closest taxa so it can be biased towards clustering at the tips of the tree and not weight clustering further into the tree. ***Answer 3c***: All of the NRI values are negative and the majority of NTI values are negative. This indicates that the taxa are less closely related than anticipated. ***Answer 3d***: After altering the code and running the calculations again, roughly half of the NRI values are now positive and almost all of the NTI values are positive. The positive values indicate that the taxa are more closely related than anticipated which is the completely opposite interpretation of what was found when using presence-absence data.

## 5) PHYLOGENETIC BETA DIVERSITY

### A. Phylogenetically Based Community Resemblance Matrix

In the R code chunk below, do the following:
1. calculate the phylogenetically based community resemblance matrix using Mean Pair Distance, and
2. calculate the phylogenetically based community resemblance matrix using UniFrac distance.

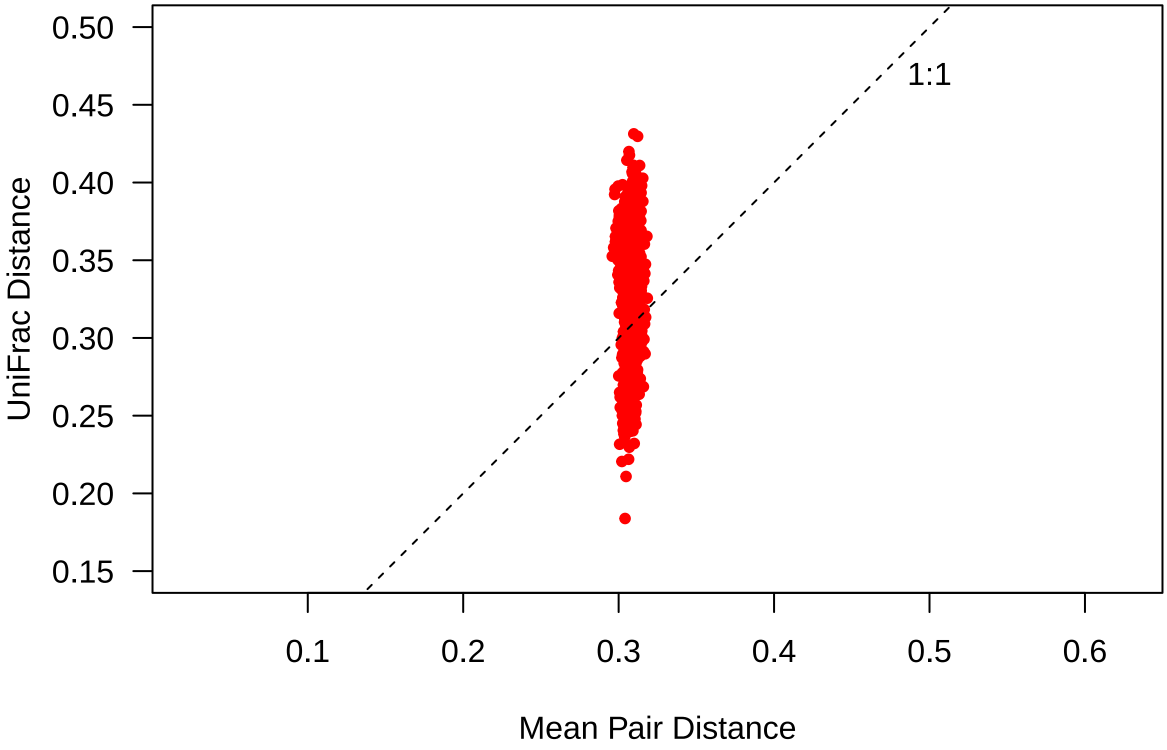
```
#Mean pairwise distance
dist.mp <- comdist(comm, phydist)
```

```
## [1] "Dropping taxa from the distance matrix because they are not present in the community data:"
## [1] "Methanosarcina"
```

```r
#Unifrac distance (Note: this takes a few minutes; be patient)
dist.uf <- unifrac(comm, phy)
```

In the R code chunk below, do the following:
1. plot Mean Pair Distance versus UniFrac distance and compare.

```r
par(mar = c(5, 5, 2, 1) + 0.1)
plot(dist.mp, dist.uf,
     pch = 20, col = "red", las = 1, asp = 1, xlim = c(0.15, 0.5), ylim = c(0.15, 0.5),
     xlab = "Mean Pair Distance", ylab = "UniFrac Distance")
abline(b = 1, a = 0, lty = 2)
text(0.5, 0.47, "1:1")
```



***Question 4***:

a. In your own words describe Mean Pair Distance, UniFrac distance, and the difference between them.
b. Using the plot above, describe the relationship between Mean Pair Distance and UniFrac distance. Note: we are calculating unweighted phylogenetic distances (similar to incidence based measures). That means that we are not taking into account the abundance of each taxon in each site.
c. Why might MPD show less variation than UniFrac?

> ***Answer 4a***: Mean pairwise distance is the mean phylogenetic distance between the two taxa that are being compared. This is calculated by adding together the branch lengths between the taxa being compared. UniFrac distance is calculated by summing the branches that are unshared by the taxa and dividing by the sum of the total number of branches in the tree. UniFrac takes into account all of the branches in the tree and focuses on how many are unshared whereas mean pairwise distance only focuses on the branches that are shared. ***Answer 4b***: UniFrac accounts for all unshared branches whereas Mean Pair Difference accounts for shared branches only. UniFrac is more senstive to the addition of more species than Mean Pair Difference which explains why there would be more variation in UniFrac as seen on the graph. ***Answer 4c***: UniFrac is more sensitive to the presense of individual species so if a few different subsets of species are swapped in and out

between calculations this would have a greater impact on UniFrac than Mean Pair Difference.

### B. Visualizing Phylogenetic Beta-Diversity

Now that we have our phylogenetically based community resemblance matrix, we can visualize phylogenetic diversity among samples using the same techniques that we used in the $\beta$-diversity module from earlier in the course.

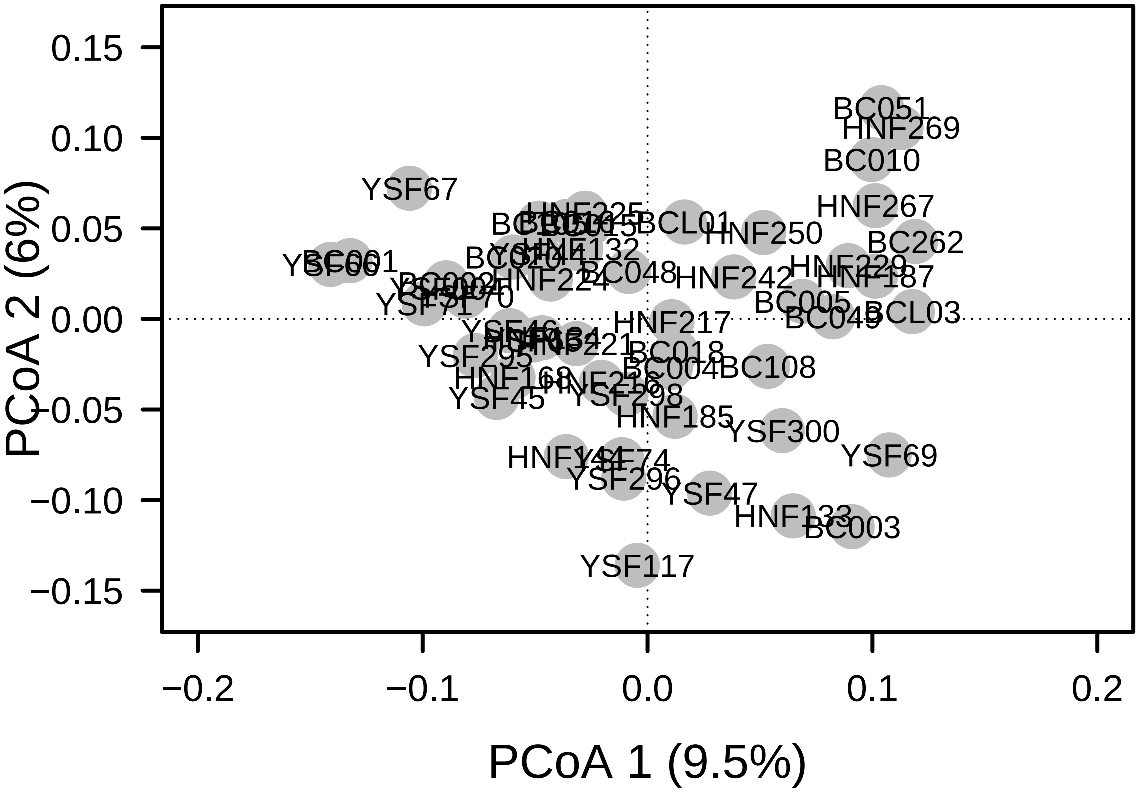
In the R code chunk below, do the following:
1. perform a PCoA based on the UniFrac distances, and
2. calculate the explained variation for the first three PCoA axes.

```r
pond.pcoa <- cmdscale(dist.uf, eig = T, k = 3)

explainvar1 <- round(pond.pcoa$eig[1] / sum(pond.pcoa$eig), 3) * 100
explainvar2 <- round(pond.pcoa$eig[2] / sum(pond.pcoa$eig), 3) * 100
explainvar3 <- round(pond.pcoa$eig[3] / sum(pond.pcoa$eig), 3) * 100
sum.eig <- sum(explainvar1, explainvar2, explainvar3)
```

Now that we have calculated our PCoA, we can plot the results.

In the R code chunk below, do the following:
1. plot the PCoA results using either the R base package or the `ggplot` package,
2. include the appropriate axes,
3. add and label the points, and
4. customize the plot.

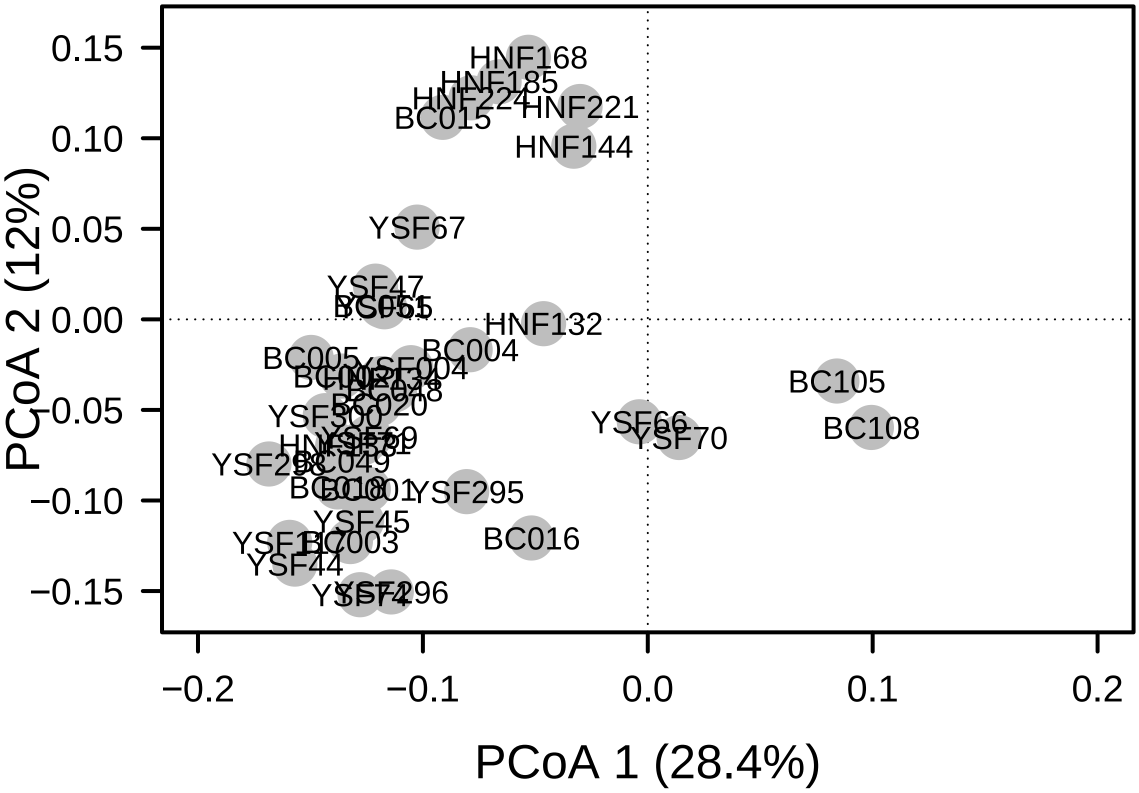
```r
#Define plot parameters
par(mar = c(5, 5, 1, 2) + 0.1)

#Initiate plot
plot(pond.pcoa$points[ ,1], pond.pcoa$points[ ,2],
     xlim = c(-0.2, 0.2), ylim = c(-0.16, 0.16),
     xlab = paste("PCoA 1 (", explainvar1, "%)", sep = ""),
     ylab = paste("PCoA 2 (", explainvar2, "%)", sep = ""),
     pch = 16, cex = 2.0, type = "n", cex.lab = 1.5, cex.axis = 1.2, axes = FALSE)

#Add Axes
axis(side = 1, labels = T, lwd.ticks = 2, cex.axis = 1.2, las = 1)
axis(side = 2, labels = T, lwd.ticks = 2, cex.axis = 1.2, las = 1)
abline(h = 0, v = 0, lty = 3)
box(lwd = 2)

#Add Points and Labels
points(pond.pcoa$points[ ,1], pond.pcoa$points[ ,2],
       pch = 19, cex = 3, bg = "gray", col = "gray")
text(pond.pcoa$points[ ,1], pond.pcoa$points[ ,2],
     labels = row.names(pond.pcoa$points))
```



In the following R code chunk: 1. perform another PCoA on taxonomic data using an appropriate measure of dissimilarity, and 2. calculate the explained variation on the first three PCoA axes.

```r
comm.db <- vegdist(comm, method = "bray")

comm.pcoa <- cmdscale(comm.db, eig = TRUE, k = 3)

explainvar1.c <- round(comm.pcoa$eig[1] / sum(comm.pcoa$eig), 3) * 100
explainvar2.c <- round(comm.pcoa$eig[2] / sum(comm.pcoa$eig), 3) * 100
explainvar3.c <- round(comm.pcoa$eig[3] / sum(comm.pcoa$eig), 3) * 100
sum.eig.c <- sum(explainvar1.c, explainvar2.c, explainvar3.c)

#Define plot parameters
par(mar = c(5, 5, 1, 2) + 0.1)

#Initiate plot
plot(comm.pcoa$points[ ,1], comm.pcoa$points[ ,2],
     xlim = c(-0.2, 0.2), ylim = c(-0.16, 0.16),
     xlab = paste("PCoA 1 (", explainvar1.c, "%)", sep = ""),
     ylab = paste("PCoA 2 (", explainvar2.c, "%)", sep = ""),
     pch = 16, cex = 2.0, type = "n", cex.lab = 1.5, cex.axis = 1.2, axes = FALSE)

#Add Axes
axis(side = 1, labels = T, lwd.ticks = 2, cex.axis = 1.2, las = 1)
axis(side = 2, labels = T, lwd.ticks = 2, cex.axis = 1.2, las = 1)
abline(h = 0, v = 0, lty = 3)
box(lwd = 2)

#Add Points and Labels
```

```r
points(comm.pcoa$points[ ,1], comm.pcoa$points[ ,2],
       pch = 19, cex = 3, bg = "gray", col = "gray")
text(comm.pcoa$points[ ,1], comm.pcoa$points[ ,2],
     labels = row.names(comm.pcoa$points))
```



***Question 5***: Using a combination of visualization tools and percent variation explained, how does the phylogenetically based ordination compare or contrast with the taxonomic ordination? What does this tell you about the importance of phylogenetic information in this system?

> ***Answer 5***: The taxonomic ordination explains significantly more variation than the phylogenetic ordination. PCoA1 in the taxonomic ordination explains 28.4% of the variation and PCoA 1 in the phylogenetic ordination only explains 9.5% of the variation. This suggested that phylogenetic information is not that important in this system.

### C. Hypothesis Testing

### i. Categorical Approach

In the R code chunk below, do the following:
1. test the hypothesis that watershed has an effect on the phylogenetic diversity of bacterial communities.

```r
#Define Environmental Category
watershed <- env$Location

#Run PERMANOVA with 'adonis()' function {vegan}
phylo.adonis <- adonis2(dist.uf ~ watershed, permutations = 999)

tax.adonis <- adonis2(vegdist(decostand(comm, method = "log"),
                              method = "bray") ~ watershed,
                      permutations = 999)
```

### ii. Continuous Approach

In the R code chunk below, do the following: 1. from the environmental data matrix, subset the variables related to physical and chemical properties of the ponds, and
2. calculate environmental distance between ponds based on the Euclidean distance between sites in the environmental data matrix (after transforming and centering using `scale()`).

```
#Define environmental variables
envs <- env[, 5:19]

#Remove redundant variables
envs <- envs[, -which(names(env) %in% c("TDS", "Salinity", "Cal_Volume"))]

#Create distance matrix for environmental variables
env.dist <- vegdist(scale(envs), method = "euclid")
```

In the R code chunk below, do the following:
1. conduct a Mantel test to evaluate whether or not UniFrac distance is correlated with environmental variation.

```
#Conduct Mantel Test ('vegan')
mantel(dist.uf, env.dist)
```

```
##
## Mantel statistic based on Pearson's product-moment correlation
##
## Call:
## mantel(xdis = dist.uf, ydis = env.dist)
##
## Mantel statistic r: 0.08433
##       Significance: 0.151
##
## Upper quantiles of permutations (null model):
##   90%   95% 97.5%   99%
## 0.107 0.138 0.170 0.202
## Permutation: free
## Number of permutations: 999
```

Last, conduct a distance-based Redundancy Analysis (dbRDA).

In the R code chunk below, do the following:
1. conduct a dbRDA to test the hypothesis that environmental variation effects the phylogenetic diversity of bacterial communities,
2. use a permutation test to determine significance, and 3. plot the dbRDA results

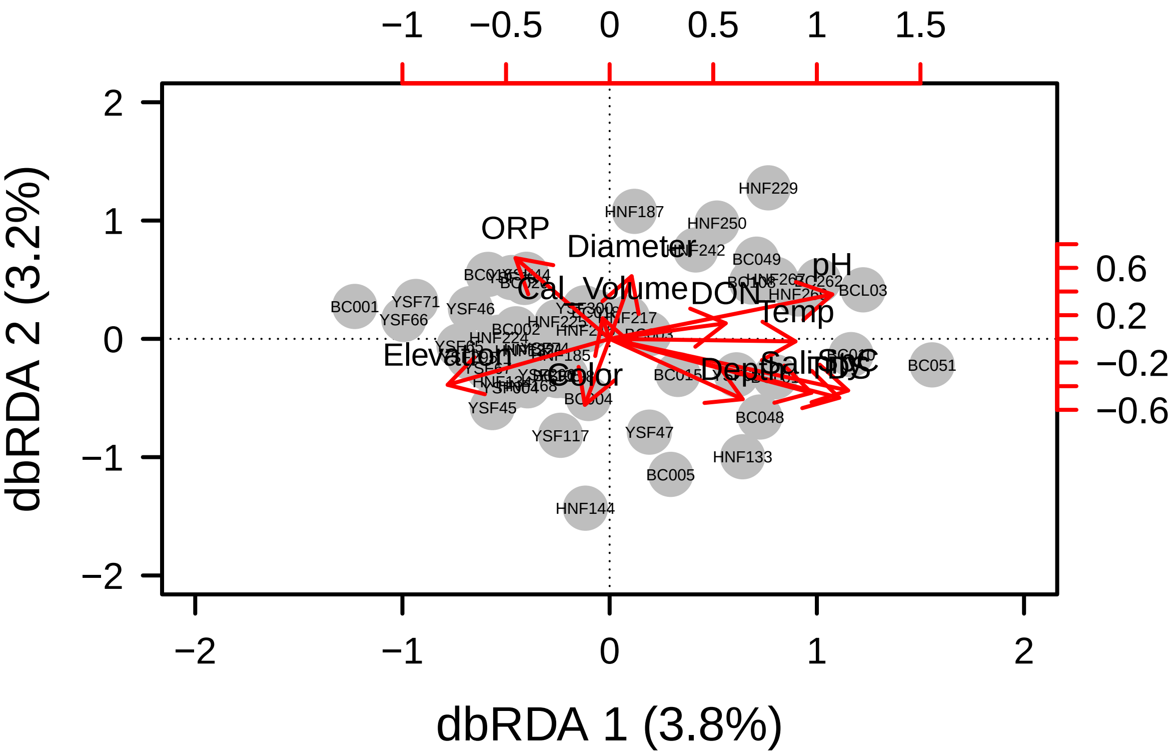
```
#Conduct dbRDA ('vegan')
ponds.dbrda <- vegan::dbrda(dist.uf ~ ., data = as.data.frame(scale(envs)))

#permutation tests: axes and environment variables
anova(ponds.dbrda, by = "axis")
```

```
## Permutation test for dbrda under reduced model
## Forward tests for axes
## Permutation: free
## Number of permutations: 999
##
## Model: vegan::dbrda(formula = dist.uf ~ Elevation + Diameter + Depth + Cal_Volume + ORP + Temp + SpC
##           Df SumOfSqs      F Pr(>F)
## dbRDA1    1  0.10324 1.9852  0.477
## dbRDA2    1  0.08592 1.6521  0.821
## dbRDA3    1  0.08171 1.5711  0.854
## dbRDA4    1  0.07321 1.4077  0.949
## dbRDA5    1  0.06591 1.2674  0.992
## dbRDA6    1  0.05049 0.9709  1.000
## dbRDA7    1  0.04671 0.8982
## dbRDA8    1  0.04175 0.8027
## dbRDA9    1  0.03606 0.6934
## dbRDA10   1  0.03302 0.6349
## dbRDA11   1  0.03078 0.5919
## dbRDA12   1  0.02921 0.5617
## Residual 39  2.02820
```

```
ponds.fit <- envfit(ponds.dbrda, envs, perm = 999)
ponds.fit
```

```
## 
## ***VECTORS
## 
##              dbRDA1   dbRDA2     r2 Pr(>r)
## Elevation  -0.86345 -0.50444 0.1084  0.064 .
## Diameter    0.06683  0.99776 0.0543  0.245
## Depth       0.68050 -0.73275 0.1213  0.050 *
## Cal_Volume -0.22122  0.97523 0.0062  0.854
## ORP        -0.48281  0.87572 0.1309  0.040 *
## Temp        0.98974 -0.14286 0.1179  0.050 *
## SpC         0.85986 -0.51052 0.2464  0.001 ***
## TDS         0.82742 -0.56158 0.2451  0.003 **
## Salinity    0.81889 -0.57395 0.1931  0.006 **
## pH          0.93813  0.34629 0.1823  0.007 **
## Color      -0.07756 -0.99699 0.0604  0.204
## DON         0.97923  0.20274 0.0467  0.314
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Permutation: free
## Number of permutations: 999
```

```
#Calculate explained variation
dbrda.explainvar1 <- round(ponds.dbrda$CCA$eig[1] /
                             sum(c(ponds.dbrda$CCA$eig, ponds.dbrda$CA$eig)), 3) * 100

dbrda.explainvar2 <- round(ponds.dbrda$CCA$eig[2] /
                             sum(c(ponds.dbrda$CCA$eig, ponds.dbrda$CA$eig)), 3) * 100


#Make dbRDA plot
#Extract scores from the dbRDA object
ponds_scores <- vegan::scores(ponds.dbrda, display = "sites")

#Define plot parameters
par(mar = c(5, 5, 4, 4) + 0.1)

#Initiate plot
plot(ponds_scores, xlim = c(-2, 2), ylim = c(-2, 2),
     xlab = paste("dbRDA 1 (", dbrda.explainvar1, "%)", sep = ""),
     ylab = paste("dbRDA 2 (", dbrda.explainvar2, "%)", sep = ""),
     pch = 16, cex = 2.0, type = "n", cex.lab = 1.5, cex.axis = 1.2, axes = FALSE)

#Add axes
axis(side = 1, labels = T, lwd.ticks = 2, cex.axis = 1.2, las = 1)
axis(side = 2, labels = T, lwd.ticks = 2, cex.axis = 1.2, las = 1)
abline(h = 0, v = 0, lty = 3)
box(lwd = 2)

#Extract site scores
wa_scores <- vegan::scores(ponds.dbrda, display = "sites")

#Add points and labels
points(wa_scores,
       pch = 19, cex = 3, col = "gray")
text(wa_scores,
     labels = rownames(wa_scores),
     cex = 0.5)

#extract environmental vectors (biplot scores)
vectors <- vegan::scores(ponds.dbrda, display = "bp")

#Add environmental vectors to the plot
arrows(0, 0, vectors[, 1] * 2, vectors[, 2] * 2,
       lwd = 2, lty = 1, length = 0.2, col = "red")

#Add labels for the environmental vectors
text(vectors[, 1] * 2, vectors[, 2] * 2, pos = 3,
     labels = row.names(vectors))

#Add axes for the vectors
axis(side = 3, lwd.ticks = 2, cex.axis = 1.2, las = 1, col = "red", lwd = 2.2,
     at = pretty(range(vectors[ ,1]) * 2),
     labels = pretty(range(vectors[, 1]) * 2))

axis(side = 4, lwd.ticks = 2, cex.axis = 1.2, las = 1, col = "red", lwd = 2.2,
     at = pretty(range(vectors[ ,2]) * 2),
     labels = pretty(range(vectors[, 2]) * 2))
```



***Question 6***: Based on the multivariate procedures conducted above, describe the phylogenetic patterns of $\beta$-diversity for bacterial communities in the Indiana ponds.

> ***Answer 6***: Based on the results of the Mantel test, UniFrac distance and environmental distances do not have a statistically significant correlation, indicating that environmental factors are not exerting a strong influence on phylogenetic patterns. This suggests that other variables are likely contributing to the patterns seen in the phylogenetic diversity. The permutation test found that depth, temp, SpC, TDS, Salinity, and pH had a significant correlation with the phylogenetic patterns in this dataset, however, the dbRDA plot showed that the first two axes only account for 3.8% and 3.2% of the variation respectively. This supports the findings from the Mantel test that show that environmental variables are not the main factor controlling the observed phylogenetic patterns in diversity.

## SYNTHESIS

***Question 7***: Ignoring technical or methodological constraints, discuss how phylogenetic information could be useful in your own research. Specifically, what kinds of phylogenetic data would you need? How could you use it to answer important questions in your field? In your response, feel free to consider not only phylogenetic approaches related to phylogenetic community ecology, but also those we discussed last week in the PhyloTraits module, or any other concepts that we have not covered in this course.

> ***Answer 7***: I study mycorrhizal fungi and phylogenetic data would be expecially usful in answering questions about ectomycorrhizal fungi given that they did not evolve from a single common ancestor and therefore likely exhibit a great deal of genetic variation as well as trait variation. Under the MANE hypothesis we currently group all ectomycorrhizae as a single functional group which may be over simplifying predictions. To answer this question I would use DNA sequences and data on functional traits to develop trait correlations with phylogeny to potentially improve our predictions of the role of different mycorrhizae in modulating nutrient cycling.