Phylogenetic Diversity - Communities

Student Name; Z620: Quantitative Biodiversity, Indiana University

01 March, 2017

OVERVIEW

Complementing taxonomic measures of α- and β-diversity with evolutionary information yields insight into a broad range of biodiversity issues including conservation, biogeography, and community assembly. In this assignment, 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 α- and β-diversity
3. evaluate the contribution of phylogeny to spatial patterns of biodiversity

Directions:

1. Change "Student Name" on line 3 (above) with your name.
2. Complete as much of the assignment as possible during class; what you do not complete in class will need to be done 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. When you are done, **Knit** the text and code into a PDF file.
7. After Knitting, please submit the completed assignment by creating a **pull request** via GitHub. Your pull request should include this file *PhyloCom_assignment.Rmd* and the PDF output of `Knitr` (*PhyloCom_assignment.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 `/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
clr = function() {
  ENV = globalenv()
  ll = ls(envir = ENV)
  ll = ll[ll != "clr"]
  rm(list = ll, envir = ENV)
}
```

```
getwd()
setwd("/Users/bhbeidler/GitHub/QB2017_Beidler/Week7-PhyloCom")

package.list = c('picante', 'ape', 'seqinr', 'vegan', 'fossil', 'simba')
for (package in package.list) {if (!require(package, character.only = TRUE, quietly = TRUE)) { install.
  library(package, character.only = TRUE)
} }

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

2) DESCRIPTION OF DATA

We will revisit the data that was used in the Spatial Diversity module. As a reminder, in 2013 we sampled ~ 50 forested ponds located in Brown County State Park, Yellowwood State Park, and Hoosier National Forest in southern Indiana. See the handout for a further description of this week's dataset.

3) LOAD THE DATA

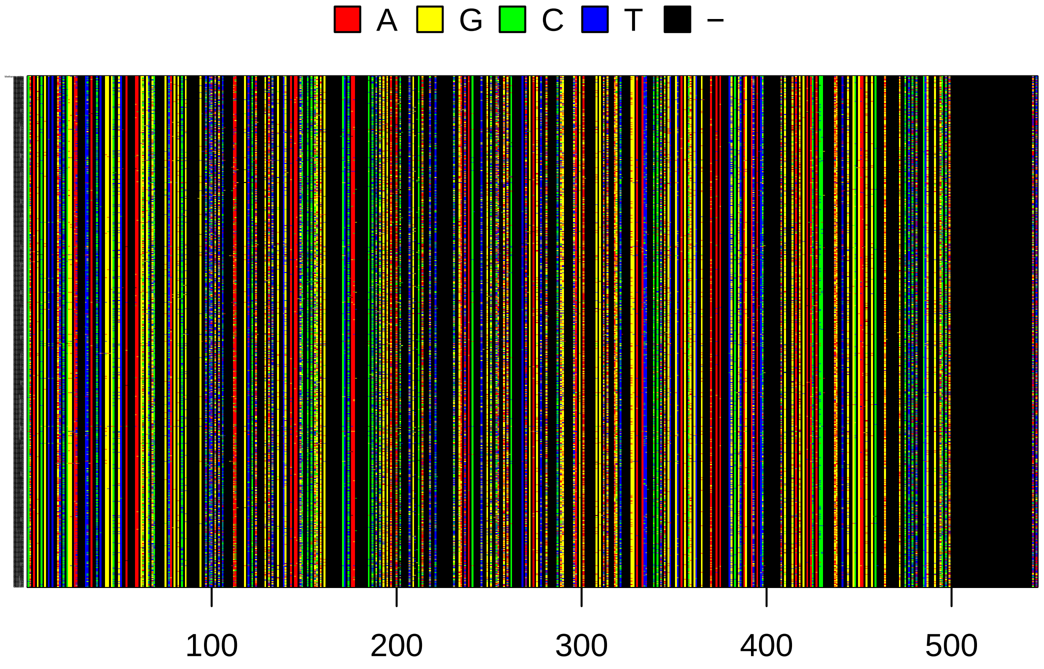
In the R code chunk below, do the following:
1. load the environmental data for the Brown County ponds (*20130801_PondDataMod.csv*),
2. load the site-by-species matrix using the `read.otu()` function,
3. subset the data to include only DNA-based identifications of bacteria,
4. rename the sites by removing 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.

```
# 1. load the environmental data for the Brown County ponds (*20130801_PondDataMod.csv*)
env = read.table("data/20130801_PondDataMod.csv", sep = ",", header = TRUE)
env = na.omit(env)

# 2. Load Site-by-Species Matrix
comm = read.otu(shared = "./data/INPonds.final.rdp.shared", cutoff = "1")

# 3. subset the data to include only DNA-based identifications of bacteria
# 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))

# 4. rename the sites by removing extra characters
comm = comm[rownames(comm) %in% env$Sample_ID, ]

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

# 6. load the taxonomic data using the `read.tax()` function from the source-code file.
tax = read.tax(taxonomy = "./data/INPonds.final.rdp.1.cons.taxonomy")
```

Next, in the R code chunk below, do the following:
1. load the FASTA alignment for the bacterial operational taxonomic units (OTUs),

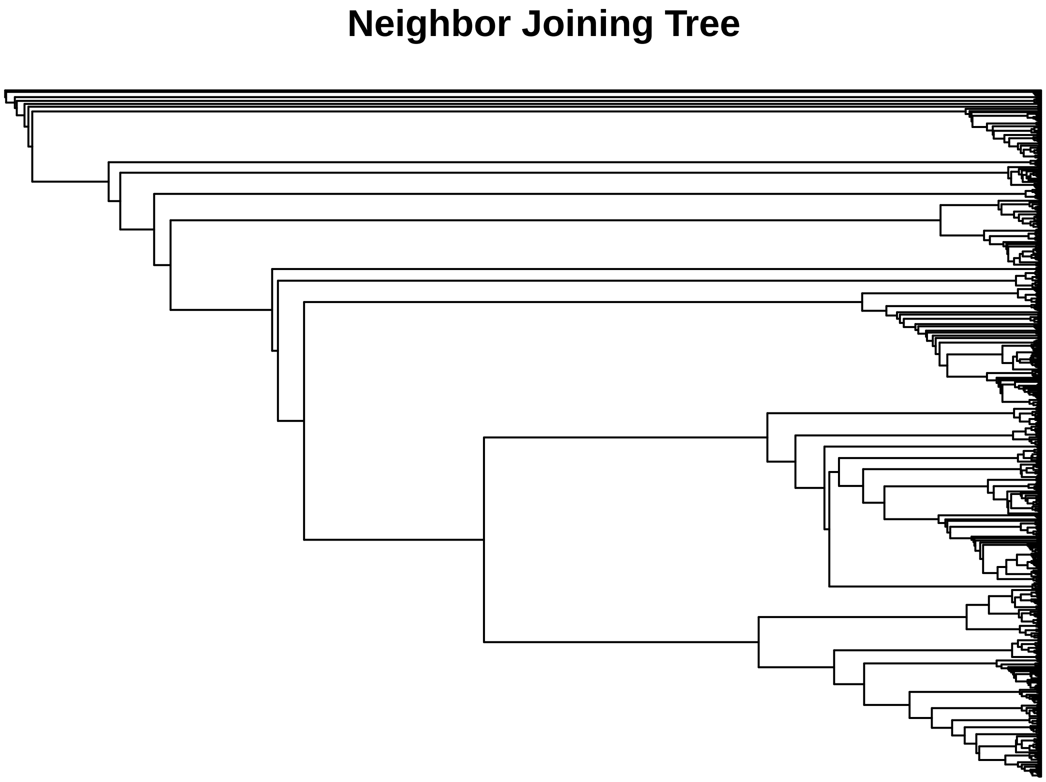
2. rename the OTUs by removing everything before the tab (\t) and after the bar (|),
3. import the *Methanosarcina* outgroup FASTA file,
4. convert both FASTA files into the DNAbin format and combine using `rbind()`,
5. visualize the sequence alignment,
6. using the alignment (with outgroup), pick a 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.

```
# 1. load the FASTA alignment for the bacterial operational taxonomic units (OTUs)
ponds.cons = read.alignment(file = "./data/INPonds.final.rdp.1.rep.fasta", format = "fasta")

# 2. rename the OTUs by removing everything before the tab (\\t) and after the bar (|),
ponds.cons$nam =  gsub("\\|.*$", "", gsub("^.*?\t", "", ponds.cons$nam))

# 3. import the *Methanosarcina* outgroup FASTA file
outgroup = read.alignment(file = "./data/methanosarcina.fasta", format = "fasta")

# 4. convert both FASTA files into the DNAbin format and combine using `rbind()`
DNAbin = rbind(as.DNAbin(outgroup), as.DNAbin(ponds.cons))

# 5. visualize the sequence alignment
image.DNAbin(DNAbin, show.labels=T, cex.lab = 0.05, las = 1)
```



```
# 6. using the alignment (with outgroup), pick a DNA substitution model, and create a phylogenetic dist
seq.dist.jc = dist.dna(DNAbin, model = "JC", pairwise.deletion = FALSE)

# 7. using the distance matrix above, make a neighbor joining tree
phy.all = bionj(seq.dist.jc)

# 8. remove any tips (OTUs) that are not in the community data set
```

```
phy = drop.tip(phy.all, phy.all$tip.label[!phy.all$tip.label %in% c(colnames(comm), "Methanosarcina")])

# Id the outgroup
outgroup = match("Methanosarcina", phy$tip.label)

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

# 9. plot the rooted tree
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

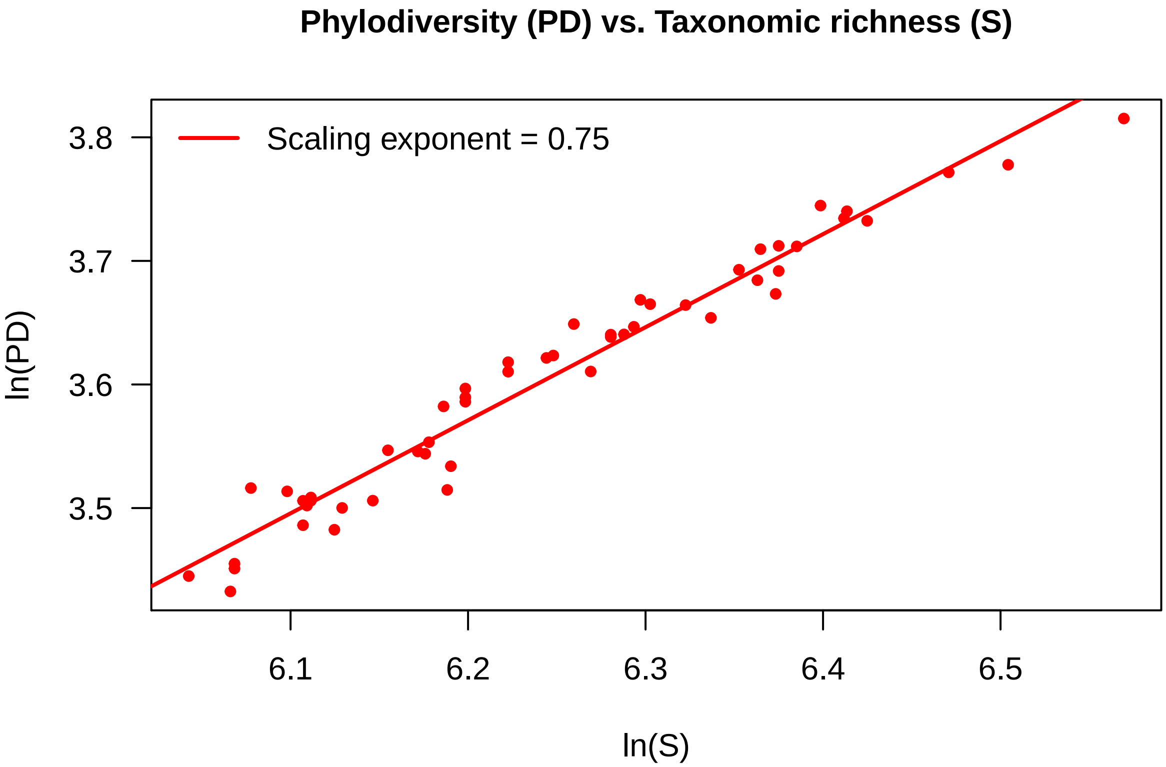
A. Faith's Phylogenetic Diversity (PD)

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.

```r
# 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
# 2. add the trend line
fit = lm('log(pd$PD) ~ log(pd$S)')
abline(fit, col = "red", lw = 2)

# 3. calculate the scaling exponent.
exponent = round(coefficients(fit)[2], 2)
legend("topleft", legend=paste("Scaling exponent = ", exponent, sep = ""),
       bty = "n", lw = 2, col = "red")
```


Phylodiversity (PD) vs. Taxonomic richness (S)

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

> ***Answer 1a***: Taxonomic richness is equal to the number of different species at a site or within a sample. PD sums the branch lengths for each species found in a sample, thus PD covaries with richness. As richness increases so does diversity. In other words, PD is the phylogenetic counterpart to richness- the sum of the number of tree segements instead of the sum of the number of species found at a site or within a sample.

> ***Answer 1b***: The relationship between PD and richness is a power relationship. PD varies as a power of richness.

Answer 1c: Rapid speciation or adaaptive radiation can result in low phylogenetic diversity but high taxonomic richness.

Answer 1d: PD scales to the 3/4 power of richness. So we can draw conclusions about the phylogenetic diversity of a sample based off of the richness value (as richness increaes, PD increases by a factor of 0.75)

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.

```
# 1. estimate the standardized effect size of PD using the `richness` randomization method.
ses.pd = ses.pd(comm[1:2,], phy, null.model = "richness", runs = 25, include.root = FALSE)
# Richness null model - Randomize community data matrix abundances within samples (maintains sample spe

# Using 2 other null models
# Frequncy
#Randomize community data matrix abundances within species (maintains species occurence frequency)
ses.pd_f = ses.pd(comm[1:2,], phy, null.model = "frequency", runs = 25, include.root = FALSE)

# independentswap
# Randomize community data matrix with the independent swap algorithm (Gotelli 2000) maintaining specie
ses.pd_is = ses.pd(comm[1:2,], phy, null.model = "independentswap", 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: H0: Mean phylogenetic diversity does not differ from that of a randomly assembled community (taxa are more or less similar than what you would expect by chance- PD = 0). H1: Mean PD is greater than that of a randomly assembled community (the sample is more phylogenetically diverse than what is expected by chance).

Answer 2b: For both the richness and independentswap null models, the observed mean PD values did not differ from the random mean PD values ($P>0.05$) for both samples. However, mean PD is significantly higher than the null expectation ($P=0.04$) for the BCOO2 sample when the frequency null model is used. Of the 3 the null models used, the frequency null model does not maintain sample species richness which may be more important than maintaining species occurrence frequency.

B. Phylogenetic Dispersion Within a Sample

Another way to assess phylogenetic α-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.

```
# 1. calculate the phylogenetic resemblance matrix for taxa in the Indiana ponds data set.
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.

```
# 1. Calculate the NRI for each site in the Indiana ponds data set.
# Estimate standardized effect size of NRI via randomization for presence absence data
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"

# Estimate standardized effect size of NRI via randomization for abundance data
ses.mpd.ab = ses.mpd(comm, phydist, null.model = "taxa.labels", abundance.weighted = TRUE, runs = 25)

# Calculate NRI
NRI.ab = 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 for presence absence data
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"

# Estimate Standardized Effect Size of NRI via Randomization for abundance data
ses.mntd.ab = ses.mntd(comm, phydist, null.model = "taxa.labels", abundance.weighted = TRUE, runs = 25)
# Calculate NTI
NTI.ab = 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: To calculate the Nearest Relatedness Index (NRI), first you calculate the average pairwise branch lengths between taxa in your community to get the mean phylogenetic distance

(MPD) for that sample. Next, you repeat the procedure for a randomized community or sample. Then you see how the MPD matches up with the MPD you would expect from a randomly assembled community by subtracting the random MPD from the MPD and dividing by standard deviation for the random MPD values.

Answer 3b: To calculate the Nearest Taxon Index (NTI) you perform a series of steps similar to the calculation of NRI. However instead of comparing the MPD between observed and random communities for a given sample, instead you compare the mean phylogenetic distance between all taxa and their closest phylogenetic neighbor. The NTI weights more based on the tips, because it only considers the most closely related taxa- which is a divergence from the NRI.

Answer 3c: The majority of the NRI and NTI values in the dataset are negative, meaning that most of the samples are phylogenetically overdispersed. In other words taxa are less related than by chance and nearest taxa are more distantly related than expexted.

Answer 3d: This does not affect the interpretation of the NRI and NTI values.

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.

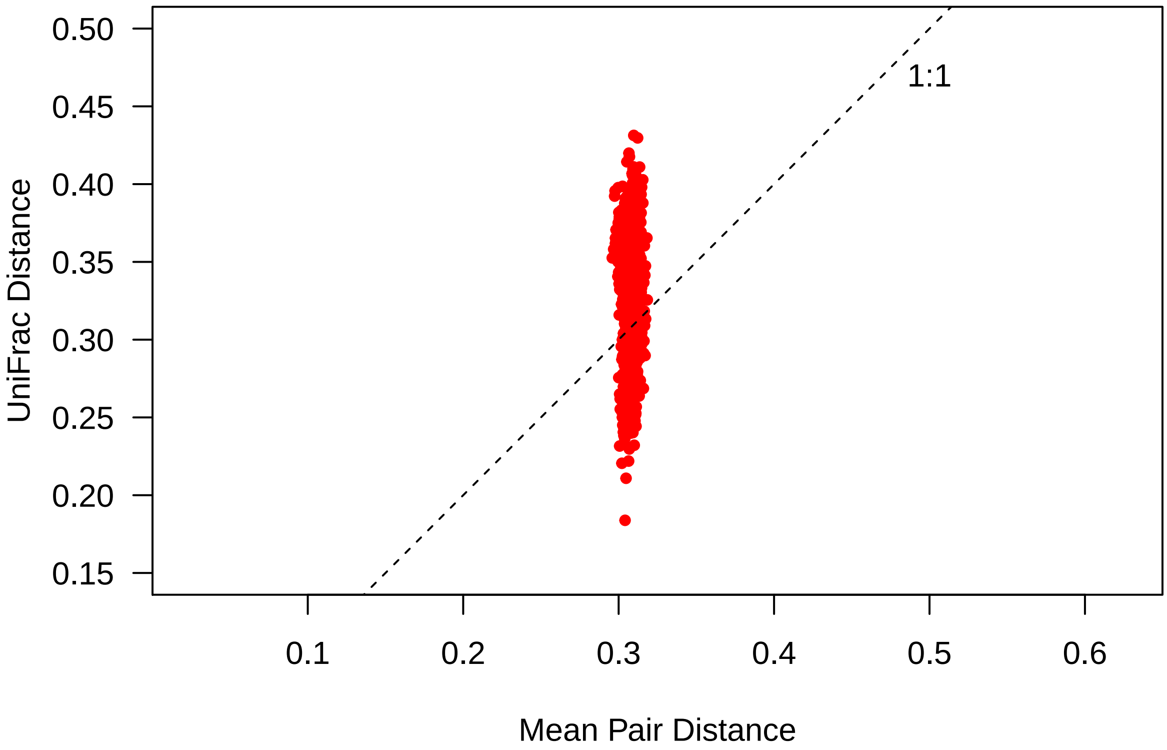
```
# 1. calculate the phylogenetically based community resemblance matrix using Mean Pair Distance
# Mean Pairwise Distance
dist.mp = comdist(comm, phydist)
```

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

```
# 2. calculate the phylogenetically based community resemblance matrix using UniFrac distance.
# 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.

```
# 1. plot Mean Pair Distance versus UniFrac distance and compare.
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
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 Pair Distance is the mean phylogenetic distance or average branch length between pairs of taxa. UniFrac distance is calculated as the shared branch length between taxa divded by the total branch length (shared + unshared). In other words, the proportion of total branch length which is not shared between taxa. Unlike Mean Pair Distance, UniFrac compares pairs of taxa to the entire tree.
>
> ***Answer 4b***: All of the Mean Pair Distance are close in value; they range from 0.29 to 0.32 (Stdev = 0.003) Whereas, the UniFrac distances are more variable and range from 0.18 to 0.43 (Stdev = 0.037). There appears to be no relationship between Mean Pair Distance and UniFrac distance (likely because they are on such different scales).
>
> ***Answer 4c***:MPD may show less variation because the distance is not compared to that of un-related taxa.

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 β-diversity module from earlier in the course.

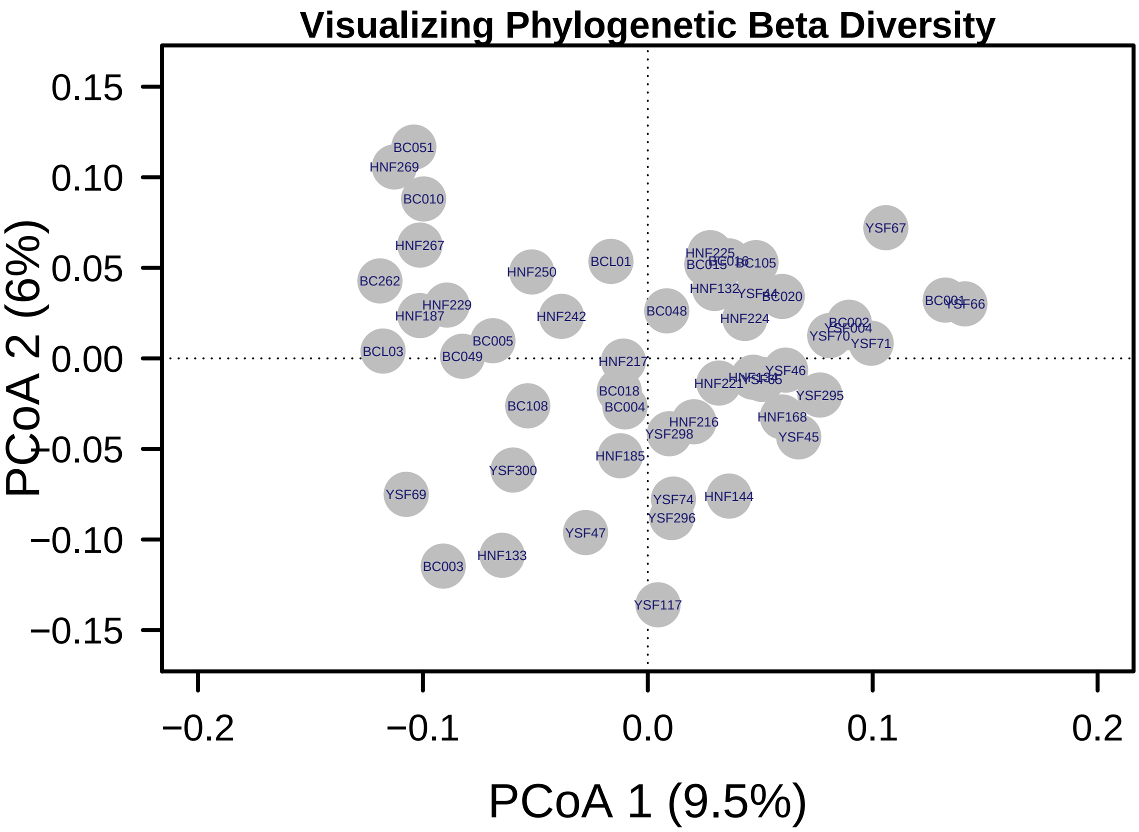
In the R code chunk below, do the following:
1. perform a PCoA based on the UniFrac distances, and
2. calculate the explained variation for the first three PCoA axes.

```r
# 1. perform a PCoA based on the UniFrac distances
pond.pcoa = cmdscale(dist.uf, eig = T, k = 3)

# 2. calculate the explained variation for the first three PCoA axes
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.

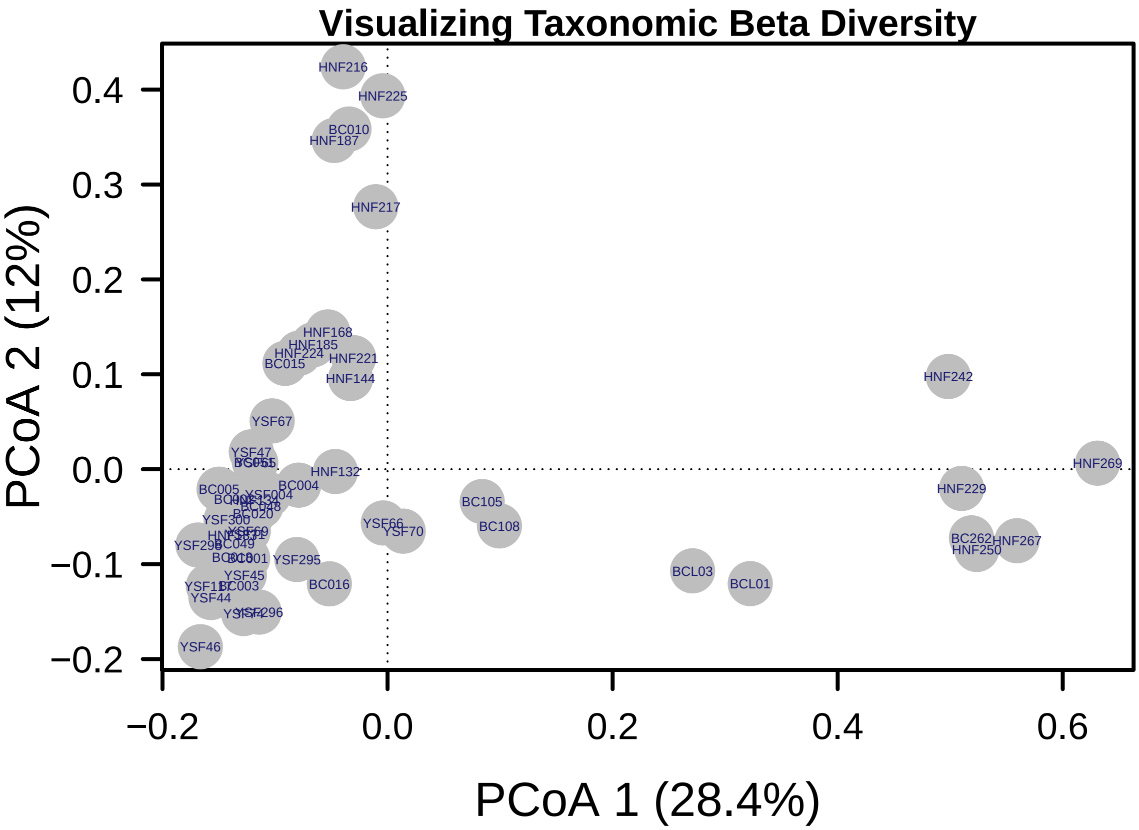
In the R code chunk below, do the following:
1. plot the PCoA results using either the R base package or the `ggplot` package,
2. include the appropriate axes,
3. add and label the points, and
4. customize the plot.

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

# 1. plot the PCoA results
plot(pond.pcoa$points[ ,1], pond.pcoa$points[ ,2], xlim = c(-0.2, 0.2), ylim = c(-.16, 0.16),main = "Vi

# 2. 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)
# 3. Add Points & 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], cex = 0.45, col = "midnightblue", labels = row.names
```



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.

```
# 1. perform another PCoA on taxonomic data using an appropriate measure of dissimilarity
bact.db = vegdist(comm, method = "bray")
# Perform a Principal Coordinates Analysis to visualize beta-diversity
bact.pcoa = cmdscale(bact.db, eig = TRUE, k = 3)

# 2. calculate the explained variation on the first three PCoA axes.
explainvar1 = round(bact.pcoa$eig[1] / sum(bact.pcoa$eig), 3) * 100
explainvar2 = round(bact.pcoa$eig[2] / sum(bact.pcoa$eig), 3) * 100
explainvar3 = round(bact.pcoa$eig[3] / sum(bact.pcoa$eig), 3) * 100
sum.eig = sum(explainvar1, explainvar2, explainvar3)

# Plot the PCOA for taxonomic data
# Define Plot Parameters
par(mar = c(5, 5, 1, 2) + 0.1)

plot(bact.pcoa$points[ ,1],bact.pcoa$points[ ,2], main = "Visualizing Taxonomic Beta Diversity",xlab = p

# 2. 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)
# 3. Add Points & Labels
points(bact.pcoa$points[ ,1], bact.pcoa$points[ ,2], pch = 19, cex = 3, bg = "gray", col = "gray")
text(bact.pcoa$points[ ,1], bact.pcoa$points[ ,2], cex = 0.45, col = "midnightblue", labels = row.names
```



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 phylogenetic ordination differs from the taxonomic ordination in that points are more clustered in the taxonomic ordination. Additionally, the PCoA axes explain more of the variation for the taxonomic ordination. This tells us that phylogenetic information may not be as important as environmental or site level variables for structuring microbial communities 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
# 1. test the hypothesis that watershed has an effect on the phylogenetic diversity of bacterial commun
# Define Environmental Category
watershed = env$Location
# Run PERMANOVA with `adonis()` Function {vegan}
adonis(dist.uf ~ watershed, permutations = 999)
```

```
##
## Call:
```

```
## adonis(formula = dist.uf ~ watershed, permutations = 999)
##
## Permutation: free
## Number of permutations: 999
##
## Terms added sequentially (first to last)
##
##           Df SumsOfSqs  MeanSqs F.Model     R2 Pr(>F)
## watershed  2   0.13316 0.066579  1.2679 0.0492   0.02 *
## Residuals 49   2.57305 0.052511         0.9508
## Total     51   2.70621                  1.0000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# We can compare to PERMANOVA results based on taxonomy
adonis(vegdist(
        decostand(comm, method = "log"),
         method = "bray") ~ watershed,
        permutations = 999)
```

```
##
## Call:
## adonis(formula = vegdist(decostand(comm, method = "log"), method = "bray") ~      watershed, permuta
##
## Permutation: free
## Number of permutations: 999
##
## Terms added sequentially (first to last)
##
##           Df SumsOfSqs  MeanSqs F.Model      R2 Pr(>F)
## watershed  2   0.16601 0.083003  1.5689 0.06018   0.01 **
## Residuals 49   2.59229 0.052904         0.93982
## Total     51   2.75829                  1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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()`).

```
# 1. 1. from the environmental data matrix, subset the variables related to physical and chemical prope
# Define environmental variables
envs = env[, 5:19]
# Remove redudnant variables
envs = envs[, -which(names(envs) %in% c("TDS", "Salinity", "Cal_Volume"))]

# 2. calculate environmental distance between ponds based on the Euclidean distance between sites in th
env.dist = vegdist(scale(envs), method = "euclid")
# 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.

```
# 1. conduct a Mantel test to evaluate whether or not UniFrac distance is correlated with environmental
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: 0.1604
##       Significance: 0.06
##
## Upper quantiles of permutations (null model):
##   90%   95% 97.5%   99%
## 0.126 0.171 0.200 0.240
## 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

```
# 1. conduct a dbRDA to test the hypothesis that environmental variation effects the phylogenetic diver
ponds.dbrda = vegan::dbrda(dist.uf ~ ., data = as.data.frame(scale(envs)))

# 2. use a permutation test to determine significance
anova(ponds.dbrda, by = "axis")
```

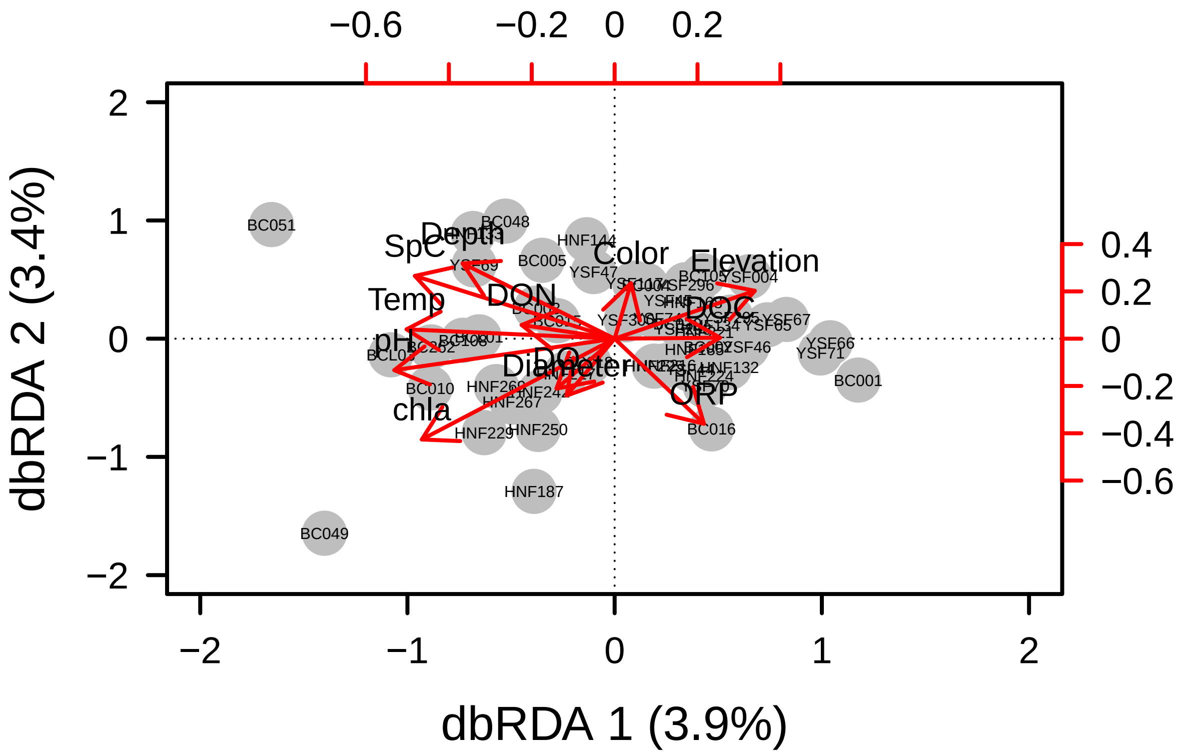
```
## Permutation test for dbrda under reduced model
## Marginal tests for axes
## Permutation: free
## Number of permutations: 999
##
## Model: vegan::dbrda(formula = dist.uf ~ Elevation + Diameter + Depth + ORP + Temp + SpC + DO + pH + C
##          Df SumOfSqs      F Pr(>F)
## dbRDA1    1  0.10566 2.0152  0.002 **
## dbRDA2    1  0.09258 1.7658  0.001 ***
## dbRDA3    1  0.07555 1.4409  0.041 *
## dbRDA4    1  0.06677 1.2735  0.089 .
## dbRDA5    1  0.05666 1.0807  0.356
## dbRDA6    1  0.05293 1.0095  0.454
## dbRDA7    1  0.04750 0.9059  0.646
## dbRDA8    1  0.03941 0.7517  0.908
## dbRDA9    1  0.03775 0.7201  0.927
## dbRDA10   1  0.03280 0.6256  0.977
## dbRDA11   1  0.02876 0.5485  0.997
## dbRDA12   1  0.02501 0.4770  0.998
## Residual 39  2.04482
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

```
##
## ***VECTORS
##
##             dbRDA1   dbRDA2     r2 Pr(>r)
## Elevation  0.77670  0.62986 0.0959  0.082 .
## Diameter  -0.27972 -0.96008 0.0541  0.271
## Depth     -0.63137  0.77548 0.1756  0.006 **
## ORP        0.41879 -0.90808 0.1437  0.025 *
## Temp      -0.98250  0.18628 0.1523  0.016 *
## SpC       -0.77101  0.63682 0.2087  0.005 **
## DO        -0.39318 -0.91946 0.0464  0.314
## pH        -0.96210 -0.27270 0.1756  0.014 *
## Color      0.06353  0.99798 0.0464  0.307
## chla      -0.60392 -0.79704 0.2626  0.010 **
## DOC        0.99847 -0.05526 0.0382  0.356
## DON       -0.91633  0.40042 0.0339  0.426
## ---
## 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

# 3. plot the dbRDA results
# Define plot parameters
par(mar = c(5, 5, 4, 4) + 0.1)
# Initiate plot
plot(scores(ponds.dbrda, display = "wa"),
     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)

# Add points & labels
points(scores(ponds.dbrda, display = "wa"),
pch = 19, cex = 3, bg = "gray", col = "gray")
```

```
text(scores(ponds.dbrda, display = "wa"),
labels = row.names(scores(ponds.dbrda, display = "wa")), cex = 0.5)
# Add environmental vectors
vectors = scores(ponds.dbrda, display = "bp")
arrows(0, 0, vectors[,1] * 2, vectors[, 2] * 2,
lwd = 2, lty = 1, length = 0.2, col = "red")
text(vectors[,1] * 2, vectors[, 2] * 2, pos = 3, labels = row.names(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])))
axis(side = 4, lwd.ticks = 2, cex.axis = 1.2, las = 1, col = "red", lwd = 2.2, at = pretty(range(vector
```



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

> ***Answer 6***: Variation in the relatedness of bacterial taxa within pond communities can be explained in part by environmental variables. The following environmental variables significantly influence the phylogenetic diversity of bacterial communities in Indiana ponds: depth, oxidation reduction potential, temperature, specific conductivity of water, pH and chlorophyll a content. More similar bacterial species are found at similar depths, pH and temperature ranges - likely due to similarities in physiology/ pH and temperature tolerance among species.

6) SPATIAL PHYLOGENETIC COMMUNITY ECOLOGY

A. Phylogenetic Distance-Decay (PDD)

First, calculate distances for geographic data, taxonomic data, and phylogenetic data among all unique pair-wise combinations of ponds.

In the R code chunk below, do the following:
1. calculate the geographic distances among ponds,
2. calculate the taxonomic similarity among ponds,
3. calculate the phylogenetic similarity among ponds, and
4. create a dataframe that includes all of the above information.

```
# 1. calculate the geographic distances among ponds
long.lat = as.matrix(cbind(env$long, env$lat))
coord.dist = earth.dist(long.lat, dist = TRUE)

# 2. calculate the taxonomic similarity among ponds
bray.curtis.dist = 1 - vegdist(comm)

# 3. calculate the phylogenetic similarity among ponds
unifrac.dist = 1 - dist.uf
# Transform all distances into list format:
unifrac.dist.ls = liste(unifrac.dist, entry = "unifrac")
bray.curtis.dist.ls = liste(bray.curtis.dist, entry = "bray.curtis")
coord.dist.ls = liste(coord.dist, entry = "geo.dist")
env.dist.ls = liste(env.dist, entry = "env.dist")

# 4. create a dataframe that includes all of the above information
df = data.frame(coord.dist.ls, bray.curtis.dist.ls[, 3], unifrac.dist.ls[, 3], env.dist.ls[, 3])
names(df)[4:6] = c("bray.curtis", "unifrac", "env.dist")
```

Now, let's plot the DD relationships:
In the R code chunk below, do the following:
1. plot the taxonomic distance decay relationship,
2. plot the phylogenetic distance decay relationship, and
3. add trend lines to each.

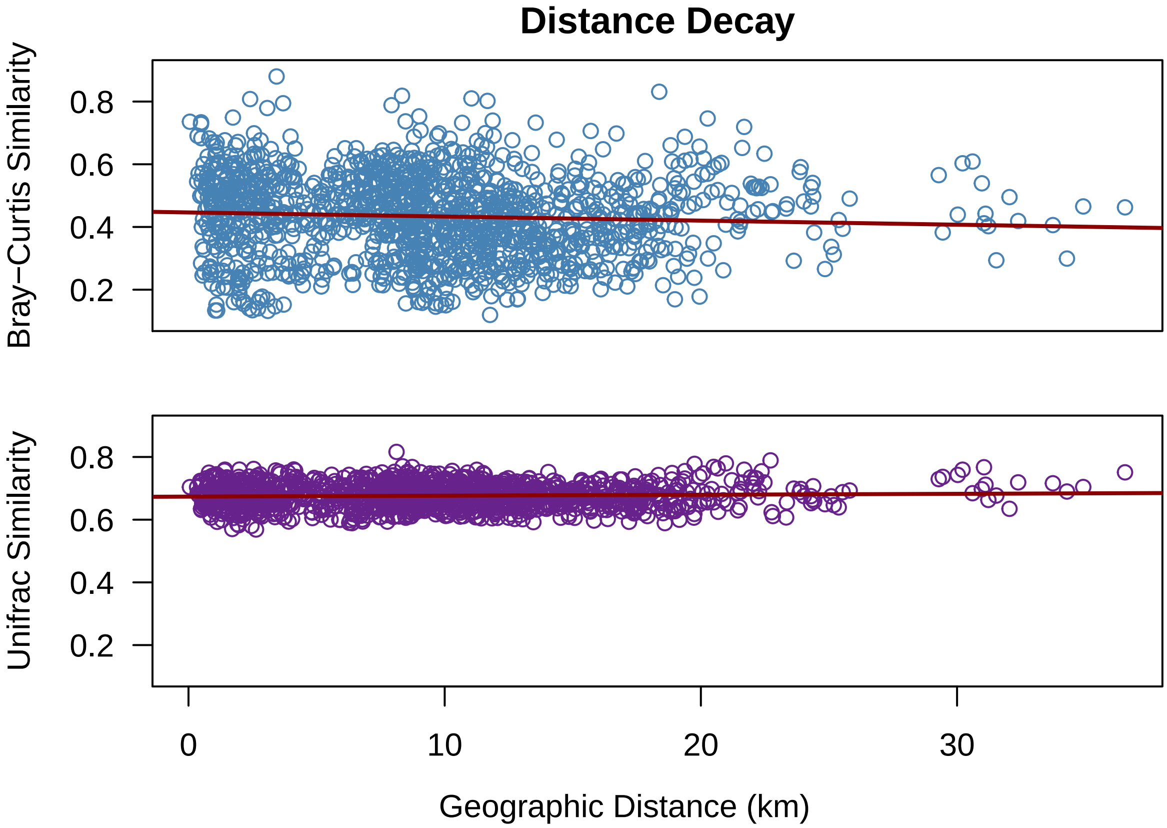
```
# 1. plot the taxonomic distance decay relationship
# Set initial plot parameters
par(mfrow=c(2, 1), mar = c(1, 5, 2, 1) + 0.1, oma = c(2, 0, 0, 0))
# Make plot for taxonomic DD
plot(df$geo.dist, df$bray.curtis, xlab = "", xaxt = "n", las = 1, ylim = c(0.1, 0.9), ylab="Bray-Curtis
# Regression for taxonomic DD
DD.reg.bc = lm(df$bray.curtis ~ df$geo.dist)
summary(DD.reg.bc)
```

```
##
## Call:
## lm(formula = df$bray.curtis ~ df$geo.dist)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.31151 -0.08843  0.00315  0.09121  0.43817
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.4463453  0.0066883  66.735   <2e-16 ***
## df$geo.dist -0.0013051  0.0005864  -2.226   0.0262 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1303 on 1324 degrees of freedom
## Multiple R-squared:  0.003728,   Adjusted R-squared:  0.002975
## F-statistic: 4.954 on 1 and 1324 DF,  p-value: 0.0262
```

```
abline(DD.reg.bc , col = "red4", lwd = 2)
# New plot parameters
par(mar = c(2, 5, 1, 1) + 0.1)

# 2. plot the phylogenetic distance decay relationship
plot(df$geo.dist, df$unifrac, xlab = "", las = 1, ylim = c(0.1, 0.9), ylab = "Unifrac Similarity", col =

# 3. add trend lines to each
DD.reg.uni = lm(df$unifrac ~ df$geo.dist)
summary(DD.reg.uni)
```

```
##
## Call:
## lm(formula = df$unifrac ~ df$geo.dist)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.105629 -0.027107 -0.000077  0.026761  0.140215
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.6735186  0.0019206 350.677   <2e-16 ***
## df$geo.dist 0.0002976  0.0001684   1.767   0.0774 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03741 on 1324 degrees of freedom
## Multiple R-squared:  0.002354,   Adjusted R-squared:  0.0016
## F-statistic: 3.124 on 1 and 1324 DF,  p-value: 0.07738
```

```
abline(DD.reg.uni, col = "red4", lwd = 2)
mtext("Geographic Distance (km)", side = 1, adj = 0.55, line = 0.5, outer = TRUE)
```



In the R code chunk below, test if the trend lines in the above distance decay relationships are different from one another.

```
diffslope(df$geo.dist, df$unifrac, df$geo.dist, df$bray.curtis)
```

```
##
## Is difference in slope significant?
## Significance is based on 1000 permutations
##
## Call:
## diffslope(x1 = df$geo.dist, y1 = df$unifrac, x2 = df$geo.dist,      y2 = df$bray.curtis)
##
## Difference in Slope: 0.001603
## Significance: 0.004
##
## Empirical upper confidence limits of r:
##      90%      95%    97.5%      99%
## 0.000704 0.000914 0.001087 0.001350
```

Question 7: Interpret the slopes from the taxonomic and phylogenetic DD relationships. If there are differences, hypothesize why this might be.

> ***Answer 7***: The slopes significantly differed (P = 0.007) in both sign and magnitude for the taxonomic and phylogenetic DD relationships. Taxonomic DD relationship (slope = -0.0013, R2=0.003). Phylogenetic DD relationship (slope = 0.0003, R2=0.002). The slopes differed by 0.001603. The taxonomic DD relationship reflects spatial autocorrelation in that communities

that are closer to one another in space should be more similar than those that are further from one another. Phylogenetic diversity does not seem to be spatially autocorrelated- resulting in a difference in the slopes.

B. Phylogenetic diversity-area relationship (PDAR)

i. Constructing the PDAR

In the R code chunk below, write a function to generate the PDAR.

```r
PDAR = function(comm, tree){
areas = c()
diversity = c()
num.plots = c(2, 4, 8, 16, 32, 51)
  for (i in num.plots){
areas.iter = c()
diversity.iter = c()
for (j in 1:10){
pond.sample = sample(51, replace = FALSE, size = i)
area = 0
sites = c()
for (k in pond.sample) {
  area = area + pond.areas[k]
  sites = rbind(sites, comm[k, ])
}
areas.iter = c(areas.iter, area)

# Calculate PSV or other phylogenetic alpha-diversity metric
psv.vals = psv(sites, tree, compute.var = FALSE)
psv = psv.vals$PSVs[1]
diversity.iter = c(diversity.iter, as.numeric(psv)) }
diversity = c(diversity, mean(diversity.iter))
areas = c(areas, mean(areas.iter))
print(c(i, mean(diversity.iter), mean(areas.iter)))
}
# Return vectors of areas (x) and diversity (y)
return(cbind(areas, diversity)) }
```

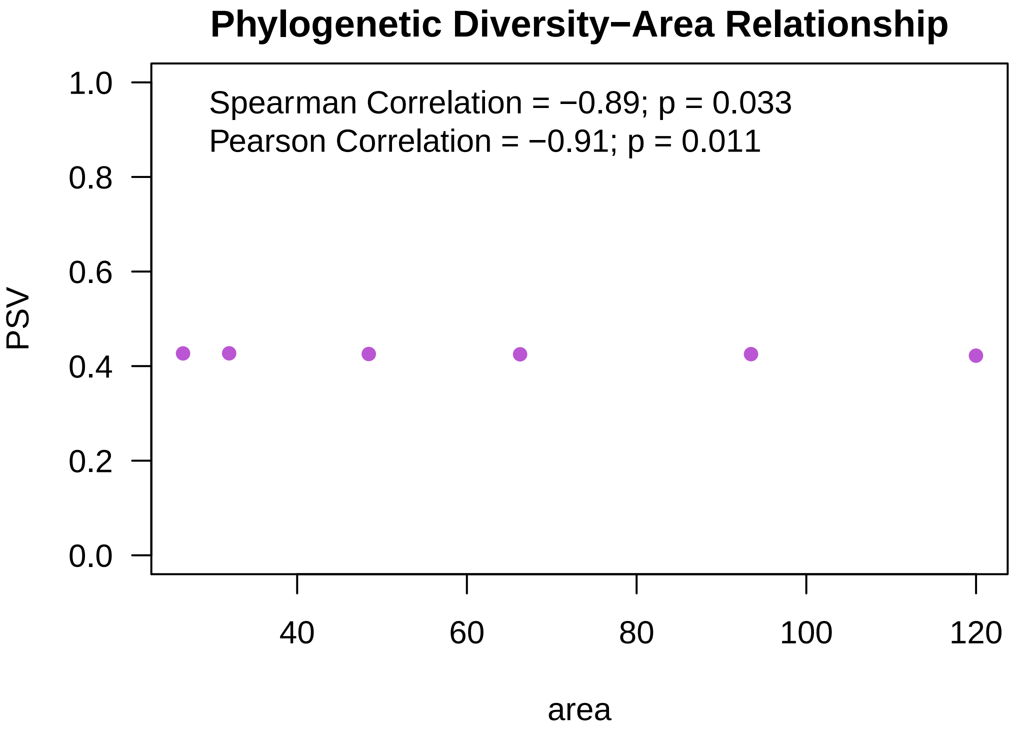
ii. Evaluating the PDAR

In the R code chunk below, do the following:
1. calculate the area for each pond,
2. use the `PDAR()` function you just created to calculate the PDAR for each pond,
3. calculate the Pearson's and Spearman's correlation coefficients,
4. plot the PDAR and include the correlation coefficients in the legend, and
5. customize the PDAR plot.

```r
# 1. calculate the area for each pond
pond.areas = as.vector(pi * (env$Diameter/2)^2)

# 2. use the `PDAR()` function you just created to calculate the PDAR for each pond
pdar = PDAR(comm, phy)
```

```
## [1]   2.0000000   0.4268573 705.1108361
## [1]    4.0000000    0.4270825 1023.1138401
## [1]    8.0000000    0.4256234 2346.5835729
## [1]   16.0000000    0.4249472 4391.7863085
## [1]   32.0000000    0.4253336 8738.8229858
## [1] 5.100000e+01 4.221028e-01 1.439763e+04
```

```
pdar = as.data.frame(pdar)
pdar$areas = sqrt(pdar$areas)
# Calculate Pearson's correlation coefficient
Pearson = cor.test(pdar$areas, pdar$diversity, method = "pearson")
P = round(Pearson$estimate, 2)
P.pval = round(Pearson$p.value, 3)

# 3. calculate the Pearson's and Spearman's correlation coefficients
Spearman = cor.test(pdar$areas, pdar$diversity, method = "spearman")
rho = round(Spearman$estimate, 2)
rho.pval = round(Spearman$p.value, 3)

# 4. plot the PDAR and include the correlation coefficients in the legend & customize the PDAR plot
plot.new()
par(mfrow=c(1, 1), mar = c(5, 5, 2, 5) + 0.1, oma = c(2, 0, 0, 0))
plot(pdar[, 1], pdar[, 2], ylab = "PSV", xlab ="area",  ylim = c(0, 1), main = "Phylogenetic Diversity-A
legend("topleft", legend= c(paste("Spearman Correlation = ", rho, "; p = ", rho.pval, sep = ""),paste("P
```



Question 8: Compare your observations of the microbial PDAR and SAR in the Indiana ponds? How might you explain the differences between the taxonomic (SAR) and phylogenetic (PDAR)?

> ***Answer 8***: The SAR shows that species are being discovered at a rate of 0.144 with increasing area. Unlike species discovery, phylogenetic diversity does not change with increasing area (P>0.1). It is difficult to identify the ecological mechanisms that give rise to SAR- soely with taxonomic

information. If there is a relationship between phylogenetic diveristy- this could be caused by repulsion, environmental filtering or competition. SAR can be caused by niche, random or neutral processes. Whereas, PDAR can be caused by macroevolutionary processes including speciation, adaptive radiation and extinction. Unlike PDAR, SAR can never have a negaitve slope.

SYNTHESIS

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 Synthesis: Recently I have become interested in the functional traits of soil fungi and how they influence ecosystem processes like decomposition. While there are number of papers that call for the study of fungal traits and organize the traits into various frameworks, fungi are hard to culture in the lab and difficult to measure in the field. Phylogeny can be used as a proxy for functional traits which are difficult to measure. Certain fungal traits including mycelial exploration type (rhizomorphic vs diffuse) and melanization, differ among fungal guilds (saprotrophs, ectomycorrhizal and arbuscular mycorrhizal fungi). Phylogenetic information could be useful in informing fungal community structure and mapping on traits that could be correlated with soil processes, such as carbon sequestration. I am also interested in how fungal community structure varies across time and space (different depths in the soil). Other questions related to the importance of niche processes in structuring fungal communities could be addressed with the phylogenetic approaches discussed in this lesson- including looking to see whether species traits are overdispersed or clustered.