Phylogenetic Diversity - Communities

Savannah Bennett; Z620: Quantitative Biodiversity, Indiana University

28 February, 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.

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

```
## [1] "C:/Users/Savannah/GitHub/QB2017_Bennett/Week7-PhyloCom"
```

```
setwd("C:/Users/Savannah/GitHub/QB2017_Bennett/Week7-PhyloCom/")
```

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.

```
require(seqinr)
```

```
## Loading required package: seqinr
```

```
require(picante)
```

```
## Loading required package: picante
## Loading required package: ape
##
## Attaching package: 'ape'
## The following objects are masked from 'package:seqinr':
##
##     as.alignment, consensus
## Loading required package: vegan
## Loading required package: permute
##
## Attaching package: 'permute'
## The following object is masked from 'package:seqinr':
##
##     getType
## Loading required package: lattice
## This is vegan 2.4-2
## Loading required package: nlme
##
## Attaching package: 'nlme'
## The following object is masked from 'package:seqinr':
##
##     gls
```

```
require(ape)
require(vegan)
require(fossil)
```

```
## Loading required package: fossil
## Loading required package: sp
## Loading required package: maps
## Loading required package: shapefiles
## Loading required package: foreign
##
## Attaching package: 'shapefiles'
## The following objects are masked from 'package:foreign':
##
##     read.dbf, write.dbf
```

```
require(simba)
```

```
## Loading required package: simba
## This is simba 0.3-5
##
## Attaching package: 'simba'
## The following object is masked from 'package:picante':
##
##     mpd
## The following object is masked from 'package:stats':
##
##     mad
```

```
require(ade4)
```

```
## Loading required package: ade4
##
## Attaching package: 'ade4'
## The following object is masked from 'package:vegan':
##
##     cca
```

```
require(gplots)
```

```
## Loading required package: gplots
##
## Attaching package: 'gplots'
## The following object is masked from 'package:stats':
##
##     lowess
```

```
require(indicspecies)
```

```
## Loading required package: indicspecies
```

```
require(BiodiversityR)
```

```
## Loading required package: BiodiversityR
## Loading required package: tcltk
```

```
## BiodiversityR 2.8-0: Use command BiodiversityRGUI() to launch the Graphical User Interface and to lea
```

```
#Load source code
source("./bin/MothurTools.R")
```

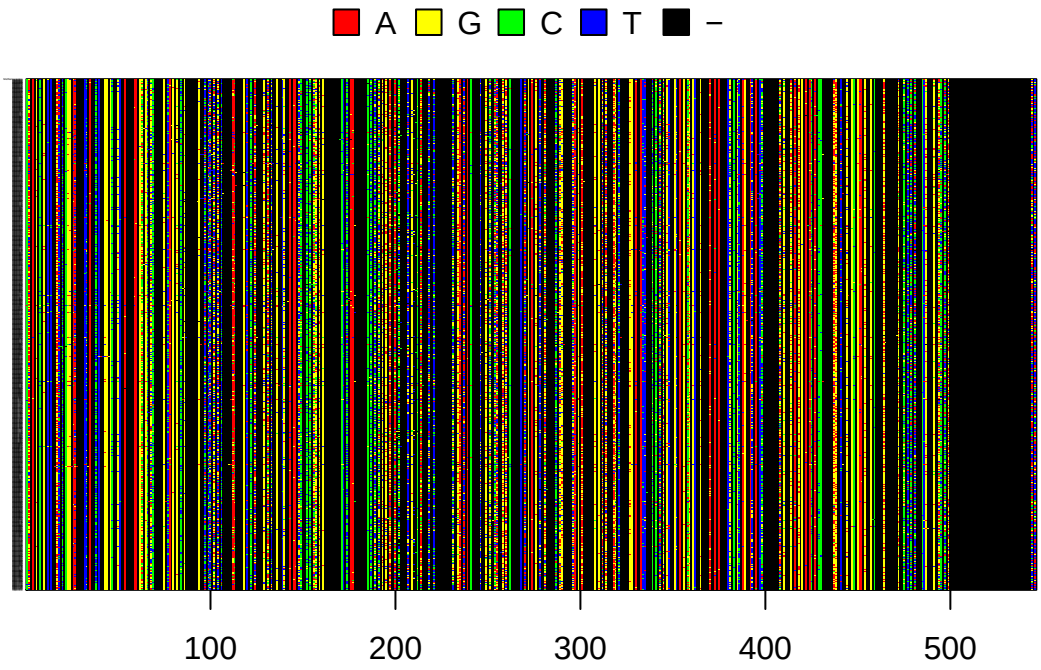
```
## Loading required package: reshape
```

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


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

#Select DNA 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]

#Import taxonomic data with 'read.tax()' function
tax <- read.tax(taxonomy = "./data/INPonds.final.rdp.1.cons.taxonomy")
```

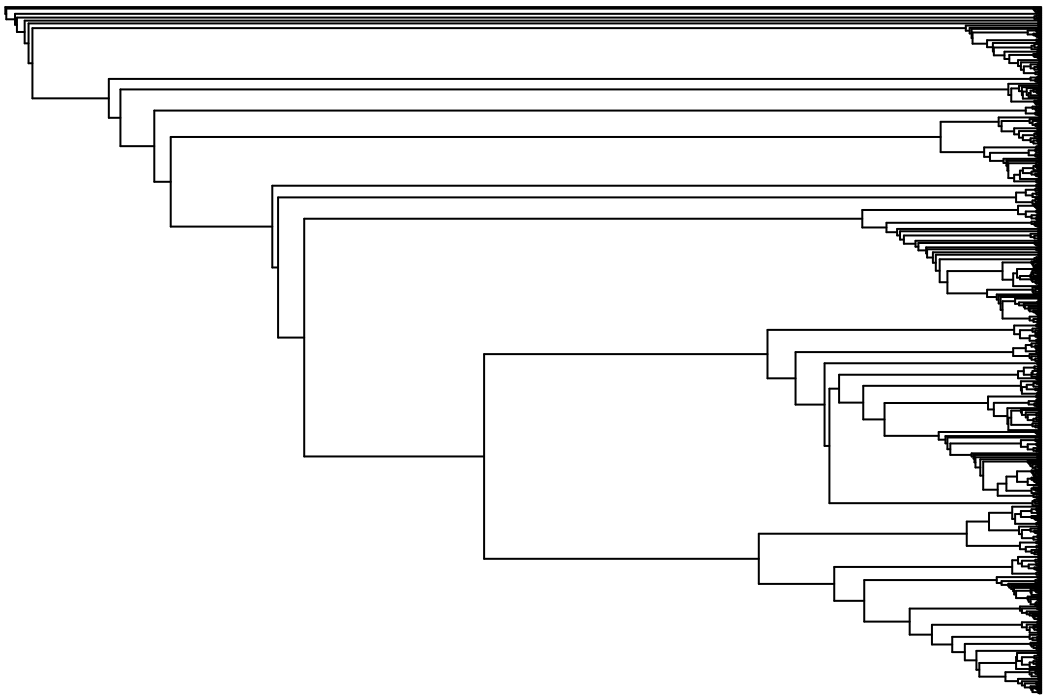
Next, in the R code chunk below, do the following:
1. load the FASTA alignment for the bacterial operational taxonomic units (OTUs),
2. rename the OTUs by removing everything before the tab (\t) and after the bar (|),
3. import the *Methanosarcina* outgroup FASTA file,
4. convert both FASTA files into 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.

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

#Rename OTUs in the FASTA file
ponds.cons$nam <- gsub("\\|.*$", "", gsub("^.*?\t", "", 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 tips of zero occurence 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 =
```



4) PHYLOGENETIC ALPHA DIVERSITY

A. Faith's Phylogenetic Diversity (PD)

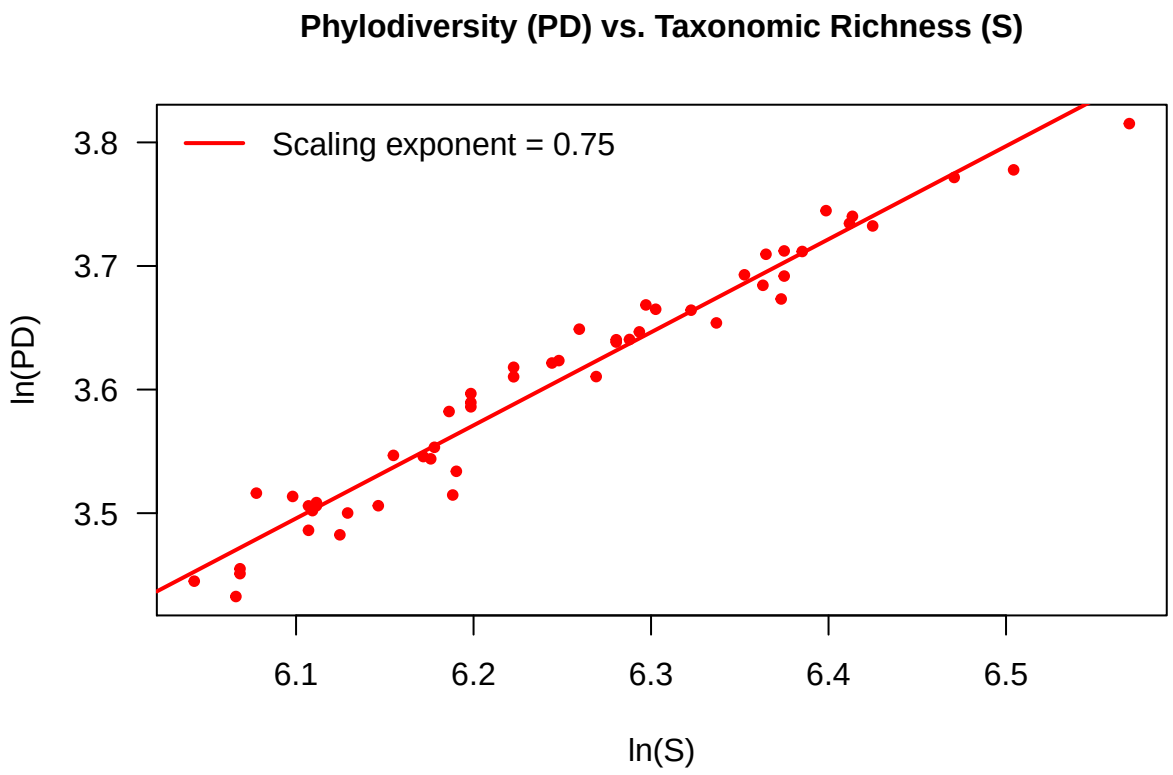
In the R code chunk below, do the following:
1. calculate Faith's D using the `pd()` function.

```
#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, 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***: This metric should relate to taxonomic richness because PD is calculated by adding the branch lengths in a sample, so with higher richness PD would be expected to increase as well.
>
> ***Answer 1b***: As taxonomic richness increases, phylodiversity increases as well. There is a positive linear relationship between these two variables.
>
> ***Answer 1c***: These two estimates might deviate from one another if the species being examined are very phylogenetically distant or have a restricted evolutionary history.
>
> ***Answer 1d***: The scaling exponent is 0.75, which suggests that there is a relatively strong relationship between richness and phylodiversity.

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 fo PD via randomization ('picante')
ses.pd <- ses.pd(comm[1:2,], phy, null.model = "richness", runs = 25,
                 include.root = FALSE)
```

```
#Run 'ses.pd()' function with taxa.labels
ses.pd1 <- ses.pd(comm[1:2,], phy, null.model = "taxa.labels", runs = 25,
                  include.root = FALSE)

#Run 'ses.pd()' function with frequency
ses.pd2 <- ses.pd(comm[1:2,], phy, null.model = "frequency", 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 alternative hypothesis would be that the patterns observed differ from the null expectation. The taxa.labels null model shuffles taxa labels across the tips of the phylogeny. The richness null model maintains sample species richness, while the frequency null model maintains species frequencies.

Answer 2b: The PD values were the same with each null model, and this could relate to the number of randomizations performed with each model. A large number of randomizations could yield differences in the PD values, and if a large number of randomizations was not performed, then the values could be the same.

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.

```
#Create phylogenetic resemblance matrix
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 standardized 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"
```

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)

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

```
#Rerun code so that NRI and NTI are calculated with abundance data

#Estimate standardized effect size of NRI via randomization ('picante')
ses.mpd1 <- ses.mpd(comm, phydist, null.model = "taxa.labels",
                    abundance.weighted = TRUE, runs = 25)

#Calculate NRI
NRI1 <- as.matrix(-1 * ((ses.mpd1[,2] - ses.mpd1[,3]) / ses.mpd1[,4]))
rownames(NRI) <- row.names(ses.mpd1)
colnames(NRI) <- "NRI"

#Estimate standardized effect size of NRI via randomization {picante}
ses.mntd1 <- ses.mntd(comm, phydist, null.model = "taxa.labels",
                      abundance.weighted = TRUE, runs = 25)

#Calcualte NTI
NTI1 <- as.matrix(-1 * ((ses.mntd1[,2] - ses.mntd1[,3]) / ses.mntd1[,4]))
rownames(NTI) <- row.names(ses.mntd1)
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: NRI is a method used to test whether a sample is clustered or overdispersed, and it uses the mean phylogenetic distance to make this determination. More specifically, it compares the observed mean phylogenetic distance values to average mean phylogenetic distance values from a null model.

Answer 3b: NTI is another analysis to test for phylogenetic clustering and overdispersion. Unlike NRI, NTI uses the average phylogenetic distance between all taxa in a particular sample and the closest related neighbor.

Answer 3c: The NRI values ranged from -4.55 to -0.78. All of the NRI values were negative, which suggests that the sample is overdispersed. In other words, the taxa are less related to one another than in the null model. The NTI values, on the other hand, ranged from -1.92 to 1.02. The majority of the NTI values were negative, which suggests overdispersion, but some were positive, which suggests clustering.

Answer 3d: The NRI values become very low negative and very low positive numbers, whereas when abundance.weighted = FALSE, they were all negative numbers. This may change the interpretation of the analysis from being overdispersed to being clustered. The NTI values all become positive when they are calculated using abundance data. Therefore, when abundance data is used, the data would be interpreted as clustered instead of overdispersed.

5) PHYLOGENETIC BETA DIVERSITY

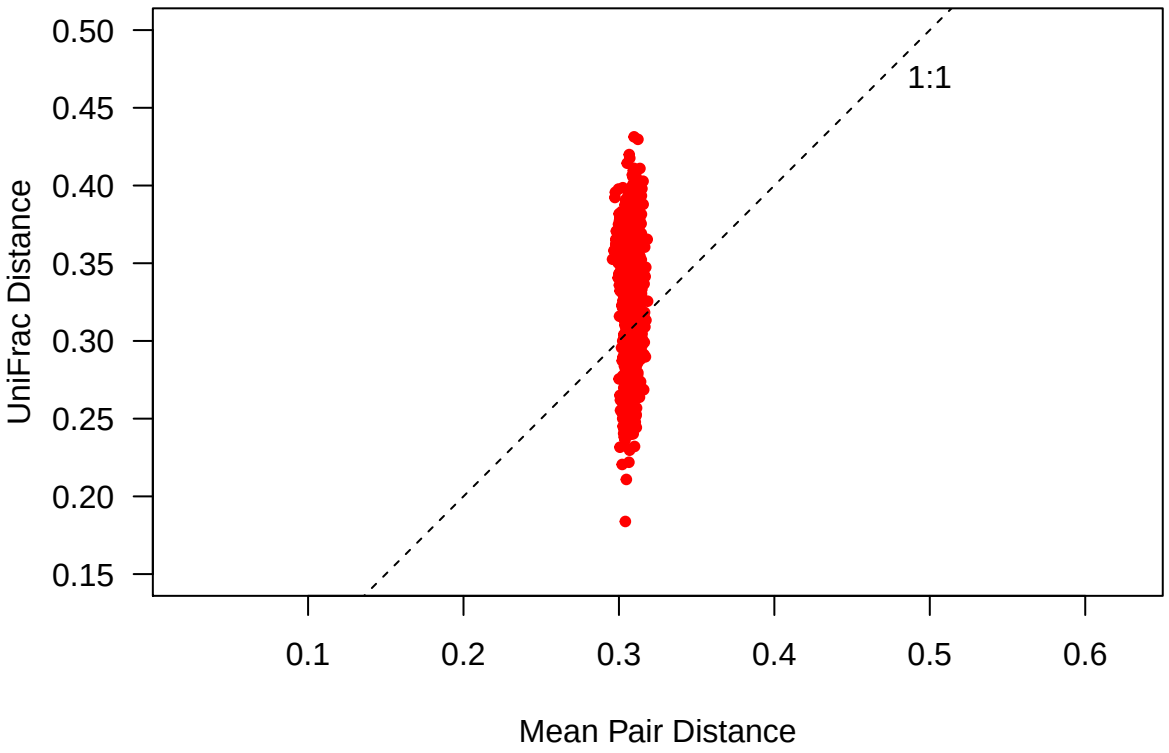
A. Phylogenetically Based Community Resemblance Matrix

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)
```

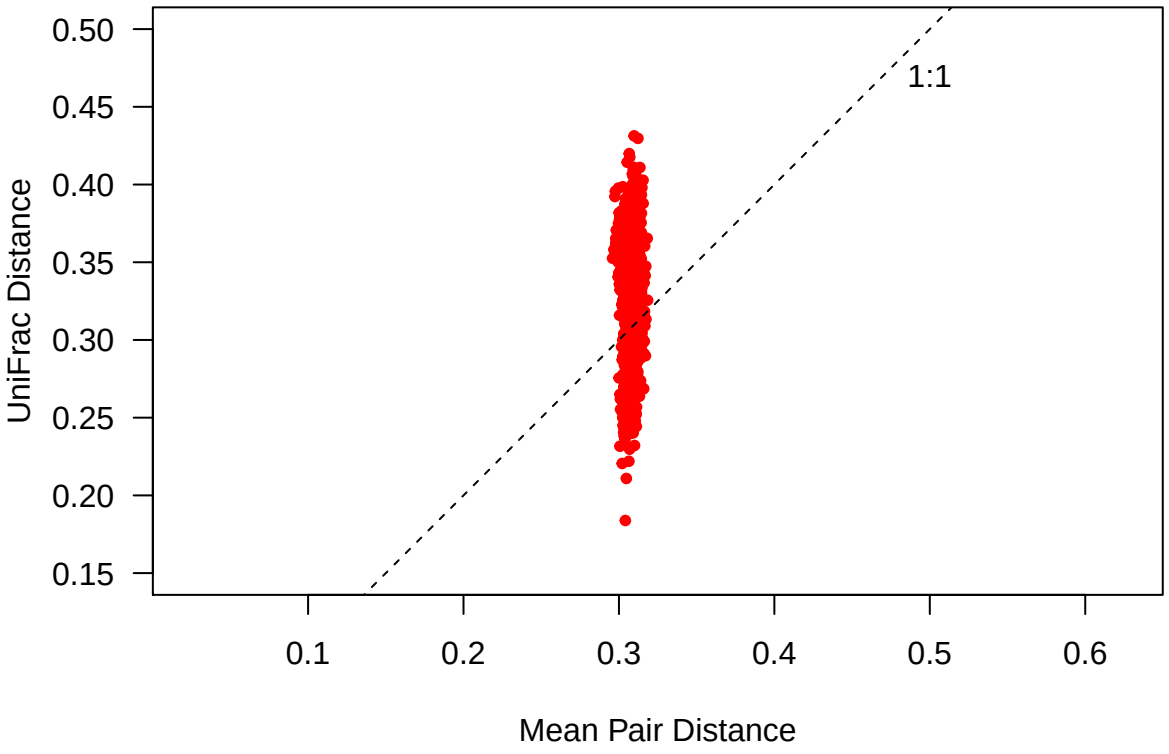
```
## [1] "Dropping taxa from the distance matrix because they are not present in the community data:"
## [1] "Methanosarcina"
```

```
#UniFrac distance
dist.uf <- unifrac(comm, phy)

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")
```



In the R code chunk below, do the following:
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),
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 and UniFrac distance are both measures of calculating phylogenetic distances using a community resemblance matrix. Mean pairwise distance calculates the average phylogenetic distance between two taxa. UniFrac is measured using the sum of unshared branch lengths and the total number of shared and unshared branch lengths. In other words, the two indices are calculated differently, but they both explore phylogenetic distance.

Answer 4b: Mean pair distance is consistently around 0.3, while UniFrac distance varies from around 0.2 to about 0.45. Therefore, the plot depicts a horizontal line because there is only a small range of mean pair distance values.

Answer 4c: Mean pair distance might show less variation than UniFrac because it is only looking at phylogenetic distance between pairs of taxa, whereas UniFrac is looking at the total number of shared and unshared branches, as well as the number of unshared branches, which can be more

variable.

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.

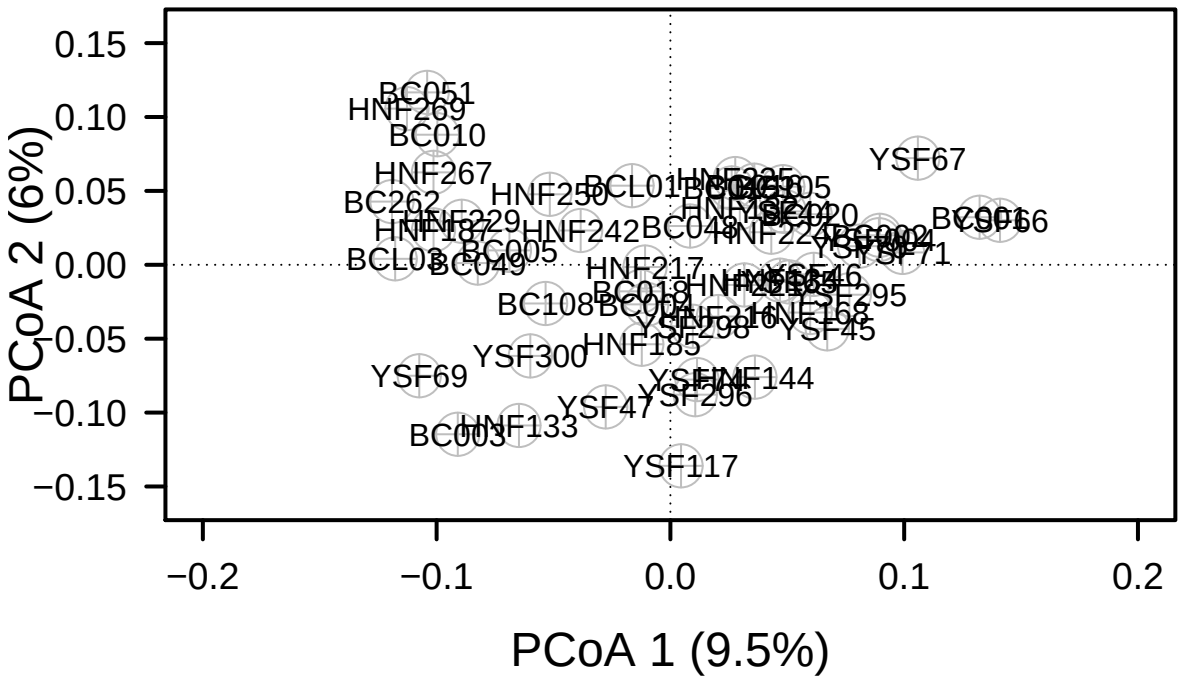
In the R code chunk below, do the following:
1. perform a PCoA based on the UniFrac distances, and
2. calculate the explained variation for the first three PCoA axes.

```
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)

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

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.

```
#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 = 10, cex = 3, bg = "gray", col = "gray")
text(pond.pcoa$points[ ,1], pond.pcoa$points[ ,2],
     labels = row.names(pond.pcoa$points))
```



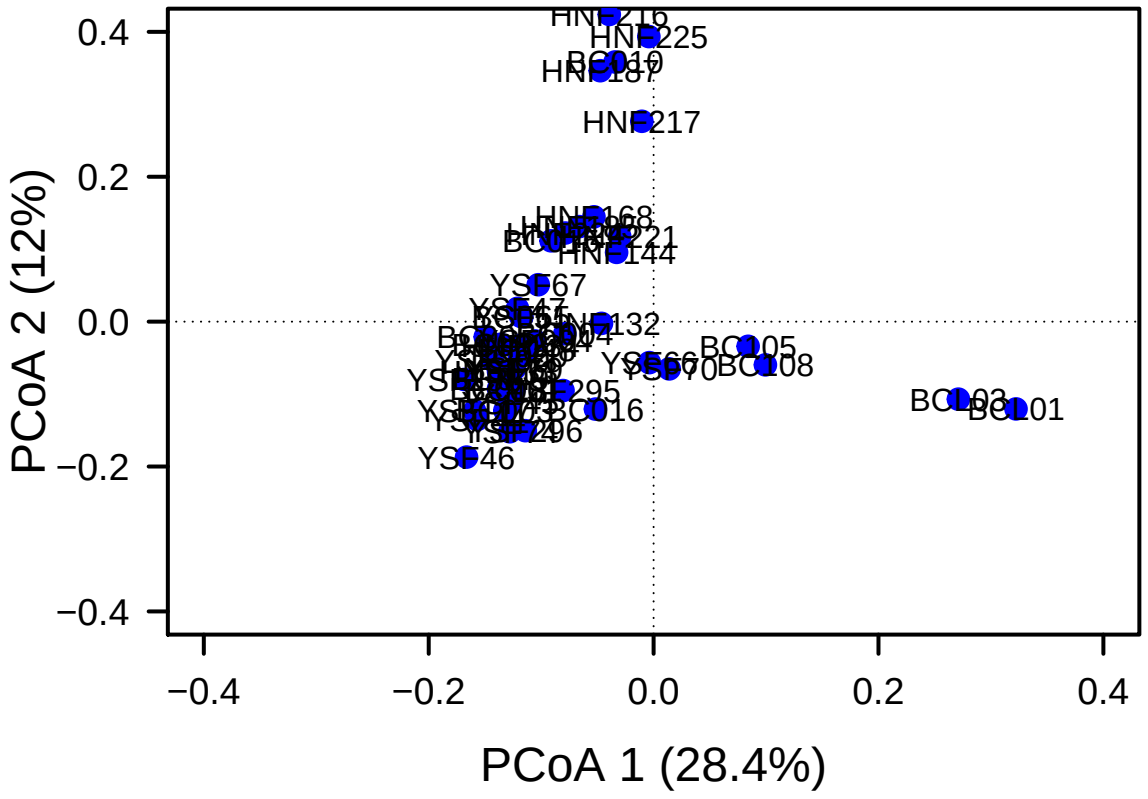
In the following R code chunk: 1. perform another PCoA on taxonomic data using an appropriate measure of dissimilarity, and 2. calculate the explained variation on the first three PCoA axes.

```
#Bray Curtis
ponds.db <- vegdist(comm, method = "bray")

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

explainvar1.db <- round(ponds.pcoa.db$eig[1] / sum(ponds.pcoa.db$eig), 3) * 100
explainvar2.db <- round(ponds.pcoa.db$eig[2] / sum(ponds.pcoa.db$eig), 3) * 100
explainvar3.db <- round(ponds.pcoa.db$eig[3] / sum(ponds.pcoa.db$eig), 3) * 100
33
```

```
## [1] 33
```

```
sum.eig.ponds.db <- sum(explainvar1.db, explainvar2.db, explainvar3.db)

#Ordination Plot (Bray-Curtis)
par(mar = c(5,5,1,2) + 0.1)

#Initiate Plot
plot(ponds.pcoa.db$points[ ,1], ponds.pcoa.db$points[ ,2],
xlab = paste("PCoA 1 (", explainvar1.db, "%)", sep = ""),
ylab = paste("PCoA 2 (", explainvar2.db, "%)", sep = ""),
pch = 16, cex = 2.0, type = "n", cex.lab = 1.5, cex.axis = 1.2, axes = FALSE, ylim = c(-.4,.4), xlim = c

#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(ponds.pcoa.db$points[,1], ponds.pcoa.db$points[,2],
pch = 19, cex = 1.5, bg = "gray", col = "blue")

text(ponds.pcoa.db$points[,1], ponds.pcoa.db$points[,2],
labels = row.names(ponds.pcoa.db$points))
```




```
ponds.db <- add.spec.scores(ponds.pcoa.db,comm,method = "pcoa.scores")
```

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 shows one main cluster of points in the center of the plot, and two additional clusters of points (one at the top of the plot, and a second towards the righthand corner). The phylogenetic based ordination only has one cluster of points in the center of the plot. These differences can tell you whether your clustering is related to phylogeny. It can show you whether there are clusters of closely related species on the plot.

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.

```
# 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  0.025 *
## 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( # create a distance matrix on
    decostand(comm, method = "log"), # log-transformed relative abundances
    method = "bray") ~ watershed, # using Bray-Curtis dissimilarity metric
  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  0.007 **
## 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()`).

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

```
# Remove redudnant variables
envs <- envs[, -which(names(envs) %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: 0.1604
##       Significance: 0.063
##
## Upper quantiles of permutations (null model):
##   90%   95% 97.5%   99%
## 0.135 0.169 0.199 0.234
## 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 environmental variables
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 + (
##          Df SumOfSqs      F Pr(>F)
## dbRDA1    1  0.10566 2.0152  0.005 **
## dbRDA2    1  0.09258 1.7658  0.002 **
## dbRDA3    1  0.07555 1.4409  0.031 *
## dbRDA4    1  0.06677 1.2735  0.100 .
## dbRDA5    1  0.05666 1.0807  0.317
## dbRDA6    1  0.05293 1.0095  0.462
## dbRDA7    1  0.04750 0.9059  0.617
## dbRDA8    1  0.03941 0.7517  0.899
```

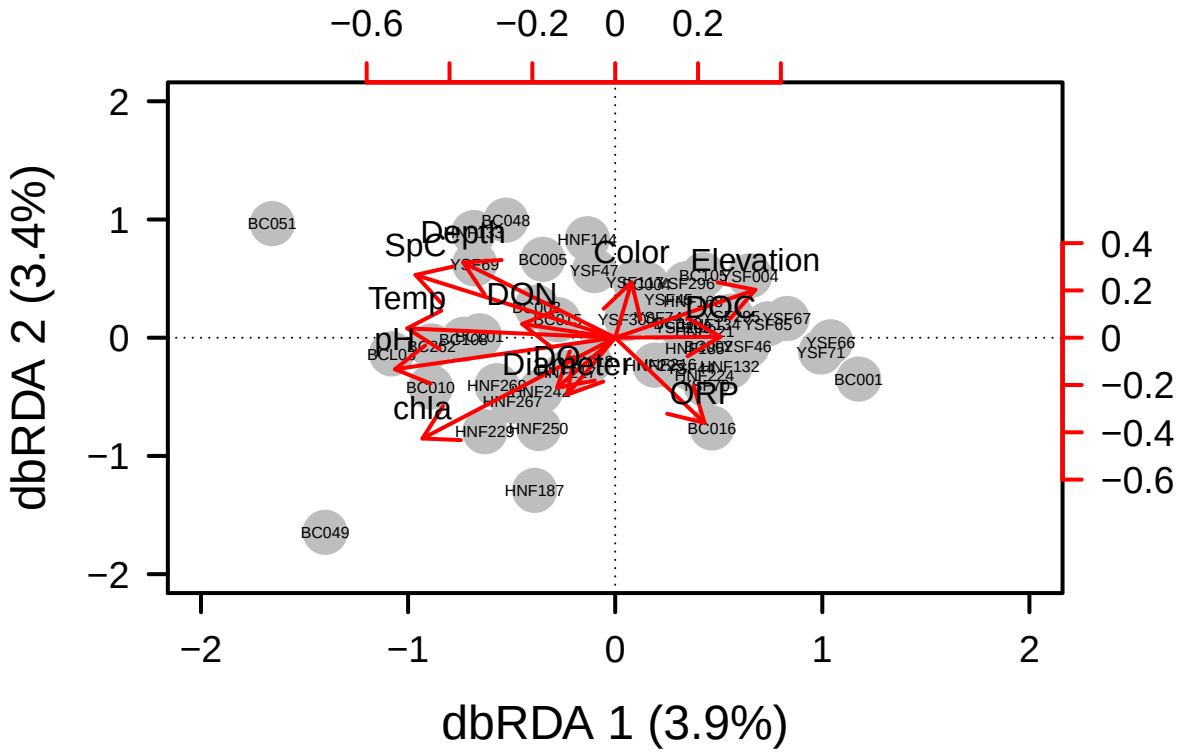
```
## dbRDA9    1  0.03775 0.7201  0.925
## dbRDA10   1  0.03280 0.6256  0.985
## dbRDA11   1  0.02876 0.5485  0.997
## dbRDA12   1  0.02501 0.4770  1.000
## Residual 39  2.04482
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

```
##
## ***VECTORS
##
##             dbRDA1   dbRDA2     r2 Pr(>r)
## Elevation  0.77670  0.62986 0.0959  0.090 .
## Diameter  -0.27972 -0.96008 0.0541  0.238
## Depth     -0.63137  0.77548 0.1756  0.013 *
## ORP        0.41879 -0.90808 0.1437  0.026 *
## Temp      -0.98250  0.18628 0.1523  0.022 *
## SpC       -0.77101  0.63682 0.2087  0.005 **
## DO        -0.39318 -0.91946 0.0464  0.300
## pH        -0.96210 -0.27270 0.1756  0.015 *
## Color      0.06353  0.99798 0.0464  0.314
## chla      -0.60392 -0.79704 0.2626  0.008 **
## DOC        0.99847 -0.05526 0.0382  0.375
## DON       -0.91633  0.40042 0.0339  0.408
## ---
## 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

# 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")
#row.names(vectors) <- c("Temp", "DO", "chla", "DON")
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(vectors[, 2])) * 2, labels = pretty(range(vectors[, 2])))
```



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***: The permanova indicates that watershed does have a significant effect on the phylogenetic diversity of bacterial communities. Environmental variables such as depth (p=0.012), oxidation reduction potential (p=0.021), temperature (p=0.016), specific conductivity of the water (p=0.004), pH (p=0.015), and chlorophyll a concentrations (p=0.012) significantly impacted phylogenetic bacterial community diversity. The dbRDA plot shows which taxonomic groups are more related/affected by each environmental variable.

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.

```
# Geographic distances (kilometers) among ponds
long.lat <- as.matrix(cbind(env$long, env$lat))
coord.dist <- earth.dist(long.lat, dist = TRUE)

# Taxonomic similarity among ponds (Bray-Curits distance)
bray.curtis.dist <- 1 - vegdist(comm)

# Phylogenetic similarity among ponds (UniFrac)
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")

# Create a data frame from the lists of distances
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.

```
# 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 Similarity",
     main = "Distance Decay", col = "SteelBlue")

# Regression for taxonomic DD
DD.reg.bc <- lm(df$bray.curtis ~ df$geo.dist)
summary(DD.reg.bc)
```

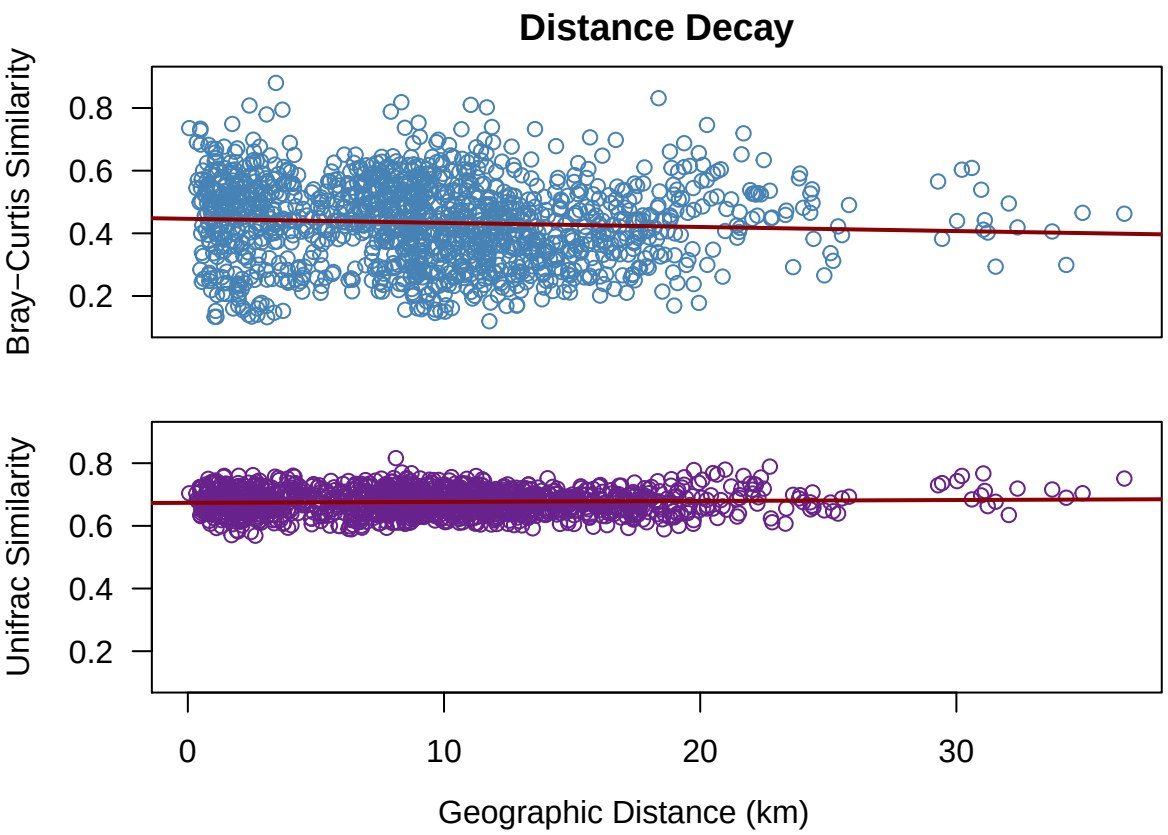
```
##
## Call:
## lm(formula = df$bray.curtis ~ df$geo.dist)
##
```

```
## 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)

# Make plot for phylogenetic DD
plot(df$geo.dist, df$unifrac, xlab = "", las = 1, ylim = c(0.1, 0.9),
     ylab = "Unifrac Similarity", col = "darkorchid4")

# Regression for phylogenetic DD
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)

# Add x-axis label to plot
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: 0.002
##
## Empirical upper confidence limits of r:
##      90%      95%    97.5%      99%
## 0.000735 0.000970 0.001131 0.001399
```

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 of the taxonomic and phylogenetic distance decay relationships are significantly different (p=0.001). If phylogenetic diversity is not spatially correlated, or is only slightly spatially correlated, then the slopes of taxonomic and phylogenetic DD relationships could differ.

B. Phylogenetic diversity-area relationship (PDAR)

i. Constructing the PDAR

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

```
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)
18
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(cbind(areas, diversity))
}
```

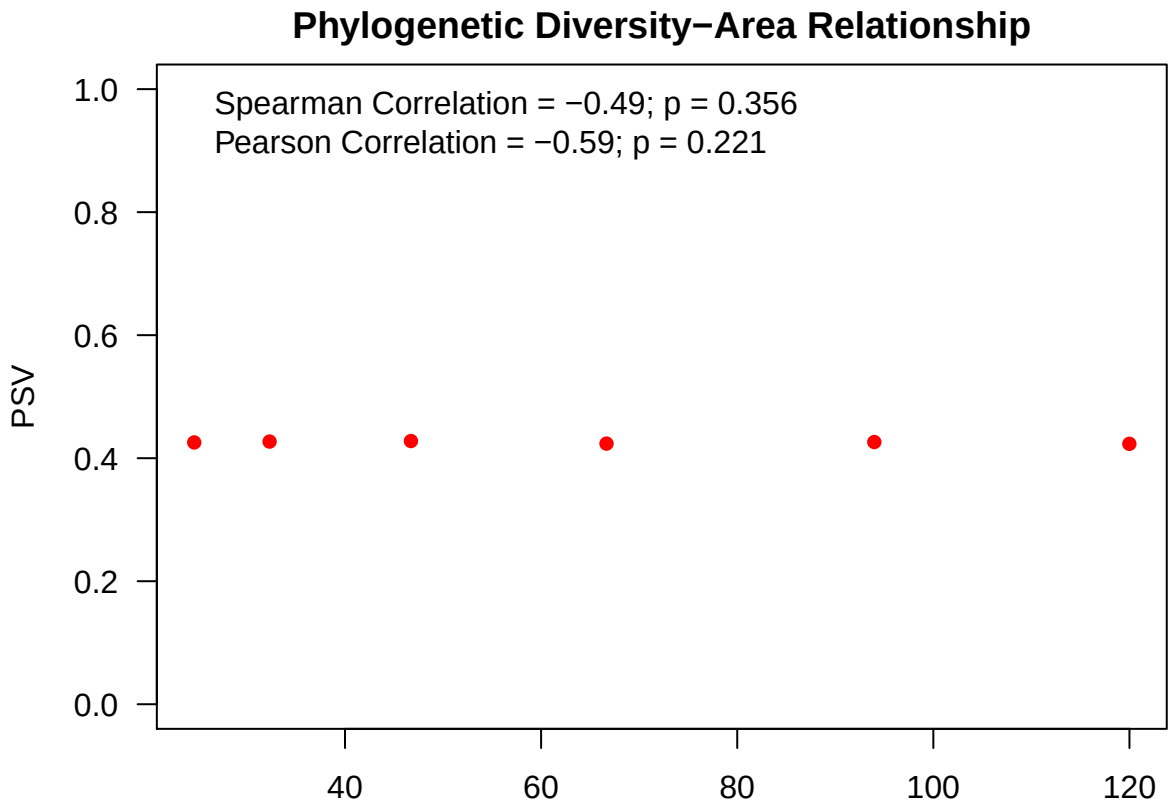
ii. Evaluating the PDAR

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.

```
# Calculate areas for ponds: find areas of all ponds
pond.areas <- as.vector(pi * (env$Diameter/2)^2)

# Compute the PDAR
pdar <- PDAR(comm, phy)
```

```
## [1]   2.0000000   0.4255957 606.3862870
## [1]    4.0000000    0.4270544 1043.8679866
## [1]    8.0000000    0.4278896 2183.1728456
## [1]   16.0000000    0.4236983 4444.7512044
## [1]   32.0000000    0.4262406 8831.9366505
## [1] 5.100000e+01 4.233449e-01 1.439763e+04
```

```r
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)

# Calculate Spearman's correlation coefficient
Spearman <- cor.test(pdar$areas, pdar$diversity, method = "spearman")
rho <- round(Spearman$estimate, 2)
rho.pval <- round(Spearman$p.value, 3)

# Plot the PDAR
plot.new()
par(mfrow=c(1, 1), mar = c(1, 5, 2, 1) + 0.1, oma = c(2, 0, 0, 0))
plot(pdar[, 1], pdar[, 2], xlab = "Area", ylab = "PSV", ylim = c(0, 1),
     main = "Phylogenetic Diversity-Area Relationship",
     col = "red", pch = 16, las = 1)

legend("topleft", legend= c(paste("Spearman Correlation = ", rho, "; p = ", rho.pval, sep = ""),
    paste("Pearson Correlation = ", P, "; p = ", P.pval, sep = "")),
    bty = "n", col = "red")
```



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 has a positive slope, which indicates that species richness increases as sample size/area increases. In other words, the number of different species that can be sampled increases with sampling area. The PDAR has a very low slope as most of the phylogenetic species variability values are around 0.4. These graphs indicate that even though richness increases linearly with area, phylogenetic diversity does not vary greatly with area. This could suggest that although there are many species present, they are not phylogenetically distant from one another.

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.

I am broadly interested in studying the role of herbivory in plant-soil feedbacks, and these two forces shape the relationships between native and invasive plants. When doing feedback experiments, I plan to sample the rhizosphere communities associated with my plants, as soil communities are the primary drivers of plant-soil feedbacks. I would like to compare the soil communities of the native and invasive plants that I am using to better understand the mechanisms behind plant-soil feedbacks in my study system. Phylogenetic information would be useful because it would help me understand these microbial communities. More specifically, I could use the analyses in these exercises to determine the extent to which phylogeny and environmental variables shape microbial community structure associated with different treatment groups (herbivory, no herbivory) and plant species/functional groups. I could also examine phylogenetic diversity within and among treatment groups and species.