Phylogenetic Diversity - Communities

Venus Kuo ; Z620: Quantitative Biodiversity, Indiana University

25 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.

```
# Set working directory #
rm(list = ls())
getwd()
```

```
## [1] "C:/Users/Venus/Github/QB2017_Kuo/Week7-PhyloCom"
```

```
setwd("C:/Users/Venus/Github/QB2017_Kuo/Week7-PhyloCom/")
#setwd("/Users/vkuo/GitHub/QB2017_Kuo/Week7-PhyloCom//")

# Require or install packages #
package.list <- c('picante', 'ape', 'seqinr', 'vegan', 'fossil', 'simba')
for (package in package.list) {
  if (!require(package, character.only=T, quietly=T)) {
    install.packages(package)
    library(package, character.only=T)
  } }
```

```
## This is vegan 2.4-1
```

```
##
## 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
```

```
## 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
```

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

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

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.

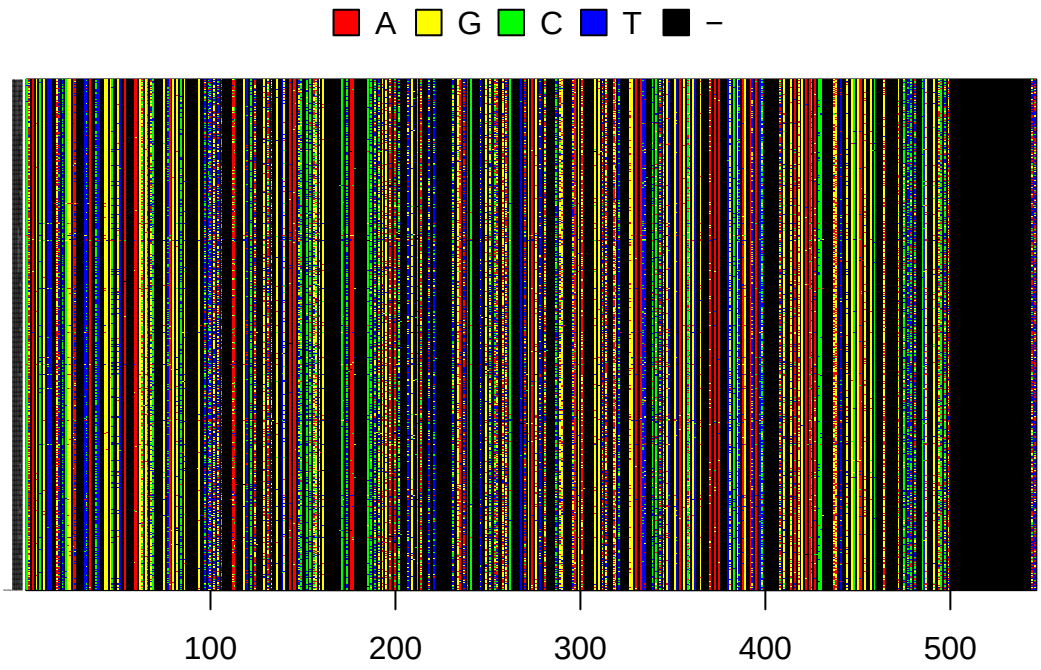
3) LOAD THE DATA

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.

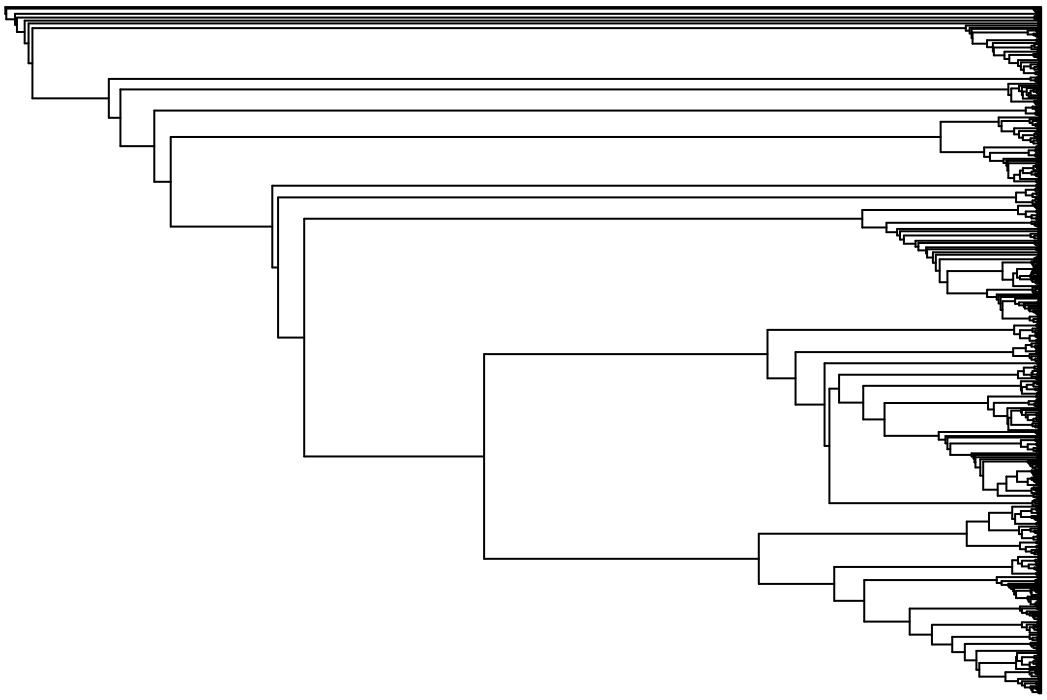
```
# Load Enviromental data for Brown County Ponds #
env <- read.table("data/20130801_PondDataMod.csv", sep = ",", header = TRUE)
env <- na.omit(env)
# Load site-by-species matrix #
comm <- read.otu(shared = "./data/INPonds.final.rdp.shared", cutoff = "1")
# Subset data to include only DNA-based identification #
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 unnecessary OTUs and sites in the site-by-species #
comm <- comm[rownames(comm) %in% env$Sample_ID, ]
# Remove zero abundance OTUs from data set #
comm <- comm[ , colSums(comm) > 0]
# Load the taxonomic data using the read.tax() #
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.

```
# Load the FASTA alignment for the bacterial operational taxonomic units #
ponds.cons <- read.alignment(file = "./data/INPonds.final.rdp.1.rep.fasta", format = "fasta")
# Rename the OTUs by removing everything before the tab and after the bar #
ponds.cons$nam <- gsub("\\|.*$", "", gsub("^.*?\t", "", ponds.cons$nam))
# Import the outgroup fasta file #
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)
```




```
# Use alignment with outgroup, pick DNA substitution model, and create a phylogenetic distance matrix #
seq.dist.jc <- dist.dna(DNAbin, model = "JC", pairwise.deletion = FALSE)
# Make a neighbor-joining tree file #
phy.all <- bionj(seq.dist.jc)
# Remove tips that are not in community data set #
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 #
phy <- root(phy, outgroup, resolve.root = TRUE)
# 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 =
```



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.

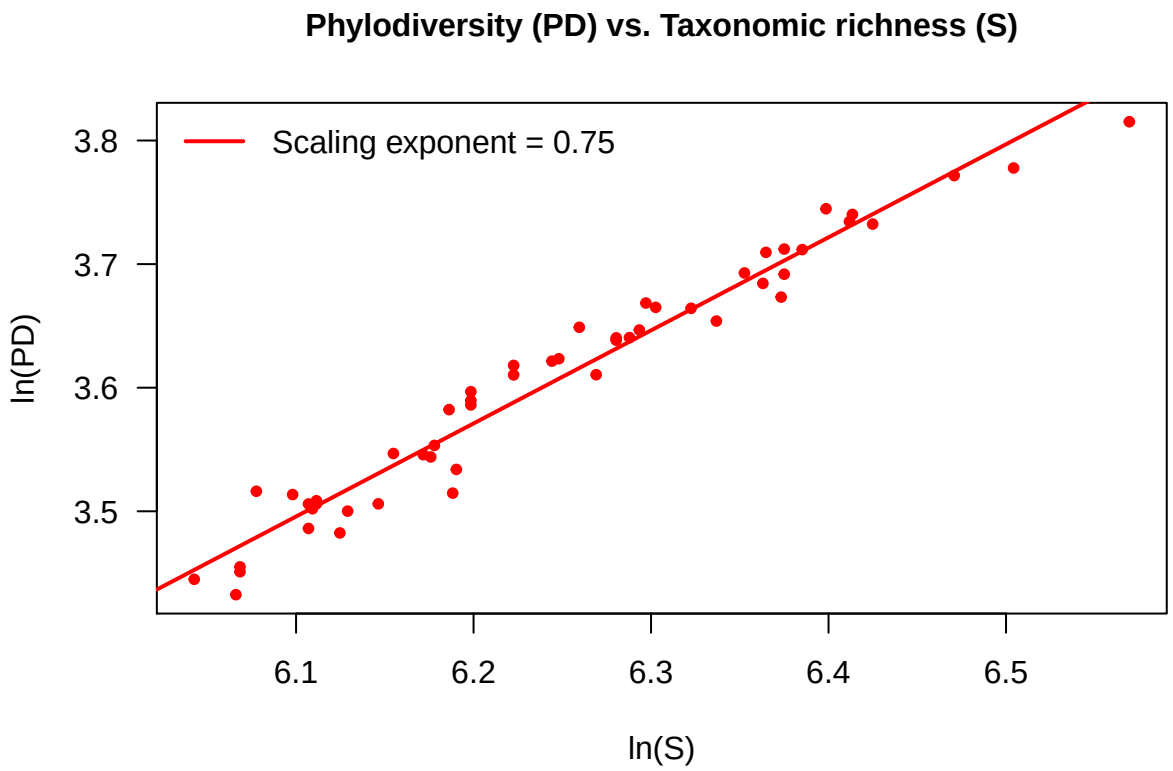
```
# Calculate Faith's D using pd() #
pd <- pd(comm, phy, include.root = FALSE)
```

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.

```
# Plot species richness versus phylogenetic diversity (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 and add trendline #
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***: Faith's diversity metric (PD) is calculated by summing the branch lengths of each species of a sample from the root to the tips of the phylogenetic tree. Basically, PD measures the total evolutionary distance between species of a sample by their corresponding tree branch lengths. PD should be related to taxonomic richness becuase higher PD corresponds to increased evolutionary distance between species of a sample, i.e. taxonomic richness, because more distantly related species indicate more richness. ***Answer 1b***: With increasing taxonomic richness, it is assumed that there is increased phylodiversity. The more distict phylogenetic groups you have, then very likely there is higher taxonomic richness. ***Answer 1c***: You would expect these two estimates of diversity to deviate from one another when there is overdispersion of species in each of the phylogenic groups. In that case, the assumptions of taxomic richness and phylodiversity falls apart. ***Answer 1d***: The PD-S scaling exponent shows that with increasing taxonomic species richness, there is a increasing evolutionarily divergent taxa.

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 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)
ses.pd.freq <- ses.pd(comm[1:2,], phy, null.model = "frequency", runs = 25,
                include.root = FALSE)
ses.pd.tax <- ses.pd(comm[1:2,], phy, null.model = "taxa.labels", runs = 25,
                include.root = FALSE)
ses.pd
```

```
##       ntaxa   pd.obs pd.rand.mean pd.rand.sd pd.obs.rank    pd.obs.z
## BC001   668 43.71912     43.76101  0.8096401          16 -0.05173276
## BC002   587 40.94334     40.12631  0.7654219          22  1.06741588
##        pd.obs.p runs
## BC001 0.6153846   25
## BC002 0.8461538   25
```

```
ses.pd.freq
```

```
##       ntaxa   pd.obs pd.rand.mean pd.rand.sd pd.obs.rank  pd.obs.z
## BC001   668 43.71912     42.16137  0.6849136          25  2.274378
## BC002   587 40.94334     42.42493  0.6848756           2 -2.163308
##         pd.obs.p runs
## BC001 0.96153846   25
## BC002 0.07692308   25
```

```
ses.pd.tax
```

```
##       ntaxa   pd.obs pd.rand.mean pd.rand.sd pd.obs.rank   pd.obs.z
## BC001   668 43.71912     44.20349  0.7007832           7 -0.6911816
## BC002   587 40.94334     39.97672  1.0459092          20  0.9241928
##        pd.obs.p runs
## BC001 0.2692308   25
## BC002 0.7692308   25
```

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: We are using the ses.pd function to generate a randomized null model to test if the phylogenetic diversity of a sample really is robust or if the patterns that we see fall apart with slight alterations of the tree or species abundance/richness. ***Answer 2b***: I chose richness, taxa.labels, and frequency null models to see if the phylogenetically diverse than expected with randomization of taxa labels across tips of phylogeny, community data matrix abundannces within samples, and abundances within species. This choice did not affect the output because the species in the sample are not very phylogenetically diverse, at lease not anymore than by random chance. So the sample species could be exhibiting overdispersion.

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.

```
# Calculate 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.

```
# Calculate the NRI for each site in the IN ponds data set #
# Estimate standardized effect size of NRI via randomization #
ses.mpd <- ses.mpd(comm, phydist, null.model = "taxa.labels",
                   abundance.weighted = TRUE, 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"
NRI
```

```
##                NRI
## BC001  -0.34834411
## BC002   0.15192781
## BC003   0.65909075
## BC004  -0.52503261
## BC005   0.16101800
## BC010   0.26262738
## BC015  -0.25828301
## BC016  -0.07348471
## BC018  -0.17167745
## BC020  -0.12289406
## BC048  -0.57803232
## BC049  -0.15048042
## BC051  -1.06447791
## BC105  -1.36695114
## BC108  -0.63625627
## BC262  -0.64789509
## BCL01  -0.86904147
## BCL03  -1.02784024
## HNF132 -0.59663139
## HNF133 -0.39857315
## HNF134 -0.01401136
## HNF144 -0.63222510
## HNF168 -0.29799456
## HNF185  0.18321197
## HNF187  0.46425429
## HNF216  0.51409460
## HNF217 -0.15612664
## HNF221 -0.36185509
## HNF224 -0.01585653
## HNF225  0.51091810
## HNF229 -0.59588922
## HNF242 -0.07173292
## HNF250 -0.59954544
## HNF267 -0.71843660
## HNF269 -0.49175660
## YSF004 -1.08528593
```

```
## YSF117  0.42305853
## YSF295 -1.18143373
## YSF296  0.82513997
## YSF298  0.80265423
## YSF300  0.12062848
## YSF44   0.12657736
## YSF45   0.42307025
## YSF46   1.41489633
## YSF47  -0.11916810
## YSF65   0.15200071
## YSF66  -0.60435553
## YSF67  -0.57274314
## YSF69  -0.59462665
## YSF70  -0.78454262
## YSF71   0.36191163
## YSF74   0.81124714
```

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 standardied effect size of NRI via randomization #
ses.mntd <- ses.mntd(comm, phydist, null.model = "taxa.labels",
                     abundance.weighted = TRUE, 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"
NTI
```

```
##               NTI
## BC001   0.8519856
## BC002   1.8860527
## BC003   0.7647258
## BC004   1.0512249
## BC005   1.6548758
## BC010   0.3269887
## BC015   0.9393324
## BC016   1.6009700
## BC018   1.4167481
## BC020   1.0620571
## BC048   1.0599943
## BC049   1.6833638
## BC051   2.0908109
## BC105   1.3539035
## BC108   1.5487913
## BC262   1.2394903
## BCL01   1.6937134
## BCL03   0.7807927
## HNF132  1.0738309
## HNF133  0.9911947
## HNF134  1.7387756
## HNF144  0.8926510
## HNF168  0.4365960
## HNF185  1.3344330
```

```
## HNF187  0.3900380
## HNF216  0.1836179
## HNF217  0.2648033
## HNF221  0.4923350
## HNF224  1.1981925
## HNF225  0.2824000
## HNF229  1.3936347
## HNF242  1.6630668
## HNF250  1.2394859
## HNF267  0.9619011
## HNF269  0.9959531
## YSF004  0.1299081
## YSF117  1.7873850
## YSF295 -1.2166356
## YSF296  1.9430704
## YSF298  1.7578870
## YSF300  1.4784052
## YSF44   0.7400580
## YSF45   0.8600285
## YSF46   1.1351381
## YSF47   0.7873021
## YSF65   1.3311145
## YSF66   1.0700575
## YSF67   1.2499872
## YSF69   1.0339028
## YSF70   0.9842460
## YSF71   0.9134274
## YSF74   0.9823376
```

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: For calculating the NRI, I am taking the average pairwise branch length between all species in a sample minus the random average pairwise distances between species over the standard deviation of the randomly generated model. ***Answer 3b***: For calculating the NTI, I am taking the sum of the minimum values of each taxon minus the randomly generated null model over the null model standard deviation. ***Answer 3c***: Most of the samples in both the NRI and NTI model are negative, indicating that sample species are more overdispersed than expected. ***Answer 3d***: Changing the abundance.weighted arguement to TRUE gave me mostly postive NTI and NRI values for the samples. It seems that species distribution is very uneven in the sample dataset, and perhaps species abundance is correlated with phylogenetic dispersion.

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.

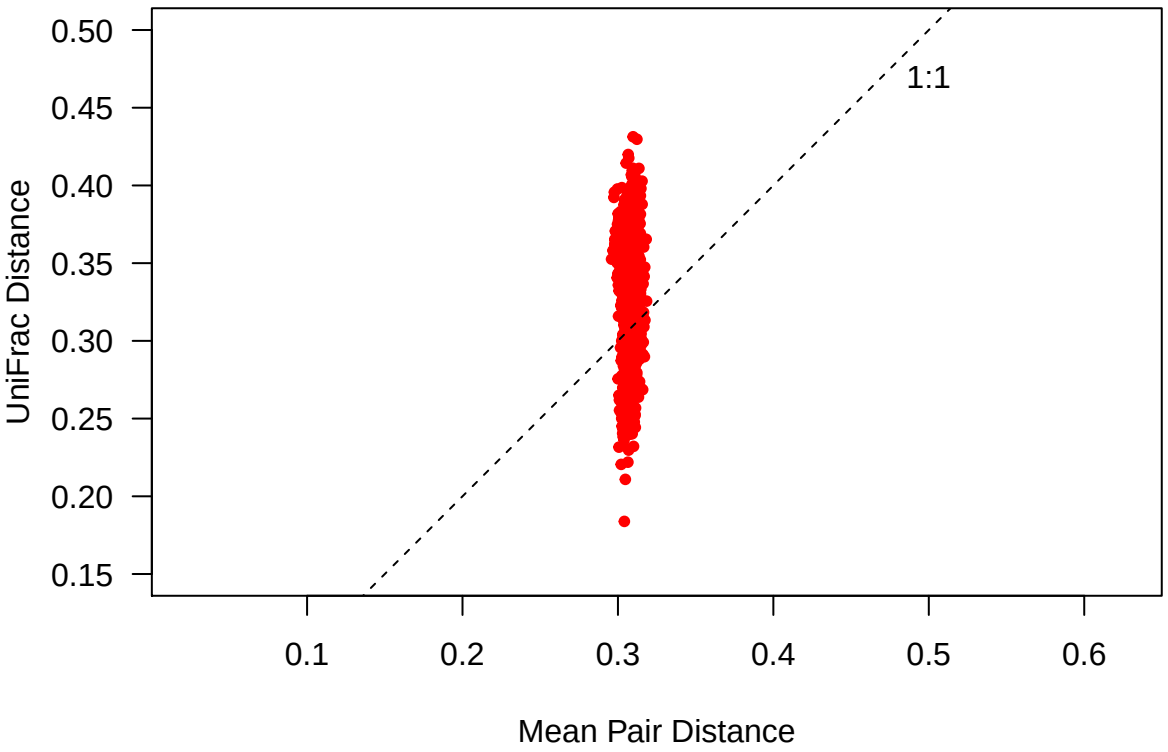
```
# Mean pairwise distance #
dist.mp <- comdist(comm, phydist)
```

```
## [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)
```

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

```
# Compare mean pair distance and unifrac distance matrices #
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 pair distance is the mean branch distance between two species on a phylogeny

tree. Unifrac distance is the sum of the unshared branch lengths divided by the total shared and unshared branch lengths in a rooted tree. It seems that Unifrac distance gives the proportion of dissimilarity between taxas while Mean pair distance gives mean distances between all species on a tree. ***Answer 4b***: Mean pair distance is 0.3 while unfrac distance varies from 0.2 to 0.45. There seems to be a large span of phylogenetic distance as indicated by the unifrac distance values. ***Answer 4c***: The mean pair distance value stays the same likely because we are using un weighted data so mean pair distance stays the same each time you calculate it.

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.

```
# Visualizing Phylogentic Beta-Diversity #
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
explainvar1
```

```
## [1] 9.5
```

```
explainvar2
```

```
## [1] 6
```

```
explainvar3
```

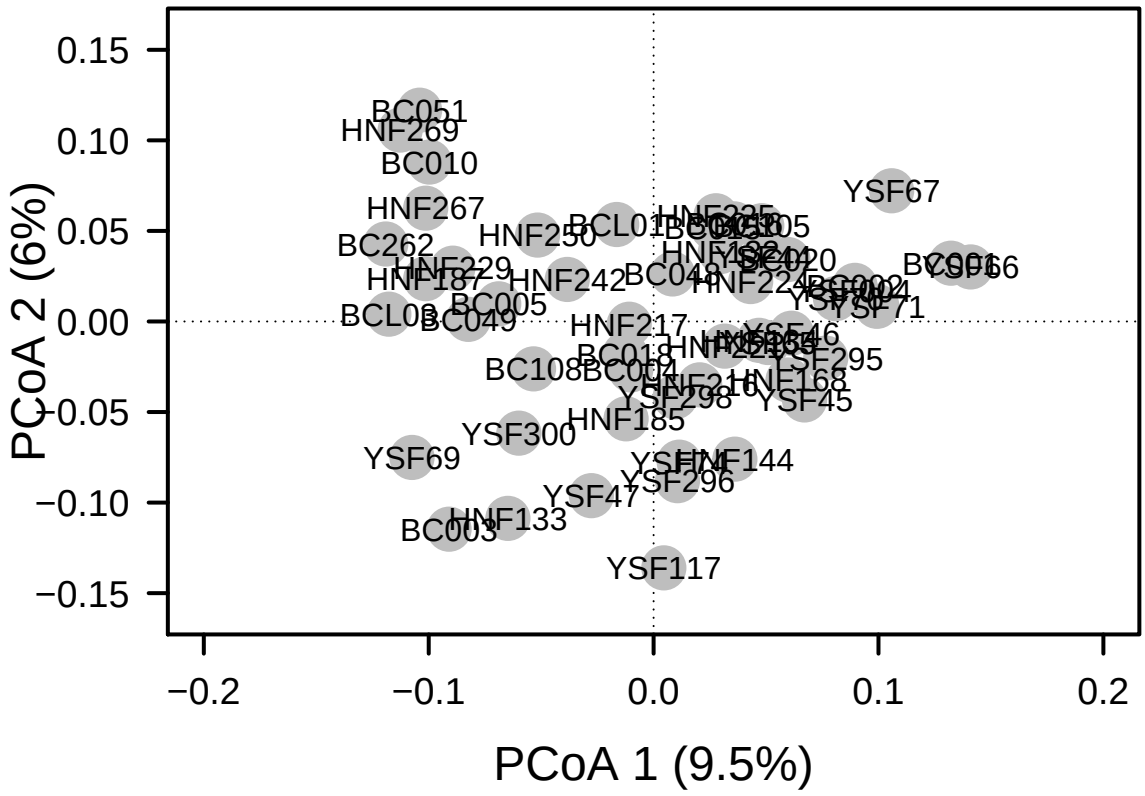
```
## [1] 5.4
```

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.

```
# Plot the PCoA results using plot function #
par(mar = c(5,5,1,2)+0.1)
# Initate plot #
plot(pond.pcoa$points[ ,1], pond.pcoa$points[ ,2],
     xlim = c(-0.2, 0.2), ylim = c(-.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.

```
pond.pcoa <- cmdscale(dist.mp, 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
explainvar1
```

```
## [1] 2.1
```

```
explainvar2
```

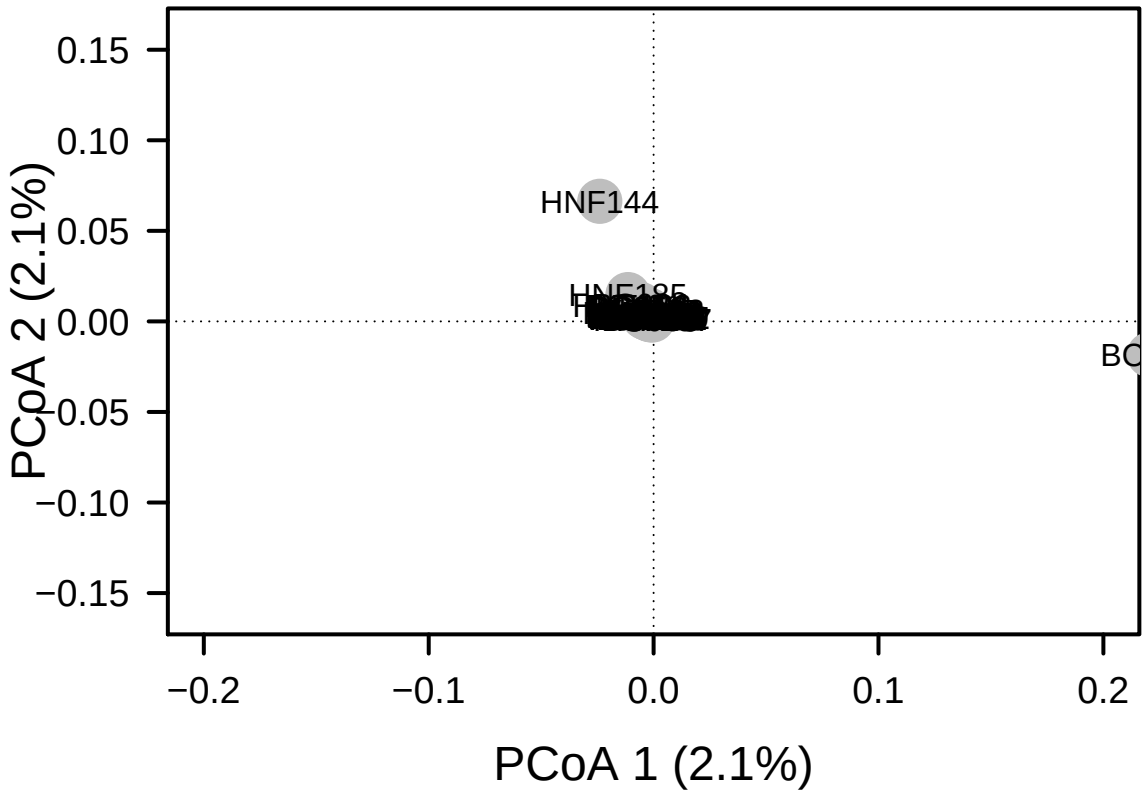
```
## [1] 2.1
```

```
explainvar3
```

```
## [1] 2.1
```

```
# Plot the PCoA results using plot function #
par(mar = c(5,5,1,2)+0.1)
# Initate plot #
plot(pond.pcoa$points[ ,1], pond.pcoa$points[ ,2],
     xlim = c(-0.2, 0.2), ylim = c(-.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))
```



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 phylogenetically based ordination (PCoA conducted with UniFrac metric) explains more of the variation in community assemblage than the taxonomic ordination (PCoA of the Mean Pairwise Distance). This implies that accounting for the phylogenetic information in this system is actually quite important, if you do not do then you would detect a signal when using these visualization tools.

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 enviromental category #
watershed <- env$Location
# Run PERMANOVA #
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  0.033 *
## 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
```

```
# Compare 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()`).

```
# Define enviromental variables #
envs <- env[, 5:19]
# Remove redudnant variables #
```

```
envs <- envs[, -which(names(envs) %in% c("TDS", "Salinity", "Cal_Volume"))]
# Create distance matrix for enviromental 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 #
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: 0.068
##
## Upper quantiles of permutations (null model):
##   90%   95% 97.5%   99%
## 0.134 0.176 0.209 0.253
## 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 #
ponds.dbrda <- vegan::dbrda(dist.uf ~ ., data = as.data.frame(scale(envs)))
# Permutation test #
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  0.003 **
## dbRDA3    1  0.07555 1.4409  0.033 *
## dbRDA4    1  0.06677 1.2735  0.096 .
## dbRDA5    1  0.05666 1.0807  0.312
## dbRDA6    1  0.05293 1.0095  0.439
## dbRDA7    1  0.04750 0.9059  0.663
## dbRDA8    1  0.03941 0.7517  0.909
## dbRDA9    1  0.03775 0.7201  0.938
## dbRDA10   1  0.03280 0.6256  0.991
## dbRDA11   1  0.02876 0.5485  0.993
```

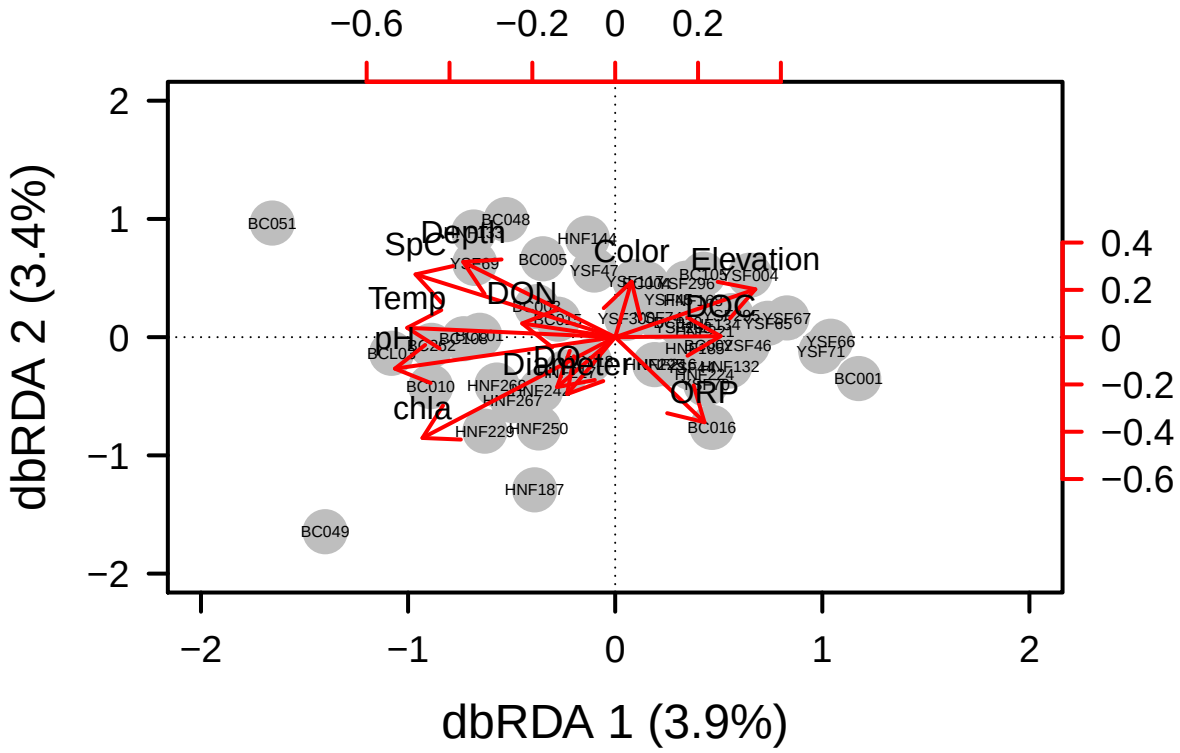
```
## dbRDA12   1  0.02501 0.4770  1.000
## Residual 39  2.04482
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

```
##
## ***VECTORS
##
##             dbRDA1   dbRDA2     r2 Pr(>r)
## Elevation  0.77670  0.62986 0.0959  0.081 .
## Diameter  -0.27972 -0.96008 0.0541  0.241
## Depth     -0.63137  0.77548 0.1756  0.010 **
## ORP        0.41879 -0.90808 0.1437  0.023 *
## Temp      -0.98250  0.18628 0.1523  0.017 *
## SpC       -0.77101  0.63682 0.2087  0.005 **
## DO        -0.39318 -0.91946 0.0464  0.294
## pH        -0.96210 -0.27270 0.1756  0.010 **
## Color      0.06353  0.99798 0.0464  0.337
## chla      -0.60392 -0.79704 0.2626  0.009 **
## DOC        0.99847 -0.05526 0.0382  0.353
## DON       -0.91633  0.40042 0.0339  0.449
## ---
## 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
# 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 and 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 =
# Add enviromental vector #
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
```

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



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 Mantel test suggests there maybe a significant correlation between UniFrac distance and enviromental variation. Further investigation of this correlation using the redundancy analysis reveals that certain enviromental variables such as lake depth, SpC, chla, and to a smaller extent ORP, Temp, and pH were correlated with the phylogenetic distance metrics. It seems that most of the beta diversity for bacterial communities in the Indiana ponds are driven by those particular variables.

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.

```
# Calculate geographic distances among ponds #
long.lat <- as.matrix(cbind(env$long, env$lat))
```

```
coord.dist <- earth.dist(long.lat, dist = TRUE)
# Taxonomic similarily among ponds #
bray.curtis.dist <- 1 - vegdist(comm)
# 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")
# Creat df that from 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.

```
# Setting inital 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
```

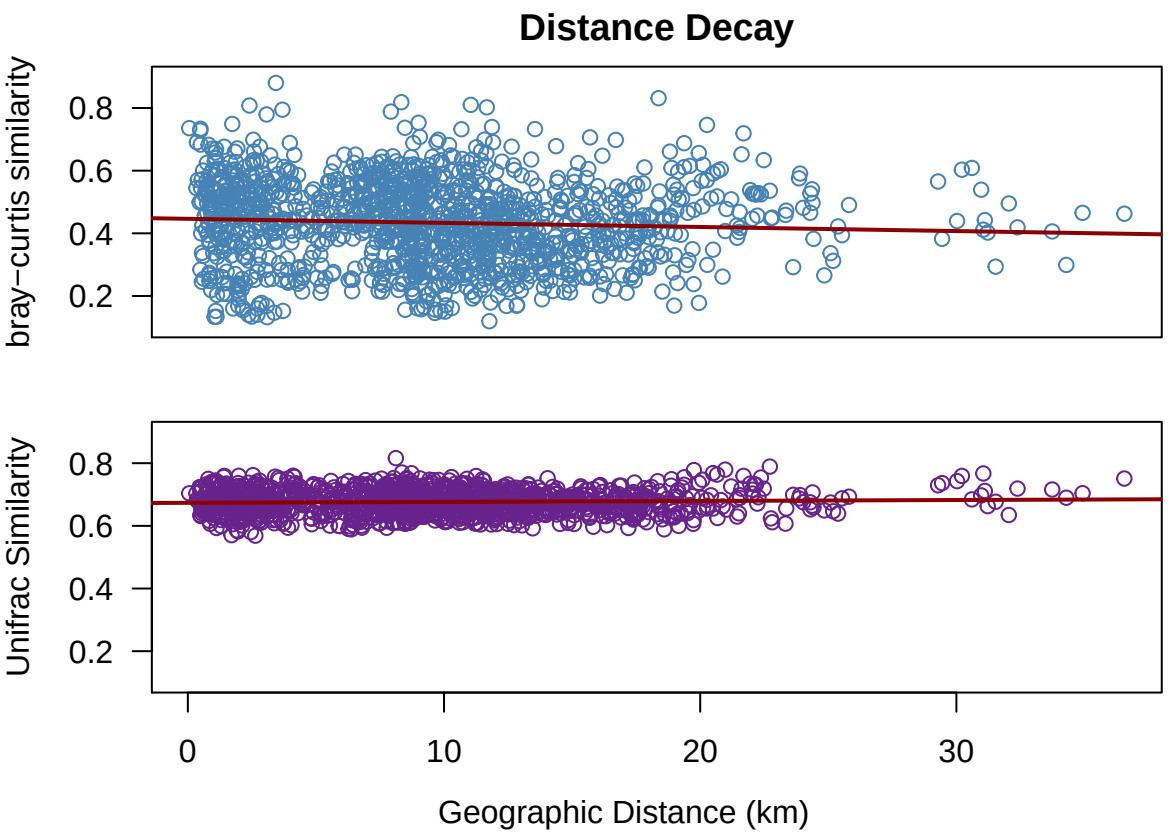
```
abline(DD.reg.bc, col="red4", lwd =2)
# new plot parameters #
par(mar = c(2,5,1,1)+0.1)
# Make plot for phylogenetic DD #
plot(df$geo.dist, df$unifrac, xlab = "", las =1, ylim =c(0.1,0.9),
     ylab = "Unifrac Similarity", col = "darkorchid4")
```

```r
# 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
```

```r
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: 0.01
##
## Empirical upper confidence limits of r:
##      90%      95%    97.5%      99%
## 0.000793 0.001011 0.001197 0.001566
```

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

> ***Answer 7***: From visualizing the taxonomic and phylogentic DD relationships appear to be relatively constant, meaning that communities remain similar to one another with increasing distance. There is a slight difference between the slop of the two DD plots (0.001603), but the two slopes are significantly different than other another likely because accounting for phylogeny results in less variation in the phylgenetic DD plot. This likely means that the bacterial communities of Indiana Lake are overdispersed or that dispersal is not limited.

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.area[k]
        sites <- rbind(sites, comm[k, ])
      }
      areas.iter <- c(areas.iter, area)
      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 the area for ponds #
pond.area <- as.vector(pi * (env$Diameter/2)^2)
# Compute the PDAR #
pdar <- PDAR(comm, phy)
```

```
## [1]   2.0000000   0.4259408 646.5397681
## [1]    4.0000000    0.4275414 1202.4697209
## [1]    8.0000000    0.4235243 2498.4497325
## [1]   16.0000000    0.4265162 4345.0551178
```

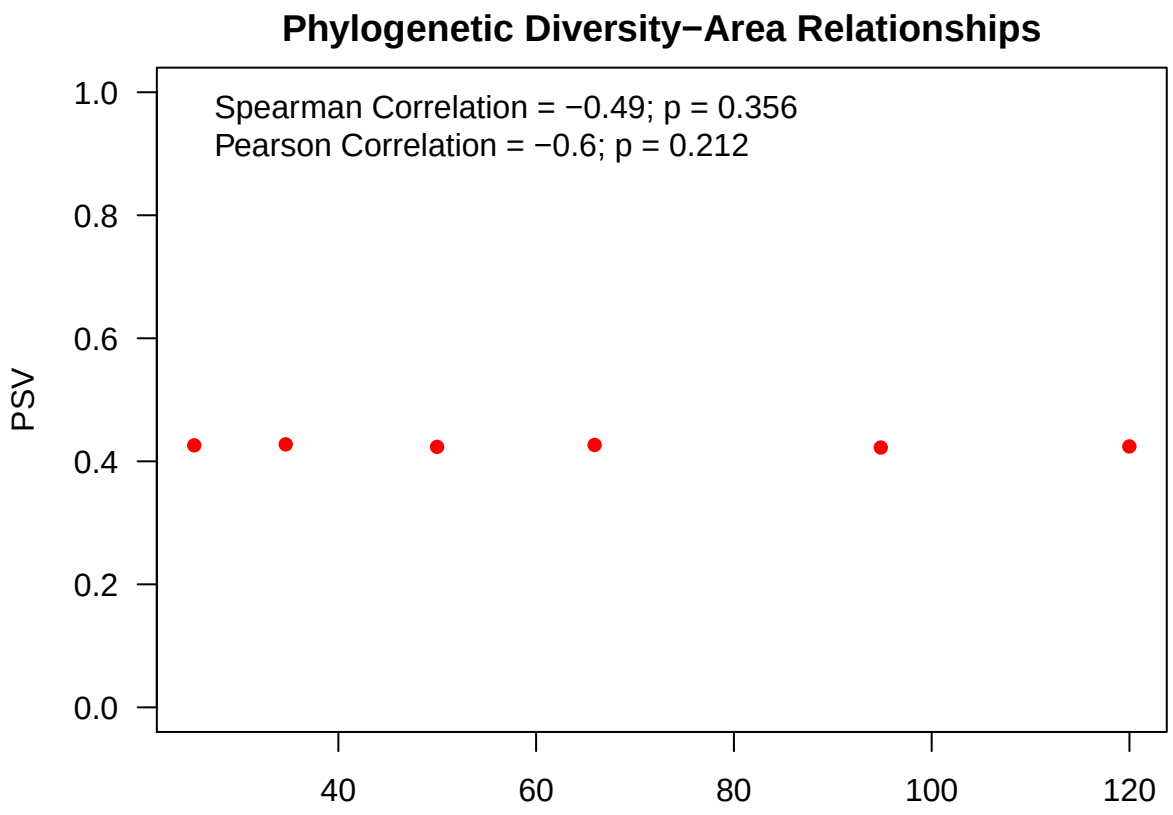
```
## [1]   32.0000000    0.4224047 8996.5223334
## [1] 5.100000e+01 4.241968e-01 1.439763e+04
```

```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 Relationships",
     col = "red", pch = 16, las = 1)
legend("topleft", legend = c(paste("Spearman Correlation = ", rho, "; p = ", rho.pval, sep = ""), paste
```



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 relationship shows that with increasing sampling effort, the log species richness increases, you are capturing more species as you sample more sites. However the PDAR shows that with increasing sampling effort, we are sampling species that are equally distantly related to the sampling effort from before. We are sampling the same amount of phylogentic diversity with each increasing sample effort.

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 currently studying a bacterial exoenzyme called the Resuscitation Promoting Factor (Rpf) which is a mechanism of resuscitation from dormancy. I have some evidence to suggest that the protein activity is general but preferrientally acts upon members of phyla Actinobacteria - Potentially creating a postitive feedback interaction effect within bacterial communities since Actinobacteria members with the Rpf gene will constitutively produce and secrete Rpf. I would need not only the phylogenetic data of bacterial communities exposed to the Rpf protein, but also the beta turn over from communities exposed to no or very small amounts of Rpf. I would expect with increasing bacterial community exposure to Rpf, the communities would become more similar.