9.Phylogenetic Diversity - Communities

Madison Brown; Z620: Quantitative Biodiversity, Indiana University

04 March, 2025

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 worksheet, you will be introduced to some commonly used methods in phylogenetic community ecology.

After completing this assignment you will know how to:

1. incorporate an evolutionary perspective into your understanding of community ecology
2. quantify and interpret phylogenetic α- and β-diversity
3. evaluate the contribution of phylogeny to spatial patterns of biodiversity

Directions:

1. In the Markdown version of this document in your cloned repo, change "Student Name" on line 3 (above) with your name.
2. Complete as much of the worksheet as possible during class.
3. Use the handout as a guide; it contains a more complete description of data sets along with examples of proper scripting needed to carry out the exercises.
4. Answer questions in the worksheet. Space for your answers is provided in this document and is indicated by the ">" character. If you need a second paragraph be sure to start the first line with ">". You should notice that the answer is highlighted in green by RStudio (color may vary if you changed the editor theme).
5. Before you leave the classroom today, it is *imperative* that you **push** this file to your GitHub repo, at whatever stage you are. This will enable you to pull your work onto your own computer.
6. When you have completed the worksheet, **Knit** the text and code into a single PDF file by pressing the `Knit` button in the RStudio scripting panel. This will save the PDF output in your '9.PhyloCom' folder.
7. After Knitting, please submit the worksheet by making a **push** to your GitHub repo and then create a **pull request** via GitHub. Your pull request should include this file *9.PhyloCom_Worksheet.Rmd* and the PDF output of `Knitr` (*9.PhyloCom_Worksheet.pdf*).

The completed exercise is due on **Wednesday, March 5th, 2025 before 12:00 PM (noon)**.

1) SETUP

Typically, the first thing you will do in either an R script or an RMarkdown file is setup your environment. This includes things such as setting the working directory and loading any packages that you will need.

In the R code chunk below, provide the code to:
1. clear your R environment,
2. print your current working directory,
3. set your working directory to your `Week7-PhyloCom/` folder,
4. load all of the required R packages (be sure to install if needed), and
5. load the required R source file.

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

```
## [1] "/cloud/project/QB2025_Brown/Week7-PhyloCom"
```

```
setwd("/cloud/project/QB2025_Brown/Week7-PhyloCom")

library(picante)
```

```
## Loading required package: ape
## Loading required package: vegan
## Loading required package: permute
## Loading required package: lattice
## Loading required package: nlme
```

```
library(ape)
library(seqinr)
```

```
##
## 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
```

```
library(vegan)
library(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
```

```
library(reshape)
library(devtools)
```

```
## Loading required package: usethis
##
## Attaching package: 'devtools'
```

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

```
library(BiocManager)
```

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

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

```
library(ineq)
library(labdsv)
```

```
## Loading required package: mgcv
```

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

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

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

```
##
## Attaching package: 'labdsv'
```

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

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

```
library(matrixStats)
```

```
##
## Attaching package: 'matrixStats'
```

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

```
library(pROC)
```

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

```
##
## Attaching package: 'pROC'
```

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

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

2) DESCRIPTION OF DATA

need to discuss data set from spatial ecology!

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

3) LOAD THE DATA

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

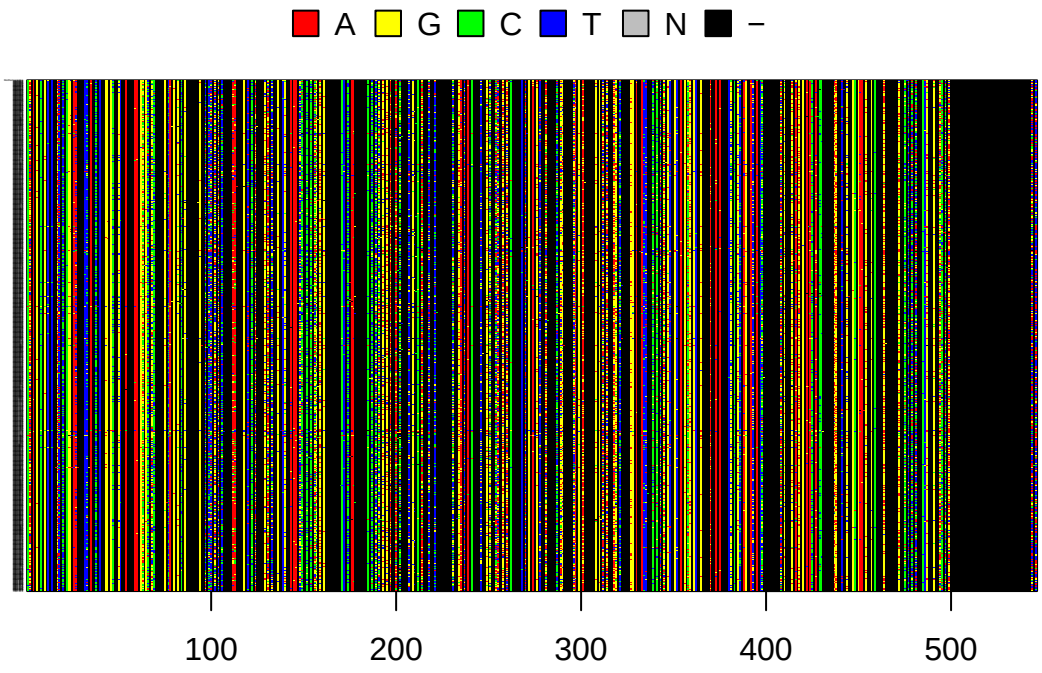
```
env <- read.table("data/20130801_PondDataMod.csv", sep = ",", header = TRUE)
env <- na.omit(env)

comm <- read.otu(shared = "./data/INPonds.final.rdp.shared", cutoff = "1")

comm <- comm[grep("*-DNA", rownames(comm)), ]

rownames(comm) <- gsub("\\-DNA", "", rownames(comm))
rownames(comm) <- gsub("\\_", "", rownames(comm))

comm <- comm[rownames(comm) %in% env$Sample_ID, ]

comm <- comm[, colSums(comm) > 0]

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

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

Next, in the R code chunk below, do the following:
1. load the FASTA alignment for the bacterial operational taxonomic units (OTUs),
2. rename the OTUs by removing everything before the tab (\t) and after the bar (|),

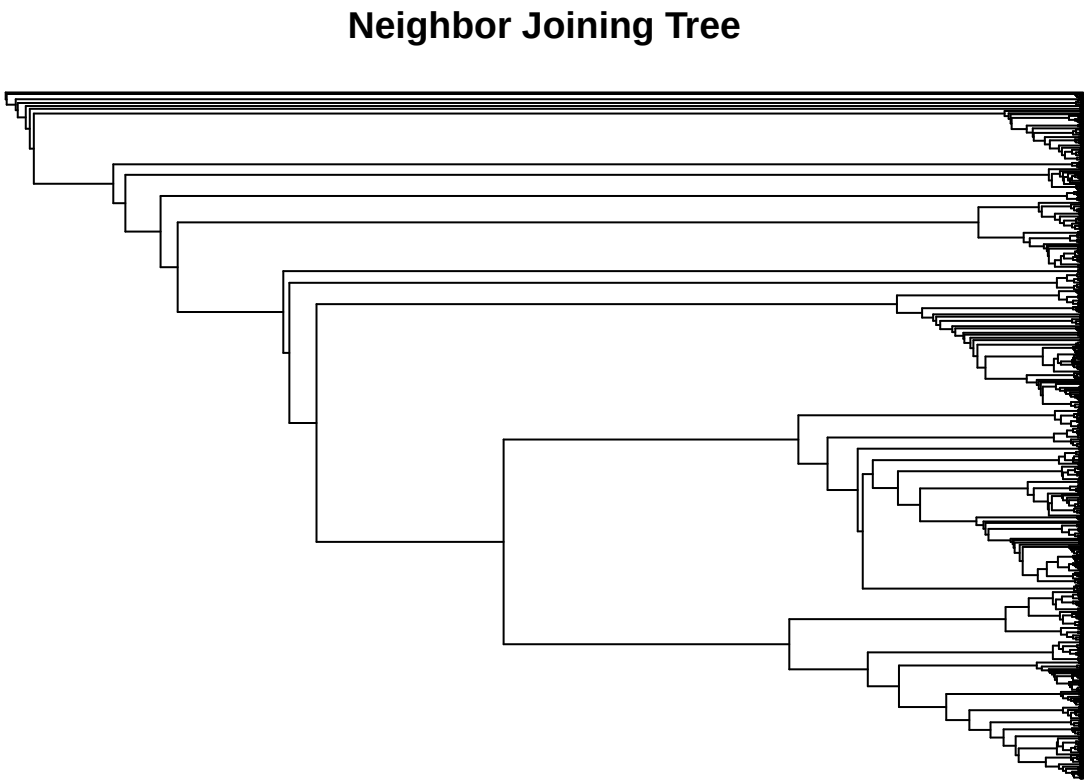
3. import the *Methanosarcina* outgroup FASTA file,
4. convert both FASTA files into the DNAbin format and combine using `rbind()`,
5. visualize the sequence alignment,
6. using the alignment (with outgroup), pick a DNA substitution model, and create a phylogenetic distance matrix,
7. using the distance matrix above, make a neighbor joining tree,
8. remove any tips (OTUs) that are not in the community data set,
9. plot the rooted tree.

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

ponds.cons$nam <- gsub(".*\t", "", ponds.cons$nam)
ponds.cons$nam <- gsub("\\|.*", "", ponds.cons$nam)

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

DNAbin <- rbind(as.DNAbin(outgroup), as.DNAbin(ponds.cons))

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




```
seq.dist.jc <- dist.dna(DNAbin, model = "JC", pairwise.deletion = FALSE)

phy.all <- bionj(seq.dist.jc)

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

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

phy <- root(phy, outgroup, resolve.root = TRUE)

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.offse
```



4) PHYLOGENETIC ALPHA DIVERSITY

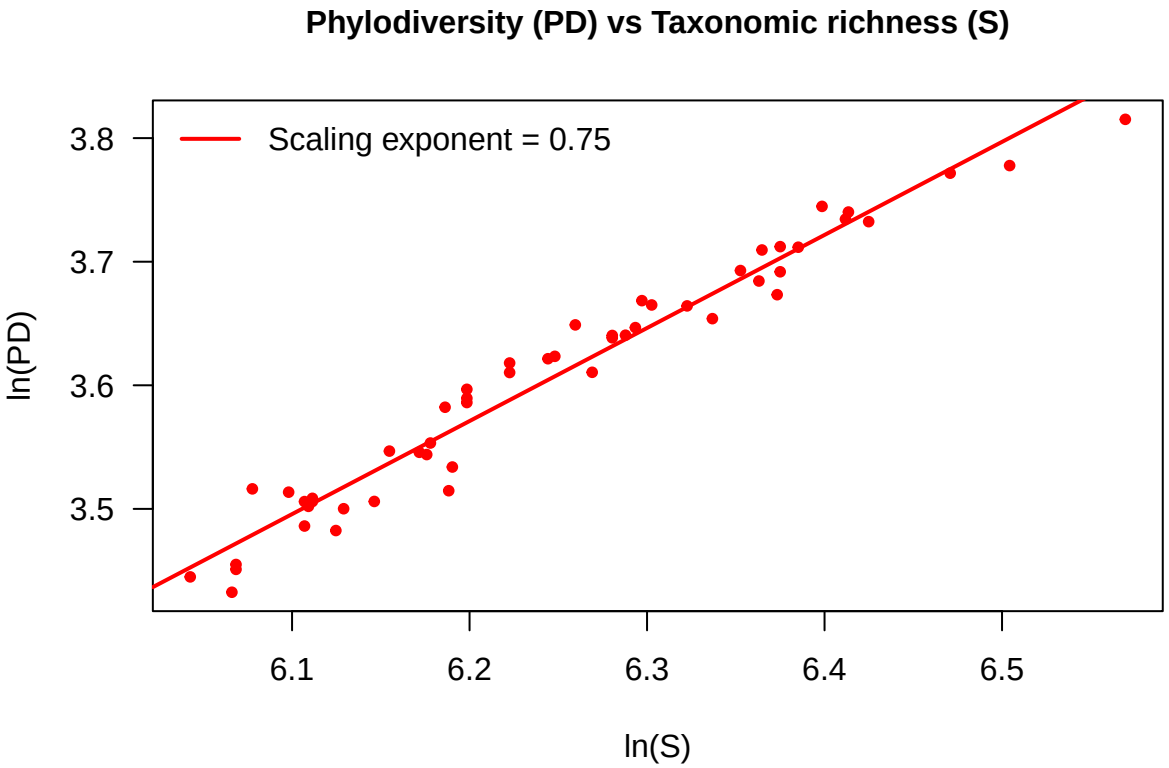
A. Faith's Phylogenetic Diversity (PD)

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

```
pd <- pd(comm, phy, include.root = FALSE)
#print(pd)
```

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.

```
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)")

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



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

> ***Answer 1a***: PD is calculated by adding the branch lengths together for each species in a phylogenetic tree. By doing this, it can help determine taxonomic richness because the amount of branch lengths present represent the amount of evolutionary diversity there is in your data. Many branches would result in a high PD, and you can then infer that many of your species are evolutionarily unique from one another. If your data has many evolutionary unique species, it is likely that taxonomic richness is also higher because you have many different species.
> ***Answer 1b***: If your data consisted of many individuals that were closely related, perhaps they have the same genus but different species, they would have a very low PD value because there would be very litte evolutionary diversity. However, taxonomic richness would still be very high becuse you would still have many unique species. ***Answer 1c***: The scaling exponent is present to illustrate a linear relationship. Without the scaling exponent, the graph would show a logarithmic plot because eventually, the amount of phylogenetic diversity would level out as taxonomic richness continues to increase.

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.

```
ses.pd <- ses.pd(comm[1:2,], phy, null.model = "richness", runs = 25,
                 include.root = FALSE)
print(ses.pd)
```

```
##       ntaxa   pd.obs pd.rand.mean pd.rand.sd pd.obs.rank   pd.obs.z  pd.obs.p
## BC001   668 43.71912     44.27643  0.8179022           6 -0.6813798 0.2307692
## BC002   587 40.94334     39.73715  0.9200584          23  1.3109959 0.8846154
##       runs
```

```
## BC001   25
## BC002   25
```

```
ses.pd2 <- ses.pd(comm[1:2,], phy, null.model = "frequency", runs = 25,
                  include.root = FALSE)
print(ses.pd2)
```

```
##       ntaxa   pd.obs pd.rand.mean pd.rand.sd pd.obs.rank  pd.obs.z   pd.obs.p
## BC001   668 43.71912     42.50956  0.6923964          26  1.746925 1.00000000
## BC002   587 40.94334     42.10181  0.6838479           2 -1.694048 0.07692308
##       runs
## BC001   25
## BC002   25
```

```
ses.pd3 <- ses.pd(comm[1:2,], phy, null.model = "sample.pool", runs = 25,
                  include.root = FALSE)
print(ses.pd3)
```

```
##       ntaxa   pd.obs pd.rand.mean pd.rand.sd pd.obs.rank   pd.obs.z  pd.obs.p
## BC001   668 43.71912     43.93947  0.6694696          10 -0.3291312 0.3846154
## BC002   587 40.94334     40.35795  0.7801118          20  0.7503850 0.7692308
##       runs
## BC001   25
## BC002   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: The null hypothesis would be that there is no difference in the observed PD compared to the expected PD. The alternative hypothesis would be that there is a difference in the observed PD via randomization. The alternative hypothesis would imply that there is a factor other than chance causing the difference in observed PD values. ***Answer 2b***: The frequencey model increased ses.pd to above 0 for pond one and decreased ses.pd to below 0 for pond two. In contrast, the sample pool model did not change the ses.pd value for pond one, but did slightly increase the ses.pd for pond two. The frequencey model affects the output compared to the richness model because you are now taking into account rare and dominant species, as opposed to just presence/absense data. The sample pool model did not result in as drastic differences. This may occur if species are well spread out; therefore, the randomized sample would result in very similiar species being chosen.

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.

```
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.

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

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.

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

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

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

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

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

abuNTI <- as.matrix(-1 * ((abuses.mntd[,2] - abuses.mntd[,3]) / abuses.mntd[,4]))
rownames(abuNTI) <- row.names(abuses.mntd)
colnames(abuNTI) <- "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 uses mean phylogenetic distance to determine how related taxa are to one another. It takes into account all of the branches in the tree to determine how distantly related species are. ***Answer 3b***: NTI uses mean nearest phylogenetic neighbor distance instead. This means it is primarily looking at the tips of the tree to determine if species are closely related to one another. ***Answer 3c***: The NRI values are all negative, which implies that overdispersion is present and species are less related to one another than initally believed. The NTI values are more variable because there are some that are negative and some that are positive. This means that there are some species who are closely related to one another and some species who evolved much longer ago and are not closely related. ***Answer 3d***: Using abundance data instead resulted in nearly all of the NTI values to be positive, along with a larger proportion of NRI values being positive. This is because rare and dominant species are no longer being treated equally, and dominant species have a larger impact than rare species do. Therefore, depending on the phylogentic history of the rare species, it will not be taken into account as much as it was before.

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.

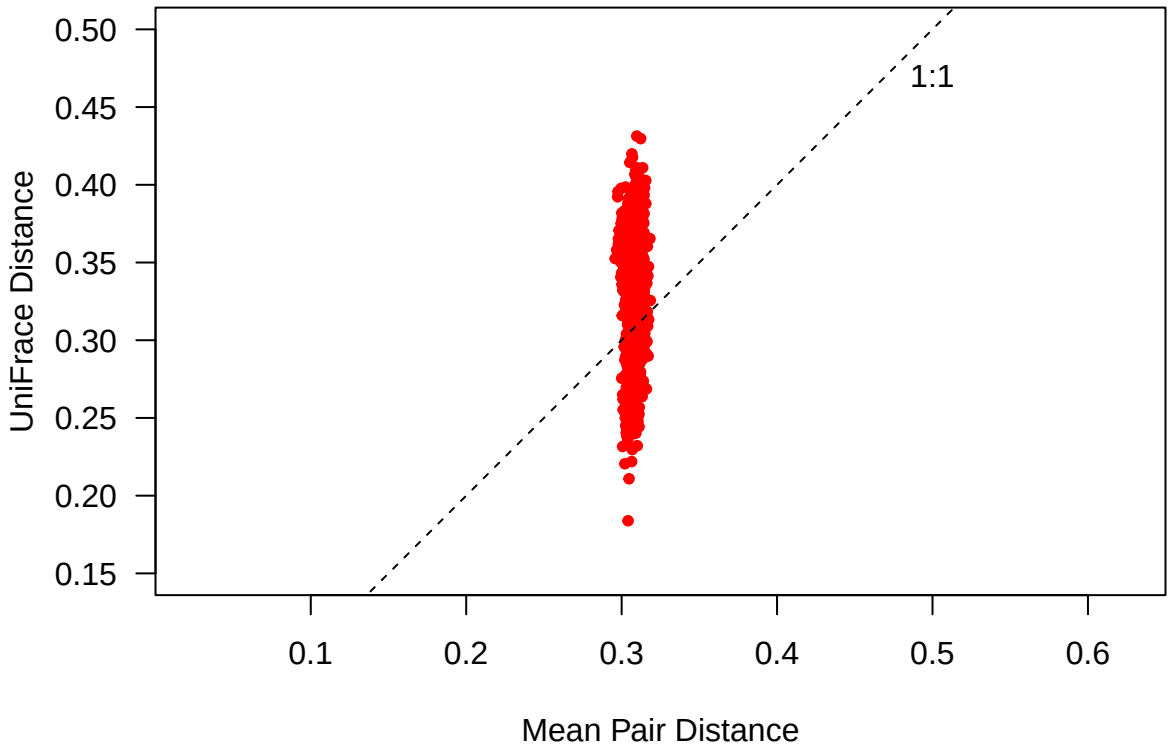
```
dist.mp <- comdist(comm, phydist)
```

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

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

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 = "UniFrace 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 analyzes all of the pairs of species in a sample and calculates

the average phylogentic distance each pair is from one another. UniFrac distance also determines phylogentic distances but by weighting the branch lengths. The different between the two is that UniFrac takes into account the length of branches that is shared and not shared and uses that to get a quantitative value of phylogenetic distance for each pair. Mean pair distance results in a singular value for the entire sample. ***Answer 4b***: The plot above shows that on average, pairs within the sample have a mean pairwise distance of about 0.3. However, there are pairs that are more closely related with a UniFrac values beginning at ~0.17 and pairs more distantly related with UniFrac values up to ~0.43. ***Answer 4c***: MPD shows less variation than UniFrac because it results in an average, singular value. All pairs in the sample will not have that exact phylogenetic difference, some taxa will be more closely related while some will be more distantly related. It is just important to understand that when using MPD and knowing that there is some variation.

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.

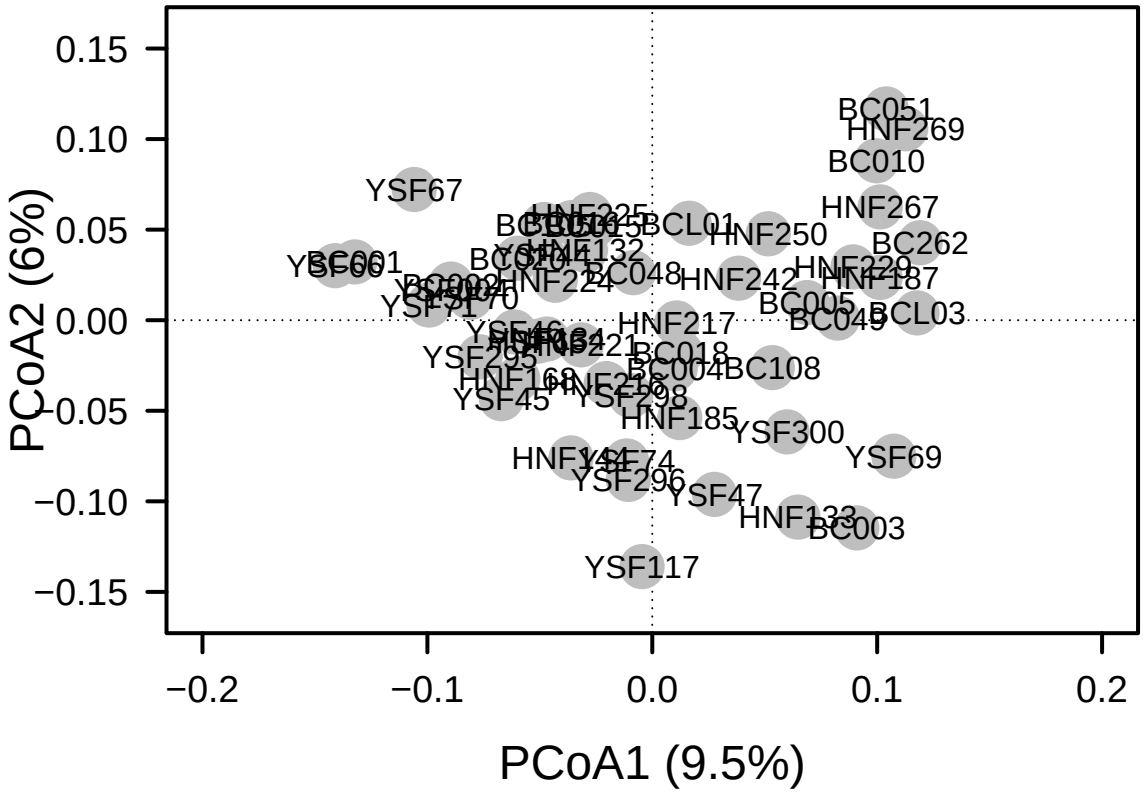
In the R code chunk below, do the following:
1. perform a PCoA based on the UniFrac distances, and
2. calculate the explained variation for the first three PCoA axes.

```
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)
```

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.

```
par(mar = c(5, 5, 1, 2) + 0.1)

plot(pond.pcoa$points[ ,1], pond.pcoa$points[ ,2],
     xlim = c(-0.2, 0.2), ylim = c(-.16, 0.16),
     xlab = paste("PCoA1 (", explainvar1, "%)", sep = ""),
     ylab = paste("PCoA2 (", explainvar2, "%)", sep = ""),
     pch = 16, cex = 2.0, type = "n", cex.lab = 1.5, cex.axis = 1.2, axes = FALSE)

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)

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.db <- vegdist(dist.uf, method = "bray", upper = TRUE, diag = TRUE)

pond_pcoa <- cmdscale(pond.db, k = 3, eig = TRUE)

explaindbvar1 <- round(pond_pcoa$eig[1] / sum(pond_pcoa$eig), 3) * 100
explaindbvar2 <- round(pond_pcoa$eig[2] / sum(pond_pcoa$eig), 3) * 100
explaindbvar3 <- round(pond_pcoa$eig[3] / sum(pond_pcoa$eig), 3) * 100
dbsum.eig <- sum(explaindbvar1, explaindbvar2, explaindbvar3)
```

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***: One of the primary differences is that phylogentic based ordination takes into account evolutionary history while taxonomic ordination does not. Taxonomic ordination only provides insight into what species are present and if there are any influential species. However, it does not give you any information on how closely or distantly related those species are. Phylogenetic based ordination will provide that insight and allows you to know if there are many closely-related species living in an area, then historically there may be an underlying reason for that. It is important to include phylogenetic information to gain a better understanding of how evolutionary history played a role in the current taxonmic richness of an area.

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.

```
watershed <- env$Location

phylo.adonis <- adonis2(dist.uf ~ watershed, permutations = 999)

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

ii. Continuous Approach

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

```
envs <- env[, 5:19]

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

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.

```
mantel(dist.uf, env.dist)
```

```
##
## Mantel statistic based on Pearson's product-moment correlation
##
## Call:
## mantel(xdis = dist.uf, ydis = env.dist)
##
## Mantel statistic r: 0.08433
##       Significance: 0.161
##
## Upper quantiles of permutations (null model):
##   90%   95% 97.5%   99%
## 0.115 0.148 0.175 0.206
## 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

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

anova(ponds.dbrda, by = "axis")
```

```
## Permutation test for dbrda under reduced model
## Forward tests for axes
## Permutation: free
## Number of permutations: 999
```

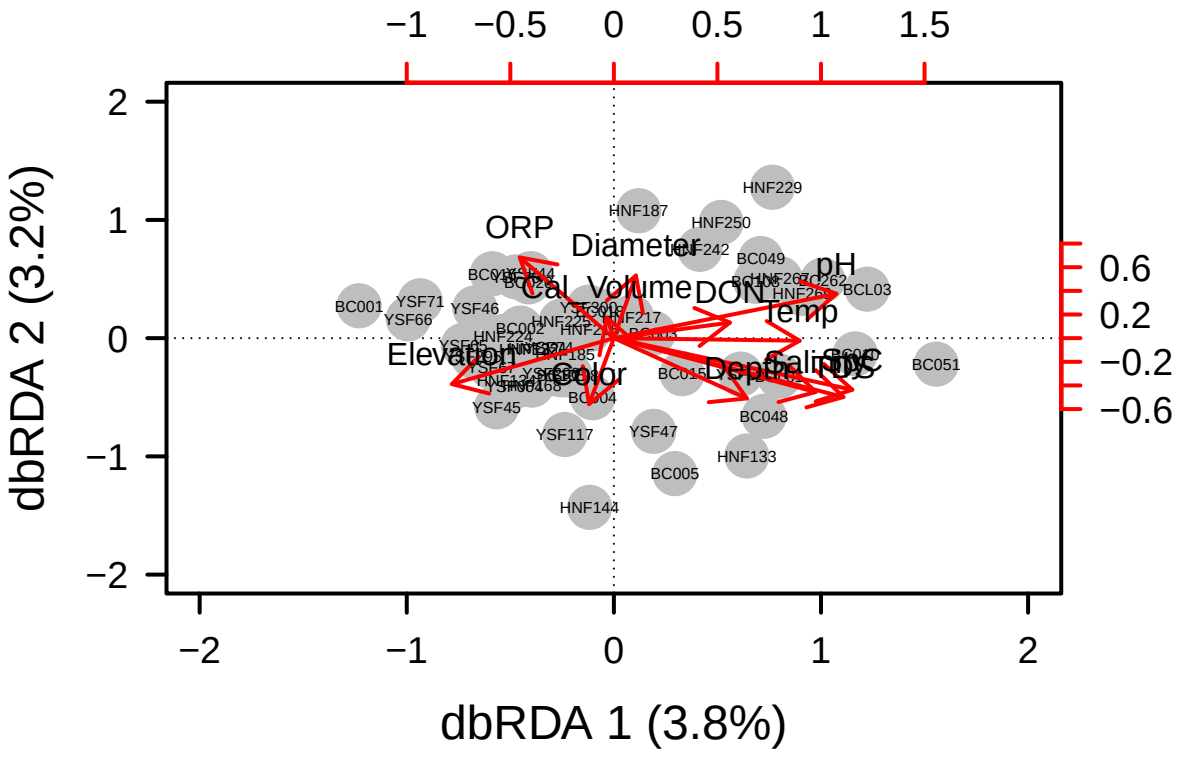
```
##
## Model: vegan::dbrda(formula = dist.uf ~ Elevation + Diameter + Depth + Cal_Volume + ORP + Temp + SpC
##          Df SumOfSqs      F Pr(>F)
## dbRDA1    1  0.10324 1.9852  0.478
## dbRDA2    1  0.08592 1.6521  0.818
## dbRDA3    1  0.08171 1.5711  0.859
## dbRDA4    1  0.07321 1.4077  0.953
## dbRDA5    1  0.06591 1.2674  0.989
## dbRDA6    1  0.05049 0.9709  1.000
## dbRDA7    1  0.04671 0.8982
## dbRDA8    1  0.04175 0.8027
## dbRDA9    1  0.03606 0.6934
## dbRDA10   1  0.03302 0.6349
## dbRDA11   1  0.03078 0.5919
## dbRDA12   1  0.02921 0.5617
## Residual 39  2.02820
```

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

```
##
## ***VECTORS
##
##              dbRDA1   dbRDA2     r2 Pr(>r)
## Elevation  -0.86345 -0.50444 0.1084  0.061 .
## Diameter    0.06683  0.99776 0.0543  0.280
## Depth       0.68050 -0.73275 0.1213  0.040 *
## Cal_Volume -0.22122  0.97523 0.0062  0.868
## ORP        -0.48281  0.87572 0.1309  0.028 *
## Temp        0.98974 -0.14286 0.1179  0.044 *
## SpC         0.85986 -0.51052 0.2464  0.003 **
## TDS         0.82742 -0.56158 0.2451  0.003 **
## Salinity    0.81889 -0.57395 0.1931  0.004 **
## pH          0.93813  0.34629 0.1823  0.005 **
## Color      -0.07756 -0.99699 0.0604  0.202
## DON         0.97923  0.20274 0.0467  0.312
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Permutation: free
## Number of permutations: 999
```

```
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

ponds_scores <- vegan::scores(ponds.dbrda, display = "sites")

par(mar = c(5, 5, 4, 4) + 0.1)

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

```r
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)

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

points(wa_scores,
       pch = 19, cex = 3, col = "gray")

text(wa_scores,
     labels = rownames(wa_scores),
     cex = 0.5)

vectors <- vegan::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 = rownames(vectors))

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

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



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

> ***Answer 6***: Based on the dbRDA plot above, the small percentages on the x and y axis represent that very little of the phylogenetic patterns for the bacterial communities can be explained by the environmental factors. This could occur for many reasons, especially when there are many confounding variables. The procedures above indicate that there are other factors impacting the beta diversity of the communities. The phylogenetic pcoa showed that phylogeney explained more variance in the beta-diversity than the environmental factors did.

SYNTHESIS

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

> ***Answer 7***: My thesis is focused on how the gut microbiome of red colobus is impacted by human disturbances. Therefore, it will be vital for me to analyze the phylogeny of the gut bacteria found in the individuals and see how it differs based on if they live in edge or interior forests. I think it will be very helpful to analyze the phylogentic patterns of beta diversity along with taxonomic richness. By understanding how many species are present, along with what species are present, I can gain a better understanding as to how the gut microbiome is impacted. Once the gut bacteria is identified, the evolutionary history of them can be explored. This can provide insight into functionality and also an estimate of how long that bacteria has been living in that species/individual and could provide insight into some of these questions: Has that type of bacteria alwyas been present in red colobus or did it recently appear? Is it only present in certain individuals, if so, why? Do environmental factors play a role in the presense/absense of certain bacteria species?