9.Phylogenetic Diversity - Communities

Jocelyn Huang; 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_Huang/Week7-PhyloCom"
```

```
setwd("/cloud/project/QB2025_Huang/Week7-PhyloCom")

package.list <-c("picante", "ape", "seqinr", "vegan", "fossil", "reshape",
                 "devtools", "BiocManager", "ineq", "labdsv", "matrixStats",
                 "pROC")
for(package in package.list){
  if(!require(package, character.only = TRUE, quietly = TRUE)){
    install.packages(package, repos = "http://cran.us.r-project.org")
    library(package, character.only = TRUE)
    }
}
```

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

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

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

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

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

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

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

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

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

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

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

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

```
## This is labdsv 2.1-0
```

```
## convert existing ordinations with as.dsvord()

##
## Attaching package: 'labdsv'

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

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

##
## Attaching package: 'matrixStats'

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

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

##
## Attaching package: 'pROC'

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

```
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.

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

```
# Subset the data to include only DNA-based
comm <- read.otu(shared = "./data/INPonds.final.rdp.shared", cutoff = "1")
comm <- comm[grep("*-DNA", rownames(comm)), ]

# Rename sites by removing extra characters
rownames(comm) <- gsub("\\-DNA", "", rownames(comm))
rownames(comm) <- gsub("\\_", "", rownames(comm))

# Remove unnecessary OTUs
comm <- comm[rownames(comm) %in% env$Sample_ID, ] # sites not included
comm <- comm[, colSums(comm) > 0] # zero-abundance

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.

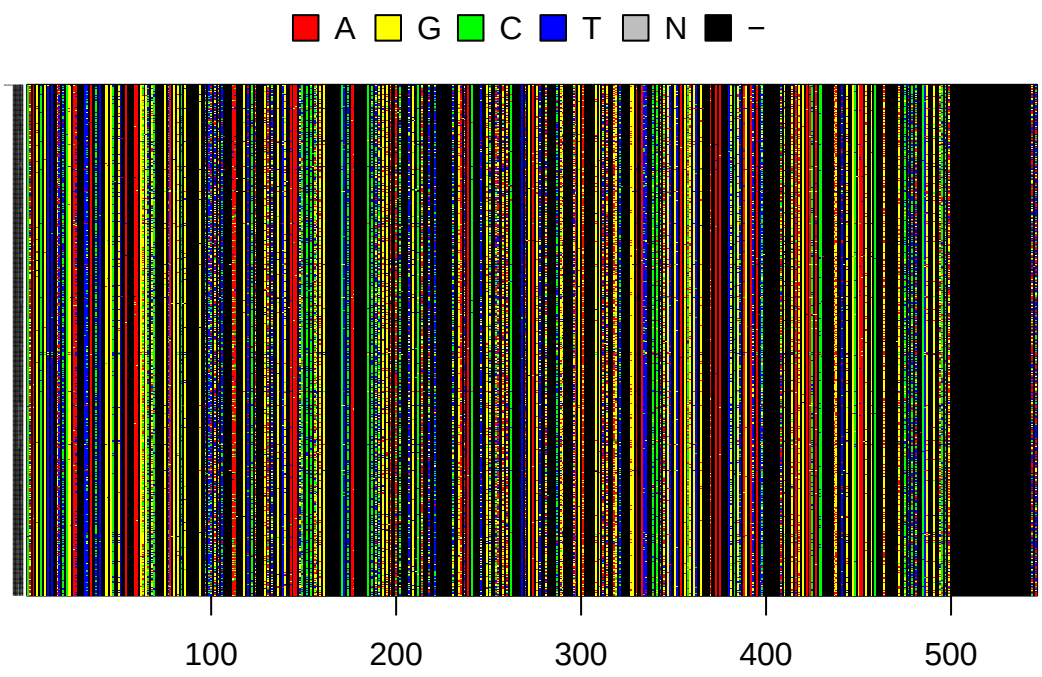
```
# Load FASTA for bacterial OTUs:
ponds.cons <- read.alignment(file = "./data/INPonds.final.rdp.1.rep.fasta",
                             format = "fasta")

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

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

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

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

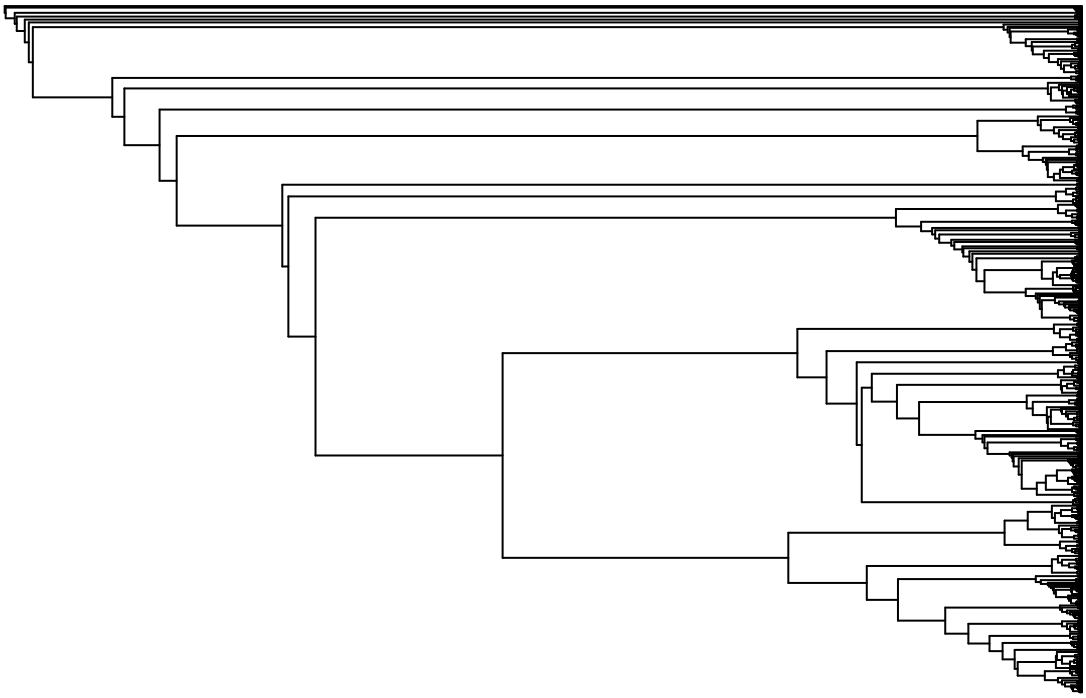



```
# Make distance matrix with Juke Cantor "JC" model:
seq.dist.jc <- dist.dna(DNAbin, model = "JC", pairwise.deletion = F)

# Neighbor-joining tree:
phy.all <- bionj(seq.dist.jc)

# Drop zero-occurring OTUs:
phy <- drop.tip(phy.all, phy.all$tip.label[!phy.all$tip.label %in%
                c(colnames(comm),"Methanosarcina")])

# Plot the tree:
outgroup <- match("Methanosarcina", phy$tip.label)
phy <- root(phy, outgroup, resolve.root = T)
par(mar = c(1,1,2,1) + 0.1)
plot.phylo(phy, main = "Neighbor Joining Tree", "phylogram",
           show.tip.label = F, use.edge.length = F,
           direction = "right", cex = 0.6, label.offset = 1)
```



4) PHYLOGENETIC ALPHA DIVERSITY

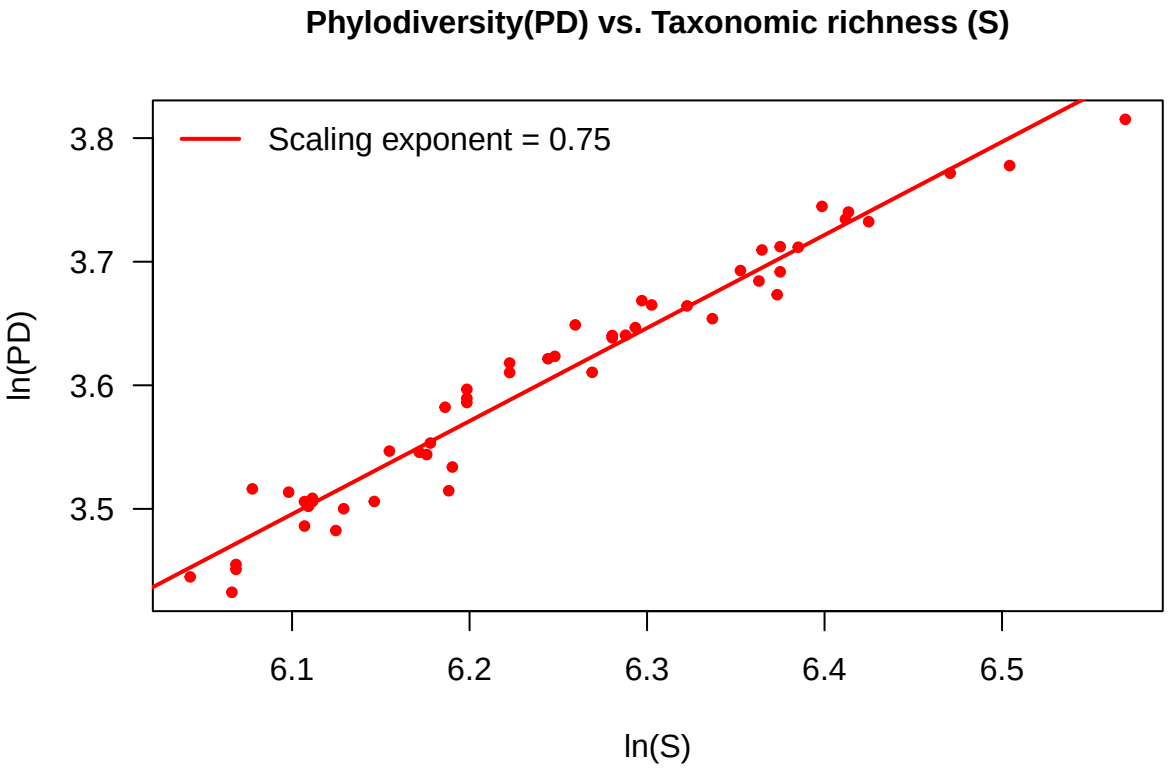
A. Faith's Phylogenetic Diversity (PD)

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

```
pd <- pd(comm, phy, include.root = F)
```

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 S versus PD:
par(mar = c(5,5,4,1) + 0.1)
plot(log(pd$SR), log(pd$PD), pch = 20, col = "red", las = 1,
     xlab = "ln(S)", ylab = "ln(PD)", cex.main = 1,
     main = "Phylodiversity(PD) vs. Taxonomic richness (S)")
# Add trend line:
fit <- lm(log(pd$PD) ~ log(pd$SR))
abline(fit, col = "red", lw = 2)
# Scaling exponent:
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***:
> Since PD is calculated based on presence-absence (occurance) data of each OTUs, it does not take into account the abundance of each taxa. Relating the phylogenetic diversity data with observed richness data help us compare whether and how the existence of rare taxon would affect the phylogenetic diversity.
>
> ***Answer 1b***:
> The two estimates would deviate from one another greatly if a lot of taxon in the studied sample/site are rare (having extremely low abundance) or very few are super abundant.
>
> ***Answer 1c***:
> Scaling exponent is the slope/coefficient of the fitted correlation between the phylogenetic diversity in log scale and the observed richness in log scale. A scaling exponent less than 1 means that phylogenetic diversity is going to level off eventually with increased observed richness. On the other hand, scaling exponent score tells us how phylogenetic diversity differ (or is going to differ) from richness because of the extend of rareness of taxon.

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 = F)

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

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

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:
When using null model "richness", the null hypothesis is that species richness within samples are the same (that, each species have the same abundance in a sample). The altnerative hypothesis is that each species have different abundance in a sample.

Answer 2b:
Using different null models means different null hypothesis is being tested. For example, if I replace the null model "richness" with "frequency," the null model for this is that all species occur at the same frequency accross the samples, which will definetly affect the output.

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 = T, runs = 25)

# Calculate NRI:
NRI <- as.matrix(-1 * ((ses.mpd$mpd.obs - ses.mpd$mpd.rand.mean)/ses.mpd$mpd.rand.sd))
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 = T, runs = 25)

NTI <- as.matrix(-1 * (ses.mntd$mntd.obs - ses.mntd$mntd.rand.mean)/ses.mntd$mntd.rand.sd)
rownames(NTI) <- row.names(ses.mntd)
colnames(NTI) <- "NTI"
```

Question 3:

a. In your own words describe what you are doing when you calculate the NRI.
b. In your own words describe what you are doing when you calculate the NTI.

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

Answer 3a:
I first randomanmize the average pairwise branch length between taxon, and calculate difference between the observed mean and the randomized mean between taxon branch length. I eventually divide this by the standard deviation of mean phylogenetic distance.

Answer 3b:
I first calculate the mean nearst phylogenetic neighbor distance, which is the mean distance between each taxa and their closet phylogenetic neighbor. Then, I calculate the difference between the observed mnnd and randomized mnnd and divide this by the standard deviation.

Answer 3c:
Since NRI for every samples are negative, this means all samples of microbes in Indiana ponds are overdispersed (less related to each other than expected by chance). Most NTI of the samples are also negative, which means only a few samples have taxa whose nearest taxa are closely related to them. The difference between NRI and NTI probably suggested that microbes in Indiana ponds are likly to be phylogenetically dissimilar with their neighoring community and that there might be a large extend of migration in and out of the ponds.

Answer 3d:
After changing the presence-absence based data to abundance based data, even though both NRI and NTI of all samples remain small for most samples, most NTI are positive, which means nearest taxa are more closely related to each other than expected. NRI, on the other hand, even if many samples are still negative, there are some that are positive and relatively large, indicating that the community of the sample is phylogenetically clustered (taxa are closely related to each other). This might suggests that abundant (unrare) taxon play important role in the community composition and they are likely more closely related to each other.

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.

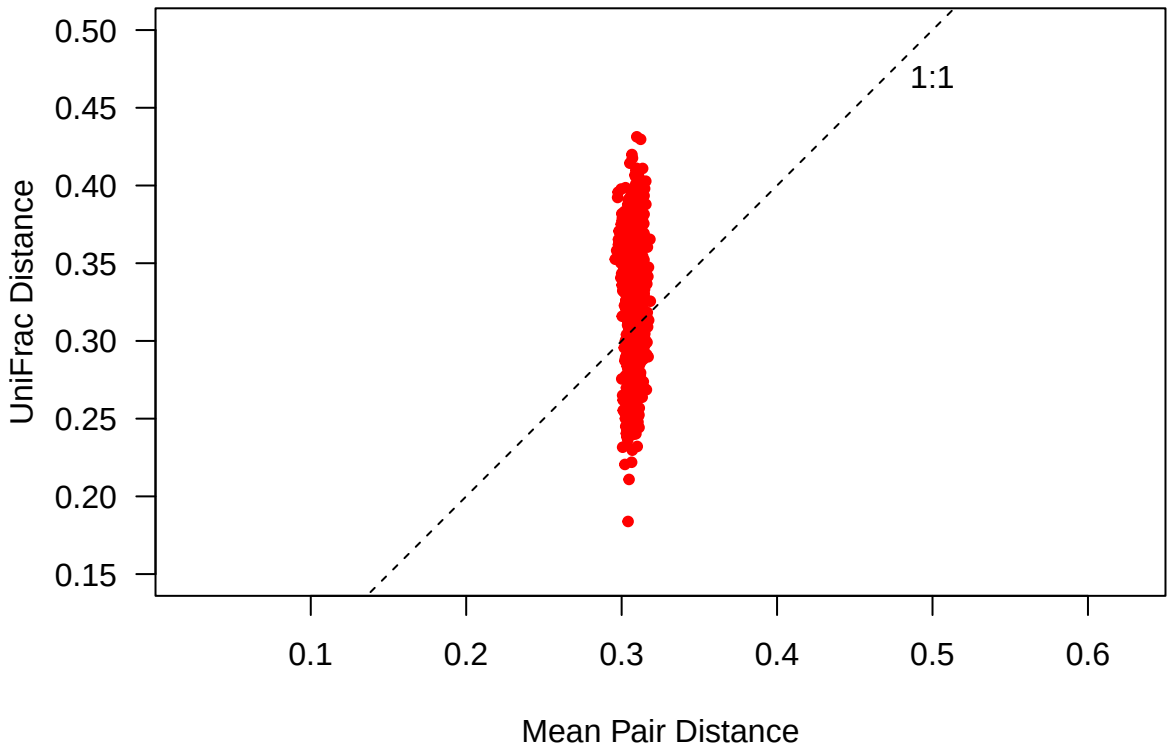
```
dist.mp <- comdist(comm, phydist) #mean pair distance
```

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

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

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: Unifrac calculates the sum of all branches length of each two samples and then calculate the proportion of branches that are not shared by the two samples. On the other hand, Mean Pair Distance calculates the mean distance of all paired taxa between two samples. In comparison, Unifrac is a fraction (0-1) while MPD is a discret number representing the mean phylogenetic distance (branch length difference) between two samples.

Answer 4b: Since we are only counting the presence-abence of each taxa between samples, mean pair distance remains relatively the same at around 0.3 even though unifrac value fluctuate from 0.15 to 0.45. This means incidence of taxa are about the same, but the taxa in each samples are different. Since fraction of unshared branched is smaller than 0.5, this probably means that many samples have a similar community composition.

Answer 4c: MPD shows less variation because based on the incidence data, each samples may have the similar composition of taxon, making the pairwise phylogenetic difference more or less similar between each two sample pairs.

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.

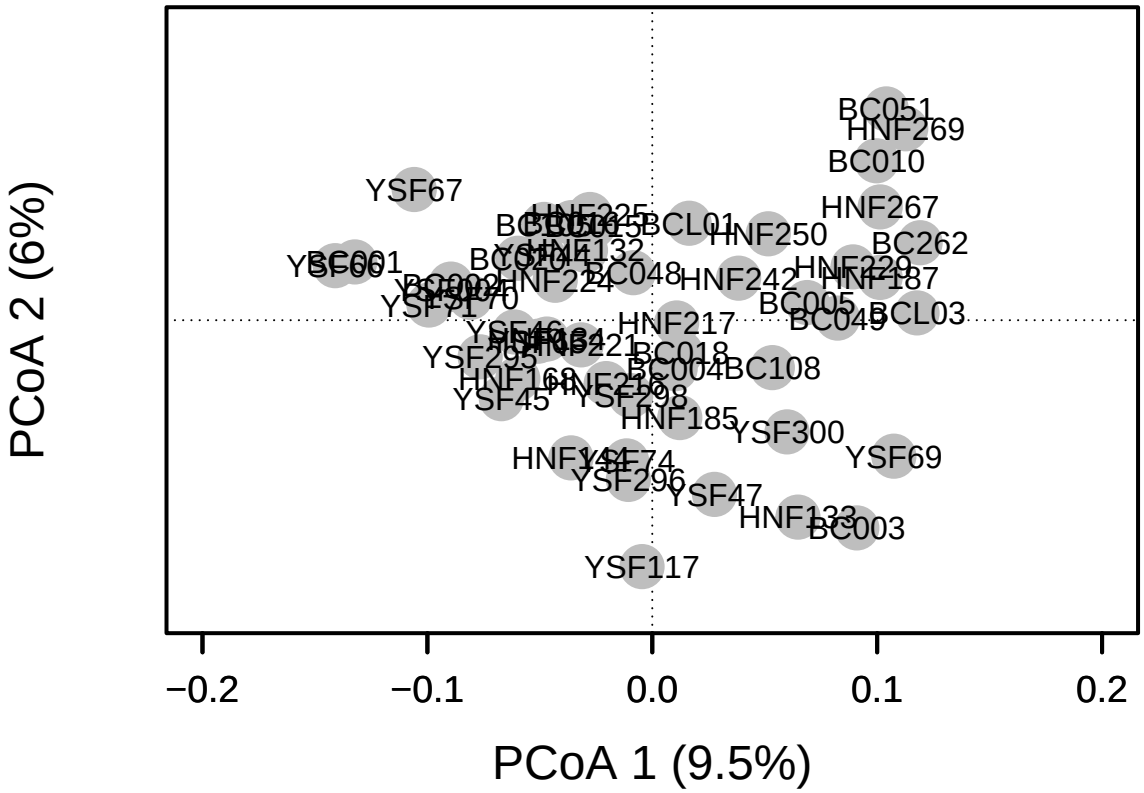
```
# Perform a PCoA based on UniFrac distance:
pond.pcoa <- cmdscale(dist.uf, eig = T, k = 3)

# Calculate the explained variation for the first three axes:
explainvar1 <- round(pond.pcoa$eig[1] / sum(pond.pcoa$eig), 3)*100
explainvar2 <- round(pond.pcoa$eig[2] / sum(pond.pcoa$eig), 3)*100
explainvar3 <- round(pond.pcoa$eig[3] / sum(pond.pcoa$eig), 3)*100
sum.eig <- sum(explainvar1, explainvar2, explainvar3)
```

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

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

```
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(-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 = F)
axis(side = 1, labels = T, lwd.ticks = 2, cex.axis = 1.2, las = 1)
axis(side = 1, 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(comm, method = "bray")
pond.pcoa.tx <- cmdscale(pond.db, eig = T, k = 3)

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

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 first two axes of the PCoA on phylogenetic data explaines 9.5% and 6% of the variance, while that on taxanomic data explains 28.4% and 12% of the variance. This shows that phylogenetic information is likley less emphasized in this system (closely related taxon likely doesn't contribute much to the ecological community composition).

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 #define env category
phylo.adonis <- adonis2(dist.uf ~ watershed, permutation = 999)
phylo.adonis
```

```
## Permutation test for adonis under reduced model
## Permutation: free
## Number of permutations: 999
##
## adonis2(formula = dist.uf ~ watershed, permutations = 999)
##          Df SumOfSqs     R2      F Pr(>F)
## Model     2  0.13316 0.0492 1.2679  0.024 *
## Residual 49  2.57305 0.9508
## Total    51  2.70621 1.0000
## ---
## 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()`).

```
envs <- env[, 5:19] #subset
envs <- envs[, -which(names(envs) %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: 0.1604
##       Significance: 0.054
##
## Upper quantiles of permutations (null model):
##   90%   95% 97.5%   99%
## 0.126 0.163 0.206 0.244
## 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
```

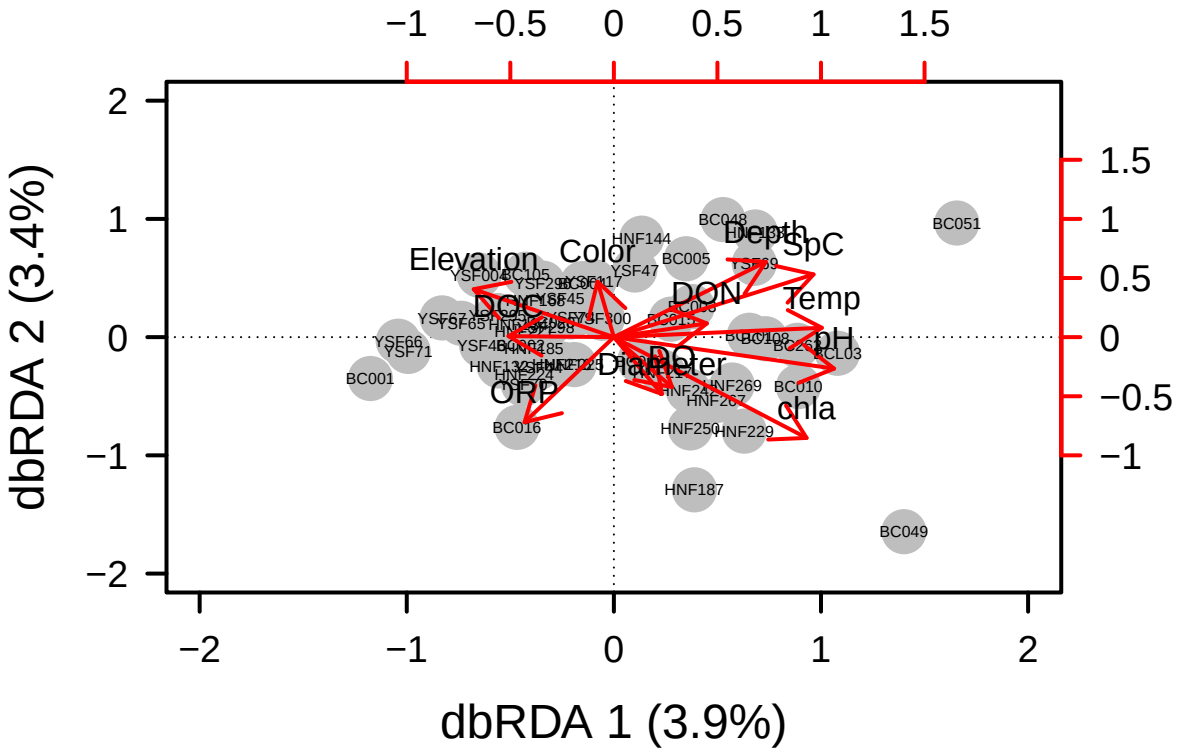
```
##
## Model: vegan::dbrda(formula = dist.uf ~ Elevation + Diameter + Depth + ORP + Temp + SpC + DO + pH + C
##          Df SumOfSqs      F Pr(>F)
## dbRDA1    1  0.10566 2.0152  0.447
## dbRDA2    1  0.09258 1.7658  0.638
## dbRDA3    1  0.07555 1.4409  0.966
## dbRDA4    1  0.06677 1.2735  0.998
## dbRDA5    1  0.05666 1.0807  1.000
## dbRDA6    1  0.05293 1.0095
## dbRDA7    1  0.04750 0.9059
## dbRDA8    1  0.03941 0.7517
## dbRDA9    1  0.03775 0.7201
## dbRDA10   1  0.03280 0.6256
## dbRDA11   1  0.02876 0.5485
## dbRDA12   1  0.02501 0.4770
## Residual 39  2.04482
```

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

# Make dbRDA plot:
ponds_score <- vegan::scores(ponds.dbrda, display = "sites")
par(mar = c(5,5,4,4)+0.1)
plot(ponds_score, 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, type = "n", cex.lab = 1.5,
     cex.axis = 1.2, axes = F)
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[,1])*2), labels = pretty(range(vectors[,1])*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 result shows that difference in watershed (location) does have a significant effect on phylogenetic diveristy of bacterial commmunity (p=0.027). On the other hand, the Mantel test shows that environmental variation only has a very weak correlation (r=0.16) with the UniFrac distance matrix, which means environmental variation probably doesn't affect phylogenetic composition of bacteria in the communities. According to the dbRDA result and plot, many environmental variables, such as pH, elevation, depth, etc. seemed to contribute to the phylogenetic composition of certain samples, however none of these environmental variables seem to be significant contributors.

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***: One of my research interests is in native community and biodiversity conservation. Comparison of the phylogenetic information of the native and nonnative/exotic/invasive community could potentially be very interesting, especially in determining the pathway of plant species invasion in a given community. Phylogenetic data can be used to test whether closely related taxon are more or less likely to become established or coexist.