9.Phylogenetic Diversity - Communities

Trang Nguyen; Z620: Quantitative Biodiversity, Indiana University

05 mars, 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 5$^{\text{th}}$, 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] "C:/Users/ttran/OneDrive - Indiana University/SP25 - Quantitative Biodiversity/QB2025_Nguyen/Wee
```

```
# Load packages
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)
  }
}
```

```
## Warning: le package 'picante' a été compilé avec la version R 4.4.2
```

```
## Warning: le package 'ape' a été compilé avec la version R 4.4.2
```

```
## Warning: le package 'vegan' a été compilé avec la version R 4.4.2
```

```
## Warning: le package 'permute' a été compilé avec la version R 4.4.2
```

```
## This is vegan 2.6-8
```

```
## Warning: le package 'seqinr' a été compilé avec la version R 4.4.2
```

```
##
## Attachement du package : 'seqinr'
```

```
## L'objet suivant est masqué depuis 'package:nlme':
##
##     gls
```

```
## L'objet suivant est masqué depuis 'package:permute':
##
##     getType
```

```
## Les objets suivants sont masqués depuis 'package:ape':
##
##     as.alignment, consensus
```

```
## Warning: le package 'fossil' a été compilé avec la version R 4.4.3
## Warning: le package 'sp' a été compilé avec la version R 4.4.2

## Warning: le package 'maps' a été compilé avec la version R 4.4.2

## Warning: le package 'shapefiles' a été compilé avec la version R 4.4.2

##
## Attachement du package : 'shapefiles'

## Les objets suivants sont masqués depuis 'package:foreign':
##
##     read.dbf, write.dbf

## Warning: le package 'reshape' a été compilé avec la version R 4.4.2

## Warning: le package 'devtools' a été compilé avec la version R 4.4.2

## Warning: le package 'usethis' a été compilé avec la version R 4.4.2

##
## Attachement du package : 'devtools'

## L'objet suivant est masqué depuis 'package:permute':
##
##     check

## Warning: le package 'BiocManager' a été compilé avec la version R 4.4.3

##
## Attachement du package : 'BiocManager'

## L'objet suivant est masqué depuis 'package:devtools':
##
##     install

## Warning: le package 'labdsv' a été compilé avec la version R 4.4.3

## 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()

##
## Attachement du package : 'labdsv'
## Les objets suivants sont masqués depuis 'package:vegan':
##
##     calibrate, pca, pco, scores

## Les objets suivants sont masqués depuis 'package:stats':
##
##     density, loadings

## Warning: le package 'matrixStats' a été compilé avec la version R 4.4.2

##
## Attachement du package : 'matrixStats'

## L'objet suivant est masqué depuis 'package:seqinr':
##
##     count

## Warning: le package 'pROC' a été compilé avec la version R 4.4.2

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

##
## Attachement du package : 'pROC'

## Les objets suivants sont masqués depuis '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,

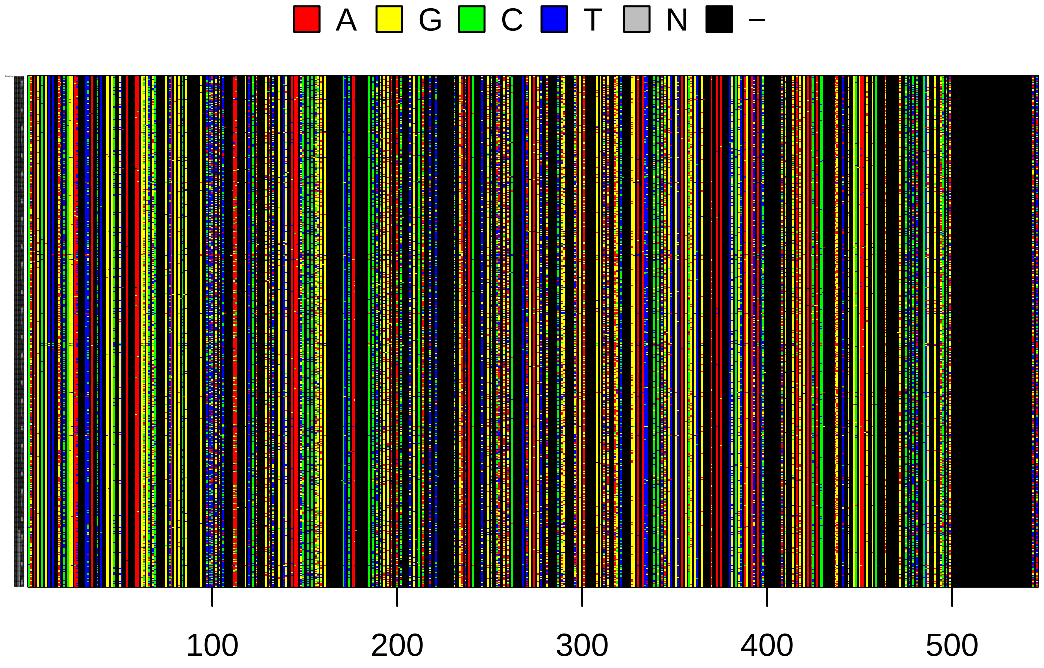
4. rename the sites by removing extra characters,
5. remove unnecessary OTUs in the site-by-species, and
6. load the taxonomic data using the `read.tax()` function from the source-code file.

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


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

# Select DNA data using `grep()`
comm <- comm[grep("*-DNA", rownames(comm)), ]

# Perform replacement of all matches with `gsub()`
rownames(comm) <- gsub("\\-DNA", "", rownames(comm))
rownames(comm) <- gsub("\\_", "", rownames(comm))

# Remove sites not in the environmental data set
comm <- comm[rownames(comm)  %in% env$Sample_ID, ]

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

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

```
## Warning in type.convert.default(as.character(x)): 'as.is' doit être spécifié
## par l'appelant ; utilisation de TRUE
## Warning in type.convert.default(as.character(x)): 'as.is' doit être spécifié
## par l'appelant ; utilisation de TRUE
## Warning in type.convert.default(as.character(x)): 'as.is' doit être spécifié
## par l'appelant ; utilisation de TRUE
## Warning in type.convert.default(as.character(x)): 'as.is' doit être spécifié
## par l'appelant ; utilisation de TRUE
## Warning in type.convert.default(as.character(x)): 'as.is' doit être spécifié
## par l'appelant ; utilisation de TRUE
## Warning in type.convert.default(as.character(x)): 'as.is' doit être spécifié
## par l'appelant ; utilisation de TRUE
```

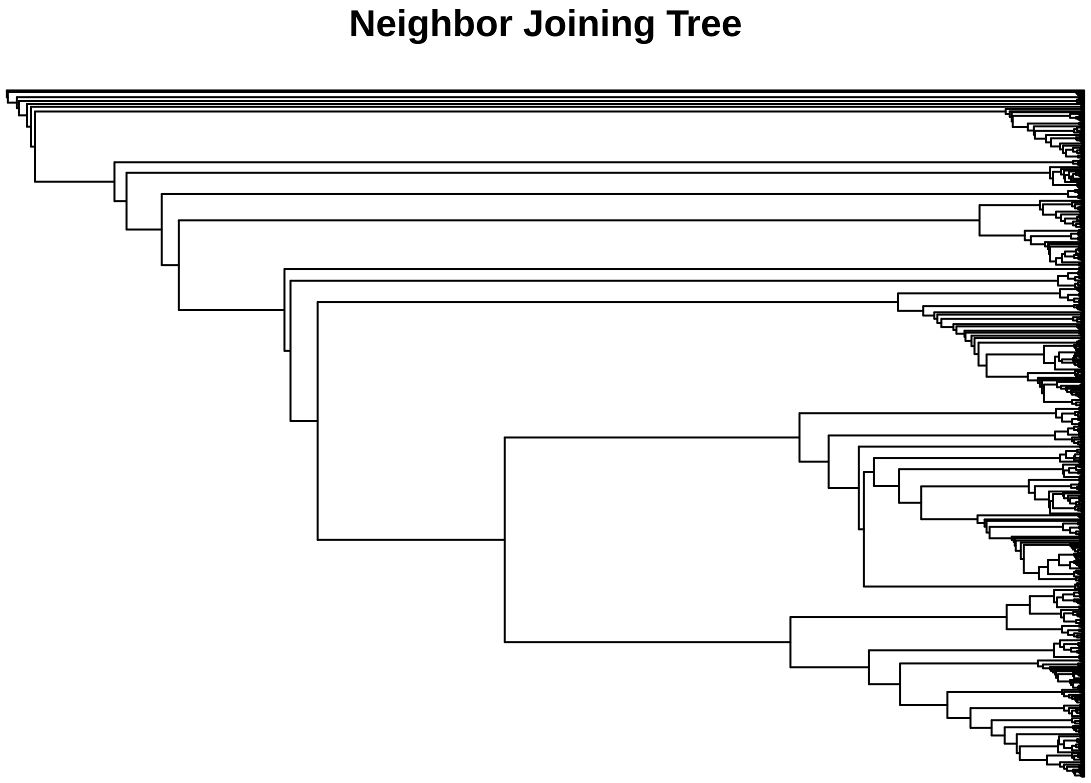
Next, in the R code chunk below, do the following:
1. load the FASTA alignment for the bacterial operational taxonomic units (OTUs),
2. rename the OTUs by removing everything before the tab (\t) and after the bar (|),
3. import the *Methanosarcina* outgroup FASTA file,
4. convert both FASTA files into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

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

# Rename OTUs in the FASTA File

# 2023, there was an issue with original gsub code,
# perhaps due to UTF8 ASCII character issue
# ponds.cons$nam <- gsub("\\|.*$", "", gsub("^.*?\t", "", ponds.cons$nam))
# new gsub() code seems to help:

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

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

# Convert alignment file to DNAbin
DNAbin <- rbind(as.DNAbin(outgroup),as.DNAbin(ponds.cons))

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



```
# Make distance matrix (`ape`)
seq.dist.jc <- dist.dna(DNAbin, model = "JC", pairwise.deletion = FALSE)

# Make a neigbor-joining tree file (`ape`)
```

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

# Drop tips of zero-occurrence OTUs (`ape`)
phy <- drop.tip(phy.all, phy.all$tip.label[!phy.all$tip.label %in%
              c(colnames(comm), "Methanosarcina")])

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

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

# Plot the rooted tree {ape}
par(mar = c(1, 1, 2, 1) + 0.1)
plot.phylo(phy, main = "Neighbor Joining Tree", "phylogram",
      show.tip.label = FALSE, use.edge.length = FALSE,
      direction = "right", cex = 0.6, label.offset = 1)
```



4) PHYLOGENETIC ALPHA DIVERSITY

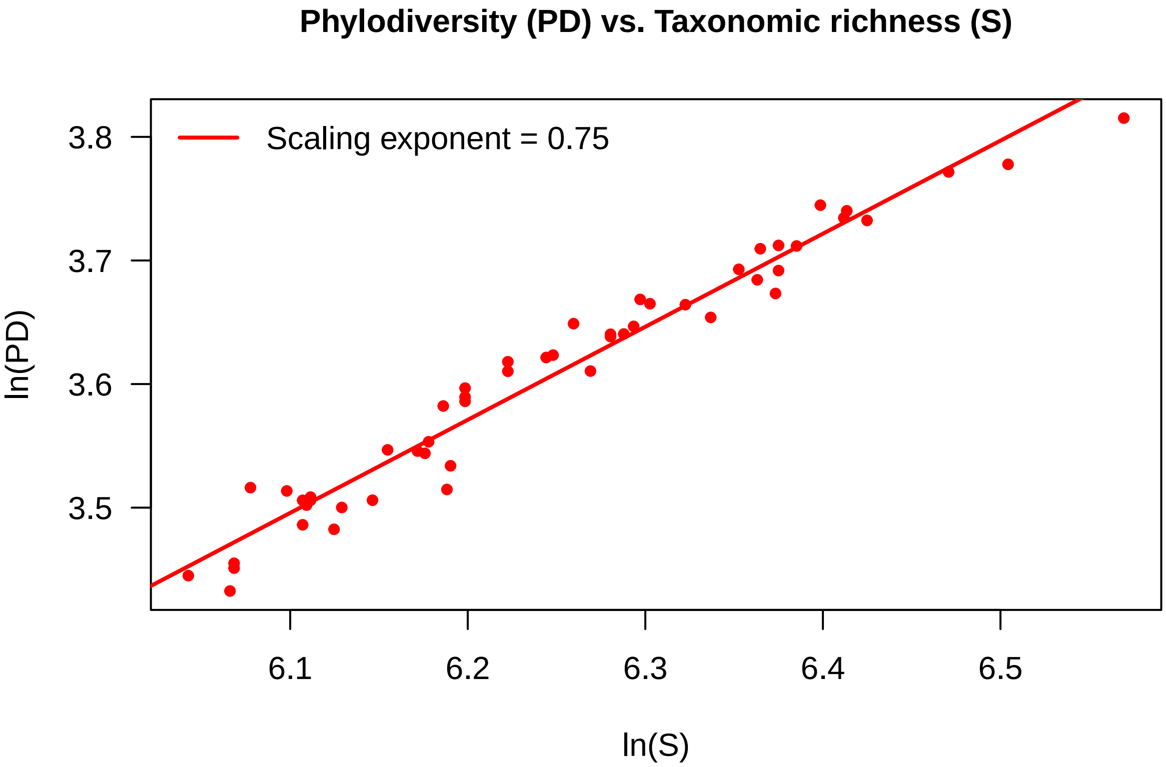
A. Faith's Phylogenetic Diversity (PD)

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

```
# Calculate PD and S {picante}
pd <- pd(comm, phy, include.root = FALSE)
```

In the R code chunk below, do the following:
1. plot species richness (S) versus phylogenetic diversity (PD),
2. add the trend line, and
3. calculate the scaling exponent.

```
# Biplot of S and PD
par(mar = c(5, 5, 4, 1) + 0.1)

plot(log(pd$S), log(pd$PD),
     pch = 20, col = "red", las = 1,
     xlab = "ln(S)", ylab = "ln(PD)", cex.main = 1,
     main="Phylodiversity (PD) vs. Taxonomic richness (S)")

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



Question 1: Answer the following questions about the PD-S pattern.
a. Based on how PD is calculated, 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***:
> Phylogenetic diversity (PD) is a measure of the total evolutionary history represented by a set of species or taxa, which takes information from the phylogenetic tree. It is related to taxonomic richness (S) because as the number of species increases (higher S), the amount of evolutionary history (or phylogenetic diversity) also tends to increase. Therefore, the more species I have in a community, the more diverse and evolutionary history the community represents, resulting in higher PD. ***Answer 1b***:
> They deviate when the species in a community are phylogenetically clustered or phylogenetically dispersed. If species are closely related (phylogenetically clustered), they will share more recent common ancestors, leading to lower PD despite a higher taxonomic richness. If species are distantly related (phylogenetically dispersed), the PD will be higher relative to the taxonomic richness. ***Answer 1c***:
> The scaling exponent in the PD-S pattern (shown by the red line in the plot) indicates the relationship between taxonomic richness (S) and phylogenetic diversity (PD). A scaling exponent of 0.75 suggests that PD increases at a slower rate than taxonomic richness.

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.

```r
# Estimate standardized effect size of PD via randomization (`picante`)
ses.pd <- ses.pd(comm[1:2,], phy, null.model = "richness", runs = 25,
                 include.root = FALSE)

ses.pd
```

```
##       ntaxa   pd.obs pd.rand.mean pd.rand.sd pd.obs.rank   pd.obs.z  pd.obs.p
## BC001   668 43.71912     43.86991  0.9281462          11 -0.1624647 0.4230769
## BC002   587 40.94334     39.78303  0.7713806          24  1.5042017 0.9230769
##       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 is when (PD) is random with respect to species richness or the structure of the community. The alternative hypothesis is that phylogenetic diversity is non-random and that the observed PD is significantly different from the expected PD under the null model. In the case of the "richness" null model, the null hypothesis is that the species richness is randomly distributed across the phylogeny, meaning that species are randomly placed within the tree's evolutionary history.

Answer 2b:
The choice of null model influences the observed ses.pd values by determining how the community is randomized for comparison with the observed community. In the case of richness null model: The community is randomized by species richness, which means the species are randomly assigned to the phylogeny, keeping the species count constant but allowing for random placement within the tree. I may find that the observed PD differs from this randomized distribution, the ses.pd value will be non-zero, indicating significant deviation from random phylogenetic distribution. In the case of other null models (e.g., "taxa.labels", "sample.pool"): The null models based on different assumptions of randomness (e.g., randomizing across species labels or using the full sample pool) will result in different expectations for the distribution of PD. These models could lead to different ses.pd values if your community deviates differently from these alternative randomizations.

B. Phylogenetic Dispersion Within a Sample

Another way to assess phylogenetic α-diversity is to look at dispersion within a sample.

i. Phylogenetic Resemblance Matrix

In the R code chunk below, do the following:
1. calculate the phylogenetic resemblance matrix for taxa in the Indiana ponds data set.

```
# Create a 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.

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

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

iii. Nearest Taxon Index (NTI)

In the R code chunk below, do the following: 1. Calculate the NTI for each site in the Indiana ponds data set.

```
# Estimate Standardized Effect Size of NRI via Randomization {picante}
ses.mntd <- ses.mntd(comm, phydist, null.model = "taxa.labels",
                     abundance.weighted = FALSE, runs = 25)

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

Question 3:

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

```r
# Calculate NRI with abundance data
ses.mpd_abundance <- ses.mpd(comm, phydist, null.model = "taxa.labels",
                             abundance.weighted = TRUE, runs = 25)

# Calculate NTI with abundance data
ses.mntd_abundance <- ses.mntd(comm, phydist, null.model = "taxa.labels",
                               abundance.weighted = TRUE, runs = 25)
```

Answer 3a:
The NRI measures the degree of phylogenetic clustering in a community. It compares the observed mean phylogenetic distance between species in a communit with the expected mean distance under a null model. If NRI > 0, this means that the community are more closely related to each other than expected by chance. If NRI < 0, this means that the community are more distantly related to each other than expected by chance.

Answer 3b:
The NTI measures how closely related the species are to their nearest neighbors within the community. NTI compares the observed distances between the closest species in the community to the expected distances under a random null model. If NTI > 0, this means that the species in the community are more distantly related to their nearest neighbors than expected by chance. If NTI < 0, this means that the species in the community are more closely related to their nearest neighbors than expected by chance.

Answer 3c:
If the NRI is significantly positive, the community is phylogenetically clustered, this means that species are more closely related than expected by chance. If NRI is negative, the community is more phylogenetically dispersed, with species being distantly related to one another. If the NTI is significantly negative, it suggests phylogenetic clustering, with species being more closely related to their nearest neighbors than expected. A positive NTI indicates phylogenetic dispersion, where the nearest neighbors are more distantly related than expected.

Answer 3d:
I thin that if we use abundance data, abundant species will have a greater influence on the calculation of phylogenetic clustering. If a species with high abundance is closely related to others, it will strongly influence whether the community is considered phylogenetically clustered. If we use presence-absence data, the NRI and NTI are not influenced by how many individuals of each species are present, only whether or not the species are present in the community.

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.

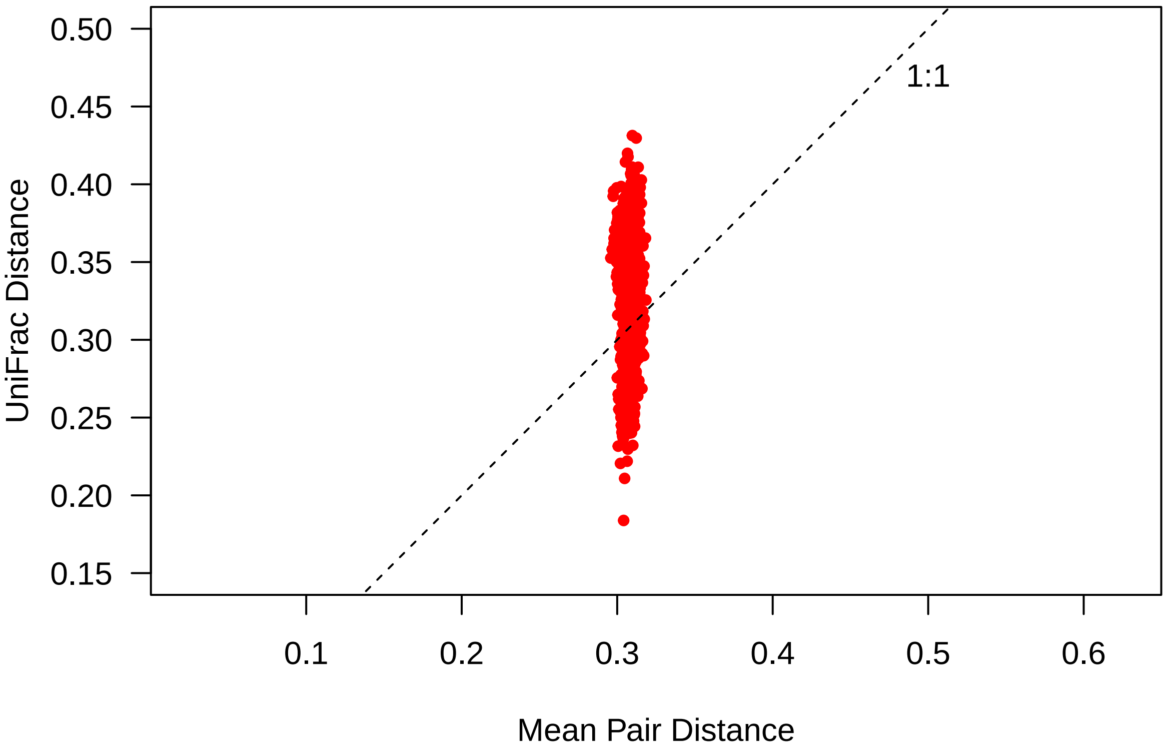
```
# Mean Pairwise Distance
dist.mp <- comdist(comm, phydist)
```

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

```
# UniFrac Distance (Note: this takes a few minutes; be patient)
dist.uf <- unifrac(comm, phy)
```

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

```
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***: MPD is a metric that measures the average phylogenetic distance between pairs of species in a community. In this case, MPD is based on the sum of the branch lengths between each pair of species in the community. It does not take into account species abundances, meaning it treats species as either present or absent. UniFrac distance is a phylogenetic distance metric that compares the phylogenetic compositions of two communities by considering their shared and unshared branches. It calculates the fraction of a phylogenetic tree's branch length that is unique to one community versus another. UniFrac can be weighted or unweighted, with the unweighted version focusing only on whether a taxon is present or absent, similar to MPD, but it incorporates a deeper view of the phylogenetic tree. UniFrac and MPD differ in that UniFrac compares entire communities to each other, while MPD focuses on pairwise distances within a single community. ***Answer 4b***: The plot comparing **Mean Pair Distance (MPD)** and **UniFrac Distance** suggests that the two metrics are generally similar but not identical. From the scatter plot, we can observe that most points are clustered along the 1:1 line, this means that that the patterns of phylogenetic distance in the two metrics are highly correlated. ***Answer 4c***: MPD tends to show **less variation** than UniFrac because MPD only measures distances between species within the same community. It treats the species in each community as a fixed set and calculates the average distance between all pairs of species without considering how they relate to other communities. On the other hand, UniFrac compares the entire phylogenetic composition of two communities, and it accounts for how the species in one community relate to the species in another. This comparison across communities allows UniFrac to capture more complex patterns and more variation, especially when comparing communities that may differ substantially in their phylogenetic compositions.

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.

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.

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.

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***:

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.

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()`).

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.

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

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***:

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***: