12.Phylogenetic Diversity - Communities

Timothy Biewer-Heisler; Z620: Quantitative Biodiversity, Indiana University

05 May, 2021

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 '12.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 *12.PhyloCom_Worksheet.Rmd* and the PDF output of `Knitr` (*12.PhyloCom_Worksheet.pdf*).

The completed exercise is due on **Monday, May 10$^{\textbf{th}}$, 2021 before 09:00 AM**.

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 `/12.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] "/Users/tbiewerh/GitHub/QB2021_Biewer-Heisler/2.Worksheets/12.PhyloCom"
```

```
setwd("~/GitHub/QB2021_Biewer-Heisler/2.Worksheets/12.PhyloCom")

package.list <- c('picante', 'ape', 'seqinr', 'vegan', 'fossil', 'reshape', 'simba')
for(package in package.list){
  if(!require(package, character.only = T, quietly = T)) {
    install.packages(package, repos = 'http://cran.us.r-project.org')
    library(package, character.only = T)
  }
}
```

```
## This is vegan 2.5-7

##
## 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
```

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

2) DESCRIPTION OF DATA

need to discuss data set from spatial ecology!

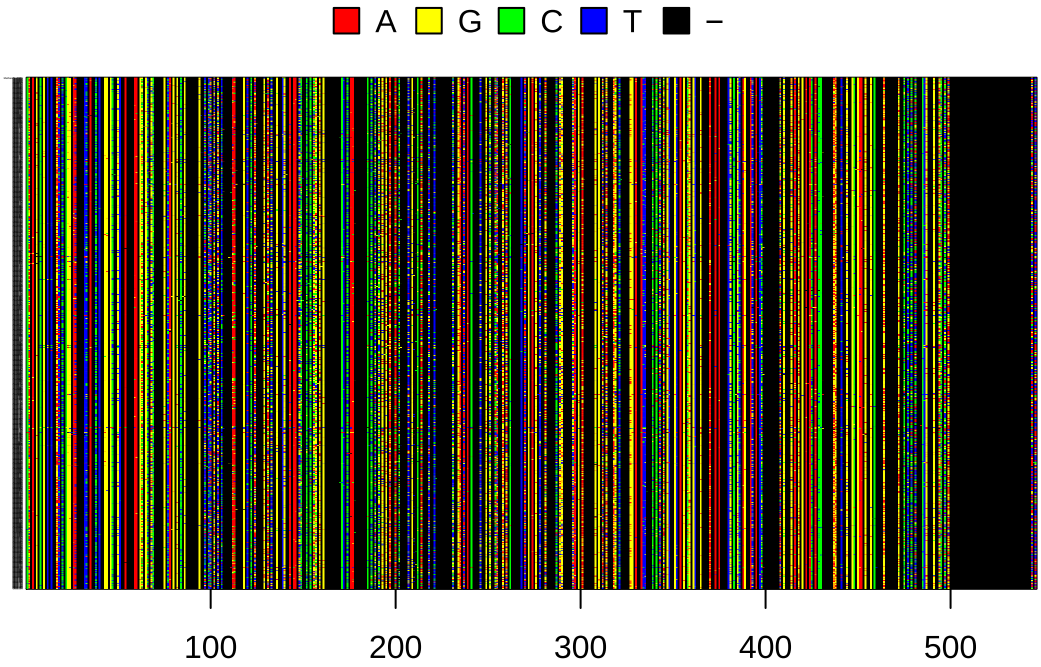
In 2013 we sampled > 50 forested ponds in Brown County State Park, Yellowood State Park, and Hoosier National Forest in southern Indiana. In addition to measuring a suite of geographic and environmental

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.

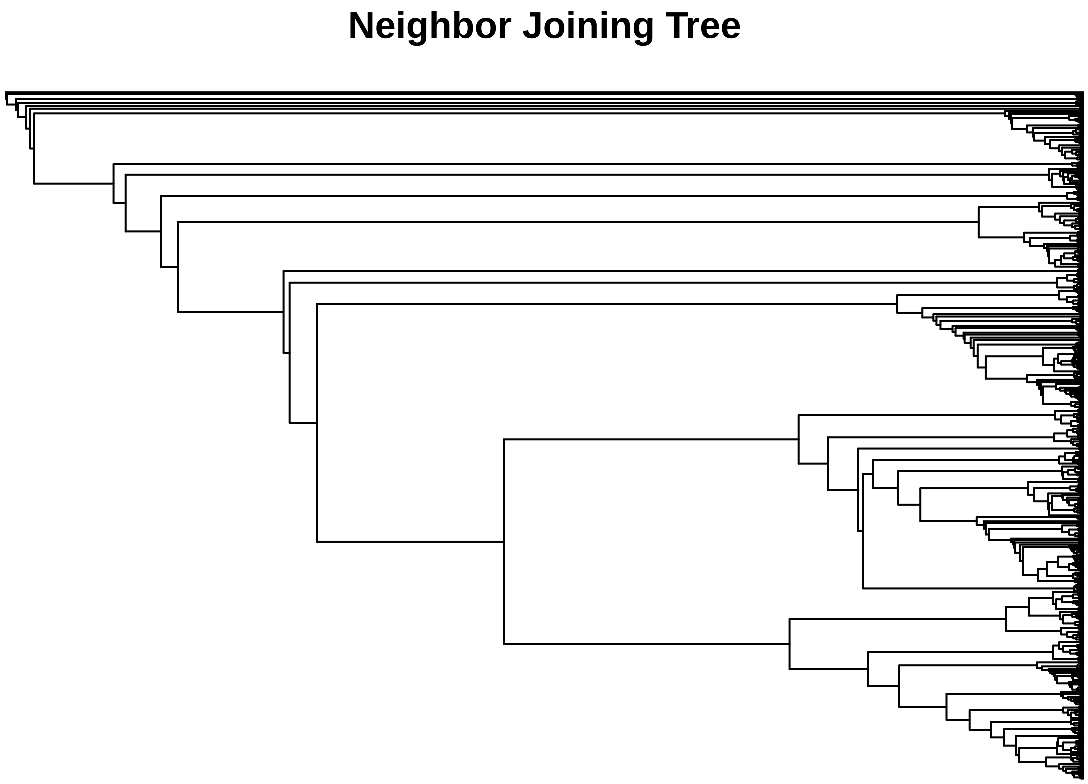
3) LOAD THE DATA

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 = T)
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")

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

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

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 = F, use.edge.length = F, di
```



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.

```
#accidentally put this in the previous chunk
```

4) PHYLOGENETIC ALPHA DIVERSITY

A. Faith's Phylogenetic Diversity (PD)

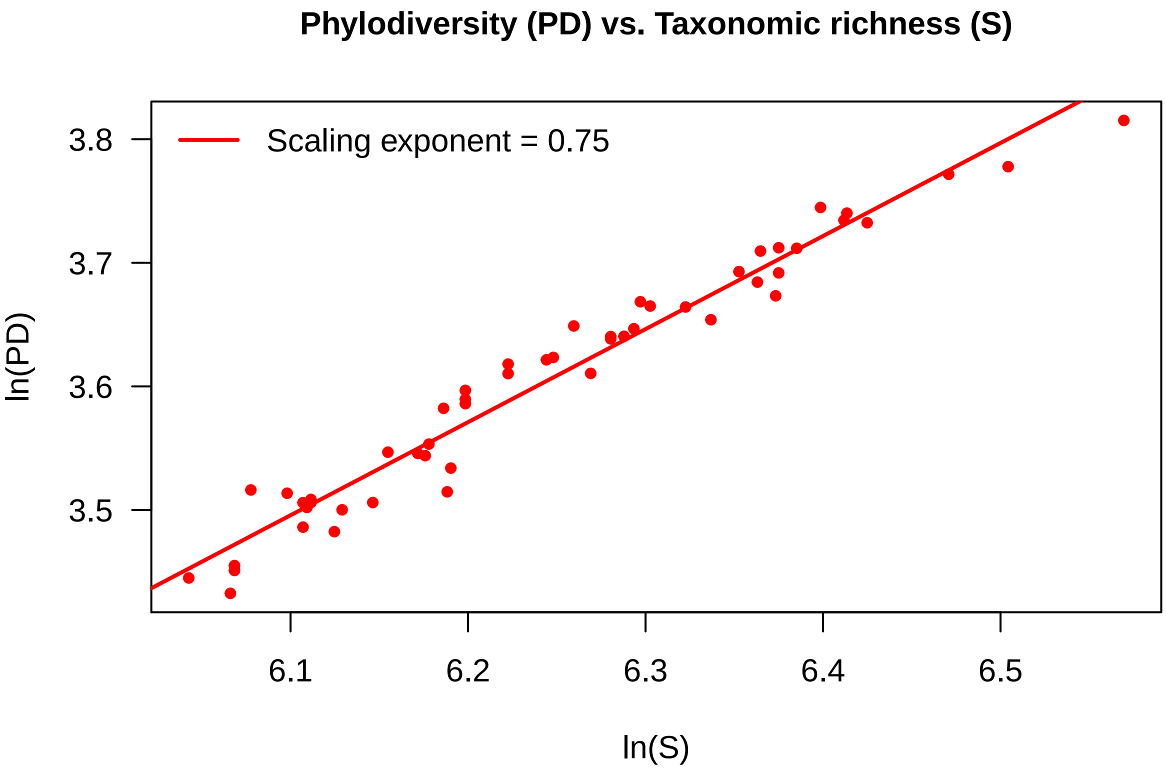
In the R code chunk below, do the following:
1. calculate Faith's D using the `pd()` function.

```
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.

```
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 =
```

```r
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 = "
```



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***:PD calculations can determine if a an assemblage contains more divergent taxa, regardless of the commonality of those species. This is similar to how one may present taxonomic richness, with a display of many species regardless of how common those species may be. ***Answer 1b***: Both taxonomic richness and phylodiversity measure species count, regardless of abundance of those species.
> ***Answer 1c***: When there hasn't been recent divergence, and all diversity is ancient, then phylodiversity would measure that differently than taxonomic richness due to the measure of branch lengths being an important factor.
> ***Answer 1d***: Phylodiversity seems to be much lower than the measure for taxonomic richness, possibly due to the extreme length of some of the branches we have.

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

```
##      ntaxa           pd.obs       pd.rand.mean     pd.rand.sd
##  Min.   :587.0   Min.   :40.94   Min.   :39.98   Min.   :0.7040
##  1st Qu.:607.2   1st Qu.:41.64   1st Qu.:40.98   1st Qu.:0.7272
##  Median :627.5   Median :42.33   Median :41.99   Median :0.7505
##  Mean   :627.5   Mean   :42.33   Mean   :41.99   Mean   :0.7505
##  3rd Qu.:647.8   3rd Qu.:43.03   3rd Qu.:43.00   3rd Qu.:0.7738
##  Max.   :668.0   Max.   :43.72   Max.   :44.01   Max.   :0.7970
##   pd.obs.rank      pd.obs.z            pd.obs.p           runs
##  Min.   : 8.00   Min.   :-0.406673   Min.   :0.3077   Min.   :25
##  1st Qu.:11.25   1st Qu.:-0.001889   1st Qu.:0.4327   1st Qu.:25
##  Median :14.50   Median : 0.402895   Median :0.5577   Median :25
##  Mean   :14.50   Mean   : 0.402895   Mean   :0.5577   Mean   :25
##  3rd Qu.:17.75   3rd Qu.: 0.807680   3rd Qu.:0.6827   3rd Qu.:25
##  Max.   :21.00   Max.   : 1.212464   Max.   :0.8077   Max.   :25
```

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

```
##      ntaxa           pd.obs       pd.rand.mean     pd.rand.sd
##  Min.   :587.0   Min.   :40.94   Min.   :39.93   Min.   :0.7933
##  1st Qu.:607.2   1st Qu.:41.64   1st Qu.:40.97   1st Qu.:0.8271
##  Median :627.5   Median :42.33   Median :42.01   Median :0.8609
##  Mean   :627.5   Mean   :42.33   Mean   :42.01   Mean   :0.8609
##  3rd Qu.:647.8   3rd Qu.:43.03   3rd Qu.:43.05   3rd Qu.:0.8947
##  Max.   :668.0   Max.   :43.72   Max.   :44.09   Max.   :0.9285
##   pd.obs.rank      pd.obs.z          pd.obs.p           runs
##  Min.   :10.00   Min.   :-0.47053   Min.   :0.3846   Min.   :25
##  1st Qu.:13.25   1st Qu.:-0.08093   1st Qu.:0.5096   1st Qu.:25
##  Median :16.50   Median : 0.30868   Median :0.6346   Median :25
##  Mean   :16.50   Mean   : 0.30868   Mean   :0.6346   Mean   :25
##  3rd Qu.:19.75   3rd Qu.: 0.69829   3rd Qu.:0.7596   3rd Qu.:25
##  Max.   :23.00   Max.   : 1.08789   Max.   :0.8846   Max.   :25
```

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

```
##      ntaxa           pd.obs       pd.rand.mean     pd.rand.sd
##  Min.   :587.0   Min.   :40.94   Min.   :42.18   Min.   :0.7477
##  1st Qu.:607.2   1st Qu.:41.64   1st Qu.:42.24   1st Qu.:0.7487
##  Median :627.5   Median :42.33   Median :42.31   Median :0.7497
##  Mean   :627.5   Mean   :42.33   Mean   :42.31   Mean   :0.7497
##  3rd Qu.:647.8   3rd Qu.:43.03   3rd Qu.:42.37   3rd Qu.:0.7507
##  Max.   :668.0   Max.   :43.72   Max.   :42.43   Max.   :0.7517
##   pd.obs.rank      pd.obs.z          pd.obs.p            runs
##  Min.   : 2.00   Min.   :-1.65647   Min.   :0.07692   Min.   :25
##  1st Qu.: 7.75   1st Qu.:-0.81487   1st Qu.:0.29808   1st Qu.:25
##  Median :13.50   Median : 0.02673   Median :0.51923   Median :25
##  Mean   :13.50   Mean   : 0.02673   Mean   :0.51923   Mean   :25
##  3rd Qu.:19.25   3rd Qu.: 0.86832   3rd Qu.:0.74038   3rd Qu.:25
##  Max.   :25.00   Max.   : 1.70992   Max.   :0.96154   Max.   :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 for taxa.labels is that the tree structure is as valid as a random tree, with the alternative hypothesis being that the tree structure is more valid as is than random. The null hypothesis for richness is that the abundance found in specific samples is effectively random, with the alternative saying that there is structure. The null hypothesis for frequency is that species abundance by site is effectively random, whereas the alternative hypothesis is that there is structure to the species abundance by site. ***Answer 2b***: Looking at the ses.pd values for richness and labels the ses.pd values of pond 2 showed greater than null expectation, whereas with frequency pond 1 showed greater than null expectation. Considering abundances within species looks at a different kind of community structure than the level at which we're viewing our phylogenetic tree which is why we may get different values than for taxa labels and richness.

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.

```r
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.

```r
ses.mpd <- ses.mpd(comm, phydist, null.model = "taxa.labels", abundance.weighted = T, 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.

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

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?

Answer 3a: We are testing for grouping or variablility in our data outside our expectations by taking average pairwise branch length between taxa in a sample. ***Answer 3b***: We are looking for similar factors as in NRI, but this time by looking at the distance between phylogenetic neighbor species. ***Answer 3c***: According to NRI everything is overdispersed, with variability in our data outside our statistical expectations. For NTI there are still many that are overdispersed, but some here actually show phylogenetic clustering. ***Answer 3d***: When running using abundance instead of presence-absence data it seems like in both cases there is much more evidence for clustering instead of overdispersion.

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.

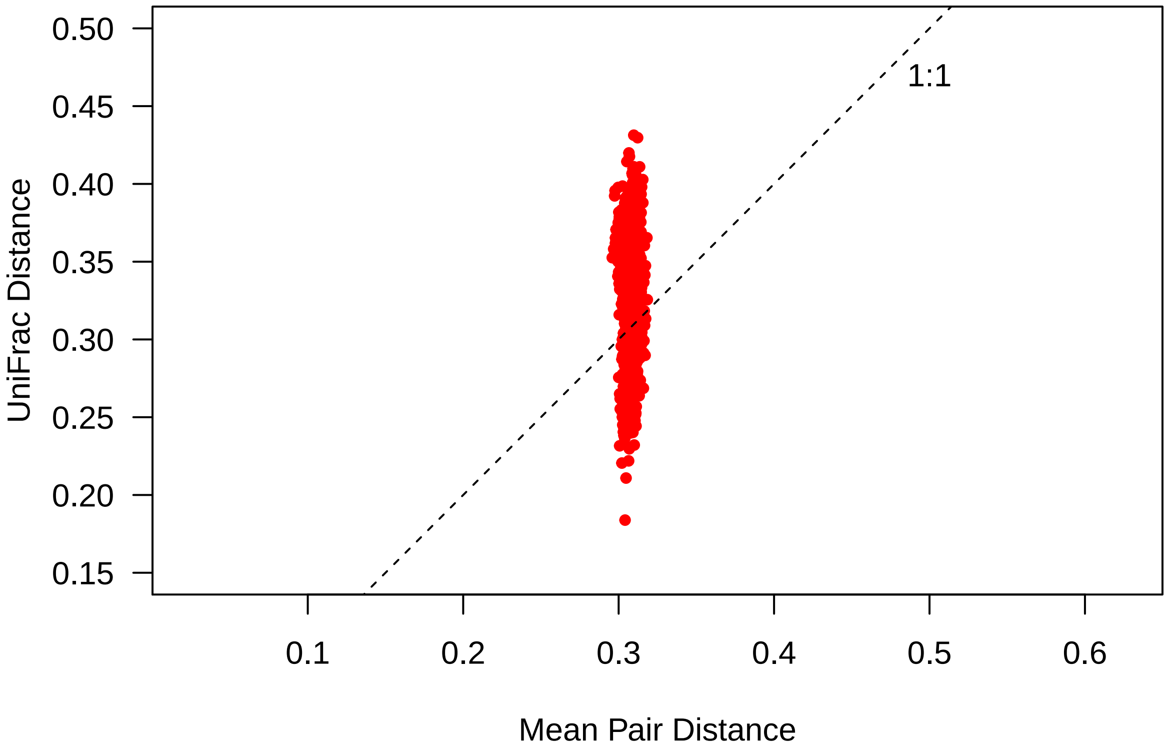
```
dist.mp <- comdist(comm, phydist)
```

```
## [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
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 uses measures of the amount of time since most recent common ancestors between pairs of taxa to estimate overall distance. Unifrac distance takes the amount of time two species have been separated from their most recent common ancestor and divides it by the total time they had a common ancestor until the root of the tree. A main difference is that unifrac is dependent on the root and mean pair distance does not. ***Answer 4b***: Mean pair distance shows much less diversity of results than UniFrac distance. ***Answer 4c***: We would get the same results from mean pair distance because it is based off of mean phylogenetic distance, which is a ratio whereas Unifrac is dependent entirely on the two samples.

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.

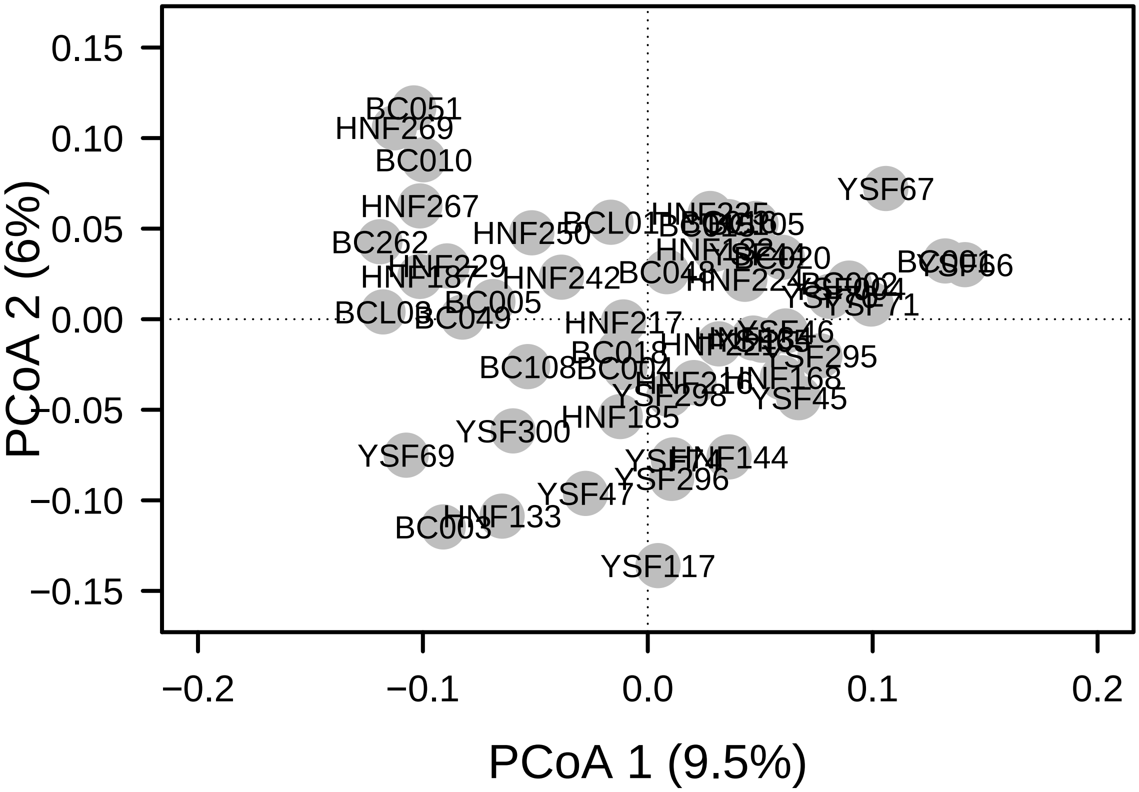
In the R code chunk below, do the following:
1. perform a PCoA based on the UniFrac distances, and
2. calculate the explained variation for the first three PCoA axes.

```
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.

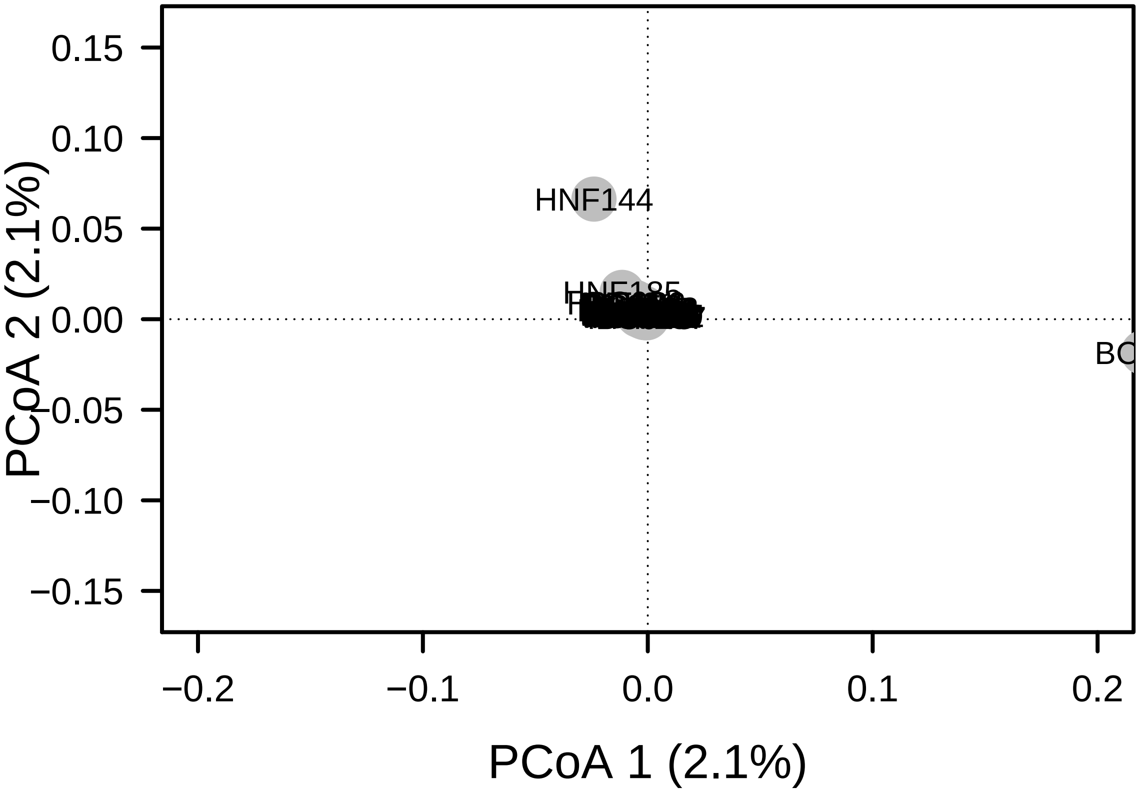
In the R code chunk below, do the following:
1. plot the PCoA results using either the R base package or the `ggplot` package,
2. include the appropriate axes,
3. add and label the points, and
4. customize the plot.

```
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("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)

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.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
sum.eig <- sum(explainvar1, explainvar2, explainvar3)

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("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)

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))
```



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***: Our new use of mp data has led to a decrease in variable explanation by our PcoAs. It seems that phylogenetic data is much more important than taxonomic data.

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
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  0.037 *
## 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
```

```
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  0.006 **
## 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()`).

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

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: 0.066
##
## Upper quantiles of permutations (null model):
##   90%   95% 97.5%   99%
## 0.129 0.171 0.195 0.218
## 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

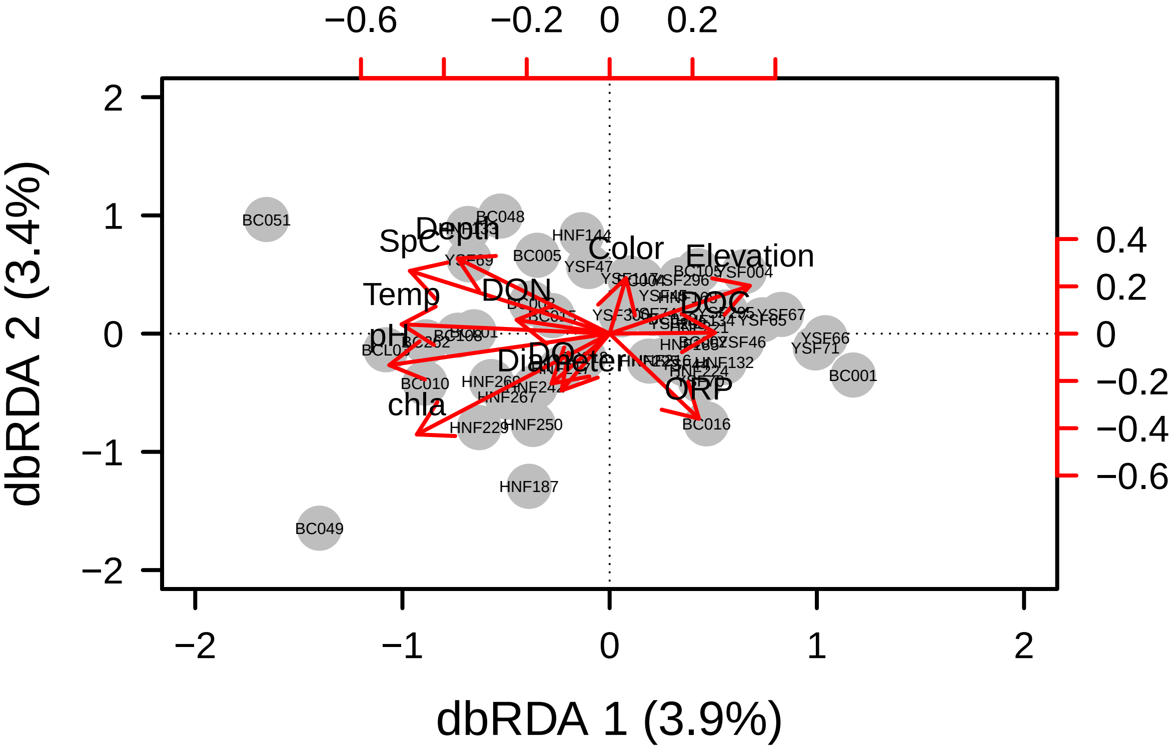
```
ponds.dbrda <- vegan::dbrda(dist.uf ~ ., data = as.data.frame(scale(envs)))
anova(ponds.dbrda, by = "axis")
```

```
## 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  0.421
## dbRDA2    1  0.09258 1.7658  0.605
## dbRDA3    1  0.07555 1.4409  0.973
## dbRDA4    1  0.06677 1.2735  0.997
## dbRDA5    1  0.05666 1.0807  1.000
## dbRDA6    1  0.05293 1.0095  1.000
## dbRDA7    1  0.04750 0.9059  1.000
## dbRDA8    1  0.03941 0.7517  1.000
## dbRDA9    1  0.03775 0.7201  1.000
## dbRDA10   1  0.03280 0.6256  1.000
## dbRDA11   1  0.02876 0.5485  1.000
## dbRDA12   1  0.02501 0.4770  0.998
## Residual 39  2.04482
```

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

```
##
## ***VECTORS
##
##             dbRDA1   dbRDA2     r2 Pr(>r)
## Elevation  0.77670  0.62986 0.0959  0.089 .
## Diameter  -0.27972 -0.96008 0.0541  0.249
## Depth     -0.63137  0.77548 0.1756  0.006 **
## ORP        0.41879 -0.90808 0.1437  0.030 *
## Temp      -0.98250  0.18628 0.1523  0.018 *
## SpC       -0.77101  0.63682 0.2087  0.003 **
## DO        -0.39318 -0.91946 0.0464  0.316
## pH        -0.96210 -0.27270 0.1756  0.012 *
## Color      0.06353  0.99798 0.0464  0.318
## chla      -0.60392 -0.79704 0.2626  0.008 **
## DOC        0.99847 -0.05526 0.0382  0.386
## DON       -0.91633  0.40042 0.0339  0.434
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Permutation: free
## Number of permutations: 999
```

```
dbrda.explainar1 <- round(ponds.dbrda$CCA$eig[1]/
                            sum(c(ponds.dbrda$CCA$eig, ponds.dbrda$CA$eig)), 3) * 100
dbrda.explainar2 <- round(ponds.dbrda$CCA$eig[2]/
                            sum(c(ponds.dbrda$CCA$eig, ponds.dbrda$CA$eig)), 3) * 100

par(mar = c(5, 5, 4, 4) + 0.1)
plot(scores(ponds.dbrda, display = "wa"), xlim = c(-2,2), ylim = c(-2,2),
     xlab = paste("dbRDA 1 (", dbrda.explainar1, "%)", sep = ""),
     ylab = paste("dbRDA 2 (", dbrda.explainar2, "%)", 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(scores(ponds.dbrda, display = "wa"),
       pch = 19, cex = 3, bg = "gray", col = "gray")
text(scores(ponds.dbrda, display = "wa"),
     labels = row.names(scores(ponds.dbrda, display = "wa")), cex = 0.5)

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



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

> ***Answer 6***: It seems that of the variables we have investigated that pH, temp, SpC, chla, depth, and elevation have the greatest affects on phylogenetic patterns. However, given the low amount of variation that is described by the dbRDA 1 and 2 it does not seem like they contribute to all that much of the variation.

6) SPATIAL PHYLOGENETIC COMMUNITY ECOLOGY

A. Phylogenetic Distance-Decay (PDD)

A distance decay (DD) relationship reflects the spatial autocorrelation of community similarity. That is, communities located near one another should be more similar to one another in taxonomic composition than distant communities. (This is analagous to the isolation by distance (IBD) pattern that is commonly found when examining genetic similarity of a populations as a function of space.) Historically, the two most common explanations for the taxonomic DD are that it reflects spatially autocorrelated environmental variables and the influence of dispersal limitation. However, if phylogenetic diversity is also spatially autocorrelated, then evolutionary history may also explain some of the taxonomic DD pattern. Here, we will construct the phylogenetic distance-decay (PDD) relationship

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.

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.

In the R code chunk below, test if the trend lines in the above distance decay relationships are different from one another.

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

> ***Answer 7***:

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.

> In some aspects phylogenetic data is vital for understanding speciation. Specifically understanding whether species are sister species or more distantly related is vital for forming conclusions about how their species separated and for what reasons. If it turns out two species aren't sister species it could nullify any conclusions about how these species would have separated and what barriers were responsible.