12.Phylogenetic Diversity - Communities

Ford Fishman; Z620: Quantitative Biodiversity, Indiana University

06 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.

```r
rm(list = ls())
getwd()
setwd("~/GitHub/QB2021_Fishman/2.Worksheets/12.PhyloCom/")
library(vegan)
```

```
## Warning: package 'vegan' was built under R version 4.0.5

## Loading required package: permute

## Loading required package: lattice

## This is vegan 2.5-7
```

```r
library(ape)
```

```
## Warning: package 'ape' was built under R version 4.0.5
```

```r
library(picante)
```

```
## Warning: package 'picante' was built under R version 4.0.5

## Loading required package: nlme
```

```r
library(fossil)
```

```
## Warning: package 'fossil' was built under R version 4.0.5

## 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
```

```r
library(reshape)
```

```
## Warning: package 'reshape' was built under R version 4.0.5
```

```r
library(seqinr)
```

```
## Warning: package 'seqinr' was built under R version 4.0.5

##
## Attaching package: 'seqinr'

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

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

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

```
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.

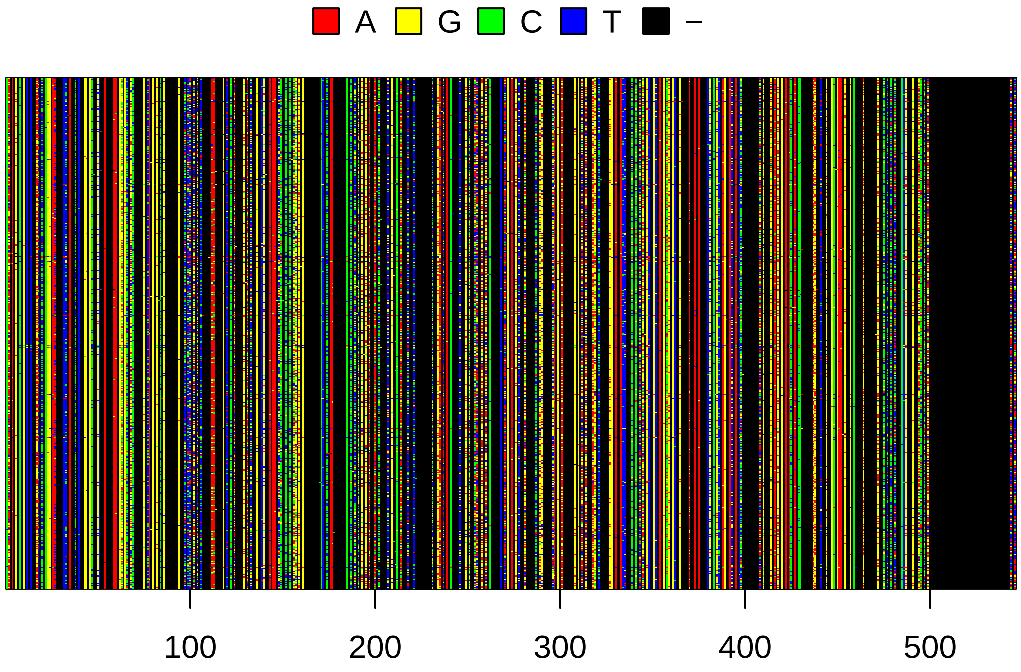
3) LOAD THE DATA

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.

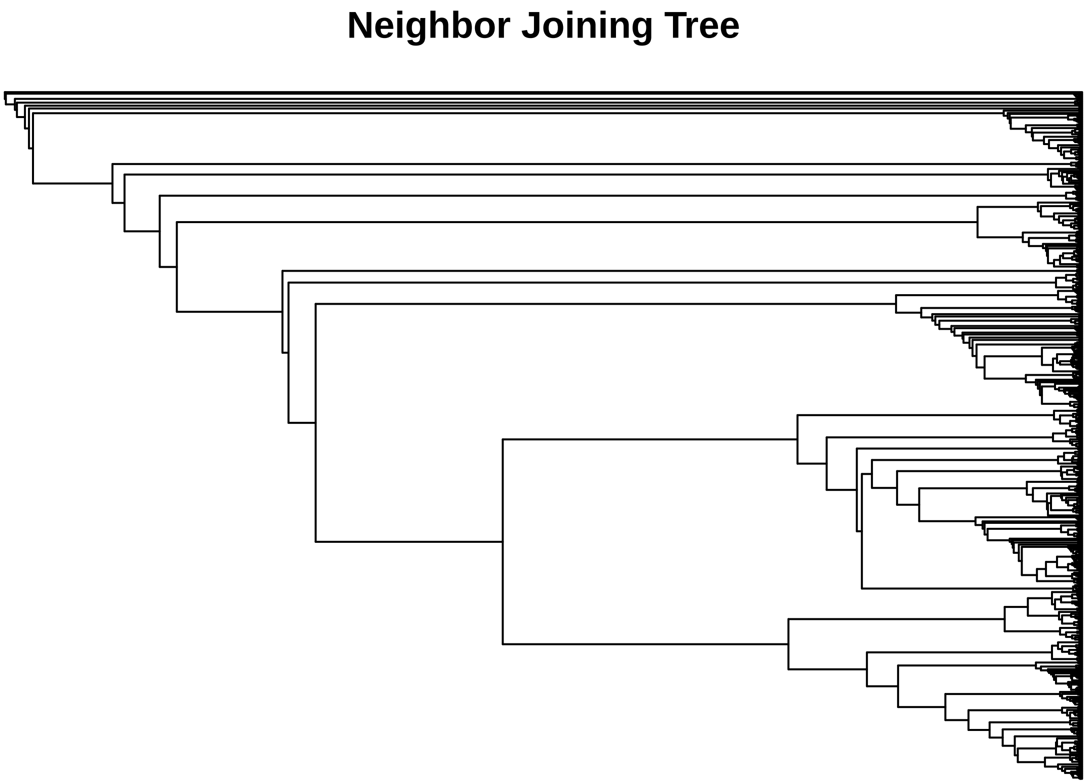
```
env <- read.table("data/20130801_PondDataMod.csv", sep = ",", header=T)
env <- na.omit(env)

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")
```

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$nam <- gsub("\\|.*$", "", 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 = 0.005, 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")]

outgroup1 <- match("Methanosarcina", phy$tip.label)

phy <- root(phy, outgroup1, 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

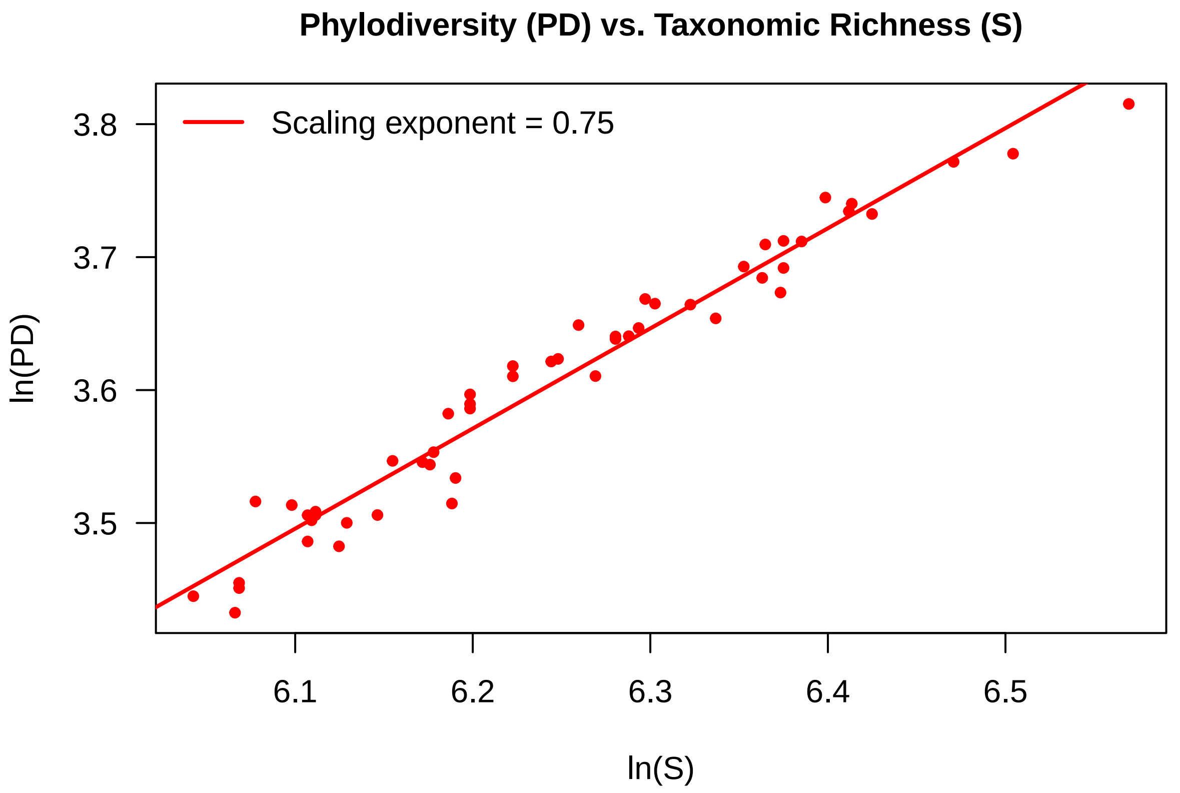
A. Faith's Phylogenetic Diversity (PD)

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.

```
par(mar=c(5,5,3,1)+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)")

fit <- lm("log(pd$PD) ~ log(pd$SR)")
abline(fit, col = "red", lw=2)
exponent <- round(coefficients(fit)[2], 2)
legend("topleft", legend=paste0("Scaling exponent = ", exponent),
        bty="n", lw=2, col="red")
```



```
# plot((pd$SR), (pd$PD))
```

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 should increase with taxonomic richness because when you increase the number of species in a community, you are likely increasing the number of highly divergent taxa. ***Answer 1b***: The natural log of richness increases linearly with the natural of PD. This is a power law relationship. ***Answer 1c***: These estimates would deviate when communities have small numbers of highly divergent taxa or large numbers of highly related taxa. ***Answer 1d***: This confirms the power law relationship between S and PD.

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))
```

```
##       ntaxa   pd.obs pd.rand.mean pd.rand.sd pd.obs.rank   pd.obs.z  pd.obs.p
## BC001   668 43.71912     44.00032  1.0026504          12 -0.2804492 0.4615385
## BC002   587 40.94334     39.94451  0.8141592          22  1.2268239 0.8461538
##       runs
## BC001   25
## BC002   25
```

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

```
##       ntaxa   pd.obs pd.rand.mean pd.rand.sd pd.obs.rank  pd.obs.z   pd.obs.p
## BC001   668 43.71912     42.19462  0.5766974          26  2.643499 1.00000000
## BC002   587 40.94334     42.38857  0.5599776           1 -2.580882 0.03846154
##       runs
## BC001   25
## BC002   25
```

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

```
##       ntaxa   pd.obs pd.rand.mean pd.rand.sd pd.obs.rank   pd.obs.z  pd.obs.p
## BC001   668 43.71912     43.94459  0.2667286           7 -0.8453191 0.2692308
## BC002   587 40.94334     40.68784  0.2801848          19  0.9119070 0.7307692
##       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 that the observed phylodiversity is due to some underlying structure in the community matrix, such as richness or frequency of species, while the alternative hypothesis is that phylodiversity is different (greater?) than expected under this assumption. ***Answer 2b***: The choice of null model appears somewhat important. The `frequency` null model led to the only significant result. This might imply that the frequency of the second sample is unrelated to its phylodiversity, but that species richness might be associated.

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 = F, runs=25)
NRI <- as.matrix(-1* (ses.mpd[,2] - ses.mpd[,3]) / ses.mpd[,4])
rownames(NRI) <- row.names(ses.mpd)
colnames(NRI) <- "NRI"
NRI
```

```
##              NRI
## BC001  -3.2717998
## BC002  -4.9448443
## BC003  -1.5984544
## BC004  -3.0344969
## BC005  -3.4782970
## BC010  -3.1609206
## BC015  -2.1624317
## BC016  -2.0666707
## BC018  -1.6905952
## BC020  -1.7074148
## BC048  -1.7237510
## BC049  -0.9784385
## BC051  -2.9230234
## BC105  -3.3885897
## BC108  -2.6678220
## BC262  -1.9897300
## BCL01  -3.6565198
## BCL03  -1.6939629
## HNF132 -3.0239234
## HNF133 -2.6744060
## HNF134 -3.3140866
## HNF144 -3.8009842
## HNF168 -1.8379915
## HNF185 -3.0523759
## HNF187  0.6847923
## HNF216 -2.1145340
## HNF217 -2.0979602
## HNF221 -2.0496850
## HNF224 -4.8998721
## HNF225 -3.5572146
## HNF229 -1.1227058
## HNF242 -1.7002627
## HNF250 -2.1157205
## HNF267 -1.2440934
## HNF269 -2.5531122
## YSF004 -3.0275548
## YSF117 -1.9377438
## YSF295 -2.2005400
## YSF296 -2.1366734
## YSF298 -3.9686821
## YSF300 -2.2778424
## YSF44  -2.0649416
## YSF45  -2.5786580
## YSF46  -1.2232850
## YSF47  -2.5951234
## YSF65  -0.6431845
## YSF66  -1.1252673
## YSF67  -2.4562221
## YSF69  -1.4051645
## YSF70  -1.3916041
## YSF71  -1.7231939
## YSF74  -2.6070013
```

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 = F, runs=25)
NTI <- as.matrix(-1* (ses.mntd[,2] - ses.mntd[,3]) / ses.mntd[,4])
rownames(NTI) <- row.names(ses.mntd)
colnames(NTI) <- "NTI"
NTI
```

```
##                 NTI
## BC001   0.20351827
## BC002  -1.43766037
## BC003  -0.31553661
## BC004  -1.55696414
## BC005  -2.25833204
## BC010  -1.40980318
## BC015  -1.89147933
## BC016  -0.53337228
## BC018  -1.06489400
## BC020  -1.21723206
## BC048  -1.70852333
## BC049  -0.85249297
## BC051  -2.06541637
## BC105  -1.49065623
## BC108  -1.66949866
## BC262  -0.29341733
## BCL01  -2.05984736
## BCL03   0.09590032
## HNF132 -0.97731472
## HNF133 -1.03787122
## HNF134 -1.18138185
## HNF144 -2.09854056
## HNF168 -1.92765212
## HNF185 -1.23416450
## HNF187 -0.19313010
## HNF216 -2.04068224
## HNF217 -2.46553485
## HNF221 -1.32586571
## HNF224 -1.80922070
## HNF225 -2.04540739
## HNF229 -0.29559653
## HNF242 -2.05155022
## HNF250 -1.13435367
## HNF267  0.78854099
## HNF269  0.27306343
## YSF004 -2.34336898
## YSF117 -1.41630766
## YSF295 -1.43223750
## YSF296 -0.26830681
## YSF298 -1.15665676
## YSF300 -1.09432877
## YSF44  -1.26368236
## YSF45  -1.45909402
## YSF46  -1.00214834
## YSF47  -1.03038131
## YSF65  -0.14631310
## YSF66   0.58647215
## YSF67  -1.40359407
## YSF69   0.02110572
## YSF70  -0.48225440
## YSF71  -0.84283036
## YSF74  -2.23808416
```

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: The NRI finds the average pairwise phylogenetic distance for sample, and compares that to a null value. The sign of the differnce between observed and expected values indicates whether overdispersion (-) or clustering (+) is indicated. ***Answer 3b***: The NTI is a similar process on the back end, but it uses the average distance to the closest phylogenetic relative in a sample instead of the average pairwise phylogenetic distance. ***Answer 3c***: The vast majority of the samples are phylogenetically overdispersed, though NRI predicts more overdispersion than NTI. More values are within 2 SD of 0, as well, though I am not sure if that necessarily means they are not significantly overdispersed.

```
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"
NRI
```

```
##                   NRI
## BC001   0.216817767
## BC002   0.592208579
## BC003   1.169037513
## BC004  -0.002543977
## BC005   0.472832972
## BC010   0.280687986
## BC015   0.260560593
## BC016   0.572564065
## BC018   0.452464528
## BC020   0.501091771
## BC048   0.273058190
## BC049   0.398272369
## BC051  -0.281758151
## BC105  -0.079720098
## BC108   0.085335882
## BC262   0.045604744
## BCL01   0.039683786
## BCL03  -0.102466643
## HNF132  0.202414969
## HNF133  0.209472521
## HNF134  0.477846074
## HNF144 -0.130356753
## HNF168  0.038635208
## HNF185  0.445404617
## HNF187  0.503145306
## HNF216  0.422906161
## HNF217  0.025541708
## HNF221 -0.162338834
## HNF224  0.381746912
## HNF225  0.371705580
## HNF229  0.079399000
## HNF242  0.203129948
## HNF250  0.048928063
## HNF267 -0.026269730
## HNF269  0.011333664
## YSF004 -0.410201820
## YSF117  0.786858087
## YSF295 -0.758304791
## YSF296  1.113892736
## YSF298  0.930489592
## YSF300  0.453860317
## YSF44   0.707629078
## YSF45   0.821555300
## YSF46   1.931818026
## YSF47   0.397230251
## YSF65   0.629518567
## YSF66  -0.368816008
## YSF67  -0.005337811
## YSF69   0.053661310
## YSF70   0.013721966
## YSF71   0.731539954
## YSF74   1.252026960
```

```
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"
NTI
```

```
##              NTI
## BC001   1.2507379
## BC002   1.7776099
## BC003   0.8343315
## BC004   1.3331164
## BC005   2.1325522
## BC010   0.6397645
## BC015   1.3505183
## BC016   1.9474777
## BC018   1.2512330
## BC020   1.2597249
## BC048   1.4746480
## BC049   1.6490347
## BC051   2.0341145
## BC105   1.9726723
## BC108   1.5341094
## BC262   1.1791794
## BCL01   1.6453934
## BCL03   0.8600219
## HNF132  1.4598375
## HNF133  1.0943056
## HNF134  1.7941154
## HNF144  1.1231933
## HNF168  0.7366733
## HNF185  1.4081635
## HNF187  0.7358220
## HNF216  0.3511728
## HNF217  0.4888804
## HNF221  0.6209245
## HNF224  2.0360371
## HNF225  0.4743988
## HNF229  1.6968079
## HNF242  1.8415605
## HNF250  1.3300293
## HNF267  0.8543402
## HNF269  0.9981824
## YSF004  0.0782584
## YSF117  1.2424137
## YSF295 -1.4639050
## YSF296  1.7912214
## YSF298  1.5127345
## YSF300  1.4445661
## YSF44   0.9713806
## YSF45   0.9765620
## YSF46   1.0795744
## YSF47   1.1316346
## YSF65   1.7592232
## YSF66   1.4942783
## YSF67   1.3797662
## YSF69   1.0062254
## YSF70   1.0176221
## YSF71   1.3419902
## YSF74   1.1675168
```

> ***Answer 3d***: This greatly alters the results. Both metrics give more positive values, and values that are smaller in magnitude, which either suggests clustering or no apparent pattern.

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.

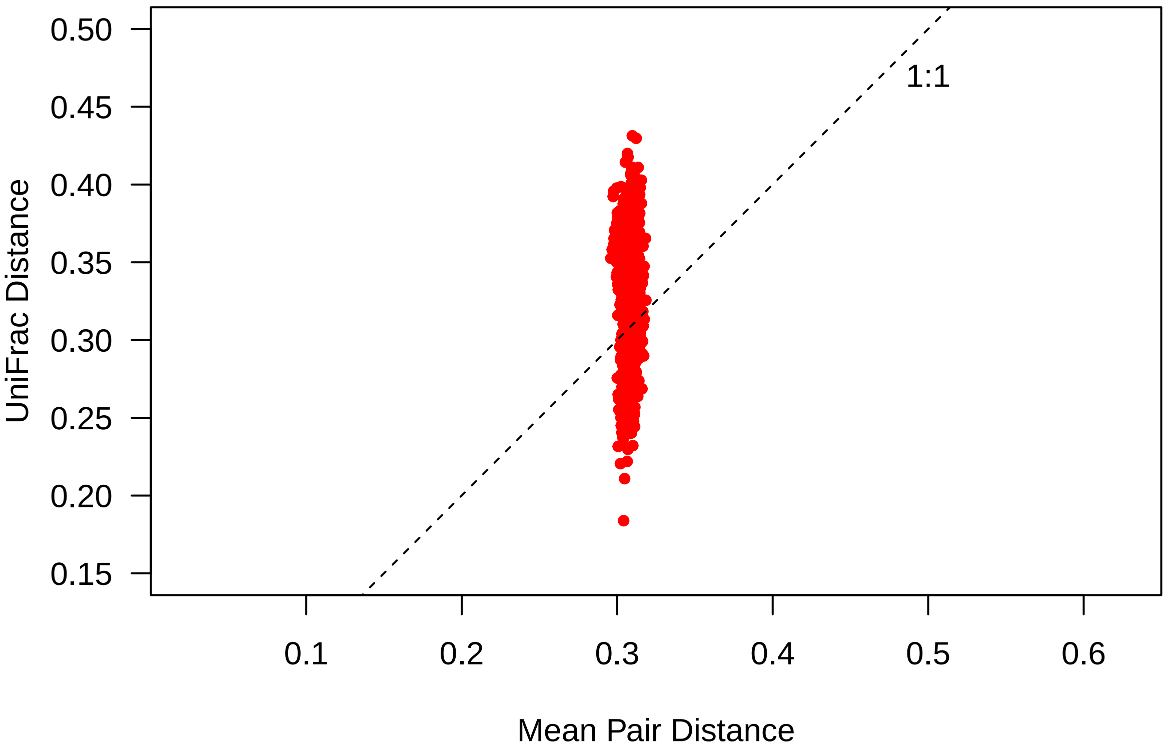
```
dist.mp <- comdist(comm, phydist)
```

```
## [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="UniFrac Distance")
abline(b=1, a=0, lty=2)
text(0.5, 0.47, "1:1")
```



Question 4:

a. In your own words describe Mean Pair Distance, UniFrac distance, and the difference between them.
b. Using the plot above, describe the relationship between Mean Pair Distance and UniFrac distance. Note: we are calculating unweighted phylogenetic distances (similar to incidence based measures). That means that we are not taking into account the abundance of each taxon in each site.
c. Why might MPD show less variation than UniFrac?

   Answer 4a: Mean pairwise distance looks at all possible pairs of tips between samples and calculates the average phylogenetic distance of all of those pairs. UniFrac considers distance on branches shared by different samples and those that are unique to specific communities. Distance here is the proportion of total distance that is unshared. These metrics thus differ in that UniFrac is scaled by total distance, while MPD is not, and UniFrac does not treat all distance between taxa as equal. ***Answer 4b***: There appears to be no variation in mean pair distance, despite large variation in UniFrac distance. Thus, there is no relation between the two. ***Answer 4c***: Perhaps if there is low variance in pairwise distances, but high variance in the amount of shared branch length? I am not exactly sure how this would physically manifest in tree, however.

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.

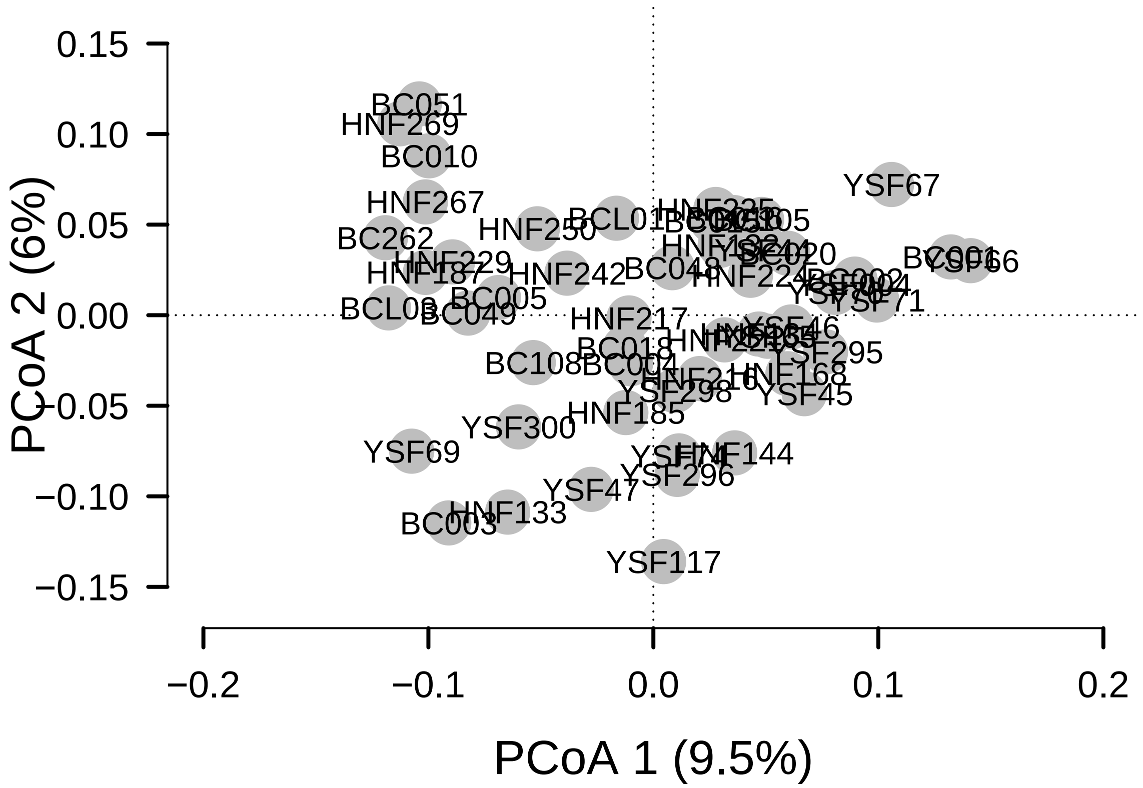
In the R code chunk below, do the following:
1. perform a PCoA based on the UniFrac distances, and
2. calculate the explained variation for the first three PCoA axes.

```
pond.pcoa <- cmdscale(dist.uf, eig=T, k=3)
eig.sum <- sum(pond.pcoa$eig)

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

sum.eig <- explainvar1+explainvar2+explainvar3
```

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

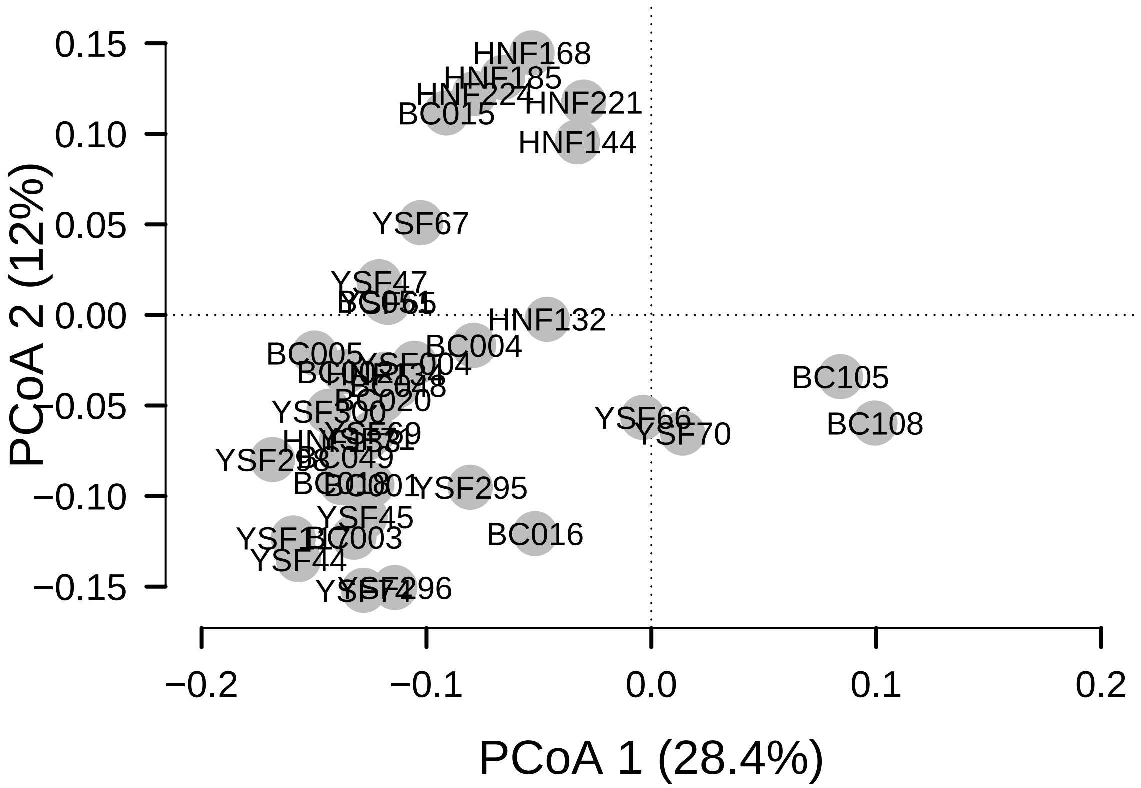
In the R code chunk below, do the following:
1. plot the PCoA results using either the R base package or the `ggplot` package,
2. include the appropriate axes,
3. add and label the points, and
4. customize the plot.

```
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=paste0("PCoA 1 (", explainvar1,"%)"),
     ylab=paste0("PCoA 2 (", explainvar2,"%)"),
     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=2, labels=T, lwd.ticks=2, cex.axis=1.2, las=1)
abline(h=0, v=0, lty=3)

points(pond.pcoa$points[,1], pond.pcoa$points[,2], pch=19, cex=3, bg="grey",col="grey")
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.

```r
pond.db <- vegdist(comm, method = "bray")
pond.pcoa <- cmdscale(pond.db, eig=T, k=3)
eig.sum <- sum(pond.pcoa$eig)

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

sum.eig <- 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(-0.16,0.16),
     xlab=paste0("PCoA 1 (", explainvar1,"%)"),
     ylab=paste0("PCoA 2 (", explainvar2,"%)"),
     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=2, labels=T, lwd.ticks=2, cex.axis=1.2, las=1)
abline(h=0, v=0, lty=3)

points(pond.pcoa$points[,1], pond.pcoa$points[,2], pch=19, cex=3, bg="grey",col="grey")
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***: There appears to be more clustering in taxonomic approach, as well as a higher proportion of total variation explained. Perhaps then in this system, based on the genes we've used to construct the phylogeny, we cannot determine the primary drivers of beta diversity.

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  0.015 *
## 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
```

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: 0.056
##
## Upper quantiles of permutations (null model):
##   90%   95% 97.5%   99%
## 0.130 0.163 0.201 0.226
## 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 <- dbrda(dist.uf ~ ., as.data.frame(scale(envs)))

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

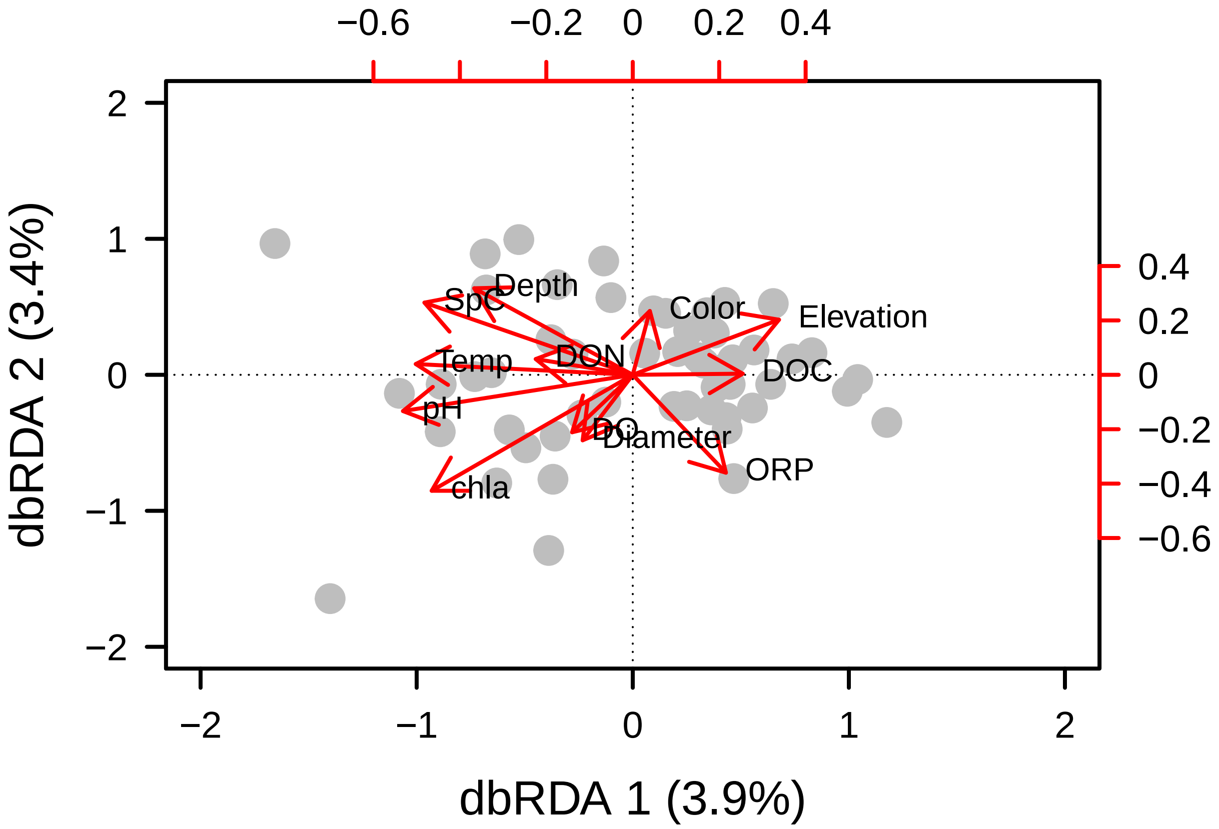
```
## Permutation test for dbrda under reduced model
## Forward tests for axes
## Permutation: free
## Number of permutations: 999
##
## Model: dbrda(formula = dist.uf ~ Elevation + Diameter + Depth + ORP + Temp + SpC + DO + pH + Color +
##           Df SumOfSqs      F Pr(>F)
## dbRDA1     1  0.10566 2.0152  0.455
## dbRDA2     1  0.09258 1.7658  0.634
## dbRDA3     1  0.07555 1.4409  0.979
## dbRDA4     1  0.06677 1.2735  1.000
## 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  1.000
## Residual 39  2.04482
```

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

```
##
## ***VECTORS
##
##             dbRDA1   dbRDA2     r2 Pr(>r)
## Elevation  0.77670  0.62986 0.0959  0.070 .
## Diameter  -0.27972 -0.96008 0.0541  0.259
## Depth     -0.63137  0.77548 0.1756  0.007 **
## ORP        0.41879 -0.90808 0.1437  0.018 *
## Temp      -0.98250  0.18628 0.1523  0.018 *
## SpC       -0.77101  0.63682 0.2087  0.006 **
## DO        -0.39318 -0.91946 0.0464  0.295
## pH        -0.96210 -0.27270 0.1756  0.007 **
## Color      0.06353  0.99798 0.0464  0.313
## chla      -0.60392 -0.79704 0.2626  0.009 **
## DOC        0.99847 -0.05526 0.0382  0.378
## DON       -0.91633  0.40042 0.0339  0.440
## ---
## 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) * 1
dbrda.explainvar2 <- round(ponds.dbrda$CCA$eig[2]/sum(c(ponds.dbrda$CCA$eig, ponds.dbrda$CA$eig)),3) * 1


vals <- scores(ponds.dbrda, display="wa")
par(mar=c(5,4,2,4) + 0.1)
plot(vals,
     xlim=c(-2, 2), ylim=c(-2, 2),
     xlab=paste0("dbRDA 1 (", dbrda.explainvar1, "%)"),
     ylab=paste0("dbRDA 2 (", dbrda.explainvar2, "%)"),
     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=2, labels=T, lwd.ticks=2, cex.axis=1.2, las=1)
abline(h=0, v=0, lty=3)
box(lwd=2)

points(vals, pch=19, cex=2, bg="grey", col="grey")

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=4, 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***: Environmental conditions appear to map onto phylogenetic diversity in these ponds, even though the total variance explained here is quite low. Chlorophyll a, pH, temperature, and elevation are large players, among others.

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 my project on rhizobia and their phages, it might be interesting to sequence the genomes of some rhizobiophages and see how phylogenetic distance maps onto environmental variables. Specifically, it would be interesting if phylogenetic diversity is associated with N-fertilization at the plots at KBS. If there are certain genomic regions in the phage genomes that return topologies that are fall along the N treatment, that would be interesting, and then we could explore the genes responsible for this. We could also consider phylogenetic clustering vs overdispersion in host range, and how that might map onto processes of antagonistic coevolution in this system.