9.Phylogenetic Diversity - Communities

Anna Lennon

05 March, 2025

OVERVIEW

Complementing taxonomic measures of α- and β-diversity with evolutionary information yields insight into a broad range of biodiversity issues including conservation, biogeography, and community assembly. In this worksheet, you will be introduced to some commonly used methods in phylogenetic community ecology.

After completing this assignment you will know how to:

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

Directions:

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

The completed exercise is due on **Wednesday, March 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.

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

```
## [1] "/cloud/project/QB2025_ALennon/Week7-PhyloCom"
```

```r
setwd("/cloud/project/QB2025_ALennon/Week7-PhyloCom")
# package.list <- c('picante', 'ape', 'seqinr', 'vegan', 'fossil',
#                   'reshape', 'devtools', 'BiocManager', 'ineq',
#                   'labdsv', 'matrixStats', 'pROC')
# for (package in package.list) {
#   if (!require(package, character.only = TRUE, quietly = TRUE)) {
#     install.packages(package, repos='http://cran.us.r-project.org')
#     library(package, character.only = TRUE)
#   }
# }
#
# install.packages("reshape")

library(picante)
```

```
## Loading required package: ape

## Loading required package: vegan

## Loading required package: permute

## Loading required package: lattice

## Loading required package: nlme
```

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

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

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

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

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

```r
library(vegan)
library(fossil)
```

```
## Loading required package: sp

## Loading required package: maps

## Loading required package: shapefiles

## Loading required package: foreign

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

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

```
## Loading required package: usethis
```

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

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

```
library(BiocManager)
```

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

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

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

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

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

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

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

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

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

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

```
library(matrixStats)
```

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

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

```
library(pROC)
```

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

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

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

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

2) DESCRIPTION OF DATA

need to discuss data set from spatial ecology!

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

3) LOAD THE DATA

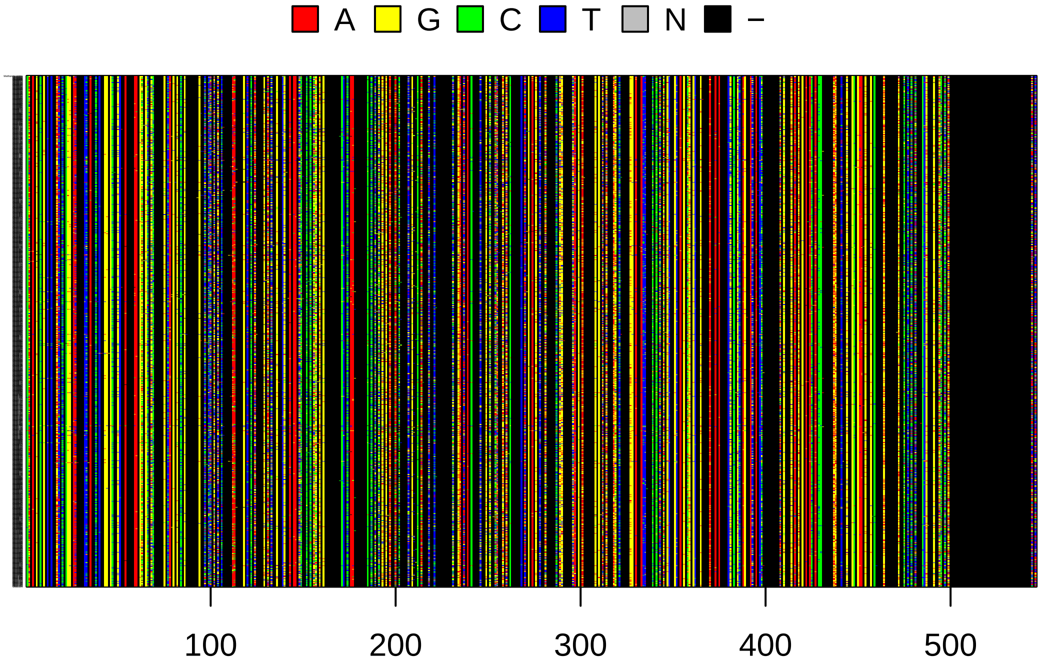
In the R code chunk below, do the following:
1. load the environmental data for the Brown County ponds (*20130801_PondDataMod.csv*),
2. load the site-by-species matrix using the `read.otu()` function,
3. subset the data to include only DNA-based identifications of bacteria,
4. rename the sites by removing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

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

#ss and cleaning
comm.ss <- read.otu(shared = "./data/INPonds.final.rdp.shared", cutoff = "1")
comm.ss <- comm.ss[grep("*-DNA", rownames(comm.ss)), ]
rownames(comm.ss) <- gsub("\\-DNA", "", rownames(comm.ss))
rownames(comm.ss) <- gsub("\\_", "", rownames(comm.ss))
comm.ss <- comm.ss[rownames(comm.ss)  %in% env$Sample_ID, ]
comm.ss <- comm.ss[ , colSums(comm.ss) > 0]

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

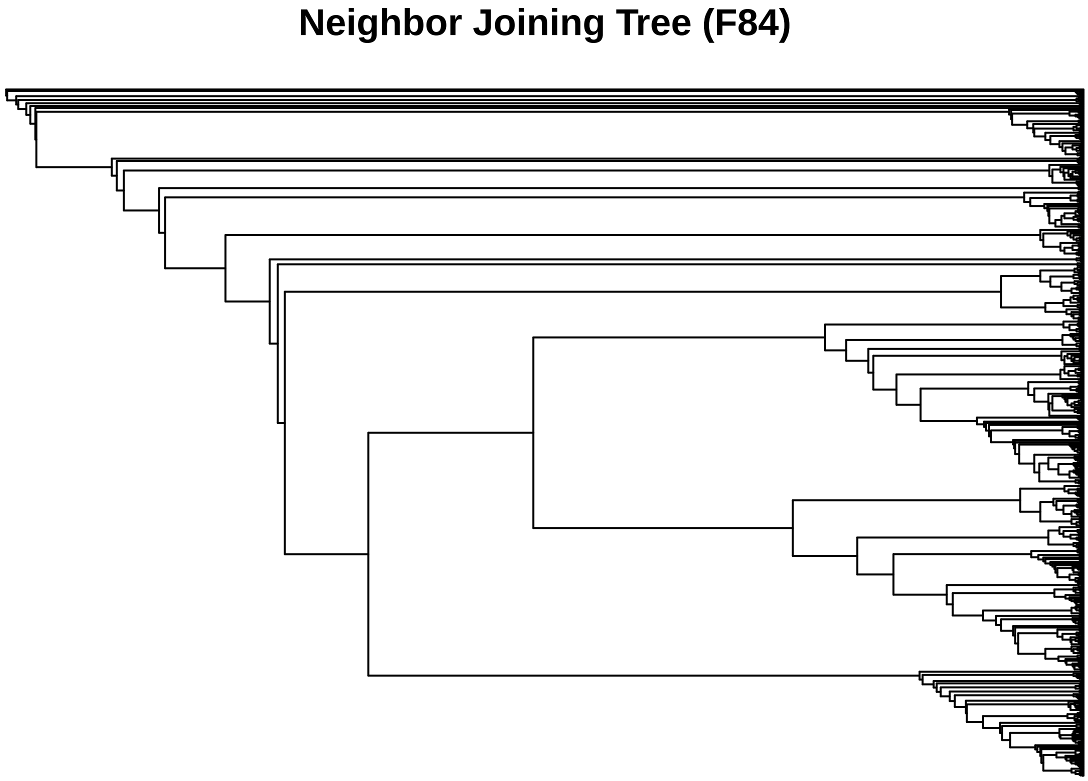
```
## Warning in type.convert.default(as.character(x)): 'as.is' should be specified
## by the caller; using TRUE
## Warning in type.convert.default(as.character(x)): 'as.is' should be specified
## by the caller; using TRUE
## Warning in type.convert.default(as.character(x)): 'as.is' should be specified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

Next, in the R code chunk below, do the following:
1. load the FASTA alignment for the bacterial operational taxonomic units (OTUs),
2. rename the OTUs by removing everything before the tab (\t) and after the bar (|),
3. import the *Methanosarcina* outgroup FASTA file,
4. convert both FASTA files into the DNAbin format and combine using `rbind()`,
5. visualize the sequence alignment,
6. using the alignment (with outgroup), pick a DNA substitution model, and create a phylogenetic distance matrix,
7. using the distance matrix above, make a neighbor joining tree,
8. remove any tips (OTUs) that are not in the community data set,
9. plot the rooted tree.

```r
ponds.cons <- read.alignment(file = "./data/INPonds.final.rdp.1.rep.fasta",
                             format = "fasta")
ponds.cons$nam <- gsub(".*\t", "", ponds.cons$nam)
ponds.cons$nam <- gsub("\\|.*", "", ponds.cons$nam)
#Visualize
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)
```



```r
#Distance Matrix
seq.dist.FC <- dist.dna(DNAbin, model = "F84", pairwise.deletion = FALSE)
phy.all <- bionj(seq.dist.FC)
phy <- drop.tip(phy.all, phy.all$tip.label[!phy.all$tip.label %in%
              c(colnames(comm.ss), "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 (F84)", "phylogram",
      show.tip.label = FALSE, use.edge.length = FALSE,
      direction = "right", cex = 0.6, label.offset = 1)
```



4) PHYLOGENETIC ALPHA DIVERSITY

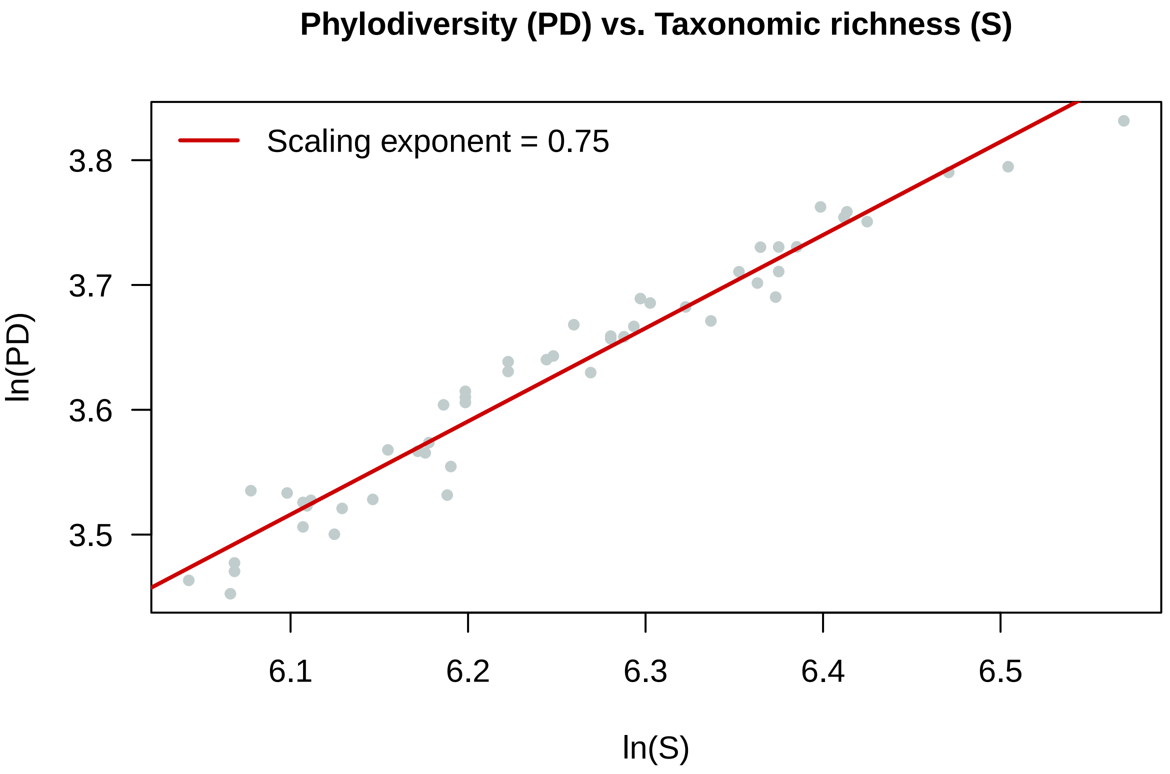
A. Faith's Phylogenetic Diversity (PD)

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

```
pd <- pd(comm.ss, 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.

```
par(mar = c(5, 5, 4, 1) + 0.1)

plot(log(pd$S), log(pd$PD),
     pch = 20, col = "azure3", las = 1,
     xlab = "ln(S)", ylab = "ln(PD)", cex.main = 1,
     main="Phylodiversity (PD) vs. Taxonomic richness (S)")

fit <- lm('log(pd$PD) ~ log(pd$S)')
abline(fit, col = "red3", lw = 2)
exponent <- round(coefficients(fit)[2], 2)
legend("topleft", legend=paste("Scaling exponent = ", exponent, sep = ""),
       bty = "n", lw = 2, col = "red3")
```



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***: Faith's D is the sum of all the branch lengths, which are the measurements of relatedness between species. Therefore, Faith's D reflects taxonomic richness present based on phylogeny. For example, with more branches, Faith's D is higher reflecting in higher taxonmic richness. As shown on the graph, Faith's D and species richness are related but the relationship levels off eventually. ***Answer 1b***:Species richness and Faith's D may diverge when comparing a small group very distantly related species which have very long branch lengths. This would result in low species richness but a high Faith's D value. ***Answer 1c***: The PD-S scaling exponet is .75 resulting in the apperance of a linear relationship. However, given that the scaling exponent is less than one, the relationship is not linear rather increases expoentntially before plateuing. Therefore, when interpreting the graph it should be noted that the relationship between PD and S is not a linear positive correlation.

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

```
##       ntaxa   pd.obs pd.rand.mean pd.rand.sd pd.obs.rank    pd.obs.z  pd.obs.p
## BC001   668 44.46687     44.51077  0.5687846          11 -0.07718791 0.4230769
## BC002   587 41.69719     40.60006  1.0680519          21  1.02722615 0.8076923
##       runs
## BC001   25
## BC002   25
```

```
ses.pd.frequency
```

```
##       ntaxa   pd.obs pd.rand.mean pd.rand.sd pd.obs.rank  pd.obs.z   pd.obs.p
## BC001   668 44.46687     43.03999  0.6666696          26  2.140307 1.00000000
## BC002   587 41.69719     43.06577  0.6707917           2 -2.040244 0.07692308
##       runs
## BC001   25
## BC002   25
```

```
ses.pd.phylo
```

```
##       ntaxa   pd.obs pd.rand.mean pd.rand.sd pd.obs.rank   pd.obs.z  pd.obs.p
## BC001   668 44.46687     44.59838  0.7921591          14 -0.1660131 0.5384615
## BC002   587 41.69719     40.26109  0.6802369          26  2.1111742 1.0000000
##       runs
## BC001   25
## BC002   25
```

```
?ses.pd
```

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 sample is as diverse as expected while the alternative is that the community more more phylogenetically diverse than expected under the null model. From ses.pd, the sample's phylogenetic diversity is being tested under a null distribution to see if it aligns or not. ***Answer 2b***: The different null models did not changed the observed ses.pd values. The three models I chose all focused on randomizing the data matrix. As a result, another model, such as taxa.labels, may result in a different PD value. ### 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.ss, phydist, null.model = "taxa.labels",
                   abundance.weighted = FALSE, runs = 25)
NRI <- as.matrix(-1 * ((ses.mpd[,2] - ses.mpd[,3]) / ses.mpd[,4]))
rownames(NRI) <- row.names(ses.mpd)
colnames(NRI) <- "NRI"
NRI
```

```
##              NRI
## BC001  -2.5800422
## BC002  -3.6828836
## BC003  -1.7993361
## BC004  -2.8342097
## BC005  -2.6903552
## BC010  -2.6036742
## BC015  -2.3460102
## BC016  -1.8947377
## BC018  -1.5732725
## BC020  -1.5524381
## BC048  -1.7228049
## BC049  -0.1128527
## BC051  -2.0366124
## BC105  -2.4746164
## BC108  -2.2743678
## BC262  -1.5241209
## BCL01  -3.4459636
## BCL03  -1.0147449
## HNF132 -3.0318479
## HNF133 -2.7021505
## HNF134 -2.8320763
## HNF144 -3.5739654
## HNF168 -1.9646459
## HNF185 -3.1829153
## HNF187  0.7394954
## HNF216 -1.7901124
## HNF217 -1.6615661
## HNF221 -1.9718169
## HNF224 -3.1634164
## HNF225 -2.2934855
## HNF229 -0.8286628
## HNF242 -1.3517015
## HNF250 -1.7725800
## HNF267 -0.8180883
## HNF269 -1.7143983
## YSF004 -3.5549170
## YSF117 -2.4047378
## YSF295 -1.6454276
## YSF296 -2.3039319
## YSF298 -3.8431369
## YSF300 -2.1724331
## YSF44  -1.2094130
## YSF45  -2.2134224
## YSF46  -1.1963969
## YSF47  -1.8906019
## YSF65  -0.3949058
## YSF66  -1.0309407
## YSF67  -2.4014393
## YSF69  -1.3358843
## YSF70  -1.5972910
## YSF71  -1.2749165
## YSF74  -2.0916482
```

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.ss, phydist, null.model = "taxa.labels",
                     abundance.weighted = FALSE, runs = 25)
NTI <- as.matrix(-1 * ((ses.mntd[,2] - ses.mntd[,3]) / ses.mntd[,4]))
rownames(NTI) <- row.names(ses.mntd)
colnames(NTI) <- "NTI"
NTI
```

```
##                NTI
## BC001   0.86177963
## BC002  -0.94598791
## BC003  -0.26932271
## BC004  -1.87945441
## BC005  -1.67279155
## BC010  -0.68306534
## BC015  -1.05102080
## BC016   0.07930508
## BC018  -0.74639491
## BC020  -0.67974236
## BC048  -1.78516289
## BC049  -0.54345397
## BC051  -1.38931505
## BC105  -0.94634830
## BC108  -0.89173556
## BC262   0.12315732
## BCL01  -1.32393827
## BCL03   0.37259173
## HNF132 -0.54707755
## HNF133 -0.94173962
## HNF134 -0.91198945
## HNF144 -2.18744497
## HNF168 -1.46515215
## HNF185 -0.72361905
## HNF187  0.06097667
## HNF216 -2.41194779
## HNF217 -1.46776428
## HNF221 -0.60201082
## HNF224 -1.79941635
## HNF225 -2.34560615
## HNF229  0.36864452
## HNF242 -1.60454357
## HNF250 -0.60992621
## HNF267  1.10896626
## HNF269  0.57346038
## YSF004 -2.02288390
## YSF117 -1.42681073
## YSF295 -1.14049850
## YSF296  0.19063946
## YSF298 -1.02264092
## YSF300 -0.94271157
## YSF44  -0.64711687
## YSF45  -1.42643276
## YSF46  -0.36180079
## YSF47  -1.26853540
## YSF65   0.39445896
## YSF66   1.41025515
## YSF67  -0.67339595
## YSF69   0.26195947
## YSF70   0.08499404
## YSF71  -0.18596598
## YSF74  -1.34256886
```

Question 3:

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

> ***Answer 3a***:NRI is calculated by taking the observed MPD and subtracting the random means generated MPD and dividing that by the standard deviation MPD generated under the null model. ***Answer 3b***:NTI is very similar to NRI but utilizes MNND, or the mean phyogenetic distance between all the taxa and their neighbor, rather than the MPD. As a result, NTI calculates the difference between observed values of MNND and the observed random means and divises it by the randomizes standard deviation of MNND. ***Answer 3c***: FOr both NRI and NTI values, a majority are negative which indicates phylogenetic overdispersion or close taxa are less related than expected. A notable exception is that the NTI value for YST66 is positive while the NRI value is negative. A positive NTI value indicates phylogenetic clustering. Additionally, there are a few values that are very close to zero displaying these samples are dispersed according to the null model. ***Answer 3d***:

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.

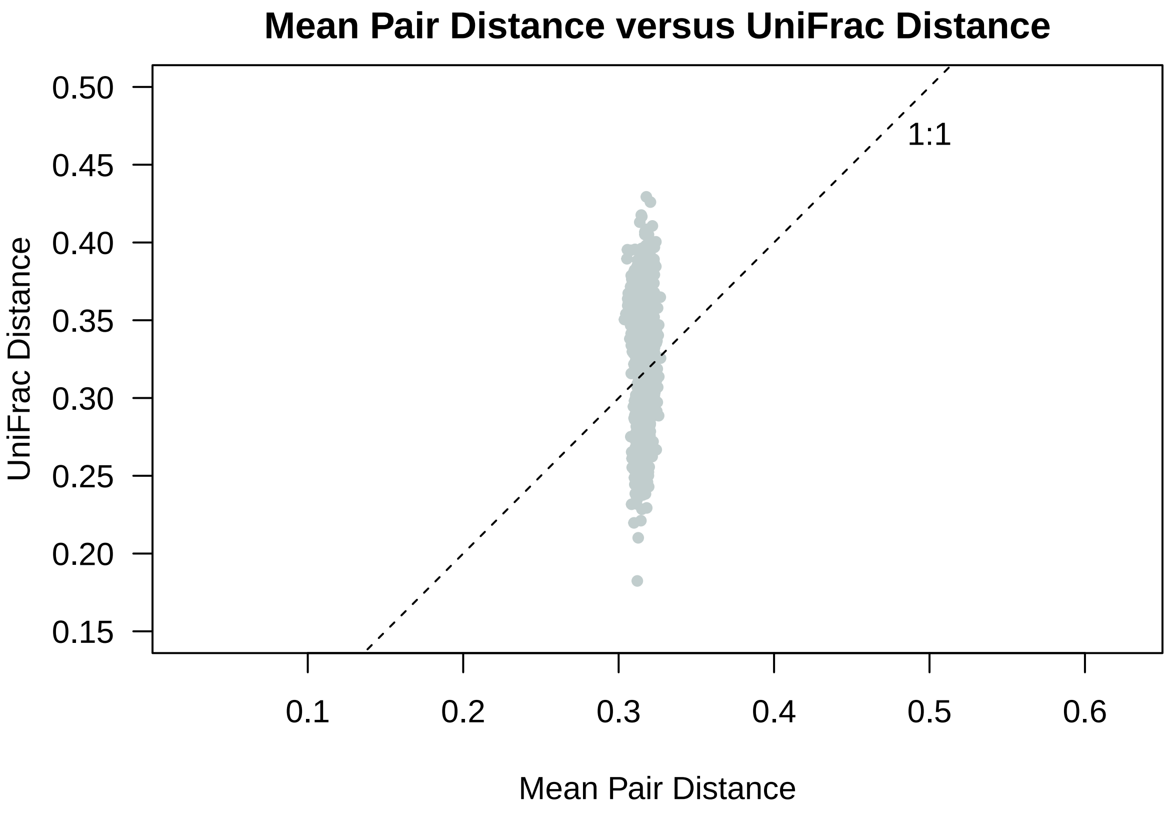
```r
# Mean Pairwise Distance
dist.mp <- comdist(comm.ss, phydist)
```

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

```r
# UniFrac
dist.uf <- unifrac(comm.ss, phy)
```

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

```r
par(mar = c(5, 5, 2, 1) + 0.1)
plot(dist.mp, dist.uf,
     pch = 20, col = "azure3", las = 1, asp = 1, xlim = c(0.15, 0.5), ylim = c(0.15, 0.5),
     xlab = "Mean Pair Distance", ylab = "UniFrac Distance", main = "Mean Pair Distance versus UniFrac
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 takes the phyolgenetic distance (ie branch length) between two taxa as a metric as the distance between the two samples in terms of similarity or dissimilarity. Comparatively, UniFrac distance calculates the summation of all the unshared branch lengths and divides it by the summation of all the tree's branch lengths to determine if samples are more similar or dissimilar. The key difference is that UniFrac also includes unshared branch lengths in its calculation compared to Mean Pair distance that only includes the shared branches. ***Answer 4b***: The graph displays a vertical clustering pattern, resulting in a linear slope of undefined meaning that Mean Pair Distance and UniFrac unweighted are completely unrelated. ***Answer 4c***: Given that mean pair distance only looks are shared branches, there will be less variation shown compared to UniFrac that incorporates other branch lengths in its model.

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.

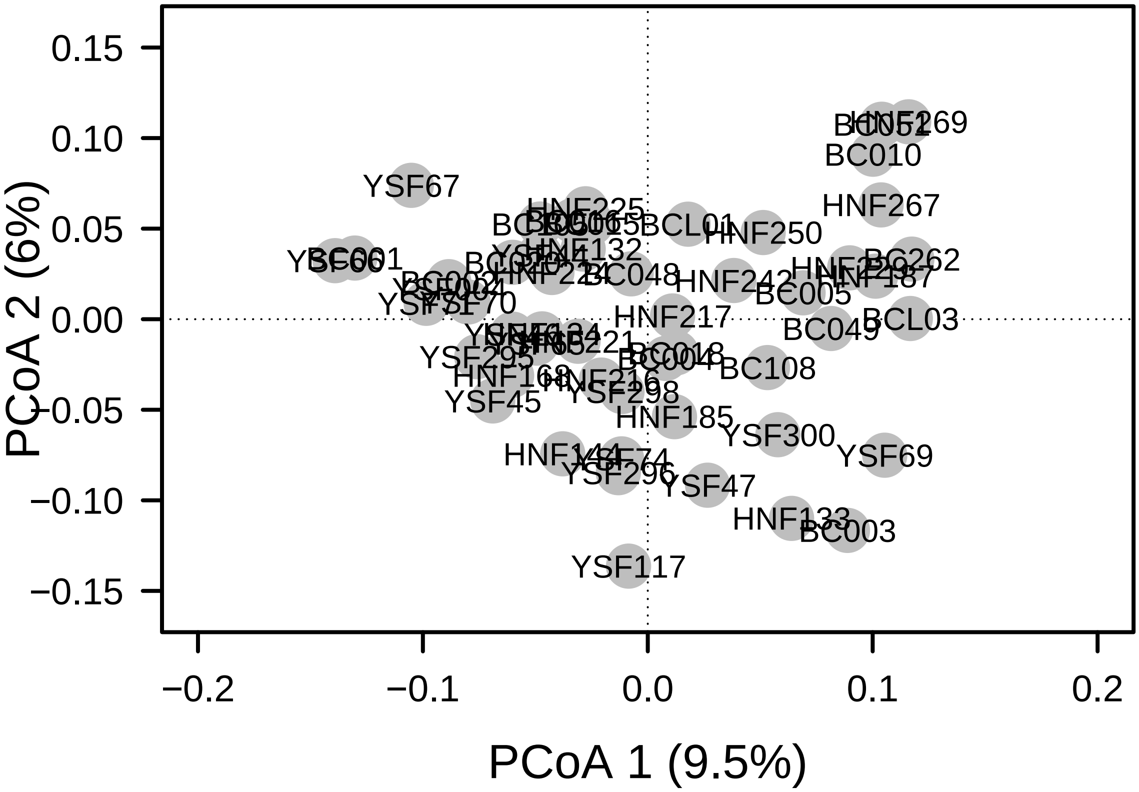
In the R code chunk below, do the following:
1. perform a PCoA based on the UniFrac distances, and
2. calculate the explained variation for the first three PCoA axes.

```
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.

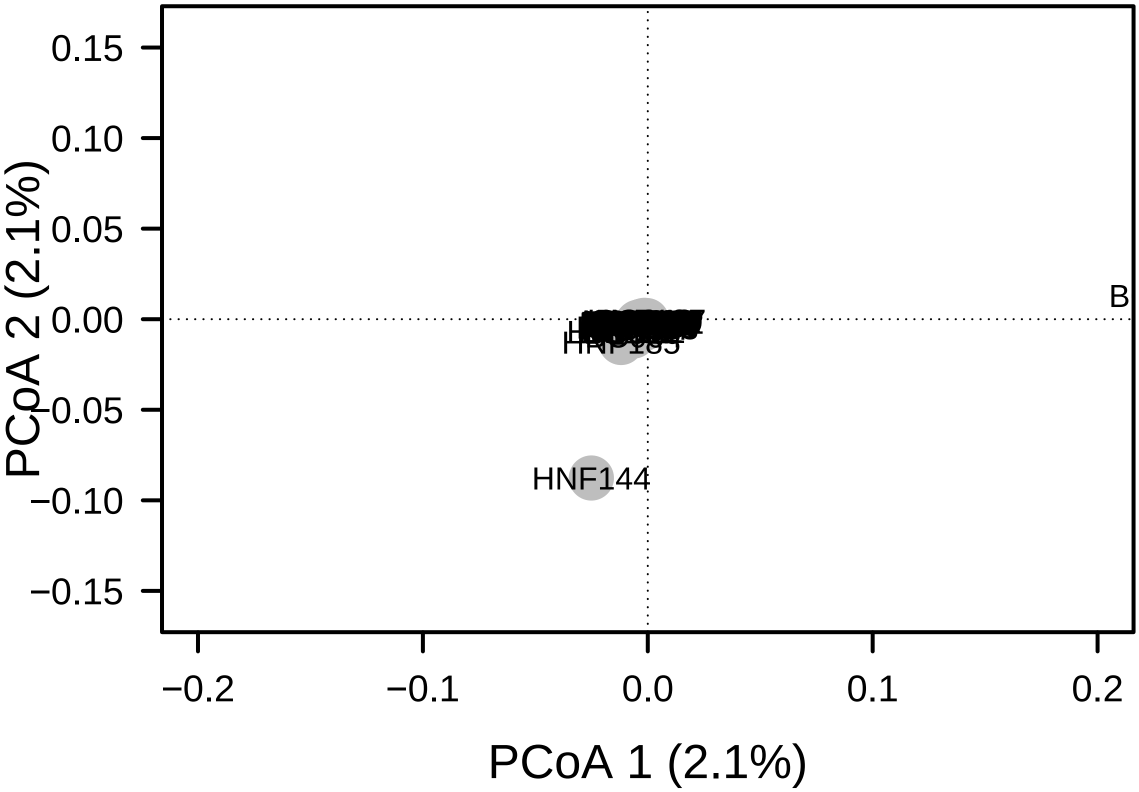
In the R code chunk below, do the following:
1. plot the PCoA results using either the R base package or the `ggplot` package,
2. include the appropriate axes,
3. add and label the points, and
4. customize the plot.

```
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.

```r
pond.pcoa.mp <- cmdscale(dist.mp, eig = T, k = 3)

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

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

# Initiate Plot
plot(pond.pcoa.mp$points[ ,1], pond.pcoa.mp$points[ ,2],
     xlim = c(-0.2, 0.2), ylim = c(-.16, 0.16),
     xlab = paste("PCoA 1 (", explainvar1.mp, "%)", sep = ""),
     ylab = paste("PCoA 2 (", explainvar2.mp, "%)", 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.mp$points[ ,1], pond.pcoa.mp$points[ ,2],
       pch = 19, cex = 3, bg = "gray", col = "gray")
text(pond.pcoa.mp$points[ ,1], pond.pcoa.mp$points[ ,2],
     labels = row.names(pond.pcoa.mp$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***: The phylogenetically based PcoA explains very little of the variation compared to taxonomic ordination. In addition the the axis labels detailing the percent variation, in both PCoA visualizations there is very little clustering patterns that can be determined from the clustering. Therefore, phylogenetic information is not essential to this system when understanding the variation observed between samples.

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

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


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

ii. Continuous Approach

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

```
envs <- env[, 5:19]
envs <- envs[, -which(names(envs) %in% c("TDS", "Salinity", "Cal_Volume"))]

#distance matrix
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: 0.158
##       Significance: 0.068
##
## Upper quantiles of permutations (null model):
##   90%   95% 97.5%   99%
## 0.136 0.167 0.191 0.239
## Permutation: free
## Number of permutations: 999
```

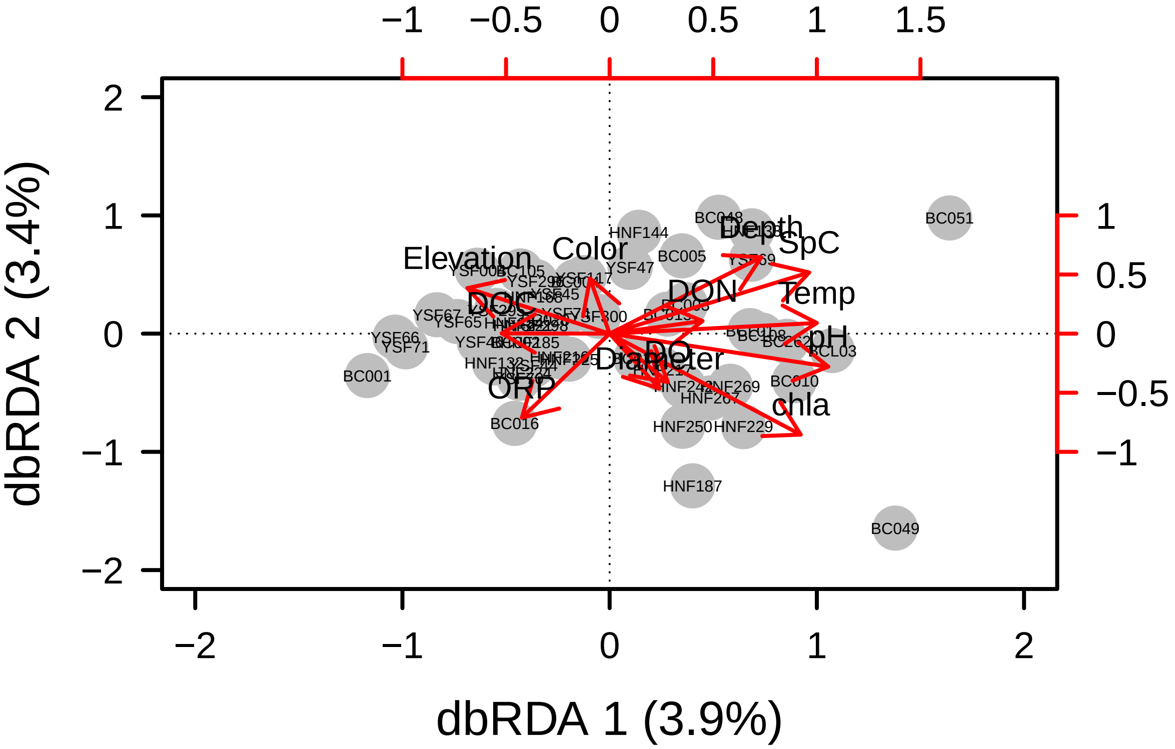
Last, conduct a distance-based Redundancy Analysis (dbRDA).

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 +
##          Df SumOfSqs      F Pr(>F)
## dbRDA1    1  0.10484 2.0210  0.439
## dbRDA2    1  0.09211 1.7755  0.614
## dbRDA3    1  0.07416 1.4295  0.973
## dbRDA4    1  0.06661 1.2840  0.998
## dbRDA5    1  0.05635 1.0861  1.000
## dbRDA6    1  0.05196 1.0016
## dbRDA7    1  0.04699 0.9058
## dbRDA8    1  0.03877 0.7474
## dbRDA9    1  0.03709 0.7150
## dbRDA10   1  0.03180 0.6130
## dbRDA11   1  0.02837 0.5469
## dbRDA12   1  0.02451 0.4725
## Residual 39  2.02322
```

```
ponds.fit <- envfit(ponds.dbrda, envs, perm = 999)
#ponds.fit
dbrda.explainvar1 <- round(ponds.dbrda$CCA$eig[1] /
                    sum(c(ponds.dbrda$CCA$eig, ponds.dbrda$CA$eig)), 3) * 100
dbrda.explainvar2 <- round(ponds.dbrda$CCA$eig[2] /
                    sum(c(ponds.dbrda$CCA$eig, ponds.dbrda$CA$eig)), 3) * 100

#Plot
ponds_scores <- vegan::scores(ponds.dbrda, display = "sites")
par(mar = c(5, 5, 4, 4) + 0.1)
plot(ponds_scores, xlim = c(-2, 2), ylim = c(-2, 2),
  xlab = paste("dbRDA 1 (", dbrda.explainvar1, "%)", sep = ""),
  ylab = paste("dbRDA 2 (", dbrda.explainvar2, "%)", sep = ""),
  pch = 16, cex = 2.0, type = "n", cex.lab = 1.5, cex.axis = 1.2, axes = FALSE)
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)
pond_scores <- vegan::scores(ponds.dbrda, display = "sites")

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

text(pond_scores,
     labels = rownames(pond_scores),
     cex = 0.5)
vectors <- vegan::scores(ponds.dbrda, display = "bp")
arrows(0, 0, vectors[,1] * 2, vectors[,2] * 2,
       lwd = 2, lty = 1, length = 0.2, col = "red")
text(vectors[,1] * 2, vectors[,2] * 2, pos = 3,
     labels = rownames(vectors))

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

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



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

> ***Answer 6***: The inital question sought to understand the importance of watershed boundaries with particular environmental variables. Phylogenetically, there seems to be a very low amount of varation explained by the dbRDA with only a rough two clusters appearing on the visualization. The environmental variables that seem to have the greatest impact on phylogenetic patterns are "SpC", "Temp", "pH", and "chla.

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***: Phylogenetic analysis is an essential portion to my own research. It would be very interesting to use phylogenetic data to understand how phylogenetic diversity in a biofilm results in increases resilience against stressors. This would require sampling two similar communities with one community exposed to an antibiotic stress and one that has not and sequencing them. From this experiment, it could be determiend how stressors shape the diversity present in a community and determine how comunal living may promote increased genetic diversity through mechanisms such as horizontal gene transfer. It would also be a very interesting element to sample these multicellular structures over time to see how interactions between indvidual organisms influences overall community composition, such as inducing increasing amount of dormant structures within a biofilm and see how diversity, both alpha and phylogenetic, changes over time.