9.Phylogenetic Diversity - Communities

Maddy Spencer; Z620: Quantitative Biodiversity, Indiana University

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_Spencer/Week7-PhyloCom"
```

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

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

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

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

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

##
## Attaching package: 'shapefiles'

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

##
## Attaching package: 'devtools'

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

##
## Attaching package: 'BiocManager'

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

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

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

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

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

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

##
## Attaching package: 'matrixStats'

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

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

##
## Attaching package: 'pROC'

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

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

2) DESCRIPTION OF DATA

need to discuss data set from spatial ecology!

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

3) LOAD THE DATA

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

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

comm <- read.otu(shared = "./data/INPonds.final.rdp.shared", cutoff = "1")
rownames(comm)
```

```
##   [1] "BC001-DNA"   "BC001-cDNA"  "BC002-DNA"   "BC002-cDNA"  "BC003-DNA"
##   [6] "BC003-cDNA"  "BC004-DNA"   "BC004-cDNA"  "BC005-DNA"   "BC005-cDNA"
##  [11] "BC010-DNA"   "BC015-DNA"   "BC015-cDNA"  "BC016-DNA"   "BC016-cDNA"
##  [16] "BC018-DNA"   "BC018-cDNA"  "BC020-DNA"   "BC020-cDNA"  "BC048-DNA"
##  [21] "BC048-cDNA"  "BC049-DNA"   "BC049-cDNA"  "BC051-DNA"   "BC051-cDNA"
##  [26] "BC105-DNA"   "BC105-cDNA"  "BC108-DNA"   "BC108-cDNA"  "BC262-DNA"
##  [31] "BC262-cDNA"  "BCL01-DNA"   "BCL01-cDNA"  "BCL03-DNA"   "BCL03-cDNA"
##  [36] "HNF132-DNA"  "HNF132-cDNA" "HNF133-DNA"  "HNF133-cDNA" "HNF134-DNA"
##  [41] "HNF134-cDNA" "HNF144-DNA"  "HNF144-cDNA" "HNF168-DNA"  "HNF168-cDNA"
##  [46] "HNF185-DNA"  "HNF185-cDNA" "HNF187-DNA"  "HNF187-cDNA" "HNF189-DNA"
##  [51] "HNF189-cDNA" "HNF190-DNA"  "HNF190-cDNA" "HNF191-DNA"  "HNF191-cDNA"
##  [56] "HNF216-DNA"  "HNF216-cDNA" "HNF217-DNA"  "HNF217-cDNA" "HNF221-DNA"
##  [61] "HNF224-DNA"  "HNF224-cDNA" "HNF225-DNA"  "HNF225-cDNA" "HNF229-DNA"
##  [66] "HNF229-cDNA" "HNF236-DNA_" "HNF236-cDNA" "HNF237-DNA"  "HNF237-cDNA"
##  [71] "HNF242-DNA"  "HNF242-cDNA" "HNF250-DNA"  "HNF250-cDNA" "HNF267-DNA"
##  [76] "HNF267-cDNA" "HNF269-DNA"  "HNF269-cDNA" "HNF279-DNA"  "HNF279-cDNA"
##  [81] "YSF004-DNA_" "YSF004-cDNA" "YSF117-DNA"  "YSF117-cDNA" "YSF295-DNA"
##  [86] "YSF295-cDNA" "YSF296-DNA"  "YSF296-cDNA" "YSF298-DNA"  "YSF298-cDNA"
##  [91] "YSF300-DNA"  "YSF300-cDNA" "YSF44-DNA"   "YSF44-cDNA"  "YSF45-DNA"
##  [96] "YSF45-cDNA"  "YSF46-DNA"   "YSF46-cDNA"  "YSF47-DNA"   "YSF47-cDNA"
## [101] "YSF65-DNA"   "YSF65-cDNA"  "YSF66-DNA"   "YSF66-cDNA"  "YSF67-DNA"
## [106] "YSF69-DNA"   "YSF69-cDNA"  "YSF70-DNA"   "YSF70-cDNA"  "YSF71-DNA"
## [111] "YSF71-cDNA"  "YSF74-DNA"   "YSF74-cDNA"
```

```
comm <- comm[grep("*-DNA", rownames(comm)), ]
rownames(comm)
```

```
##  [1] "BC001-DNA"   "BC002-DNA"   "BC003-DNA"   "BC004-DNA"   "BC005-DNA"
##  [6] "BC010-DNA"   "BC015-DNA"   "BC016-DNA"   "BC018-DNA"   "BC020-DNA"
## [11] "BC048-DNA"   "BC049-DNA"   "BC051-DNA"   "BC105-DNA"   "BC108-DNA"
## [16] "BC262-DNA"   "BCL01-DNA"   "BCL03-DNA"   "HNF132-DNA"  "HNF133-DNA"
## [21] "HNF134-DNA"  "HNF144-DNA"  "HNF168-DNA"  "HNF185-DNA"  "HNF187-DNA"
## [26] "HNF189-DNA"  "HNF190-DNA"  "HNF191-DNA"  "HNF216-DNA"  "HNF217-DNA"
## [31] "HNF221-DNA"  "HNF224-DNA"  "HNF225-DNA"  "HNF229-DNA"  "HNF236-DNA_"
## [36] "HNF237-DNA"  "HNF242-DNA"  "HNF250-DNA"  "HNF267-DNA"  "HNF269-DNA"
## [41] "HNF279-DNA"  "YSF004-DNA_" "YSF117-DNA"  "YSF295-DNA"  "YSF296-DNA"
## [46] "YSF298-DNA"  "YSF300-DNA"  "YSF44-DNA"   "YSF45-DNA"   "YSF46-DNA"
## [51] "YSF47-DNA"   "YSF65-DNA"   "YSF66-DNA"   "YSF67-DNA"   "YSF69-DNA"
## [56] "YSF70-DNA"   "YSF71-DNA"   "YSF74-DNA"
```

```
rownames(comm) <- gsub("\\-DNA", "", rownames(comm))
rownames(comm)
```

```
##  [1] "BC001"   "BC002"   "BC003"   "BC004"   "BC005"   "BC010"   "BC015"
##  [8] "BC016"   "BC018"   "BC020"   "BC048"   "BC049"   "BC051"   "BC105"
## [15] "BC108"   "BC262"   "BCL01"   "BCL03"   "HNF132"  "HNF133"  "HNF134"
## [22] "HNF144"  "HNF168"  "HNF185"  "HNF187"  "HNF189"  "HNF190"  "HNF191"
## [29] "HNF216"  "HNF217"  "HNF221"  "HNF224"  "HNF225"  "HNF229"  "HNF236_"
## [36] "HNF237"  "HNF242"  "HNF250"  "HNF267"  "HNF269"  "HNF279"  "YSF004_"
## [43] "YSF117"  "YSF295"  "YSF296"  "YSF298"  "YSF300"  "YSF44"   "YSF45"
## [50] "YSF46"   "YSF47"   "YSF65"   "YSF66"   "YSF67"   "YSF69"   "YSF70"
## [57] "YSF71"   "YSF74"
```

```
rownames(comm) <- gsub("\\_", "", rownames(comm))
rownames(comm)
```

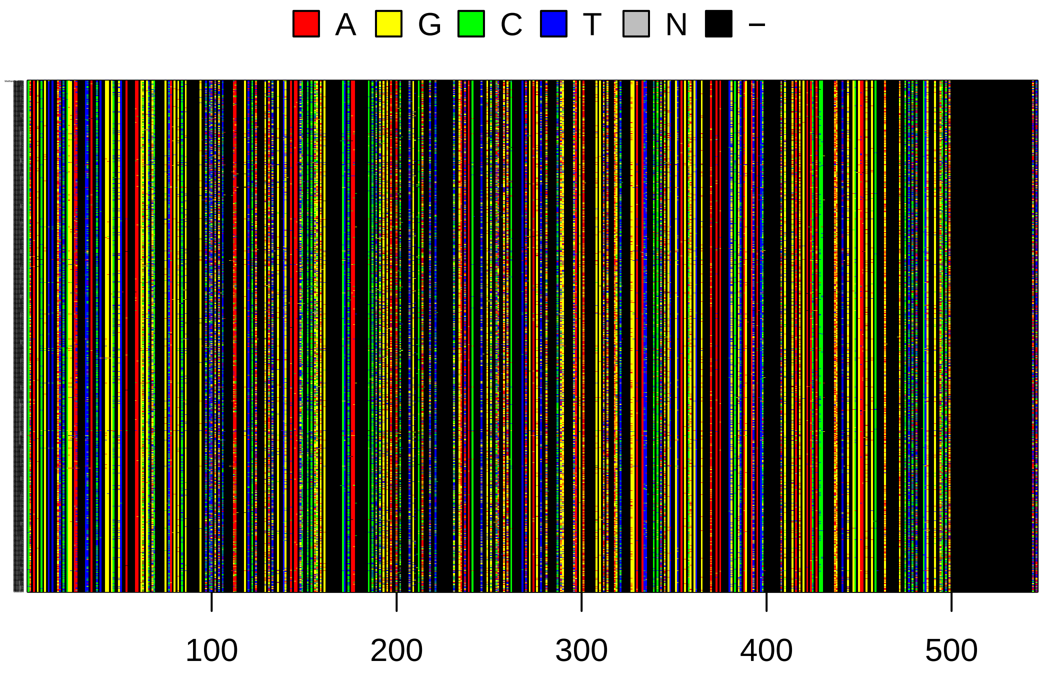
```
##  [1] "BC001"  "BC002"  "BC003"  "BC004"  "BC005"  "BC010"  "BC015"  "BC016"
##  [9] "BC018"  "BC020"  "BC048"  "BC049"  "BC051"  "BC105"  "BC108"  "BC262"
## [17] "BCL01"  "BCL03"  "HNF132" "HNF133" "HNF134" "HNF144" "HNF168" "HNF185"
## [25] "HNF187" "HNF189" "HNF190" "HNF191" "HNF216" "HNF217" "HNF221" "HNF224"
## [33] "HNF225" "HNF229" "HNF236" "HNF237" "HNF242" "HNF250" "HNF267" "HNF269"
## [41] "HNF279" "YSF004" "YSF117" "YSF295" "YSF296" "YSF298" "YSF300" "YSF44"
## [49] "YSF45"  "YSF46"  "YSF47"  "YSF65"  "YSF66"  "YSF67"  "YSF69"  "YSF70"
## [57] "YSF71"  "YSF74"
```

```
comm <- comm[rownames(comm) %in% env$Sample_ID, ]
comm <- comm[ , colSums(comm) > 0]

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

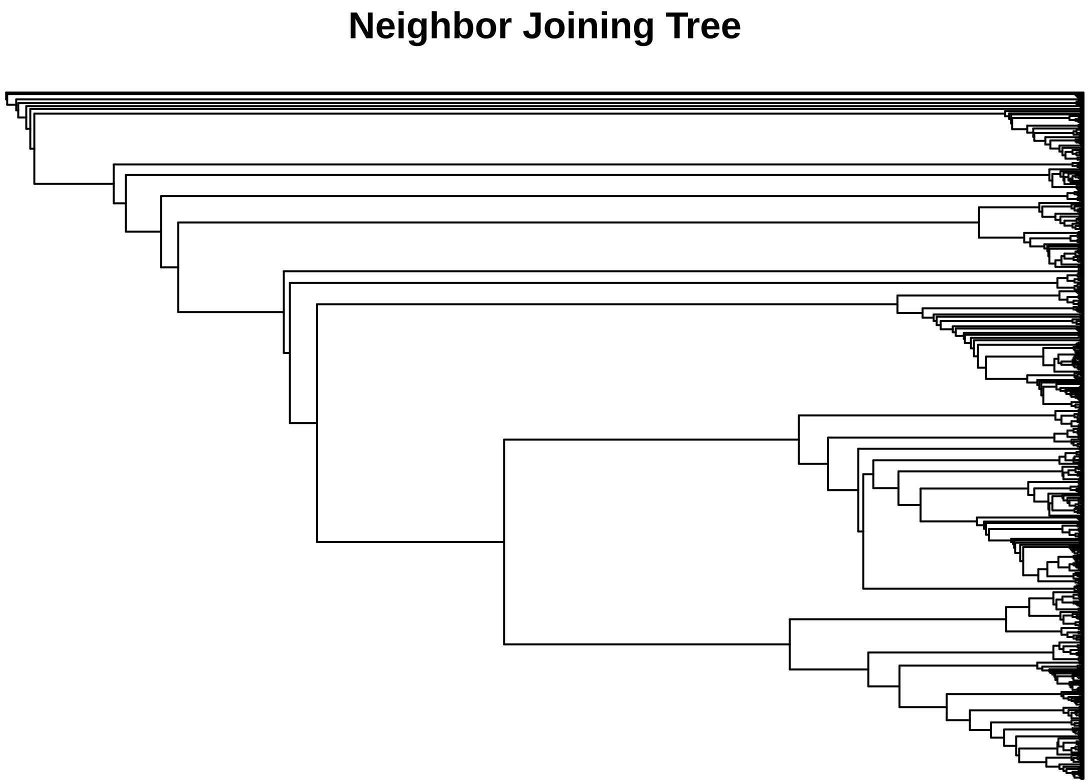
```
## Warning in type.convert.default(as.character(x)): 'as.is' should be specified
## by the caller; using TRUE
## Warning in type.convert.default(as.character(x)): 'as.is' should be specified
## by the caller; using TRUE
## Warning in type.convert.default(as.character(x)): 'as.is' should be specified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

```
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)

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 = FALSE)

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 = FALSE, use.edge.length = FALSE,
           direction = "right", cex = 0.6, label.offset = 1)
```



4) PHYLOGENETIC ALPHA DIVERSITY

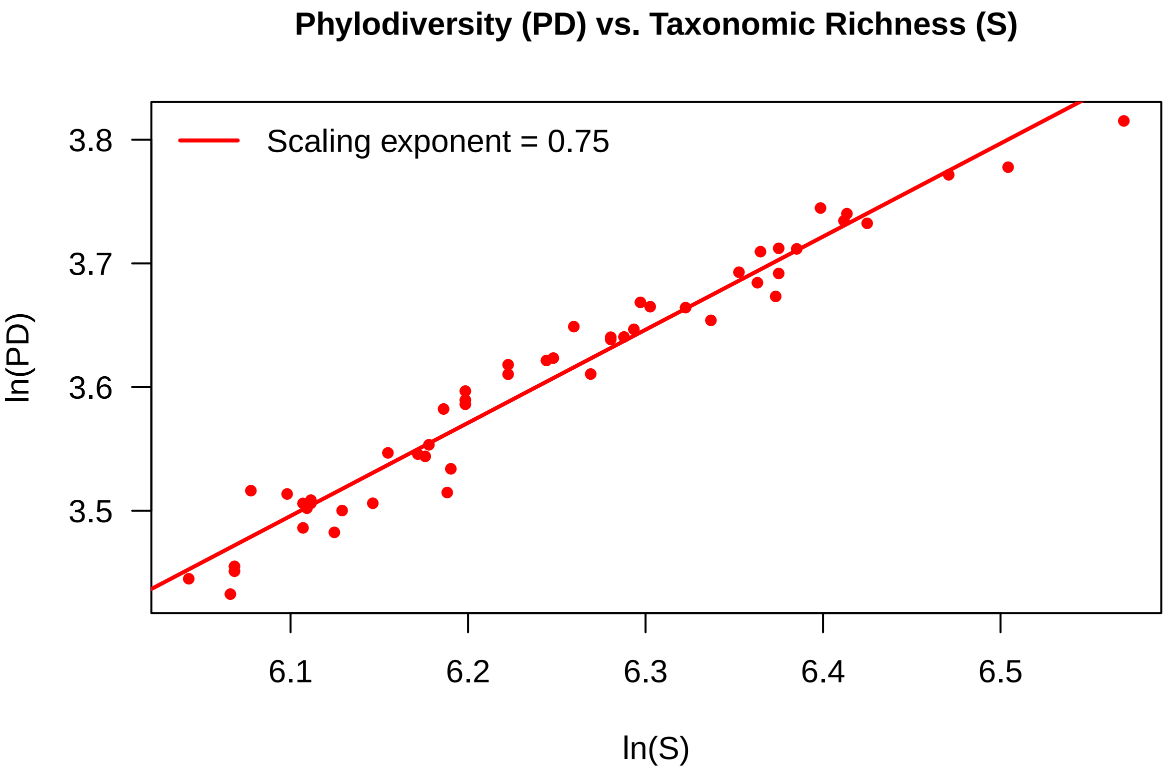
A. Faith's Phylogenetic Diversity (PD)

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

```
pd <- pd(comm, 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 = "red", las = 1,
     xlab = "ln(S)", ylab = "ln(PD)", cex.main = 1,
     main = "Phylodiversity (PD) vs. Taxonomic Richness (S)")

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



Question 1: Answer the following questions about the PD-S pattern.
a. Based on how PD is calculated, how and why should this metric be related to taxonomic 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***: PD is calculated by measuring the length of phylogenetic tree branches that are unique to species in a community. This metric is inherently tied to taxonomic richness as more taxonomic richness typically means more phylogenetic richness, unless species are very closely related. ***Answer 1b***: I would expect PD and S to deviate from eachother when there is the presence of closely related species who share most of the phylogenetic branches. ***Answer 1c***: The scaling exponent is 0.75, which means that PD and S aren't linearly related. Because the scaling exponent is less than one, the PD-S relationship would look like a log curve, where S increases with PD linearly and then plateaus out.

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.pd1 <- ses.pd(comm[1:2, ], phy, null.model = "richness", runs = 25, include.root = FALSE)
ses.pd2 <- ses.pd(comm[1:2, ], phy, null.model = "taxa.labels", runs = 25, include.root = FALSE)
ses.pd3 <- ses.pd(comm[1:2, ], phy, null.model = "phylogeny.pool", runs = 25, include.root = FALSE)
ses.pd1
```

```
##       ntaxa   pd.obs pd.rand.mean pd.rand.sd pd.obs.rank   pd.obs.z  pd.obs.p
## BC001   668 43.71912     44.06946  0.8068736          10 -0.4341852 0.3846154
## BC002   587 40.94334     39.84196  0.7746811          24  1.4217188 0.9230769
##       runs
## BC001   25
## BC002   25
```

```
ses.pd2
```

```
##       ntaxa   pd.obs pd.rand.mean pd.rand.sd pd.obs.rank   pd.obs.z  pd.obs.p
## BC001   668 43.71912     43.96299  0.9854673          12 -0.2474599 0.4615385
## BC002   587 40.94334     39.85949  0.7597446          24  1.4265932 0.9230769
##       runs
## BC001   25
## BC002   25
```

```
ses.pd3
```

```
##       ntaxa   pd.obs pd.rand.mean pd.rand.sd pd.obs.rank   pd.obs.z  pd.obs.p
## BC001   668 43.71912     43.85679  0.8215832          11 -0.1675572 0.4230769
## BC002   587 40.94334     39.61310  0.7868685          25  1.6905476 0.9615385
##       runs
## BC001   25
## BC002   25
```

```
help(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: For the richness null model, our null hypothesis is that phylodiversity (PD) varies by community abundances only due to random chance. The alternative is that PD varies based on community abundance, not due to chance alone. ***Answer 2b***: This choice did not effect ses.pd substantially, as each test has a non-signficant p-value. The PD scores stay relatively the same for BC001 and change slightly with BC002 across different null models. I would anticipate the choice of null model does effect output, as it changes what is randomized and what is feld constant.

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 = TRUE, runs = 25)

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

```
##              NRI
## BC001 0.26634503
## BC002 0.53494125
## BC003 0.86348918
## BC004 0.05283839
## BC005 0.70153425
## BC010 0.38803846
```

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 = TRUE, runs = 25)

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

```
##             NTI
## BC001 0.7607100
## BC002 1.1415991
## BC003 1.5075259
## BC004 1.0069970
## BC005 1.6481398
## BC010 0.4941916
```

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 net relatedness index uses the mean phylogenetic distance, which is basically a sum of branch lengths unique to taxa in a sample, of the observed sample and substracts an MPD value of a randomized subsample of taxa from it. This number is then divded by the standard deviation of the randomized sample to determine whether the relatedness is above or below what would be expected. If there was no difference between the MPD of the observed and randomized samples, NRI would be zero. ***Answer 3b***: NTI is functionally the sample as NRI, but rather than using MPD it uses the branch distance to the most closely related community member. Therefore, we are calculating the branch distance of the closest neighbor in a randomized sample and susbtracting this from the actual branch distance observed for the total sample. ***Answer 3c***: For both NRI and NTI, the values are below 0 for most all taxa. This means that the phylogeny is this community is overdispersed and thus the taxa present are less related to one another than would be predicted. ***Answer 3d***: By using abundance data rather than presence-absence data, all of the NRI/NTI scores are now positive. This indicates that there is phylogenetic clustering, or less phylogenetic diversity, in our samples.

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.

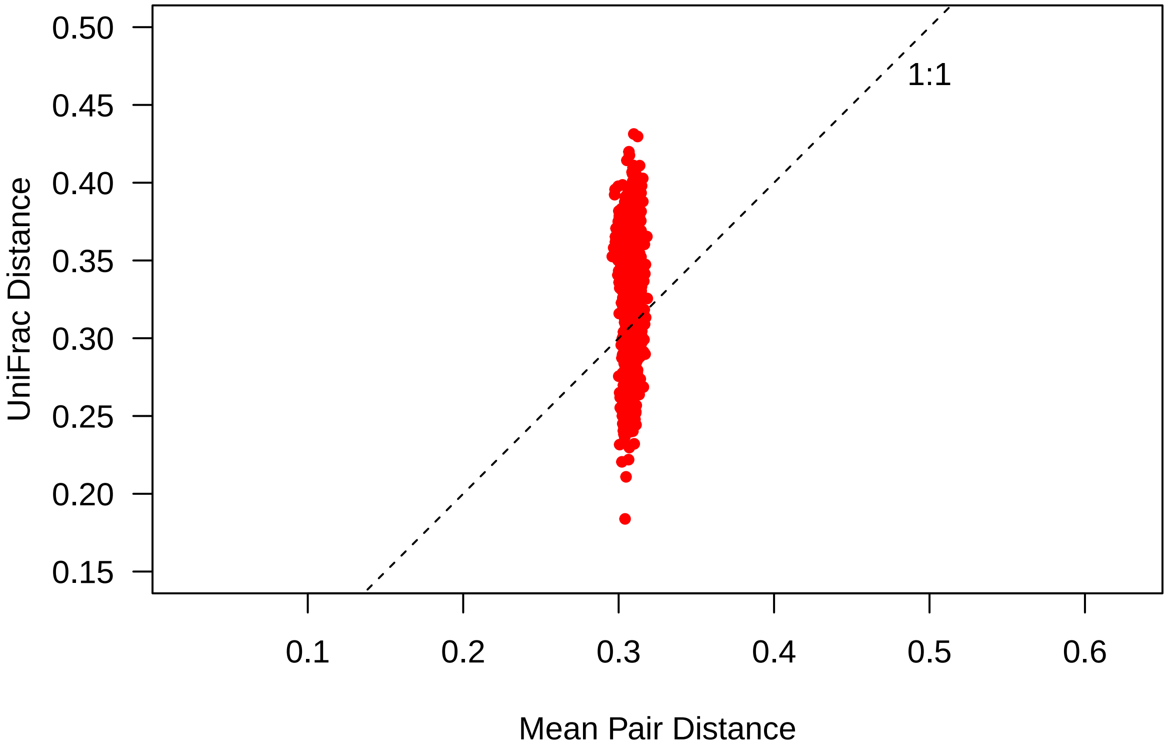
```r
dist.mp <- comdist(comm, phydist)
```

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

```r
dist.uf <- unifrac(comm, 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 = "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 Pair distance takes the difference between two taxa's mean phylogenetic distance, which calculates the average branch length between the two taxa. Unifrac functions similarly, but separates shared and unshared portions of the branches to calculate distance as the sum of unshared branches divided by the total brnach length. ***Answer 4b***: For a singular mean pair distance value, there is a wide range of unifrac values, going from 0.17 to 0.45. ***Answer 4c***: MPD may show less variation than Unifrac because MPD does not distinguish between shared and unshared branch lengths. By highlighting unshared branch length, Unifrac gives a more

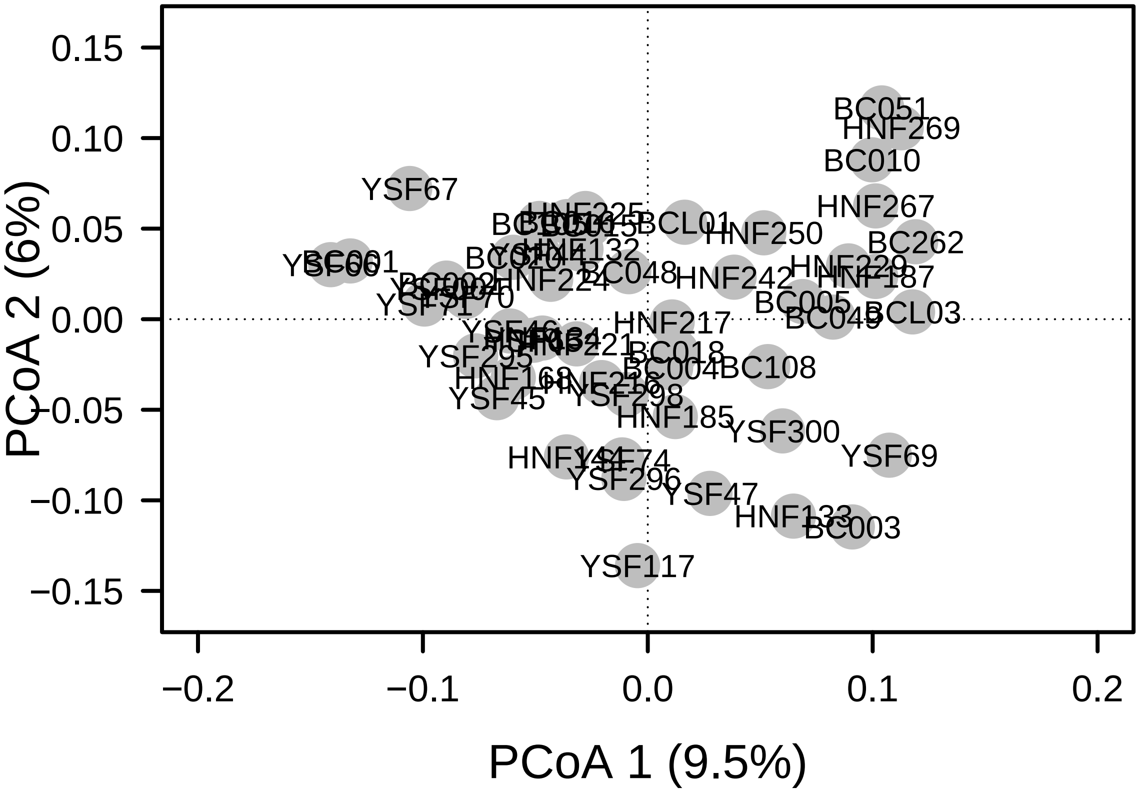
sensitive output.

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.

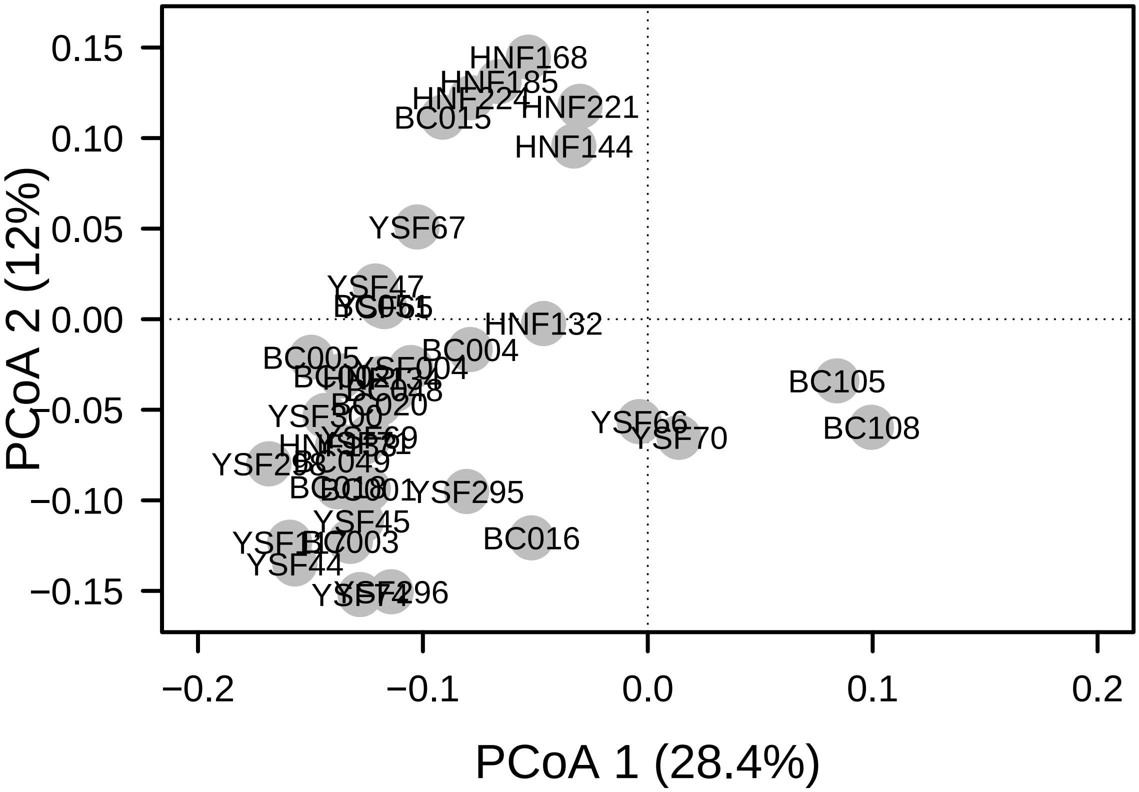
In the R code chunk below, do the following:
1. perform a PCoA based on the UniFrac distances, and
2. calculate the explained variation for the first three PCoA axes. 3. add and label the points, and
4. customize the plot.

```r
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)

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

plot(pond.pcoa$points[,1], pond.pcoa$points[,2],
     xlim = c(-0.2, 0.2), ylim = c(-0.16, 0.16),
     xlab = paste("PCoA 1 (", explainvar1, "%)", sep = ""),
     ylab = paste("PCoA 2 (", explainvar2, "%)", sep = ""),
     pch = 16, cex = 2.0, type = "n", cex.lab = 1.5, cex.axis = 1.2, axes = 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
tax.db <- vegdist(comm, method = "bray")
tax.pcoa <- cmdscale(tax.db, eig = T, k = 3)

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

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

plot(tax.pcoa$points[,1], tax.pcoa$points[,2],
     xlim = c(-0.2, 0.2), ylim = c(-0.16, 0.16),
     xlab = paste("PCoA 1 (", t.explainvar1, "%)", sep = ""),
     ylab = paste("PCoA 2 (", t.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(tax.pcoa$points[ ,1], tax.pcoa$points[ ,2],
       pch = 19, cex = 3, bg = "gray", col = "gray")
text(tax.pcoa$points[ ,1], tax.pcoa$points[ ,2],
     labels = row.names(tax.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***: In the phylogenetic PCoA there are no apparent trends in clustering/similarity. The axes also do not explain much variation, with axis 1 explaining 9.5% of the variation and 2 6%. The taxonomic PCoA is more informative in my opinion, with most of the taxa being clustered on the left side of the graph. The axes here explain more variation, with axis 1 explaining 29% and 2 12%. I believe this is an indicator that phylogeny is NOT what is explaining community trends at an abundance level and thus is most likely not effecting our test results.

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.

```r
watershed <- env$Location
phylo.adonis <- adonis2(dist.uf ~ watershed, permutations = 999)
phylo.adonis
```

```
## Permutation test for adonis under reduced model
## Permutation: free
## Number of permutations: 999
##
## adonis2(formula = dist.uf ~ watershed, permutations = 999)
##          Df SumOfSqs     R2      F Pr(>F)
## Model     2  0.13316 0.0492 1.2679  0.027 *
## Residual 49  2.57305 0.9508
## Total    51  2.70621 1.0000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

```
## Permutation test for adonis under reduced model
## Permutation: free
## Number of permutations: 999
##
## adonis2(formula = vegdist(decostand(comm, method = "log"), method = "bray") ~ watershed, permutation
##          Df SumOfSqs      R2      F Pr(>F)
## Model     2  0.16601 0.06018 1.5689  0.005 **
## Residual 49  2.59229 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(env) %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: 0.08433
##       Significance: 0.164
##
## Upper quantiles of permutations (null model):
##   90%   95% 97.5%   99%
## 0.107 0.145 0.175 0.203
## 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)))

```

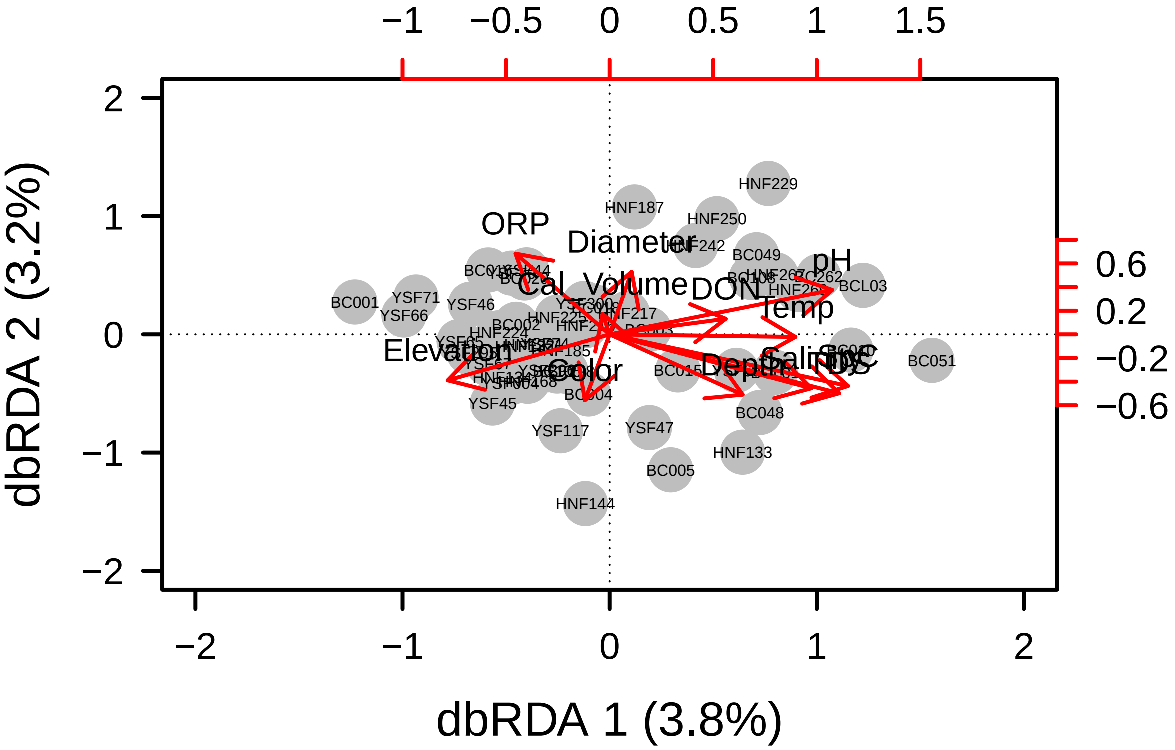
```r
anova(ponds.dbrda, by = "axis")
```

```
## 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 + Cal_Volume + ORP + Temp + SpC
##          Df SumOfSqs      F Pr(>F)
## dbRDA1    1  0.10324 1.9852  0.484
## dbRDA2    1  0.08592 1.6521  0.835
## dbRDA3    1  0.08171 1.5711  0.871
## dbRDA4    1  0.07321 1.4077  0.964
## dbRDA5    1  0.06591 1.2674  0.994
## dbRDA6    1  0.05049 0.9709  1.000
## dbRDA7    1  0.04671 0.8982
## dbRDA8    1  0.04175 0.8027
## dbRDA9    1  0.03606 0.6934
## dbRDA10   1  0.03302 0.6349
## dbRDA11   1  0.03078 0.5919
## dbRDA12   1  0.02921 0.5617
## Residual 39  2.02820
```

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

```
##
## ***VECTORS
##
##              dbRDA1   dbRDA2     r2 Pr(>r)
## Elevation  -0.86345 -0.50444 0.1084  0.059 .
## Diameter    0.06683  0.99776 0.0543  0.265
## Depth       0.68050 -0.73275 0.1213  0.037 *
## Cal_Volume -0.22122  0.97523 0.0062  0.864
## ORP        -0.48281  0.87572 0.1309  0.037 *
## Temp        0.98974 -0.14286 0.1179  0.053 .
## SpC         0.85986 -0.51052 0.2464  0.001 ***
## TDS         0.82742 -0.56158 0.2451  0.001 ***
## Salinity    0.81889 -0.57395 0.1931  0.004 **
## pH          0.93813  0.34629 0.1823  0.006 **
## Color      -0.07756 -0.99699 0.0604  0.203
## DON         0.97923  0.20274 0.0467  0.299
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Permutation: free
## Number of permutations: 999
```

```r
dbrda.explainvar1 <- round(ponds.dbrda$CCA$eig[1] / sum(c(ponds.dbrda$CCA$eig, ponds.dbrda$CA$eig)), 3)
dbrda.explainvar2 <- round(ponds.dbrda$CCA$eig[2] / sum(c(ponds.dbrda$CCA$eig, ponds.dbrda$CA$eig)), 3)

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)

wa_scores <- vegan::scores(ponds.dbrda, display = "sites")

points(wa_scores, pch = 19, cex = 3, col = "gray")
text(wa_scores, labels = rownames(wa_scores), cex = 0.5)

vectors <- vegan::scores(ponds.dbrda, display = "bp")
arrows(0, 0, vectors[,1] * 2, vectors[,2] * 2, lwd = 2, lty = 1, length = 0.2, col = "red")
text(vectors[,1] * 2, vectors[,2] * 2, pos = 3, labels = rownames(vectors))
axis(side = 3, lwd.ticks = 2, cex.axis= 1.2, las = 1, col = "red", lwd = 2.2,
     at = pretty(range(vectors[, 1]) * 2),
     labels = pretty(range(vectors[, 1]) * 2))
axis(side = 4, lwd.ticks = 2, cex.axis= 1.2, las = 1, col = "red", lwd = 2.2,
     at = pretty(range(vectors[, 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***: Phylogenetic patterns in Indiana ponds are influenced by watershed and environmental variables. Based on our PERMANOVA results, we know that what watershed the pond is located in impacts the phylogenetic beta-diversity (p-value = 0.035). This effect is even more exacerbated when looking at watershed impact on taxonomic beta-diversity (p-value = 0.004). A mantel test also informed us that environmental variations impact phylogenetic beta-diversity. Our

dbRDA test revealed which factors have a sigificant impact on phylogenetic beta-diversity: Depth (p=0.036), oxidation reduction potential (p=0.035), temperature (p=0.043), specific conductivity of water (p=0.002), total dissolved solids (p=0.001), salinity (p=0.007), and pH (p=0.008).

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***: I think that looking at phylogenetic data of different *Acinetobacter* strains that produce wax ester would be interesting. I've found that the regulation of WE synthesis is very different depending on the strain, so it would be cool to map regulators onto a phylogenetic tree and see how each strain diverged from one another. I would need phylogenetic data on each strain along with their regulator sequences. An issue in molecular biology is that we will often use model organisms to represent how regulatory pathways work and apply this as a blanket to other organisms without question. I think demonstrating that regulation can vary even on a strain level is important to the field to demonstrate the need for rigorous testing of regulation across organisms.