12.Phylogenetic Diversity - Communities

Mackenzie Caple; Z620: Quantitative Biodiversity, Indiana University

26 February, 2019

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 '8.BetaDiversity' 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*).

1) SETUP

Typically, the first thing you will do in either an R script or an RMarkdown file is setup your environment. This includes things such as setting the working directory and loading any packages that you will need.

In the R code chunk below, provide the code to:
1. clear your R environment,
2. print your current working directory,
3. set your working directory to your `/Week7-PhyloCom` folder,
4. load all of the required R packages (be sure to install if needed), and
5. load the required R source file.

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

```
## [1] "/Users/mcaple/GitHub/QB2019_Caple/2.Worksheets/12.PhyloCom"
```

```
setwd("~/GitHub/QB2019_Caple/2.Worksheets/12.PhyloCom/")

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

```
## Warning: package 'vegan' was built under R version 3.5.2
## This is vegan 2.5-4
##
## Attaching package: 'seqinr'
## The following object is masked from 'package:nlme':
##
##     gls
## The following object is masked from 'package:permute':
##
##     getType
## The following objects are masked from 'package:ape':
##
##     as.alignment, consensus
##
## Attaching package: 'shapefiles'
## The following objects are masked from 'package:foreign':
##
##     read.dbf, write.dbf
## This is simba 0.3-5
##
## Attaching package: 'simba'
## The following object is masked from 'package:picante':
##
##     mpd
## The following object is masked from 'package:stats':
##
##     mad
```

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

2) DESCRIPTION OF DATA

need to discuss data set from spatial ecology!

In 2013 we sampled > 50 forested ponds in Brown County State Park, Yellowwood 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

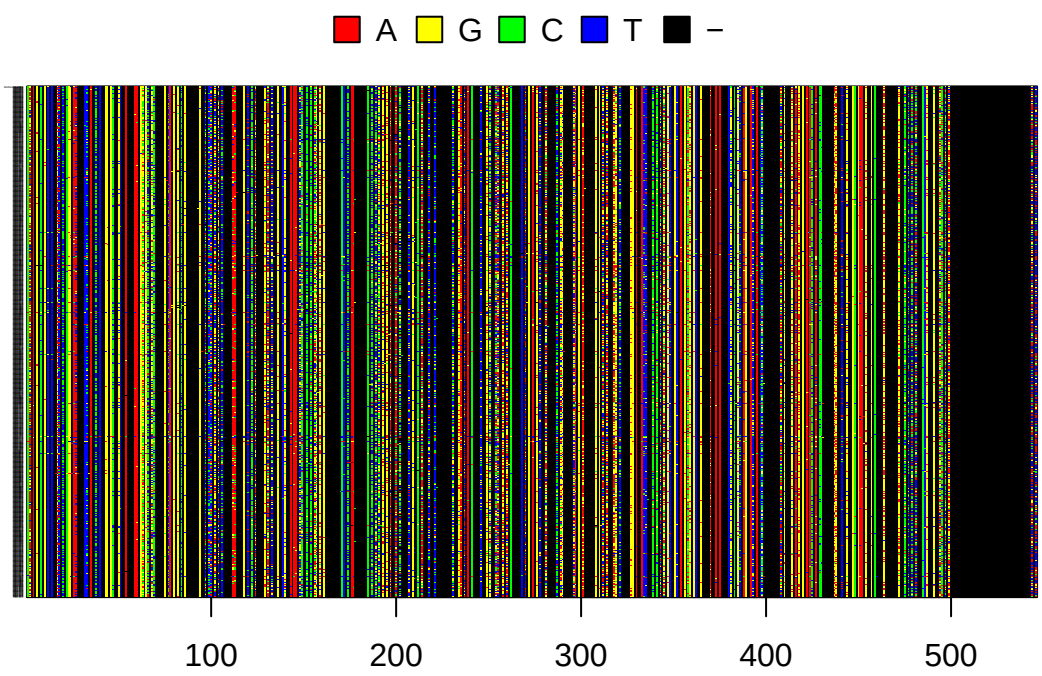
In the R code chunk below, do the following:
1. load the environmental data for the Brown County ponds (*20130801_PondDataMod.csv*),
2. load the site-by-species matrix using the `read.otu()` function,
3. subset the data to include only DNA-based identifications of bacteria,
4. rename the sites by removing 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.

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

# load taxonomic data
# load site-by-species matrix
comm <- read.otu(shared = "./data/INPonds.final.rdp.shared", cutoff = "1")
# select DNA data using 'grep()'
comm <- comm[grep("*-DNA", rownames(comm)), ]
# perform replacement of all matches with 'gsub()'
rownames(comm) <- gsub("\\-DNA", "", rownames(comm))
rownames(comm) <- gsub("\\_", "", rownames(comm))

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

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

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

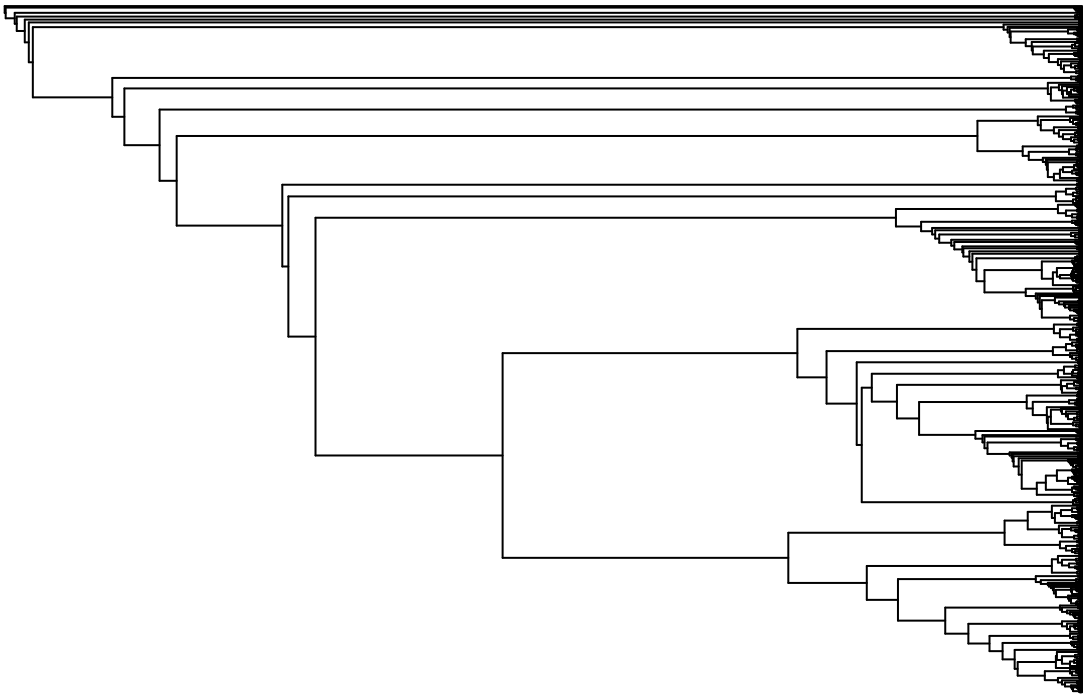
Next, in the R code chunk below, do the following:
1. load the FASTA alignment for the bacterial operational taxonomic units (OTUs),
2. rename the OTUs by removing everything before the tab (\t) and after the bar (|),
3. import the *Methanosarcina* outgroup FASTA file,
4. convert both FASTA files into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

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

# rename OTUs in the FASTA file
```

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

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

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

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




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

# make a neighbor-joining tree file ('ape')
phy.all <- bionj(seq.dist.jc)

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

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

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

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



4) PHYLOGENETIC ALPHA DIVERSITY

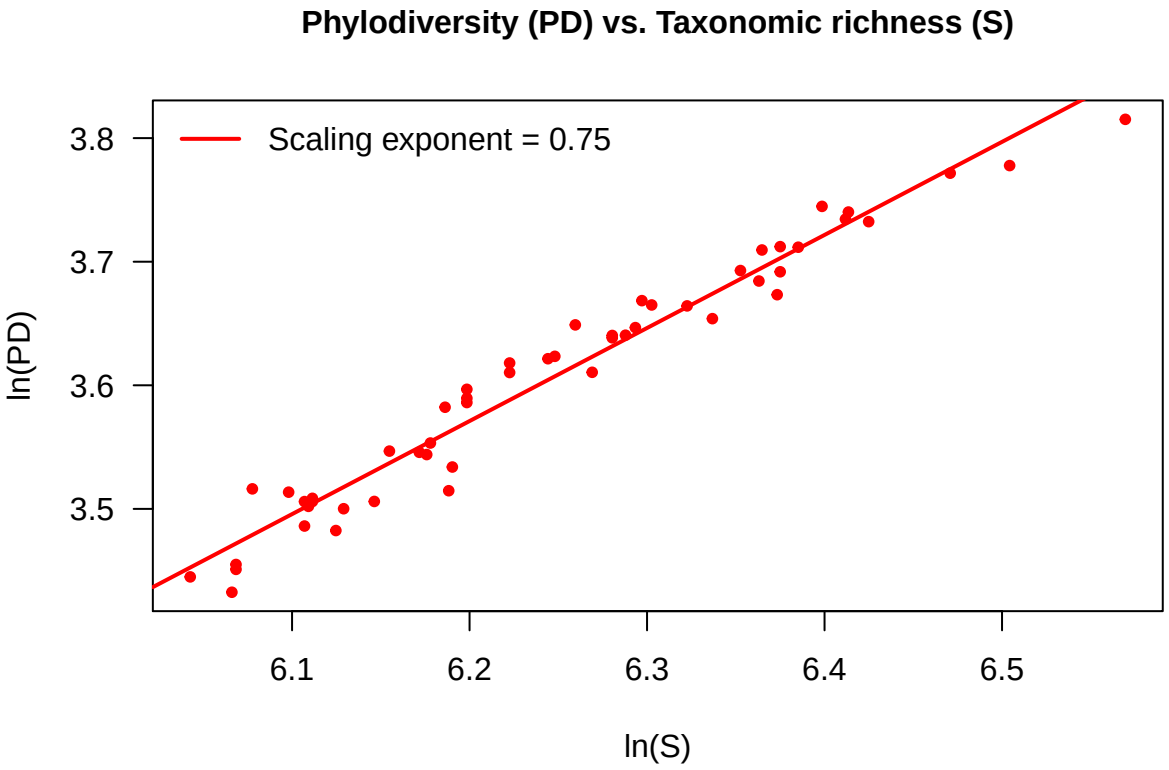
A. Faith's Phylogenetic Diversity (PD)

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

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

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

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

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

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



Question 1: Answer the following questions about the PD-S pattern.
a. Based on how PD is calculated, 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 is simply the sum of branch lengths of all the taxa in the site, so higher taxonomic richness will automatically mean a higher branch length sum, because there are more taxa present ***Answer 1b***: These two concepts are related but not quite the same. Taxonomic richness is the number of different taxa present, whereas phylodiversity takes into account not just the number of taxa, but the way in which they are related to each other (could be more or less than expected) ***Answer 1c***: If the environment is harsh, environmental filtering may lead to lower phylogenetic diversity than is expected based on taxonomic richness, whereas a large degree of competition could lead to greater phylogenetic diversity than would be expected based on taxonomic richness ***Answer 1d***: The exponent is less than 1, which means that the relationship is increasing at a decelerating rate. This indicates that phylodiversity increases less quickly the more species are added

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.

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

```
##       ntaxa   pd.obs pd.rand.mean pd.rand.sd pd.obs.rank  pd.obs.z
## BC001   668 43.71912     43.98901  0.9685602           9 -0.278647
## BC002   587 40.94334     39.84454  0.8505281          24  1.291898
##        pd.obs.p runs
## BC001 0.3461538   25
```

```
## BC002 0.9230769   25
```

Question 2: Using `help()` and the table above, run the `ses.pd()` function using two other null models and answer the following questions:

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

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

ses.pd.1
```

```
##       ntaxa   pd.obs pd.rand.mean pd.rand.sd pd.obs.rank  pd.obs.z
## BC001   668 43.71912     42.44552  0.6128009          25  2.078339
## BC002   587 40.94334     42.18047  0.6196789           2 -1.996406
##         pd.obs.p runs
## BC001 0.96153846   25
## BC002 0.07692308   25
```

```
ses.pd.2
```

```
##       ntaxa   pd.obs pd.rand.mean pd.rand.sd pd.obs.rank   pd.obs.z
## BC001   668 43.71912     44.11613  0.7618862           6 -0.5210799
## BC002   587 40.94334     39.90028  0.9851187          22  1.0588117
##        pd.obs.p runs
## BC001 0.2307692   25
## BC002 0.8461538   25
```

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: My null hypothesis was that random chance accounts for the given PD; the alternative hypothesis was that something else (e.g. environmental filtering, competition, etc) besides the null model is driving phylogenetic diversity. ***Answer 2b***: Under the taxa.labels null model I did not find any significance (the p-values were extremely large), but with the frequency null model one of the ponds did have a significant p-value; this demonstrates that the choice of null model can have a very large effect on whether significance is found.

B. Phylogenetic Dispersion Within a Sample

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

i. Phylogenetic Resemblance Matrix

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

```
# create a phylogenetic distance matrix ('picante')
phydist <- cophenetic.phylo(phy)
```

ii. Net Relatedness Index (NRI)

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

```
# estimate standardized effect size of NRI via randomization ('picante')
ses.mpd <- ses.mpd(comm, phydist, null.model = "taxa.labels",
                   abundance.weighted = FALSE, runs = 25)
# calculate NRI
NRI <- as.matrix(-1 * ((ses.mpd[ , 2] - ses.mpd[ , 3]) / ses.mpd[ , 4]))
rownames(NRI) <- row.names(ses.mpd)
colnames(NRI) <- "NRI"

# use abundance instead of presence/absence
ses.mpd.a <- ses.mpd(comm, phydist, null.model = "taxa.labels",
                   abundance.weighted = TRUE, runs = 25)
# calculate NRI
NRI.a <- as.matrix(-1 * ((ses.mpd.a[ , 2] - ses.mpd.a[ , 3]) / ses.mpd.a[ , 4]))
rownames(NRI.a) <- row.names(ses.mpd.a)
colnames(NRI.a) <- "NRI.a"
```

iii. Nearest Taxon Index (NTI)

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

```
# estimate standardized effect size of NRI via randomization {picante}
ses.mntd <- ses.mntd(comm, phydist, null.model = "taxa.labels",
                     abundance.weighted = FALSE, runs = 25)
# calculate NTI
NTI <- as.matrix(-1 * ((ses.mntd[ , 2] - ses.mntd[ , 3]) / ses.mntd[ , 4]))
rownames(NTI) <- row.names(ses.mntd)
colnames(NTI) <- "NTI"

# use abundance instead of presence/absence
ses.mntd.a <- ses.mntd(comm, phydist, null.model = "taxa.labels",
                     abundance.weighted = TRUE, runs = 25)
# calculate NTI
NTI.a <- as.matrix(-1 * ((ses.mntd.a[ , 2] - ses.mntd.a[ , 3]) / ses.mntd.a[ , 4]))
rownames(NTI.a) <- row.names(ses.mntd.a)
colnames(NTI.a) <- "NTI.a"
```

```
ses.compare <- cbind(NRI, NRI.a, NTI, NTI.a)
```

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 compares every taxon in the data set to every other taxon (pairwise) finding the distance between them, and then averages the distance. Randomization is then used to compare the NTI of the data with a null model. ***Answer 3b***: The NTI finds every taxon's closest neighbor and averages just those distances (rather than every distance as in NRI). Then, randomization is used to see if the data is over- or under-dispersed compared to a null model. ***Answer 3c***: All of the NRI values are negative, meaning that the ponds is overdispersed compared to the null model. However, some of the NTI values are positive, meaning that there are a few ponds whose NTI is clusteted compared to the null model. ***Answer 3d***: This had a pretty big effect on the outcome. Many more ponds now show as clustered instead of overdispersed,

so it;s a qualitative shift as well as a change in magnitude. In terms of magnitude, some values increased while others decreased; there didn't seem to be a pattern here. Using presence/absence or abundance data is clearly an important decision here, though.

5) PHYLOGENETIC BETA DIVERSITY

A. Phylogenetically Based Community Resemblance Matrix

In the R code chunk below, do the following:
1. calculate the phylogenetically based community resemblance matrix using Mean Pair Distance, and
2. calculate the phylogenetically based community resemblance matrix using UniFrac distance.

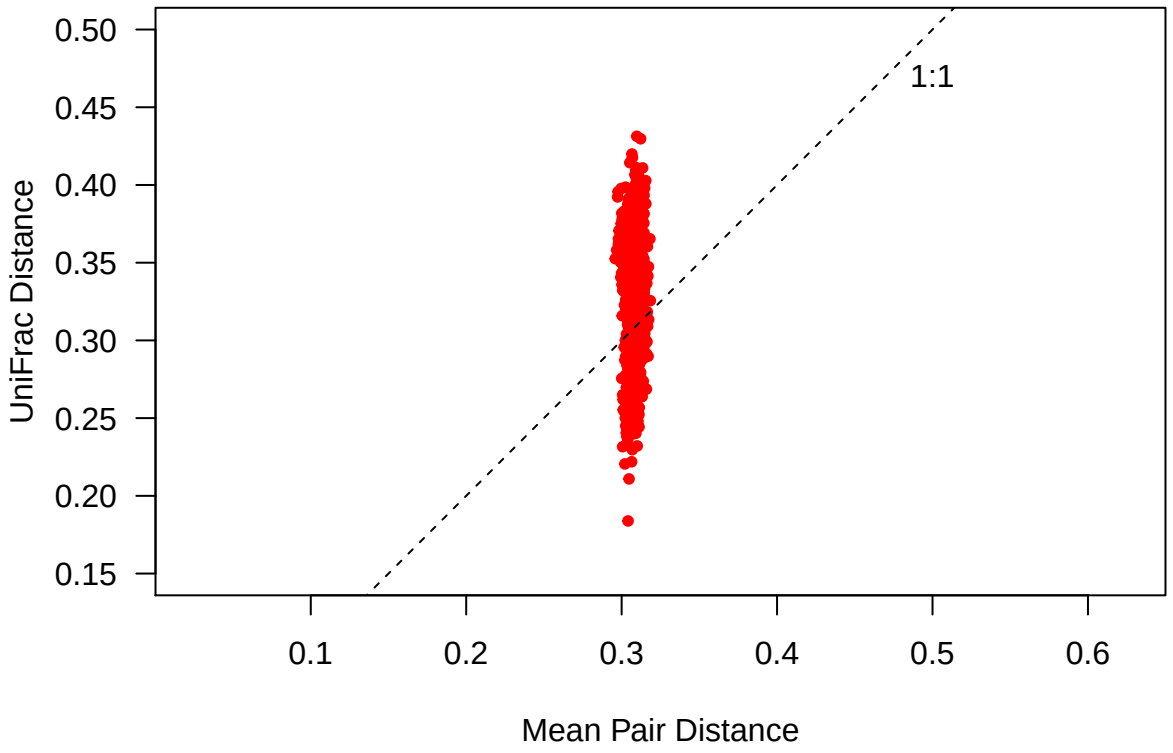
```
# mean pairwise distance
dist.mp <- comdist(comm, phydist)
```

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

```
# unifrac distance (takes a few minutes to run)
dist.uf <- unifrac(comm, phy)
```

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

```
par(mar = c(5, 5, 2, 1) + 0.1)
plot(dist.mp, dist.uf,
     pch = 20, col = "red", las = 1, asp = 1, xlim = c(0.15, 0.5), ylim = c(0.15, 0.5),
     xlab = "Mean Pair Distance", ylab = "UniFrac Distance")
abline(b = 1, a = 0, lty = 2)
text(0.5, 0.47, "1:1")
```



Question 4:

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

Answer 4a: MPD makes pairwise comparisons of every two taxa in a sample, calculates every pairwise distance, and then takes an average. UniFrac also makes pairwise comparisions, but instead of distance, it sums shared and unshared branch lengths and then finds the proportion of total branch length that is unshared. ***Answer 4b***: The MPD of every site is approximately 0.3, but the UniFrac distance ranges from about 0.17 to about 0.44. This shows that there isn't a strong relationship between the two measures– at least in this data, MPD does not change with UniFrac distance. ***Answer 4c***: In a large sample, a few outliers (taxa very unrelated compared to the rest of the community) may not affect the average pairwise distance much but may add a great deal to the proportion of unshared branch length.

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.

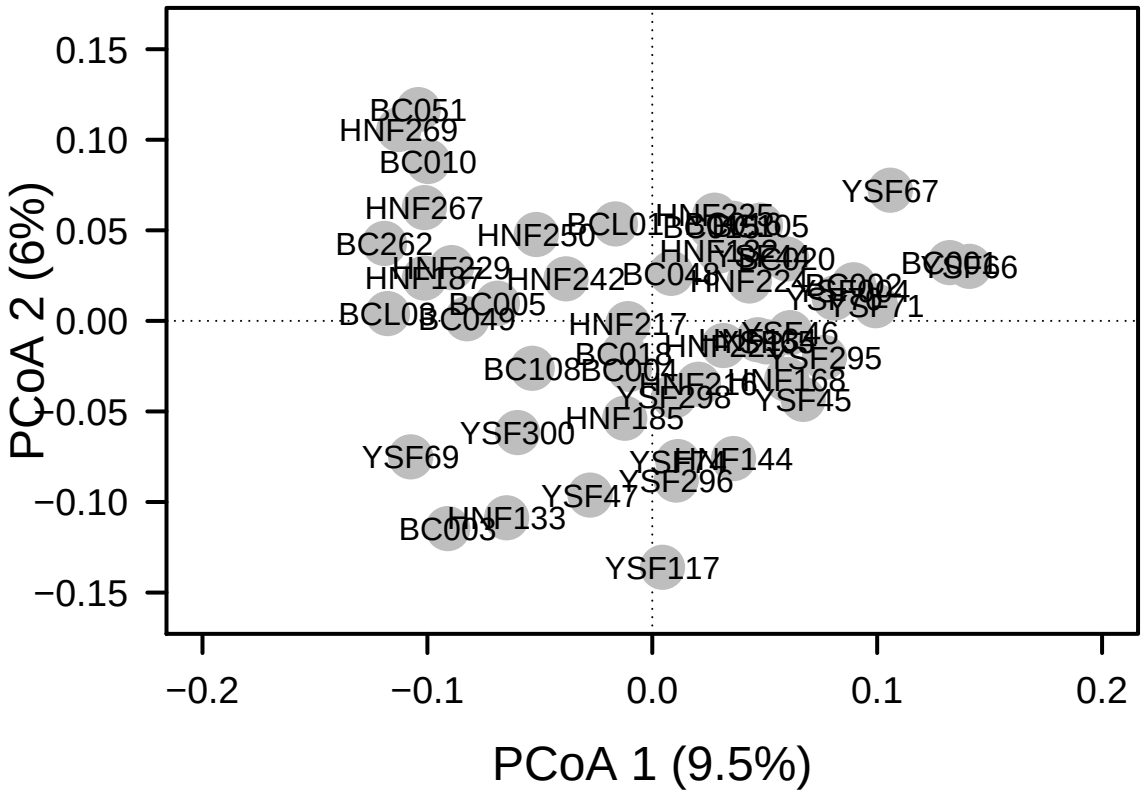
```
pond.pcoa <- cmdscale(dist.uf, eig = T, k = 3)

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.

```
# define plot parameters
par(mar = c(5, 5, 1, 2) + 0.1)

#initiate plot
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, type = "n", cex.lab = 1.5, cex.axis = 1.2, axes = FALSE)

# add axes
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)

# add points and labels
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))
```

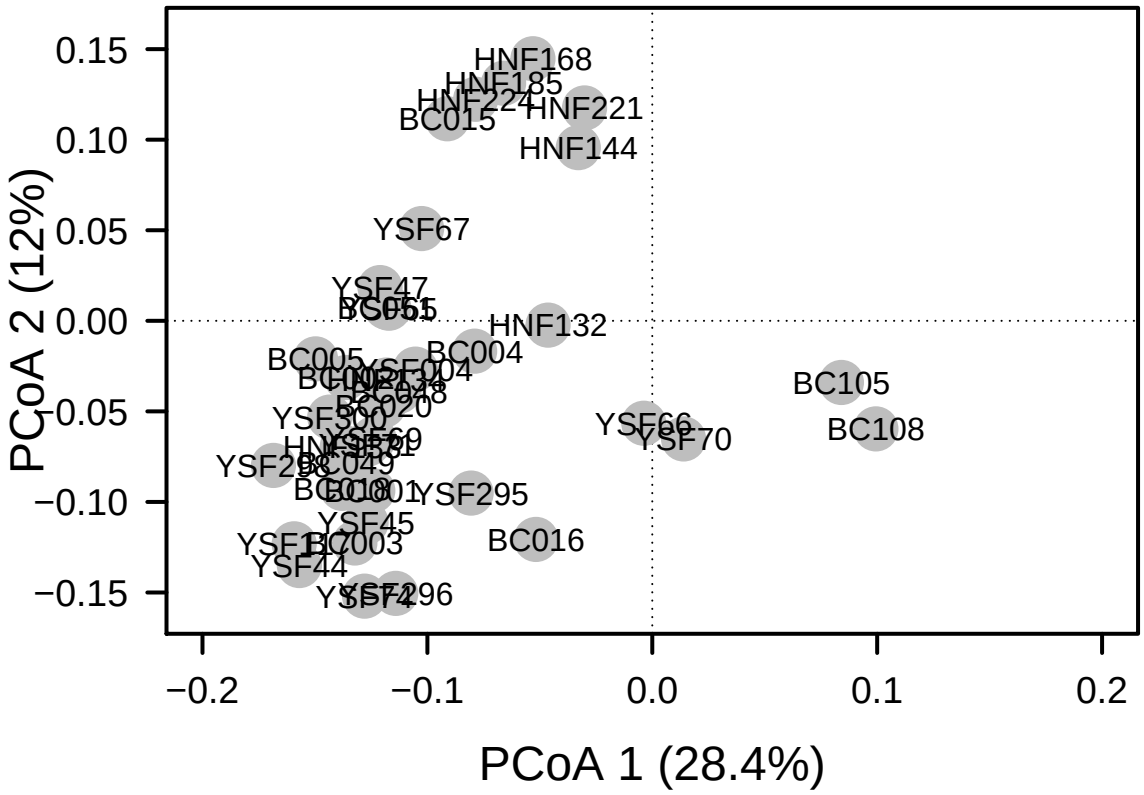


In the following R code chunk: 1. perform another PCoA on taxonomic data using an appropriate measure of dissimilarity, and 2. calculate the explained variation on the first three PCoA axes.

```
# create taxonomically-based distance matrix
pond.td <- vegdist(comm, method = "bray", binary = FALSE)

# perform PCoA
pond.td.pcoa <- cmdscale(pond.td, eig = T, k = 3)

# calculate variation explained by the first three axes
explainvar1 <- round(pond.td.pcoa$eig[1] / sum(pond.td.pcoa$eig), 3) * 100
explainvar2 <- round(pond.td.pcoa$eig[2] / sum(pond.td.pcoa$eig), 3) * 100
explainvar3 <- round(pond.td.pcoa$eig[3] / sum(pond.td.pcoa$eig), 3) * 100
sum.eig.t <- sum(explainvar1, explainvar2, explainvar3)
```

```
## plot
# define plot parameters
par(mar = c(5, 5, 1, 2) + 0.1)

#initiate plot
```

```
plot(pond.td.pcoa$points[ , 1], pond.td.pcoa$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, type = "n", cex.lab = 1.5, cex.axis = 1.2, axes = FALSE)

# add axes
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)

# add points and labels
points(pond.td.pcoa$points[ , 1], pond.td.pcoa$points[ , 2],
       pch = 19, cex = 3, bg = "gray", col = "gray")
text(pond.td.pcoa$points[ , 1], pond.td.pcoa$points[ , 2],
     labels = row.names(pond.td.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***: The PCoA run with taxonomic data explained quite a bit more of the variation than that run with phylogenetic data (49% vs 20.9%), and the sites seem to be more evenly dispersed when phylogenetic information is considered. This shows that in this system, taxonomic data is not the best proxy for phylogeny, and it's probably important to take phyl;ogeny into account if we want to make meaningful conclusions in this system.

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.

```
# define environmental category
watershed <- env$Location

# run PERMANOVA with 'adonis()' function {vegan}
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  0.029 *
## 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
```

```
# we can compare to PERMANOVA results based on taxonomy
adonis(
  vegdist(                          # create a distance matrix on
    decostand(comm, method = "log"), # log-transformed relative abundances
    method = "bray") ~ watershed,    # using Bray-Curtis dissimilarity metric
  permutations = 999)
```

```
##
## Call:
## adonis(formula = vegdist(decostand(comm, method = "log"), method = "bray") ~      watershed, permuta
##
## Permutation: free
## Number of permutations: 999
##
## Terms added sequentially (first to last)
##
##           Df SumsOfSqs  MeanSqs F.Model      R2 Pr(>F)
## watershed  2   0.16601 0.083003  1.5689 0.06018  0.009 **
## Residuals 49   2.59229 0.052904         0.93982
## Total     51   2.75829                  1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ii. Continuous Approach

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

environmental data matrix (after transforming and centering using `scale()`).

```
# define environmental variables
envs <- env[ , 5:19]

# remove redundant variables
envs <- envs[ , -which(names(envs) %in% c("TDS", "Salinity", "Cal_Volume"))]

# create distance matrix for environmental variables
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.

```
# conduct Mantel test ('vegan')
mantel(dist.uf, env.dist)
```

```
##
## Mantel statistic based on Pearson's product-moment correlation
##
## Call:
## mantel(xdis = dist.uf, ydis = env.dist)
##
## Mantel statistic r: 0.1604
##       Significance: 0.066
##
## Upper quantiles of permutations (null model):
##   90%   95% 97.5%   99%
## 0.134 0.171 0.208 0.236
## 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

```
# conduct dbRDA ('vegan')
ponds.dbrda <- vegan::dbrda(dist.uf ~ ., data = as.data.frame(scale(envs)))

# permutation tests: axes and environmental variables
anova(ponds.dbrda, by = "axis")
```

```
## Permutation test for dbrda under reduced model
## Forward tests for axes
## Permutation: free
## Number of permutations: 999
##
## Model: vegan::dbrda(formula = dist.uf ~ Elevation + Diameter + Depth + ORP + Temp + SpC + DO + pH + C
##          Df SumOfSqs      F Pr(>F)
## dbRDA1    1  0.10566 2.0152  0.428
## dbRDA2    1  0.09258 1.7658  0.606
## dbRDA3    1  0.07555 1.4409  0.962
## dbRDA4    1  0.06677 1.2735  0.999
```

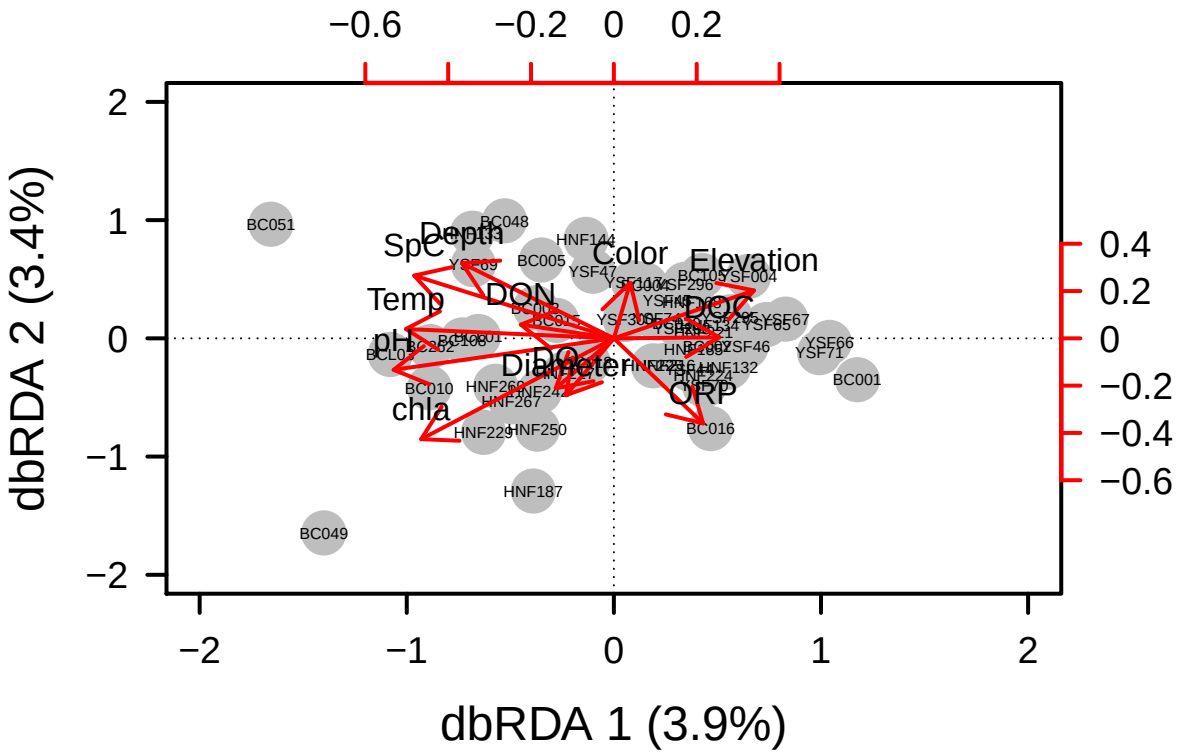
```
## dbRDA5    1  0.05666 1.0807  1.000
## dbRDA6    1  0.05293 1.0095  1.000
## dbRDA7    1  0.04750 0.9059  1.000
## dbRDA8    1  0.03941 0.7517  1.000
## dbRDA9    1  0.03775 0.7201  1.000
## dbRDA10   1  0.03280 0.6256  1.000
## dbRDA11   1  0.02876 0.5485  1.000
## dbRDA12   1  0.02501 0.4770  0.999
## Residual 39  2.04482
```

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

```
##
## ***VECTORS
##
##             dbRDA1   dbRDA2     r2 Pr(>r)
## Elevation  0.77670  0.62986 0.0959  0.093 .
## Diameter  -0.27972 -0.96008 0.0541  0.263
## Depth     -0.63137  0.77548 0.1756  0.014 *
## ORP        0.41879 -0.90808 0.1437  0.019 *
## Temp      -0.98250  0.18628 0.1523  0.024 *
## SpC       -0.77101  0.63682 0.2087  0.004 **
## DO        -0.39318 -0.91946 0.0464  0.291
## pH        -0.96210 -0.27270 0.1756  0.008 **
## Color      0.06353  0.99798 0.0464  0.311
## chla      -0.60392 -0.79704 0.2626  0.006 **
## DOC        0.99847 -0.05526 0.0382  0.408
## DON       -0.91633  0.40042 0.0339  0.446
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Permutation: free
## Number of permutations: 999
```

```
# calculate explained variation
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

# make dbRDA plot

# define plot parameters
par(mar = c(5, 5, 4, 4) + 0.1)

# initiate plot
plot(scores(ponds.dbrda, display = "wa"), 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)

# add axes
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)

# add points and labels
points(scores(ponds.dbrda, display = "wa"),
       pch = 19, cex = 3, bg = "gray", col = "gray")
text(scores(ponds.dbrda, display = "wa"),
     labels = row.names(scores(ponds.dbrda, display = "wa")), cex = 0.5)

# add environmental vectors
vectors <- scores(ponds.dbrda, display = "bp")
# row.names(vectors) <- c("Temp", "DO", "chla", "DON")
arrows(0, 0, vectors[ , 1]* 2, vectors[ , 2] * 2,
       lwd = 2, lty = 1, length = 0.2, col = "red")
text(vectors[ , 1] * 2, vectors[ , 2] * 2, pos = 3,
     labels = row.names(vectors))
axis(side = 3, lwd.ticks = 2, cex.axis = 1.2, las = 1, col = "red", lwd = 2.2,
     at = pretty(range(vectors[ , 1])) * 2, labels = pretty(range(vectors[ , 1])))
axis(side = 4, lwd.ticks = 2, cex.axis = 1.2, las = 1, col = "red", lwd  = 2.2,
     at = pretty(range(vectors[ , 2])) * 2, labels = pretty(range(vectors[ , 2])))
```



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

> ***Answer 6***: The Mantel test shows that correlations between phylogenetic beta diversity and environmental variables is trending towards significance but is not significant. Additionally, the first two dbRDA axes together explain only about 7.3% of variation in the communities. Therefore, it seems that something else besides environment seems to be structuring these communities.

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.

```
# geographic distances (kilometers) among ponds
long.lat <- as.matrix(cbind(env$long, env$lat))
coord.dist <- earth.dist(long.lat, dist = TRUE)

# taxonomic similarity among ponds (Bray-Curtis distance)
bray.curtis.dist <- 1 - vegdist(comm)

# phylogenetic similarity among ponds (UniFrac)
unifrac.dist <- 1 - dist.uf

# transform all distances into list format:
unifrac.dist.ls <- liste(unifrac.dist, entry = "unifrac")
bray.curtis.dist.ls <- liste(bray.curtis.dist, entry = "bray.curtis")
coord.dist.ls <- liste(coord.dist, entry = "geo.dist")
env.dist.ls <- liste(env.dist, entry = "env.dist")

# create a data frame from the lists of distances
df <- data.frame(coord.dist.ls, bray.curtis.dist.ls[ , 3], unifrac.dist.ls[ , 3],
                 env.dist.ls[ , 3])
names(df)[4:6] <- c("bray.curtis", "unifrac", "env.dist")
```

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.

```
# set initial plot parameters
par(mfrow = c(2, 1), mar = c(1, 5, 2, 1) + 0.1, oma = c(2, 0, 0, 0))

# make plot for taxonomic DD
plot(df$geo.dist, df$bray.curtis, xlab = "", xaxt = "n", las = 1, ylim = c(0.1, 0.9),
     ylab = "Bray-Curtis Similarity", main = "Distance Decay", col = "SteelBlue")
```

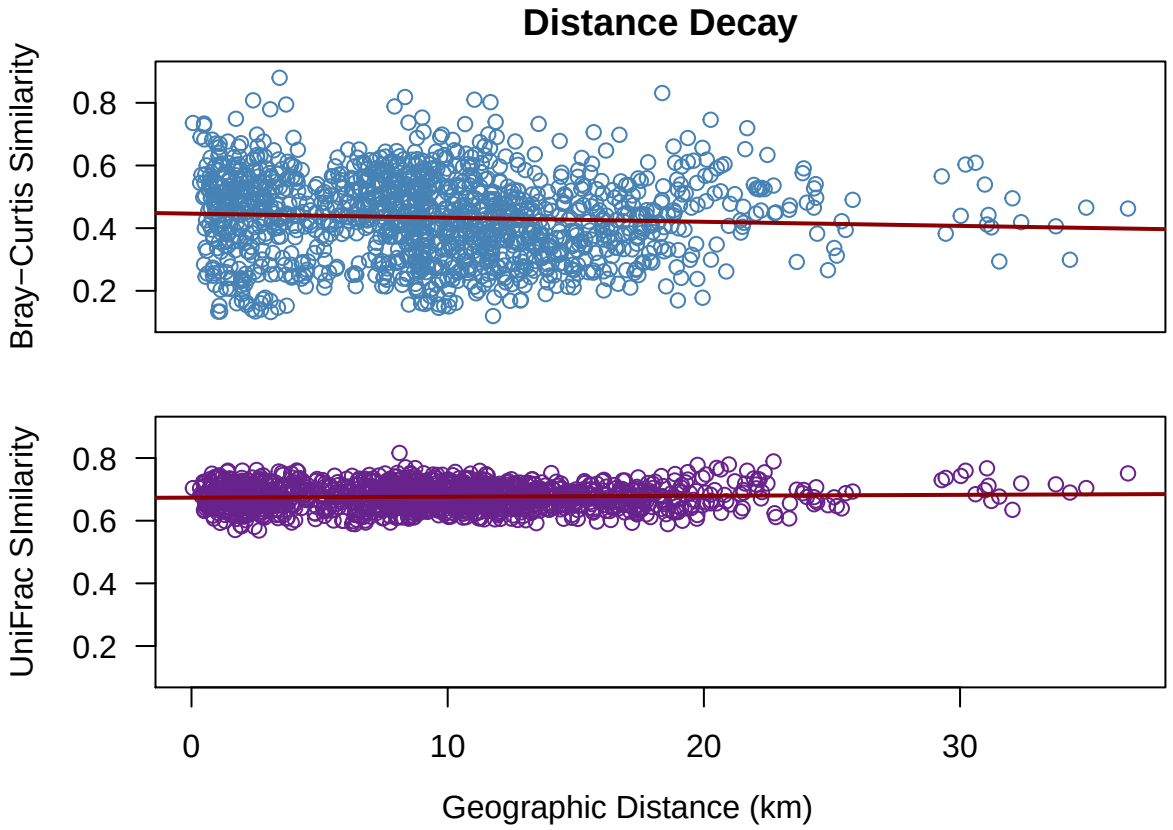
```r
# regression for taxonomic DD
DD.reg.bc <- lm(df$bray.curtis ~ df$geo.dist)
summary(DD.reg.bc)
```

```
##
## Call:
## lm(formula = df$bray.curtis ~ df$geo.dist)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.31151 -0.08843  0.00315  0.09121  0.43817
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.4463453  0.0066883  66.735   <2e-16 ***
## df$geo.dist -0.0013051  0.0005864  -2.226   0.0262 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1303 on 1324 degrees of freedom
## Multiple R-squared:  0.003728,   Adjusted R-squared:  0.002975
## F-statistic: 4.954 on 1 and 1324 DF,  p-value: 0.0262
```

```r
abline(DD.reg.bc, col = "red4", lwd = 2)

# new plot parameters
par(mar = c(2, 5, 1, 1) + 0.1)

# make plot for phylogenetic DD
plot(df$geo.dist, df$unifrac, xlab = "", las = 1, ylim = c(0.1, 0.9),
     ylab = "UniFrac SImilarity", col = "darkorchid4")

# regression for phylogenetic DD
DD.reg.uni <- lm(df$unifrac ~ df$geo.dist)
summary(DD.reg.uni)
```

```
##
## Call:
## lm(formula = df$unifrac ~ df$geo.dist)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.105629 -0.027107 -0.000077  0.026761  0.140215
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.6735186  0.0019206 350.677   <2e-16 ***
## df$geo.dist 0.0002976  0.0001684   1.767   0.0774 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03741 on 1324 degrees of freedom
## Multiple R-squared:  0.002354,   Adjusted R-squared:  0.0016
## F-statistic: 3.124 on 1 and 1324 DF,  p-value: 0.07738
```

```r
abline(DD.reg.uni, col = "red4", lwd = 2)

# add x-axis label to plot
mtext("Geographic Distance (km)", side = 1, adj = 0.55,
      line = 0.5, outer = TRUE)
```



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

```r
diffslope(df$geo.dist, df$unifrac, df$geo.dist, df$bray.curtis)
```

```
##
## Is difference in slope significant?
## Significance is based on 1000 permutations
##
## Call:
## diffslope(x1 = df$geo.dist, y1 = df$unifrac, x2 = df$geo.dist,      y2 = df$bray.curtis)
##
## Difference in Slope: 0.001603
## Significance: 0.003
##
## Empirical upper confidence limits of r:
##      90%      95%    97.5%      99%
## 0.000751 0.000985 0.001200 0.001390
```

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

Answer 7: Taxonomic similarity shows a slight but significant negative slope, which is evidence

for a distance decay. However, phylogenetic similarity does not have a significant slope, meaning that there is no relationship between distance and phylogenetic similarity. This may be because of dispersal limitation or because of environmetal variables that autocorrelate with space; it is likely not evolutionary history because then we would expect to also see a distance decay in phylogenetic similarity.

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.

I'm broadly interested in feedbacks between plant invasions and soil microbial communities, so I can think of several ways that this kind of analysis would be useful. I think it would be very interesting to compare soil microbial communities in the invaded ranges of plants with those in the plants' native ranges; a less diverse community in the invaded range may be indicative of the enemy release hypothesis, while a more diverse community may be indicative of newly acquired mutualists. In addition, it would be interesting to compare soil microbial communities of ecosystems that have been invaded by a plant vs those that have not. There have been a lot of hypotheses about soil microbes and invasions, but from what I've read it seems like none of them is universally right– similar to community structuring theory, it seems like different mechanisms may all be at play at varying strengths in different invasions, and at different ages of invasion. It would be especially interesting to compare phylogenies of invaded and native range soils, becuase that could help determine whether invading plants are acquiring new mutualists or whether their microbial mutualists are dispersing along with them. With OTU data from both invaded and native ranges, comparing the phylogenies would allow me to see if plants are invading areas with communities that are novel to them, or whether they're living in communities that are actually quite similar to the ones they left behind, whether this is due to microbial dispersion or to microbes from similar taxa already existing in the invaded range.