Phylogenetic Diversity - Communities

Alison Partee; Z620: Quantitative Biodiversity, Indiana University

01 March, 2017

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 assignment, 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. Change "Student Name" on line 3 (above) with your name.
2. Complete as much of the assignment as possible during class; what you do not complete in class will need to be done outside of class.
3. Use the handout as a guide; it contains a more complete description of data sets along with the proper scripting needed to carry out the exercise.
4. Be sure to **answer the questions** in this exercise document; they also correspond to the handout. Space for your answer is provided in this document and indicated by the ">" character. If you need a second paragraph be sure to start the first line with ">".
5. Before you leave the classroom, **push** this file to your GitHub repo.
6. When you are done, **Knit** the text and code into a PDF file.
7. After Knitting, please submit the completed assignment by creating a **pull request** via GitHub. Your pull request should include this file *PhyloCom_assignment.Rmd* and the PDF output of `Knitr` (*PhyloCom_assignment.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/flopsei/GitHub/QB2017_Partee/Week7-PhyloCom"
```

```
setwd("/Users/flopsei/GitHub/QB2017_Partee/Week7-PhyloCom")

package.list <- c('picante', 'ape', 'seqinr','vegan', 'fossil', 'simba')
```

```r
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)
  }
}
```

```
## This is vegan 2.4-2

##
## 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
```

```r
source('./bin/MothurTools.R')
```

```
## Loading required package: reshape
```

2) DESCRIPTION OF DATA

We will revisit the data that was used in the Spatial Diversity module. As a reminder, in 2013 we sampled ~ 50 forested ponds located in Brown County State Park, Yellowwood State Park, and Hoosier National Forest in southern Indiana. See the handout for a further description of this week's dataset.

3) LOAD THE DATA

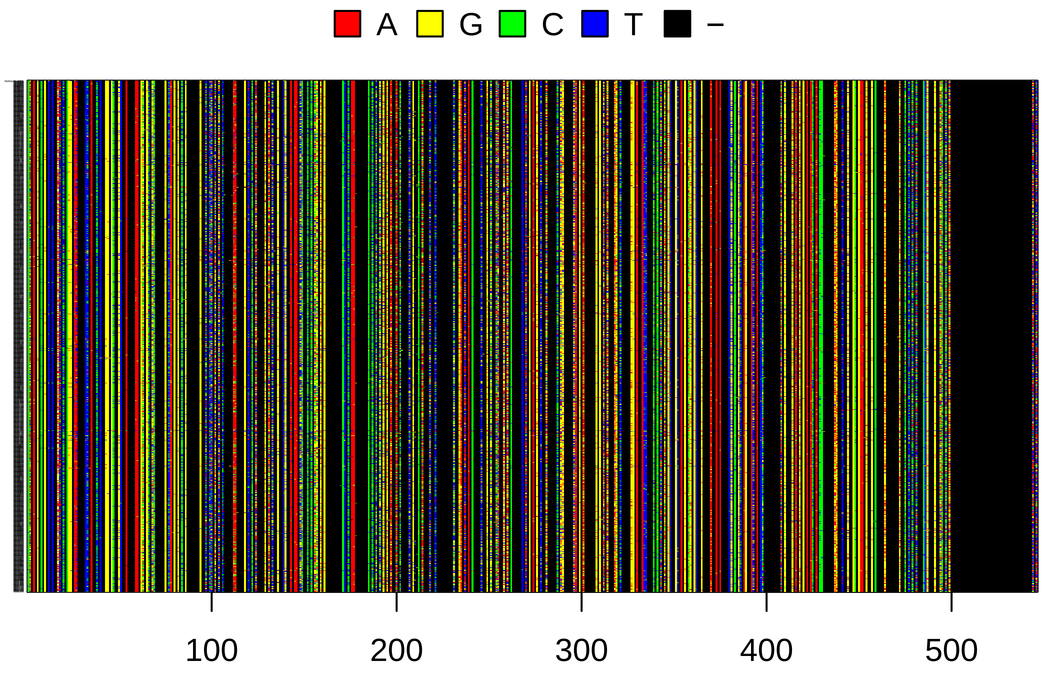
In the R code chunk below, do the following:
1. load the environmental data for the Brown County ponds (*20130801_PondDataMod.csv*),

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.

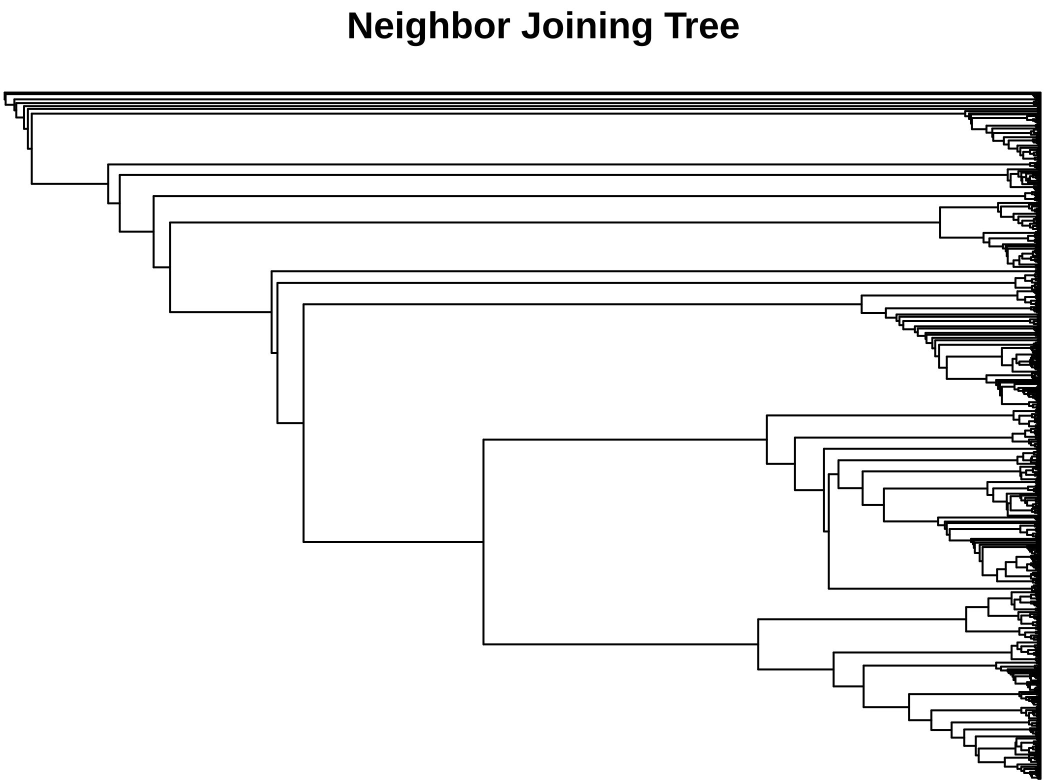
```r
env <- read.table("data/20130801_PondDataMod.csv", sep = ',', header = TRUE)
env <- na.omit(env)

# 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 the environmental data set
comm <- comm[rownames(comm) %in% env$Sample_ID, ]

# Remove zero-abundance OTUs 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.

```r
# Import the alignment file ('sequinr')
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-occurrence 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 = F
```



4) PHYLOGENETIC ALPHA DIVERSITY

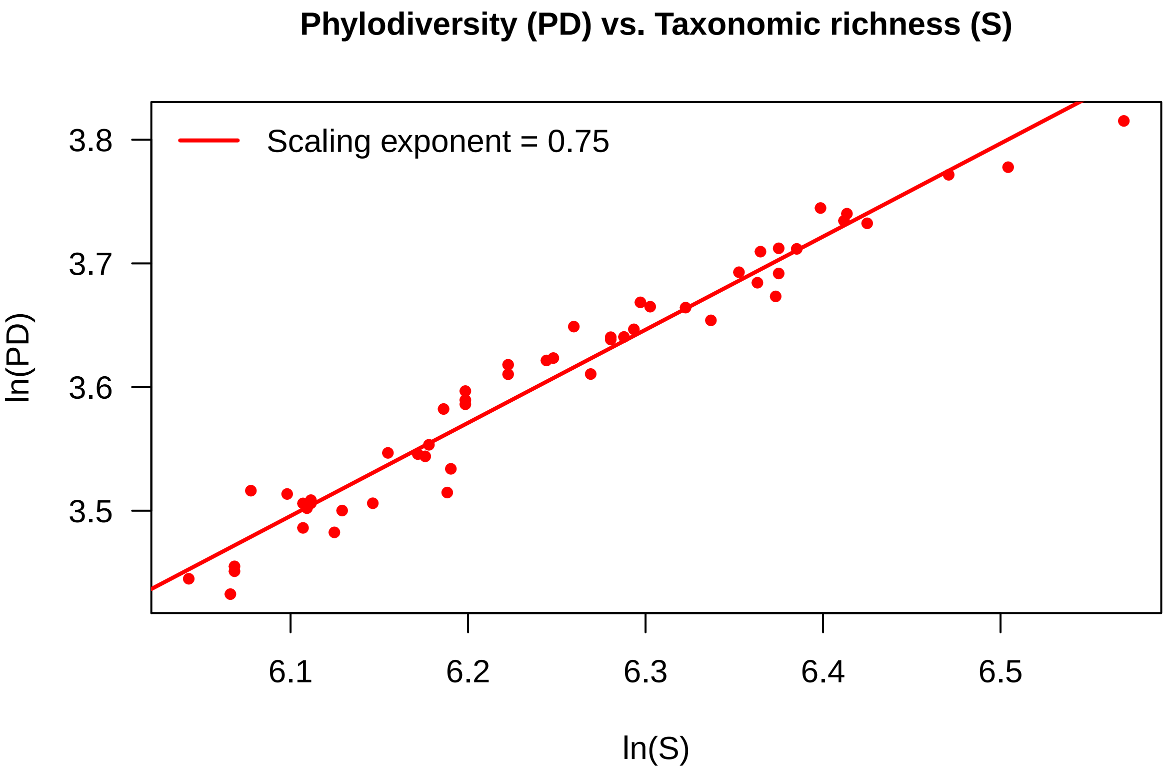
A. Faith's Phylogenetic Diversity (PD)

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

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

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

```r
# 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 should be related to taxonomic richness because it calculates diversity using only presence-absence data as opposed to abundance data. ***Answer 1b***: Taxonomic richness indicates the amount of taxa present in a sample, while phylodiversity uses both richness and relatedness to calculate a diversity measure. ***Answer 1c***: I would expect the two estimates to deviate from another when the species in the sample are not all equally related. ***Answer 1d***: The PD-D scaling exponent tells us how strongly the two measures are correlated with each other.

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 = F)
ses.pd
```

```
##       ntaxa   pd.obs pd.rand.mean pd.rand.sd pd.obs.rank   pd.obs.z
## BC001   668 43.71912     44.01279  0.8198694           8 -0.3581889
## BC002   587 40.94334     39.66075  0.7102842          26  1.8057340
##        pd.obs.p runs
## BC001 0.3076923   25
## BC002 1.0000000   25
```

```
# two other null models
ses.pd(comm[1:2,], phy, null.model = 'taxa.labels', runs = 25, include.root = F)
```

```
##       ntaxa   pd.obs pd.rand.mean pd.rand.sd pd.obs.rank  pd.obs.z
## BC001   668 43.71912     43.97769  0.8942531           8 -0.289148
## BC002   587 40.94334     39.91182  0.7082705          26  1.456387
##        pd.obs.p runs
## BC001 0.3076923   25
## BC002 1.0000000   25
```

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

```
##       ntaxa   pd.obs pd.rand.mean pd.rand.sd pd.obs.rank   pd.obs.z
## BC001   668 43.71912     43.95112  0.9731711          11 -0.2383979
## BC002   587 40.94334     39.92323  0.9518718          22  1.0716873
##        pd.obs.p runs
## BC001 0.4230769   25
## BC002 0.8461538   25
```

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

a. What are the null and alternative hypotheses you are testing via randomization when calculating `ses.pd`?
b. How did your choice of null model influence your observed ses.pd values? Explain why this choice affected or did not affect the output.

> ***Answer 2a***: When calculating ses.pd, I am using the taxa.labels null model, the richness null model, and the sample.pool null models. ***Answer 2b***: While the choice of null model did not affect the overall outcome of the observed ses.pd values, i.e. whether they were positive or negative, it did influence their values. Using richness as the null model, the site BC001 had a lower ses.pd value (in absolute value) than the value using the other two null models. This result makes sense because richness should be the most accurate null model out of the three I have chosen, meaning it makes sense that the value is lower because we expect the pd.obs and pd.rand.mean difference to be smaller. Additionally, the ses.pd value for the second site is higher than the values using the other two null models. This might be because the species in the second site are overdispersed.

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 = T, 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'
NRI
```

```
##                 NRI
## BC001   0.518788770
## BC002   0.578621616
## BC003   0.827785475
## BC004   0.112976779
## BC005   0.516939236
## BC010   0.530658642
## BC015   0.381657274
## BC016   0.609974687
## BC018   0.409597788
## BC020   0.510379818
## BC048   0.272816136
## BC049   0.319529454
## BC051  -0.176238079
## BC105   0.012082492
## BC108   0.228750431
## BC262   0.108493097
## BCL01   0.130825169
## BCL03   0.022597544
## HNF132  0.211569343
## HNF133  0.232129356
## HNF134  0.456739320
## HNF144  0.093240798
## HNF168  0.176044039
## HNF185  0.364375899
## HNF187  0.687248033
## HNF216  0.657215923
## HNF217  0.511162431
## HNF221  0.016442484
## HNF224  0.460879922
## HNF225  0.655314877
## HNF229  0.110163520
## HNF242  0.357069006
## HNF250  0.080298959
## HNF267 -0.002215779
## HNF269  0.090601678
## YSF004 -0.280975408
## YSF117  0.786846977
## YSF295 -0.774151920
## YSF296  0.892011023
## YSF298  0.909734920
## YSF300  0.379978473
## YSF44   0.614240170
## YSF45   0.685874640
## YSF46   1.198853027
## YSF47   0.494629495
## YSF65   0.614859114
## YSF66   0.219376966
## YSF67   0.214756922
## YSF69   0.183935928
## YSF70  -0.042682661
## YSF71   0.767876546
## YSF74   0.871943584
```

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 = T, 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'
NTI
```

```
##               NTI
## BC001   0.9030157
## BC002   1.9048461
## BC003   1.0899381
## BC004   1.5357628
## BC005   2.4636707
## BC010   0.5840485
## BC015   1.2952960
## BC016   2.3838188
## BC018   1.4214723
## BC020   1.2564752
## BC048   1.5319812
## BC049   2.2363031
## BC051   2.2446722
## BC105   1.6909357
## BC108   1.4595438
## BC262   1.2227120
## BCL01   1.5801689
## BCL03   1.2142771
## HNF132  1.6476846
## HNF133  1.3270581
## HNF134  1.6249185
## HNF144  0.9733421
## HNF168  1.0333628
## HNF185  1.7701414
## HNF187  0.6639979
## HNF216  0.2933652
## HNF217  0.3837620
## HNF221  1.2809659
## HNF224  1.6443787
## HNF225  0.3476905
## HNF229  1.5422846
## HNF242  1.4697452
## HNF250  1.3269778
## HNF267  1.1114787
## HNF269  1.0491502
## YSF004  0.3447645
## YSF117  1.5714501
## YSF295 -0.4214045
## YSF296  1.7639043
## YSF298  1.9200778
## YSF300  1.9010237
## YSF44   1.1105944
## YSF45   1.4400623
## YSF46   1.5239837
## YSF47   1.0993420
## YSF65   1.5750132
## YSF66   1.1984651
## YSF67   1.4023920
## YSF69   1.4876302
## YSF70   2.0316719
## YSF71   1.3011999
## YSF74   1.2672869
```

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: To calculate NRI, you find the average branch distance between pairs of taxa in a sample, and then we calculate the percent difference between our calculated average branch lengths and the average branch lengths found in a null model. ***Answer 3b***: To calculate NTI, you find the average distance between each taxa and its closest neighbor, and then we calculate the difference between that and what we would expect under a null model. ***Answer 3c***: According to both the NRI and NTI values above, most of the sites here are overdispered. However, the NTI values had less negative values than the NRI values, possibly indicating high amounts of clustering in our data. ***Answer 3d***: After allowing the indices to account for abundance data, the returned values tended to be more positive than those when using only presence-absence data. The most significant change was in the NRI values, as most of those values changed their sign.

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.

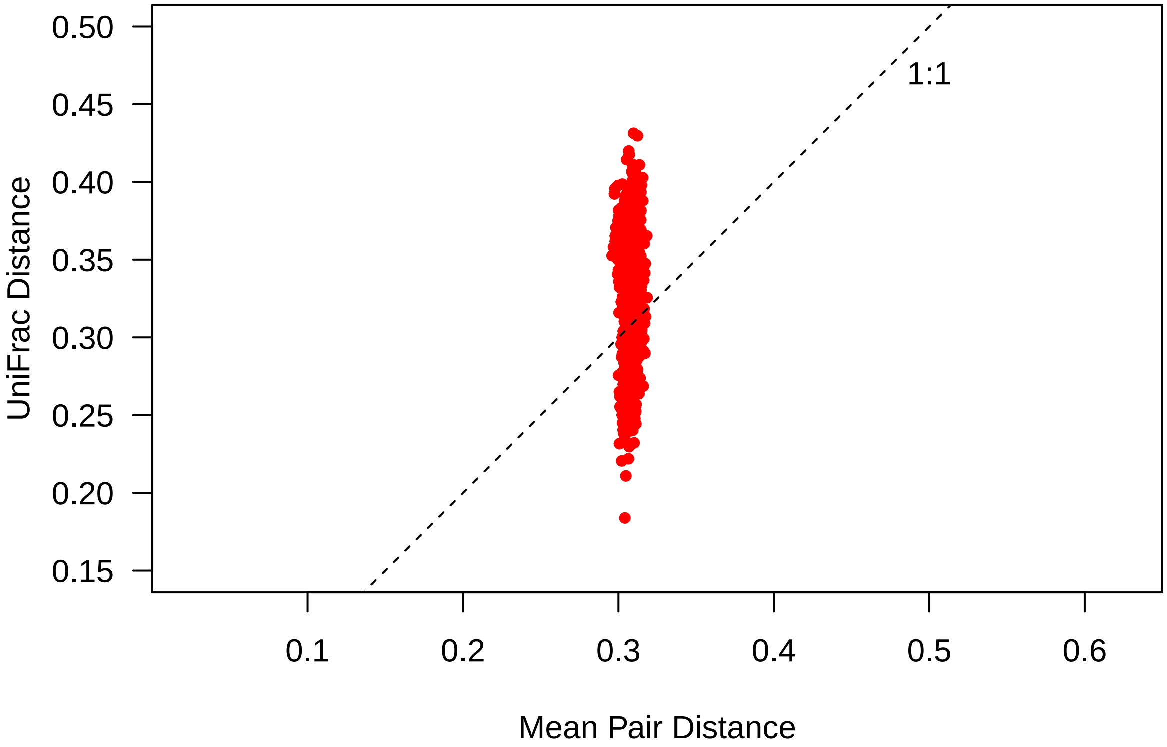
```r
# mean pairwise distance
dist.mp <- comdist(comm, phydist)
```

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

```r
# UniFrac distance (note: this takes a few minutes, be patient)
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
# compare the mean pair distance and unifrac distance matrices
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, .5), ylim = c(.15, 0.5),
abline(b = 1, a = 0, lty = 2)
text(0.5, 0.47, '1:1')
```



Question 4:

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

> ***Answer 4a***: The Mean Pair Distance, MPD, is the average branch length distance between two taxa in a tree. The UniFrac distance does not use the lengths between two taxa in its calculation, it uses the sum of the unshared branch lengths between two taxa and divides it by the sum of all branch lengths. ***Answer 4b***: Based on our plot above, UniFrac distance varies a lot compared to MPD values, which all fall around 0.3. There appears to be little correlation. ***Answer 4c***: Perhaps MPD values fall within a different, smaller range than UniFrac values, leading MPD values to have less variation.

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.

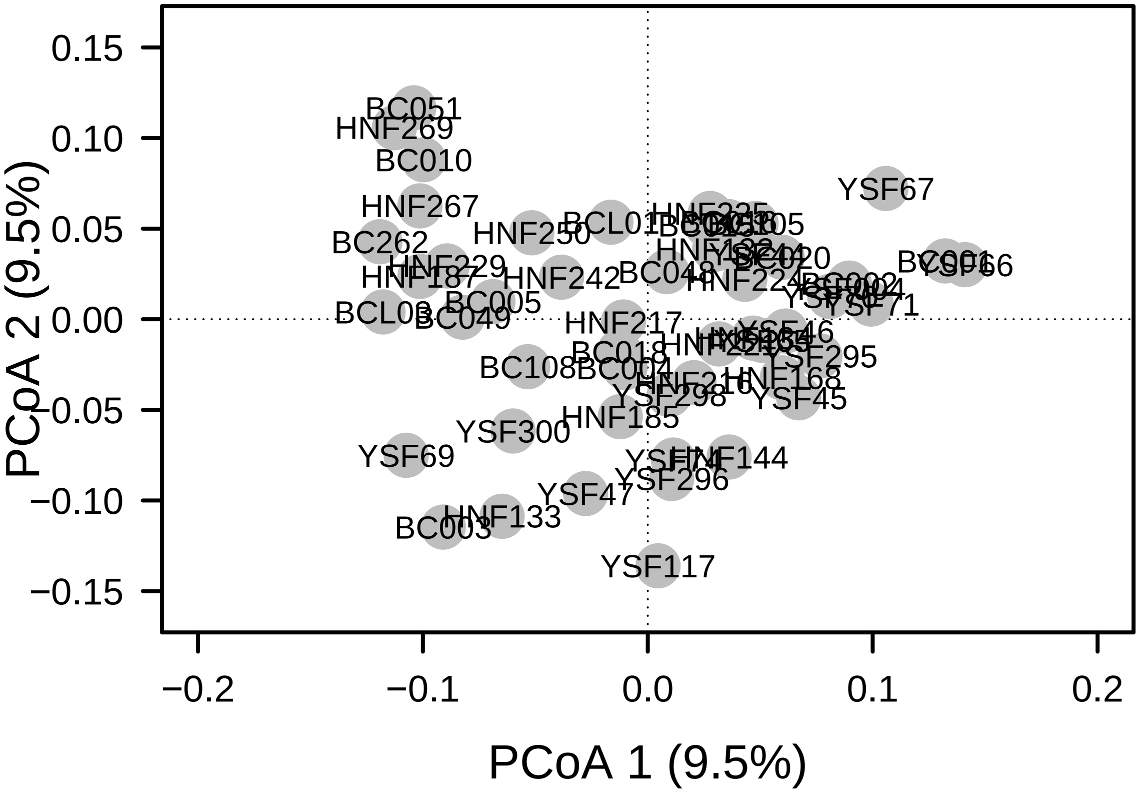
```
pond.pcoa <- cmdscale(dist.uf, eig = T, k = 3)
```

```
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.

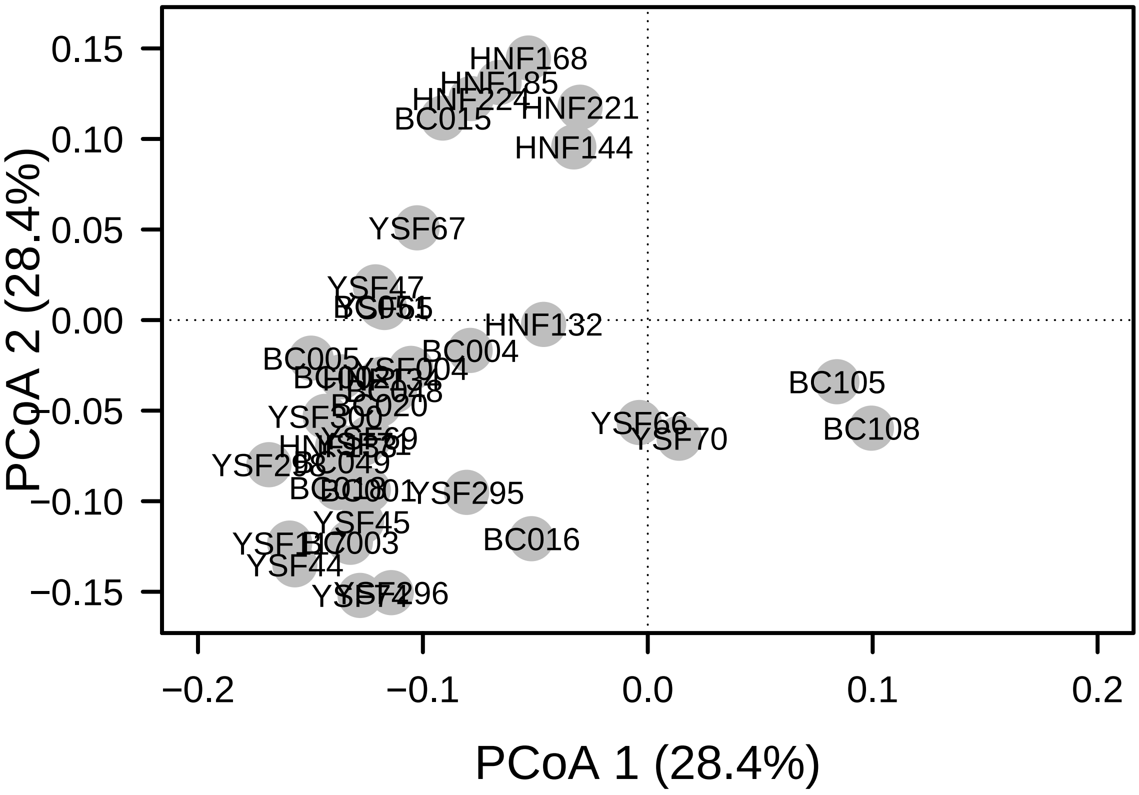
In the R code chunk below, do the following:
1. plot the PCoA results using either the R base package or the `ggplot` package,
2. include the appropriate axes,
3. add and label the points, and
4. customize the plot.

```
# 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(-.16, .16), xlab = paste

# 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.

```r
#calculate bray-curtis
pond.db <- vegdist(comm, method = 'bray')

pond.pcoa <- cmdscale(pond.db, 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)


# 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(-.16, .16), xlab = paste

# 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))
```



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 ordination contrasts sharply with the taxonomic based ordination. In the phylogenetically based ordination, the sites are much more spread out across the plot area whereas in the taxonomic based ordination, most of the sites are on the left side of the plot, possibly indicating that most of the variation is accounted for in the first axis.

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  0.031 *
## 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(decostand(comm, method = 'log'),
    method = 'bray') ~ watershed,
  permutations = 999)
```

```
##
## Call:
## adonis(formula = vegdist(decostand(comm, method = "log"), method = "bray") ~      watershed, permuta
##
## Permutation: free
## Number of permutations: 999
##
## Terms added sequentially (first to last)
##
##           Df SumsOfSqs  MeanSqs F.Model      R2 Pr(>F)
## watershed  2   0.16601 0.083003  1.5689 0.06018  0.005 **
## 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()`).

```r
# 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.

```r
# 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: 0.056
##
## Upper quantiles of permutations (null model):
##   90%   95% 97.5%   99%
## 0.127 0.163 0.204 0.248
## 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

```r
# 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
## Marginal 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  0.10566 2.0152  0.002 **
## dbRDA2    1  0.09258 1.7658  0.002 **
## dbRDA3    1  0.07555 1.4409  0.030 *
## dbRDA4    1  0.06677 1.2735  0.088 .
## dbRDA5    1  0.05666 1.0807  0.317
## dbRDA6    1  0.05293 1.0095  0.444
## dbRDA7    1  0.04750 0.9059  0.681
## dbRDA8    1  0.03941 0.7517  0.900
## dbRDA9    1  0.03775 0.7201  0.931
## dbRDA10   1  0.03280 0.6256  0.987
## dbRDA11   1  0.02876 0.5485  0.999
## dbRDA12   1  0.02501 0.4770  1.000
## Residual 39  2.04482
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

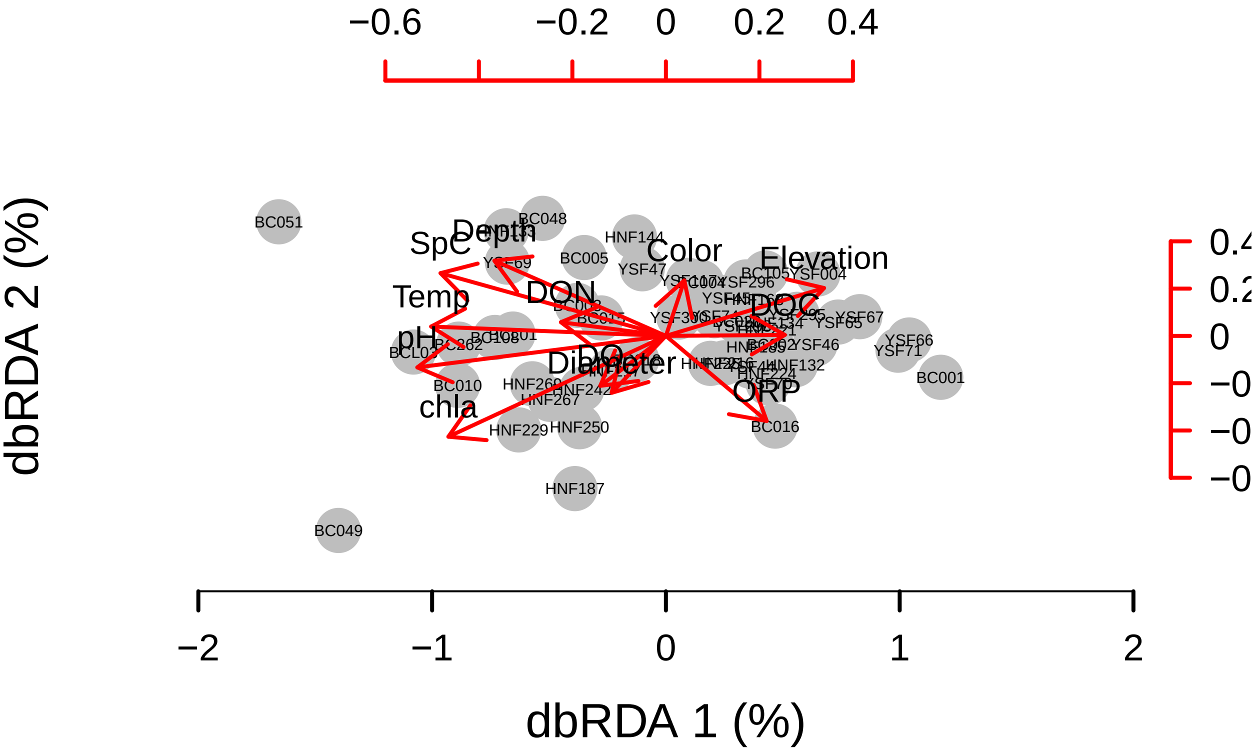
```
ponds.fit <- envfit(ponds.dbrda, envs, perm = 999)
ponds.fit
```

```
##
## ***VECTORS
##
##             dbRDA1   dbRDA2     r2 Pr(>r)
## Elevation  0.77670  0.62986 0.0959  0.094 .
## Diameter  -0.27972 -0.96008 0.0541  0.247
## Depth     -0.63137  0.77548 0.1756  0.014 *
## ORP        0.41879 -0.90808 0.1437  0.023 *
## Temp      -0.98250  0.18628 0.1523  0.020 *
## SpC       -0.77101  0.63682 0.2087  0.007 **
## DO        -0.39318 -0.91946 0.0464  0.309
## pH        -0.96210 -0.27270 0.1756  0.010 **
## Color      0.06353  0.99798 0.0464  0.308
## chla      -0.60392 -0.79704 0.2626  0.012 *
## DOC        0.99847 -0.05526 0.0382  0.363
## DON       -0.91633  0.40042 0.0339  0.418
## ---
## 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$CCS$eig[1] / sum(c(ponds.dbrda$CCA$eig, ponds.dbrda$CA$eig)), 3)
dbrda.explainvar2 <- round(ponds.dbrda$CCS$eig[2] / sum(c(ponds.dbrda$CCA$eig, ponds.dbrda$CA$eig)), 3)

# make dbRDA plot

# initiate plot
plot(scores(ponds.dbrda, display = 'wa'), xlim = c(-2, 2), ylim = c(-2, 2), xlab = paste('dbRDA 1 (', d

# add axes
axis(side = 1, labels = T, lwd.ticks = 2, cex.lab = 1.5, cex.axis = 1.2, axes = F)
```

```
## Warning in axis(side = 1, labels = T, lwd.ticks = 2, cex.lab = 1.5,
## cex.axis = 1.2, : "axes" is not a graphical parameter
```

```
# add points & 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 =

# add envitonmental 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(vector
axis(side = 4, lwd.ticks = 2, cex.axis = 1.2, las = 1, col = 'red', lwd = 2.2, at = pretty(range(vector
```



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***: Bacterial communities in these Indiana ponds are strongly driven by multiple different factors. The factors with the most influence, the ones with the longest arrows on our dbRDA, are depth, SpC, temperature, pH, chla, and elevation. Sites within similar locations (e.g., the BC sites) also tend to be grouped together, which makes sense considering they probably are more closely related to each other than sites that are farther apart.

6) SPATIAL PHYLOGENETIC COMMUNITY ECOLOGY

A. Phylogenetic Distance-Decay (PDD)

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 = T)

# 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 to 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.

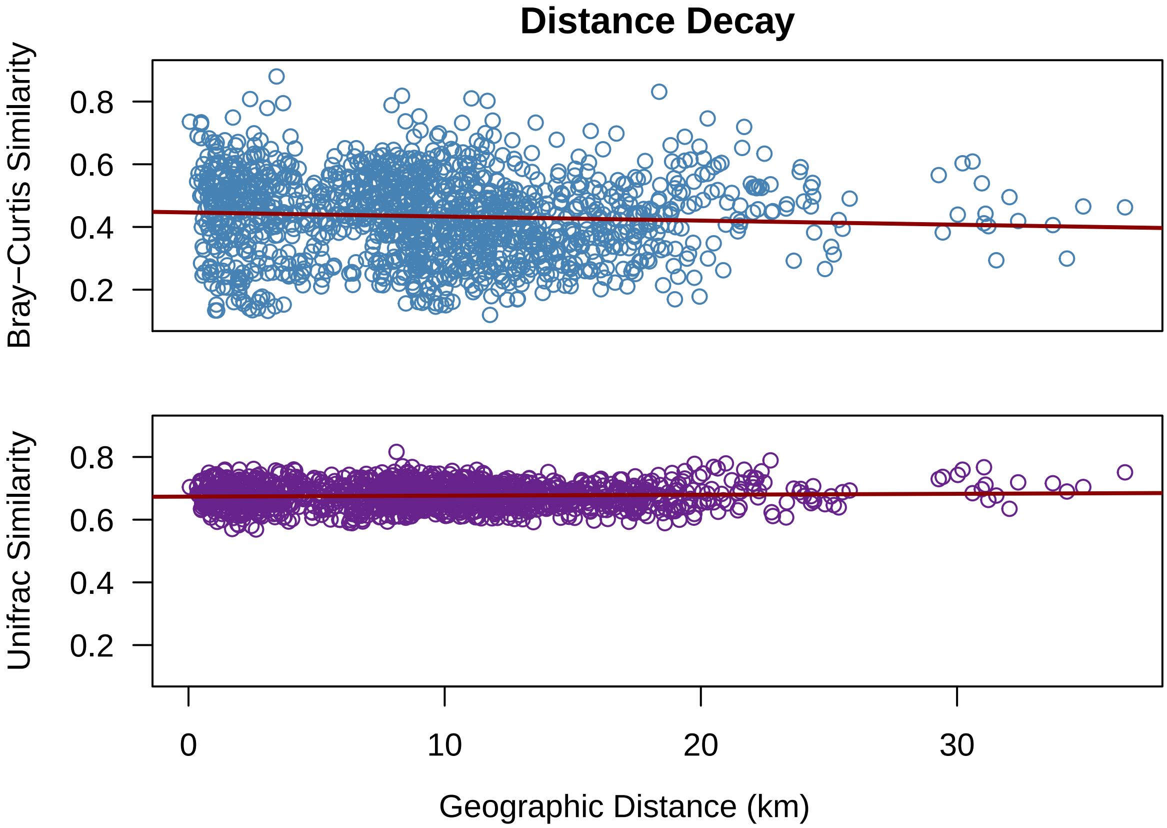
```
# set initial plot parameters
par(mfrow=c(2,1), mar = c(1, 5, 2, 1) + 0.1, oma = c(2, 0, 0, 0))

# 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')

# 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
```

```
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 =

# 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
```

```
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 = T)
```



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: 0.002
##
## Empirical upper confidence limits of r:
##      90%      95%    97.5%      99%
## 0.000773 0.001059 0.001237 0.001460
```

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

> ***Answer 7***: There are differences between the taxonomic and phylogenetic DD relationships. The taxonomic DD relationship has a negative slope, indicating that taxonomic similarity decreased with increasing distance, which makes sense because you would expect more closely related species to be closer in geographic distance to one another. The phylogenetic DD relationship had a slightly positive slope, indicating that samples become phylogenetically clustered with increasing geographic distance, since we are using Unifrac similarity and not dissimilarity. This also makes sense because we would expect the dispersion of traits to become more clustered with increading

distance because of migration limitations.

B. Phylogenetic diversity-area relationship (PDAR)

i. Constructing the PDAR

In the R code chunk below, write a function to generate the PDAR.

```r
PDAR <- function(comm, tree) {

  # create objects to hold areas and diversity
  areas <- c()
  diversity <- c()

  # create vector increasing number of plots by 2x
  num.plots <- c(2, 4, 8, 16, 32, 51)

  for (i in num.plots) {
    #create vectors to hold areas and diversity form iterations, used for means
    areas.iter <- c()
    diversity.iter <- c()

    # iterate 10 times per sample size
    for (j in 1:10) {
      # sample w/o replacement
      pond.sample <- sample(51, replace = F, size = i)

      # create variable and vector to hold accumulating area and taxa
      area <- 0
      sites <- c()

      for (k in pond.sample) { #loop through each randomly drawn pond
        area <- area + pond.areas[k] # aggregating area (roughly doubling)
        sites <- rbind(sites, comm[k,]) # and sites
      }

      # concatenate the area to areas.iter
      areas.iter <- c(areas.iter, area)
      # calculate PSV or other phylogenetic alpha-diversity metric
      psv.vals <- psv(sites, tree, compute.var = F)
      psv <- psv.vals$PSVs[1]
      diversity.iter <- c(diversity.iter, as.numeric(psv))
    }

    diversity <- c(diversity, mean(diversity.iter)) # let diversity be the mean PSV
    areas <- c(areas, mean(areas.iter)) # let areas be the average area
    print(c(i, mean(diversity.iter), mean(areas.iter))) # print as we go
  }

  # return vectors of areas (x) and diversity (y)
  return(cbind(areas, diversity))
}
```

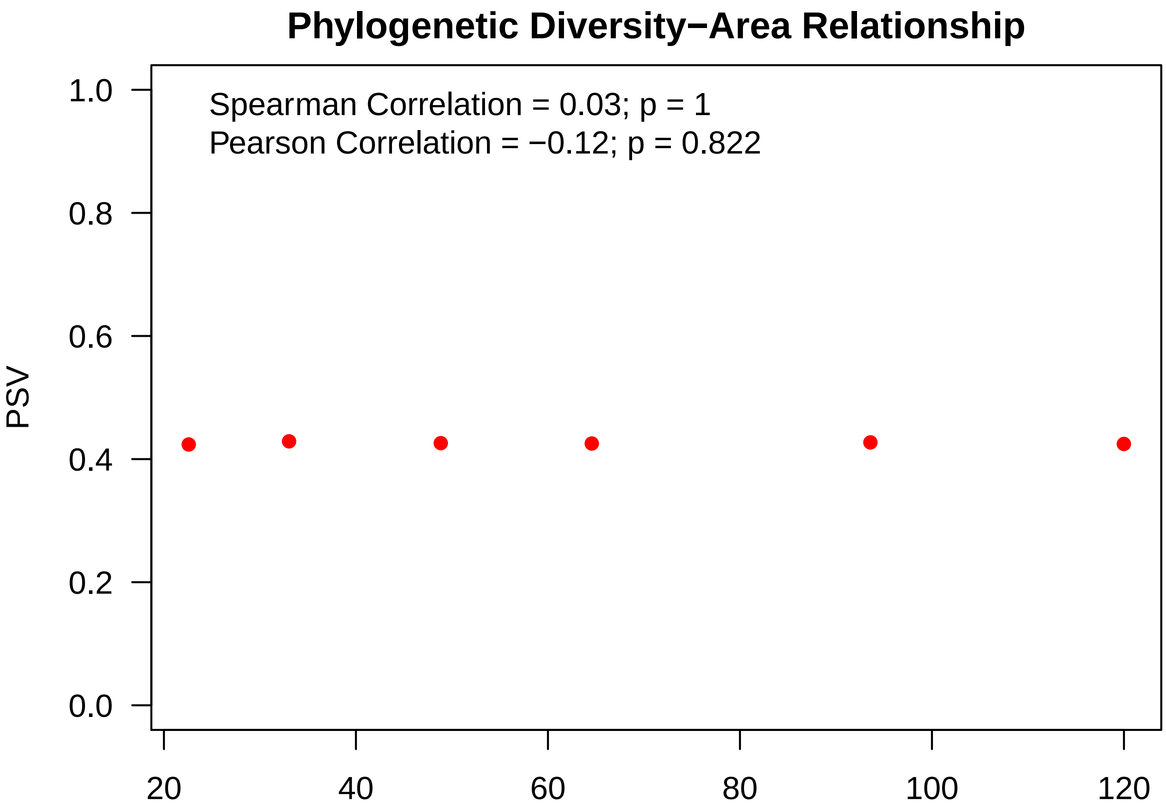
ii. Evaluating the PDAR

In the R code chunk below, do the following:
1. calculate the area for each pond,
2. use the `PDAR()` function you just created to calculate the PDAR for each pond,
3. calculate the Pearson's and Spearman's correlation coefficients,
4. plot the PDAR and include the correlation coefficients in the legend, and
5. customize the PDAR plot.

```r
# calculate areas for ponds: find areas of all ponds
pond.areas <- as.vector(pi * (env$Diameter/2)^2)

# compute the PDAR
pdar <- PDAR(comm, phy)
```

```
## [1]   2.0000000   0.4237395 510.0524899
## [1]    4.0000000    0.4288052 1091.3453056
## [1]    8.0000000    0.4258756 2385.1858926
## [1]   16.0000000    0.4253005 4168.9335066
## [1]   32.0000000    0.4270747 8759.1899310
## [1] 5.100000e+01 4.245666e-01 1.439763e+04
```

```r
pdar <- as.data.frame(pdar)
pdar$areas <- sqrt(pdar$areas)

# calculate Pearson's correlation coefficient
Pearson <- cor.test(pdar$areas, pdar$diversity, method = 'pearson')
P <- round(Pearson$estimate, 2)
P.pval <- round(Pearson$p.value, 3)

# calculate spearman's correlation coefficient
Spearman <- cor.test(pdar$areas, pdar$diversity, method = 'spearman')
rho <- round(Spearman$estimate, 2)
rho.pval <- round(Spearman$p.value, 3)

# plot the PDAR
plot.new()
par(mfrow = c(1,1), mar = c(1, 5, 2, 1) + 0.1, oma = c(2, 0, 0, 0))
plot(pdar[,1], pdar[,2], xlab = 'Area', ylab = 'PSV', ylim = c(0,1),
     main = 'Phylogenetic Diversity-Area Relationship',
     col = 'red', pch = 16, las = 1)

legend('topleft', legend = c(paste('Spearman Correlation = ', rho, '; p = ', rho.pval, sep = ''), paste
```



Question 8: Compare your observations of the microbial PDAR and SAR in the Indiana ponds? How might you explain the differences between the taxonomic (SAR) and phylogenetic (PDAR)?

> ***Answer 8***: In the SAR plot, richness clearly increased logarithmically with increasing area. However, in the PDAR the relationship was less strong, and it appears that with increasing geographic distance, phylogenetic species variability only minimally increases, if at all. This may be because we are sampling sites with similar community compositions but with different species that are analogous to the similar species in the other sites, and so when we continue to sample more area, we do not increase the phylogenetic species variability. The SAR should always increase because we are constantly finding new species when we sample more sites.

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.

> ***Synthesis Answer***: In our Daphnia disease ecology lab, phylogenetic information could be useful for answering evolutionary questions. We could make an eco-evolutionary experiment that measures the change in genotype frequency in response to some variable such as increased competition, introduction of predators, or food density changes. We would need the phylogenetic information of a Daphnia population during different time points in our data collection, and then we could identify why and how those genotype frequencies change.