Phylogenetic Diversity - Communities

Ashwini R; Z620: Quantitative Biodiversity, Indiana University

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] "C:/Users/DELL/GitHub/QB2017_Ramesh/Week7-PhyloCom"
```

```
#setwd("~ /Users/DELL/GitHub/QB2017_Ramesh/Week7-PhyloCom")
package.list <- c('picante', 'ape', 'seqinr', 'vegan', 'fossil', 'simba','reshape')
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
```

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.

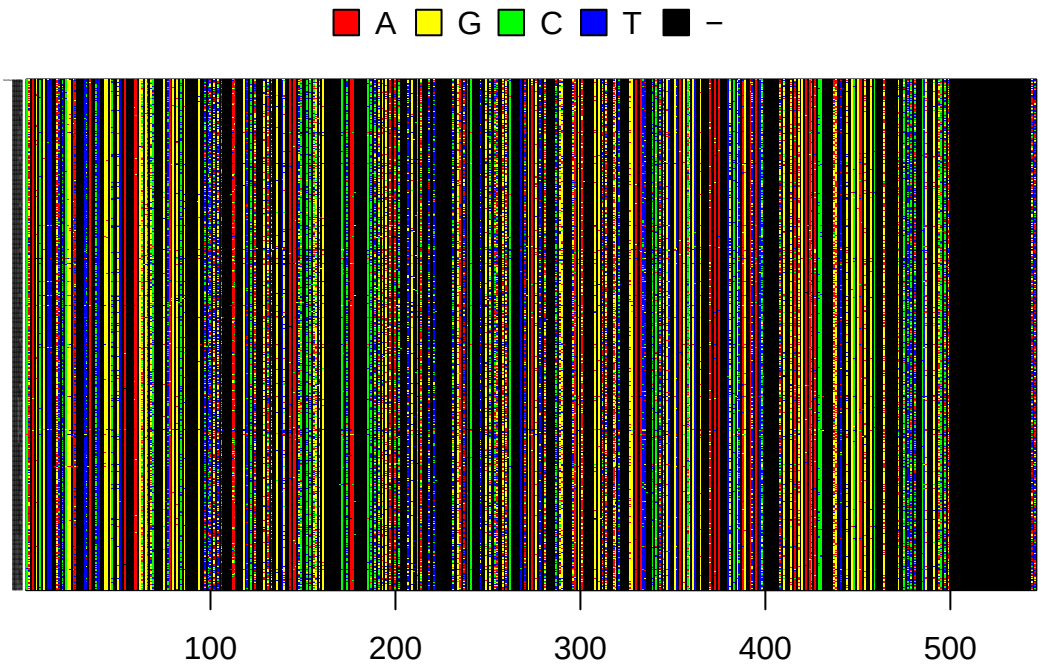
3) LOAD THE DATA

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.

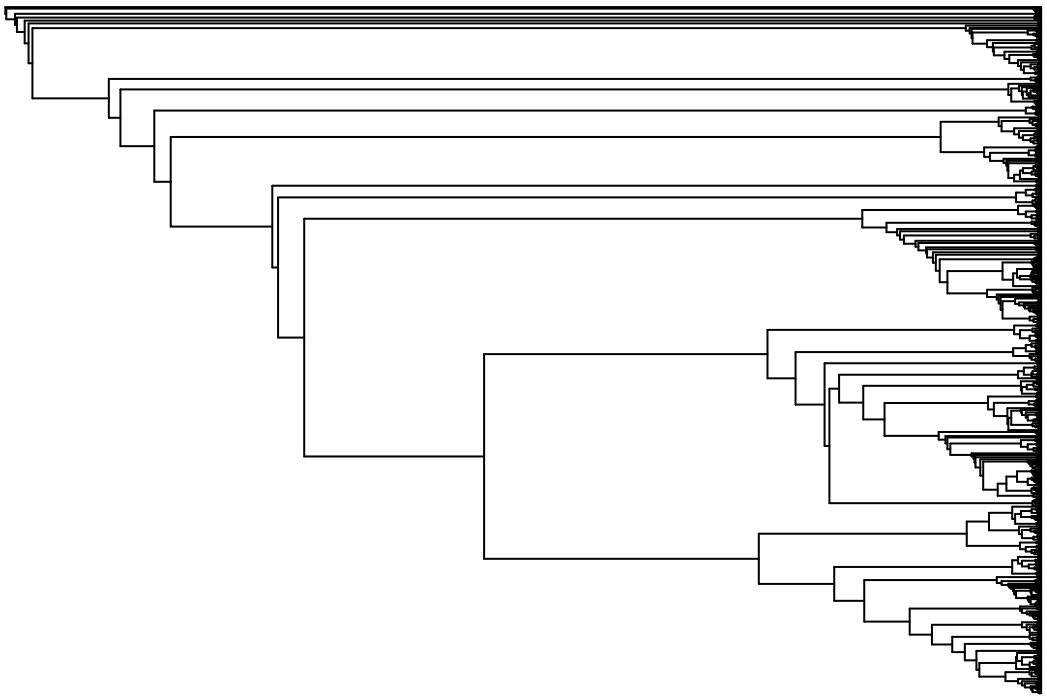
```
# Import OTU Site-by-Species Matrix
read.otu <- function(shared = " ", cutoff = "0.03"){
  matrix <- read.table(shared, header=T, fill=TRUE, comment.char="", sep="\t")
  matrix.cutoff <- subset(matrix, matrix$label == cutoff)
  matrix.out    <- as.matrix(matrix.cutoff[1:dim(matrix.cutoff)[1],4:(3+mean(matrix.cutoff$numOtus))])
  row.names(matrix.out) <- matrix.cutoff$Group
  return(matrix.out)
  }

# Import Taxonomy Information
read.tax <- function(taxonomy = " ", format = "rdp"){
  tax_raw <- read.delim(taxonomy)                  # load genus-level data
  if (format == "rdp"){
    tax <- cbind(OTU = tax_raw[,1],colsplit(tax_raw[,3], split="\\;",
                names=c("Domain","Phylum","Class","Order","Family","Genus")))
    for (i in 2:7){
      tax[,i] <- gsub("\\(.*$", "", tax[,i])
    }
  } else {
    stop("This funciton currently only works for RDP taxonomy")
  }
  return(tax)
}
env <- read.table("./data/20130801_PondDataMod.csv", sep = ",", header = TRUE)
env <- na.omit(env)
comm <- read.otu(shared = "./data/INPonds.final.rdp.shared", cutoff = "1")
comm <- comm[grep("*-DNA", rownames(comm)), ]
rownames(comm) <- gsub("\\-DNA", "", rownames(comm))
rownames(comm) <- gsub("\\_", "", rownames(comm))
comm <- comm[rownames(comm) %in% env$Sample_ID, ]
comm <- comm[ , colSums(comm) > 0]
tax <- read.tax(taxonomy = "./data/INPonds.final.rdp.1.cons.taxonomy")
```

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

```
ponds.cons <- read.alignment(file = "./data/INPonds.final.rdp.1.rep.fasta", format = "fasta")
ponds.cons$nam <- gsub("\\|.*$", "", gsub("^.*?\t", "", ponds.cons$nam))
outgroup <- read.alignment(file = "./data/methanosarcina.fasta", format = "fasta") # Convert alignment
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) # Make a neigbor-joining tree
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) # Root the tree {ape}
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 = F
```



4) PHYLOGENETIC ALPHA DIVERSITY

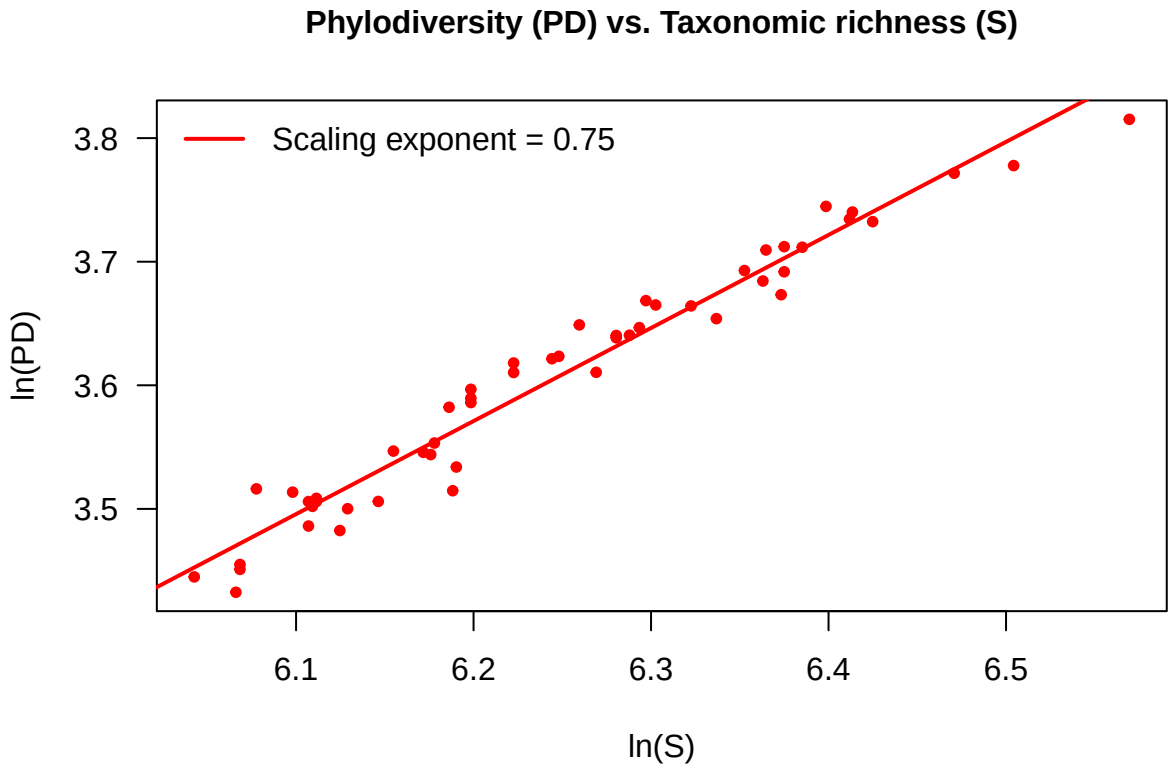
A. Faith's Phylogenetic Diversity (PD)

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

```
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)")
# 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 measures the distance of each branch length from the root to the tip of the phylo tree for each species. Thus, indirectly it accounts for number of species present in a community ***Answer 1b***: It is a power law exponent relationship, where species richness (S) = PD ^ (scaling exponent) or ln(S) ~ ln(PD) ***Answer 1c***: When all species have the same exact branch length, the slope would a vertical line parallel to the y-axis, thus deviating from the norm. ***Answer 1d***: The scaling exponent determines the directionality and magnitude of the relationship between PD and S.

i. Randomizations and Null Models

In the R code chunk below, do the following:
1. estimate the standardized effect size of PD using the `richness` randomization method.

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

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***: Through the randomization function, one is simply creating a null distribution (and

not relying on a pre-defined distribution) i.e. "What are the standardized effect sizes distribution would one expect for this species by random chance alone?" against the actual observed effect sizes in the current phylo trees. ***Answer 2b***: For pond 1 alone, the choice of null model "richness" did effect the output. We see that the sample is phylogenetically less diverse than expected under NULL distribution (ses.pd < 0). The null model "richness" randomizes community abundances while maintaining the species richness. This showed that with sufficient shuffling of abundances, the observed phylo pattern and its corresponding abundances are not strongly related. Thus, the abundance is not governed by the underlying phylo patterns and the species themselves may be less diverse than expected.

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.

```
#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"
```

iii. Nearest Taxon Index (NTI)

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

```
ses.mtnd <- ses.mntd(comm,phydist, null.model ="taxa.labels", abundance.weighted = FALSE, runs =25)

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

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: One is asking "What is the average pairwise distance between branch length of each taxa and what information can it give us about the clustering of taxa relative to the null?" ***Answer 3b***: One is asking "What is the mean nearest phylogenetic distance between all taxa and what information can it give us about the clustering of taxa at the tips (than near the branch lengths) relative to the null?" ***Answer 3c***: NRI analysis indicates that taxa are less related to one another than expected under the null model (i.e. over dispersion) and NTI indicates that a majority of the species exhibit terminal over dispersion.

Answer 3d:

```
#NRI
ses.mpd <- ses.mpd(comm, phydist, null.model = "taxa.labels", abundance.weighted = TRUE, 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"

ses.mtnd <- ses.mntd(comm,phydist, null.model ="taxa.labels", abundance.weighted = TRUE, runs =25)

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

The NTI and NRI analysis indicates that accounting for abundance of species in the community affects the way one views phylogenetic structure in the community. Taxonomic groups are more clustered than expected by chance, in contrast to the results of NRI/NTI sans abundance values. Biologically, one can think of it this way. Assume that there are a wide number of small rodents in the community. These rodent species have evolutionarily utilized different niche spaces, leading to lesser inter-specific competition and similar abundance (overall: higher abundance); than say when it is compared to its top-predator, like the scavenging birds. The taxa is these species again would have similar abundance (overall: lower abundance), but different from rodent or any producer species in the community. Thus, it is phylogenetically and numerically different from other taxa.

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.

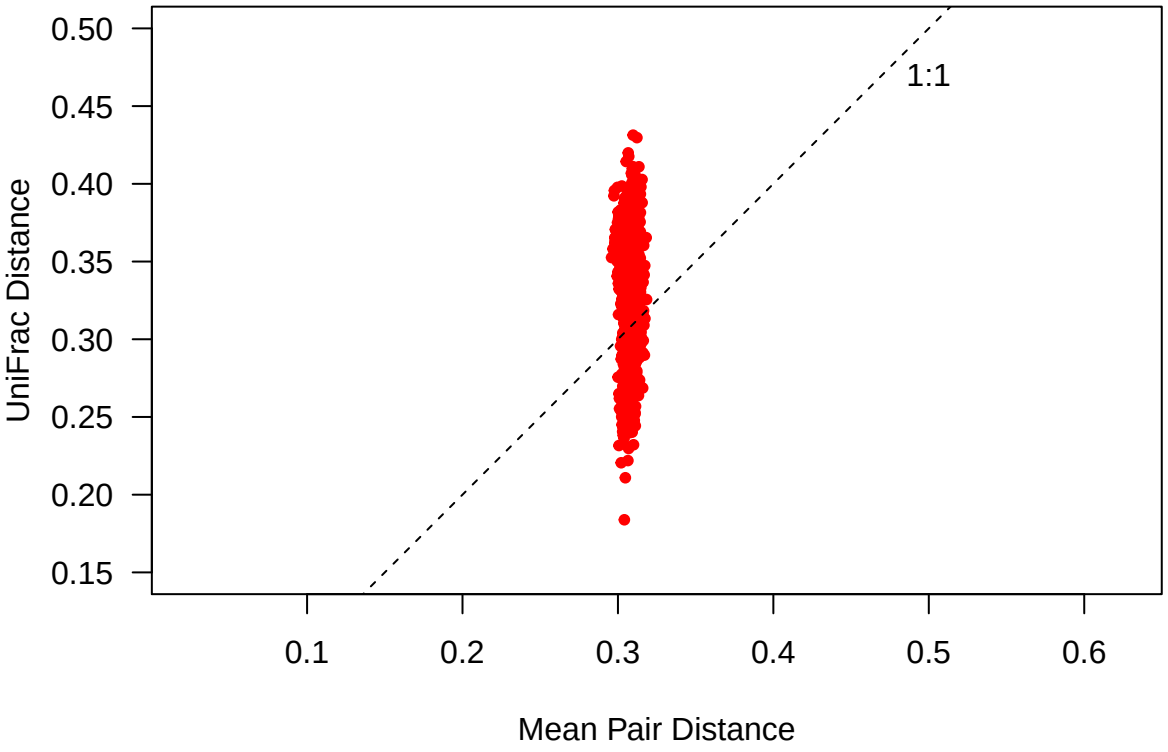
```
# Mean Pairwise Distance
dist.mp <- comdist(comm, phydist)
```

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

```
# UniFrac Distance
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)
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: Given any pair of taxa, mean pair distance estimates the average phylogenetic distance between them; while UniFrac estimates the number of unshared branch relative to the length of the entire tree. While MPD is only concerned about the pair of taxa in question, UniFrac is dependent on all the taxa between the two focal pairs. ***Answer 4b***: The above graph indicates that for a very limited range of MPD value (centered around ~0.3) there is a large range of associated UniFrac values. ***Answer 4c***: Since UniFrac displays all the unshared branches between the two focal branches, one can expect huge variation in points, largest when the two focal branches are the most distant relatives on the branch and smallest (or zero) when the two focal branches are the closest relatives to one another. Meanwhile, the mean distances between any two branches would not vary too much if you are distantly related versus when you are closely related.

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:

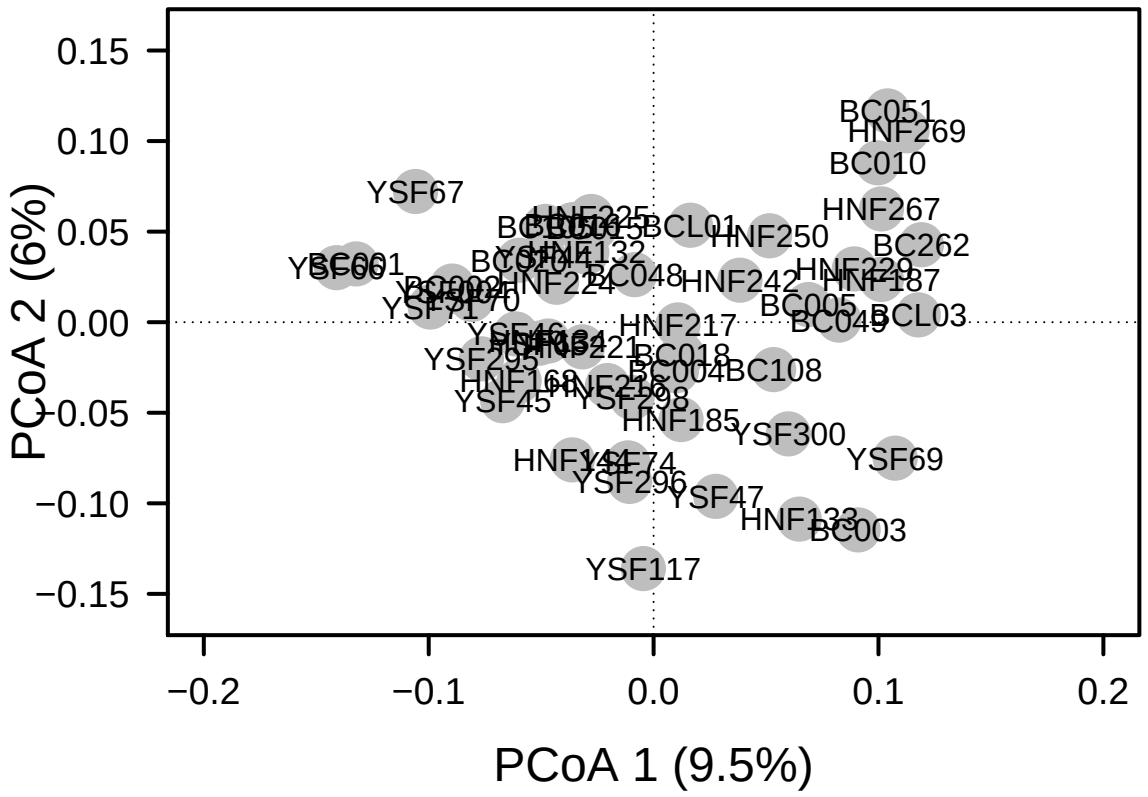
1. perform a PCoA based on the UniFrac distances, and
2. calculate the explained variation for the first three PCoA axes.

```
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.

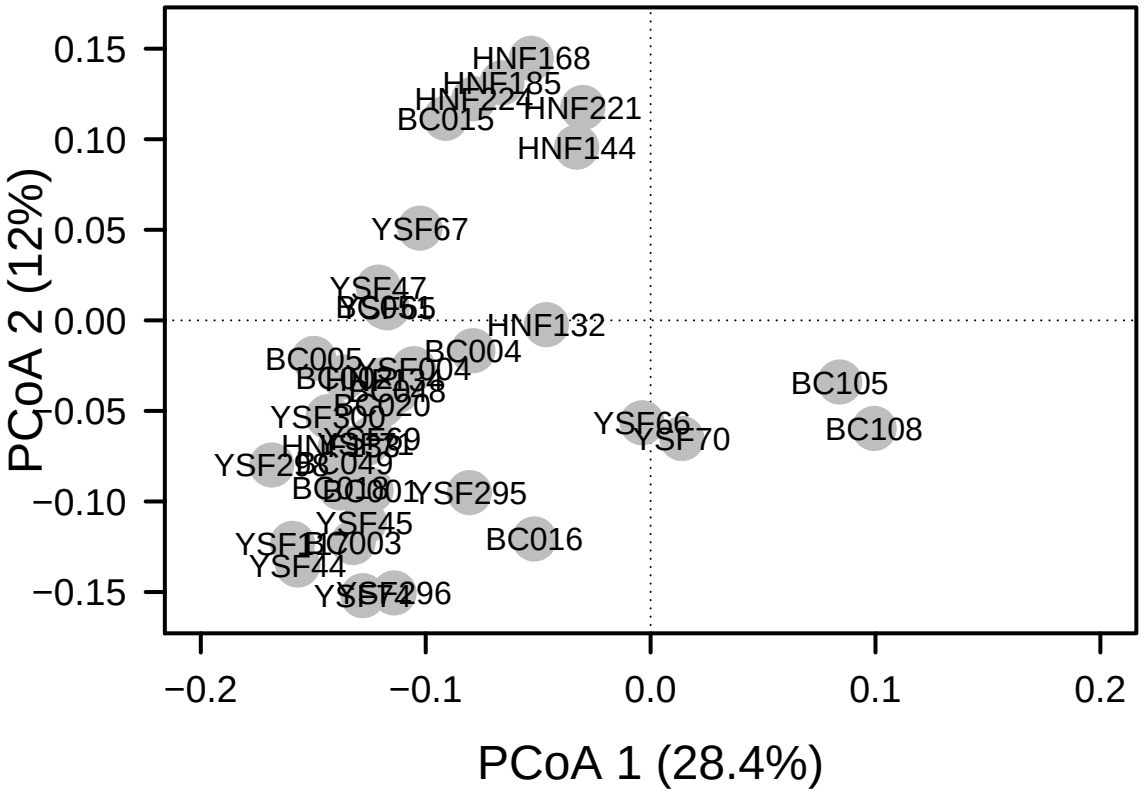
In the R code chunk below, do the following:
1. plot the PCoA results using either the R base package or the `ggplot` package,
2. include the appropriate axes,
3. add and label the points, and
4. customize the plot.

```
# 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, 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)

# 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 & 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.

```
dist.db <- vegdist(comm, method = "bray")
pond.pcoa <- cmdscale(dist.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, 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)

# 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 & 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))
```



Answer 5: Taxonomic ordination (using Bray-Curtis) indicates that ~28% of the variation in diversity between sites can be explained by this abundance-based metric. Phylogenetically based ordination indicates that ~10% of the variation can be explained by considering UniFrac index. While the former appears to have a more distinct grouping, the latter is has sites relatively spread out. This tells us that incorporating phylogenetic data can change the way one looks at relatedness among sites, making it more or less similar than that is indicated by abundance-based metric alone. However, a point to be considered is that the UniFrac that is currently used in R is unweighted i.e. it is a incidence-based measure rather than an abundance based measure and hence it may be incongruous to compare the two metric. Thus, either one can use an incidence based distance measure (like Sorensen's) or the current UniFrac index; or use abundance metric (like Bray-Curtis) and compare it to weighted UniFrac.

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   0.03 *
## 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,
perumutations = 999)
```

```
##
## Call:
## adonis(formula = vegdist(decostand(comm, method = "log"), method = "bray") ~      watershed, perumuta
##
## 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()`).

```
# Define environmental variables
envs <- env[, 5:19]
# Remove redudnant 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: 0.049
##
## Upper quantiles of permutations (null model):
##   90%   95% 97.5%   99%
## 0.126 0.158 0.197 0.254
## 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

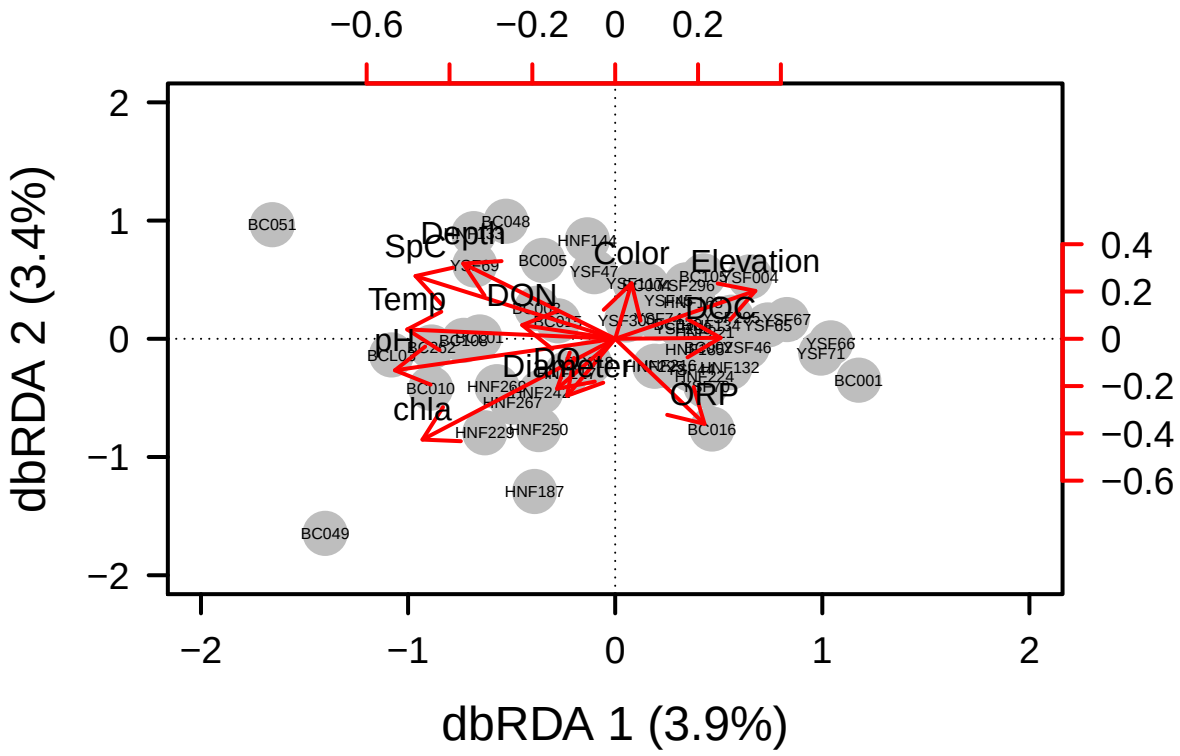
```r
# Conduct dbRDA (`vegan`)
ponds.dbrda <- vegan::dbrda(dist.uf ~ ., data = as.data.frame(scale(envs)))
# Permutation tests: axes and environmental variables
anova(ponds.dbrda, by = "axis")
```

```
## 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  0.005 **
## dbRDA3    1  0.07555 1.4409  0.028 *
## dbRDA4    1  0.06677 1.2735  0.105
## dbRDA5    1  0.05666 1.0807  0.302
## dbRDA6    1  0.05293 1.0095  0.463
## dbRDA7    1  0.04750 0.9059  0.650
## dbRDA8    1  0.03941 0.7517  0.899
## dbRDA9    1  0.03775 0.7201  0.942
## dbRDA10   1  0.03280 0.6256  0.990
## dbRDA11   1  0.02876 0.5485  0.999
## dbRDA12   1  0.02501 0.4770  0.999
## Residual 39  2.04482
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

```
##
## ***VECTORS
##
##             dbRDA1   dbRDA2     r2 Pr(>r)
## Elevation  0.77671  0.62986 0.0959  0.077 .
## Diameter  -0.27972 -0.96008 0.0541  0.275
## Depth     -0.63137  0.77548 0.1756  0.011 *
## ORP        0.41879 -0.90808 0.1437  0.027 *
## Temp      -0.98250  0.18629 0.1523  0.017 *
## SpC       -0.77101  0.63682 0.2087  0.005 **
## DO        -0.39318 -0.91946 0.0464  0.307
## pH        -0.96210 -0.27270 0.1756  0.013 *
## Color      0.06353  0.99798 0.0464  0.304
## chla      -0.60393 -0.79704 0.2626  0.007 **
## DOC        0.99847 -0.05526 0.0382  0.369
## DON       -0.91633  0.40042 0.0339  0.423
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Permutation: free
## Number of permutations: 999
```

```r
# 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 & 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 RD Analysis shows that ~ 4% of the variation in bacterial communities is explained by environmental variables. Of these, elevation, DOC, and ORP best explain these phylogenetic patterns of beta diversity for bacterial communities.

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 = TRUE)
# Taxonomic similarity among ponds (Bray-Curits distance)
bray.curtis.dist <- 1 - vegdist(comm)
# Phylogenetic similarity among ponds (UniFrac)
unifrac.dist <- 1 - dist.uf
```

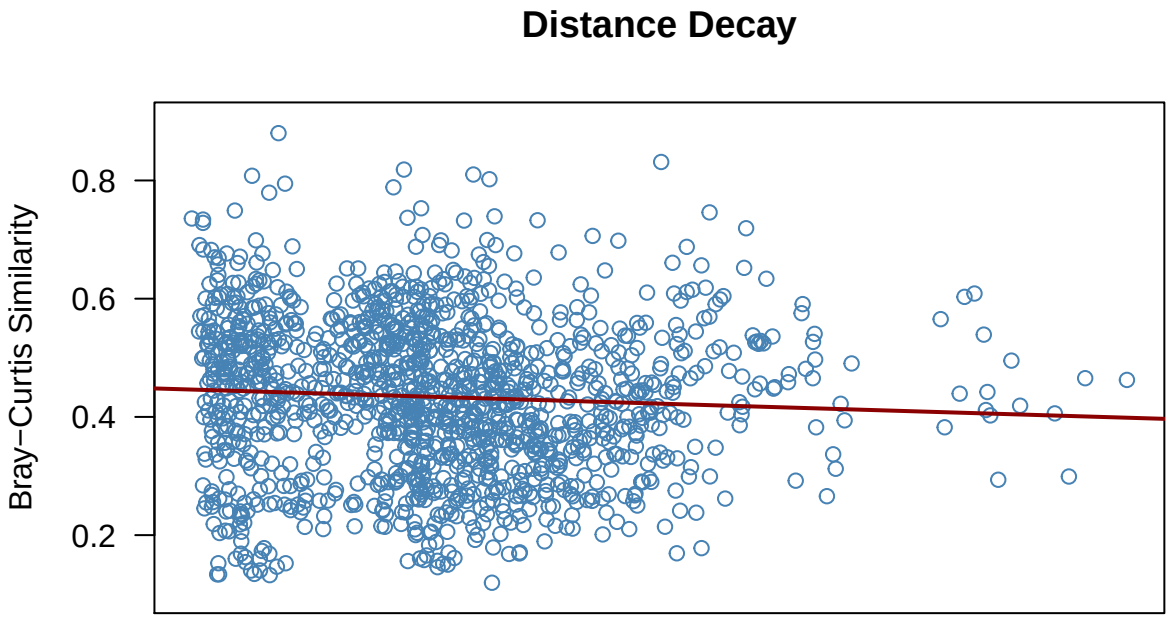
```r
# Transform all distances into list format:
unifrac.dist.ls <- liste(unifrac.dist, entry = "unifrac")
bray.curtis.dist.ls <- liste(bray.curtis.dist, entry = "bray.curtis")
coord.dist.ls <- liste(coord.dist, entry = "geo.dist")
env.dist.ls <- liste(env.dist, entry = "env.dist")
# Create a data frame 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")
# Set initial plot parameters
par(mfrow=c(2, 1), mar = c(1, 5, 2, 1) + 0.1, oma = c(2, 0, 0, 0))
```

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.

```r
# 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
# 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
```

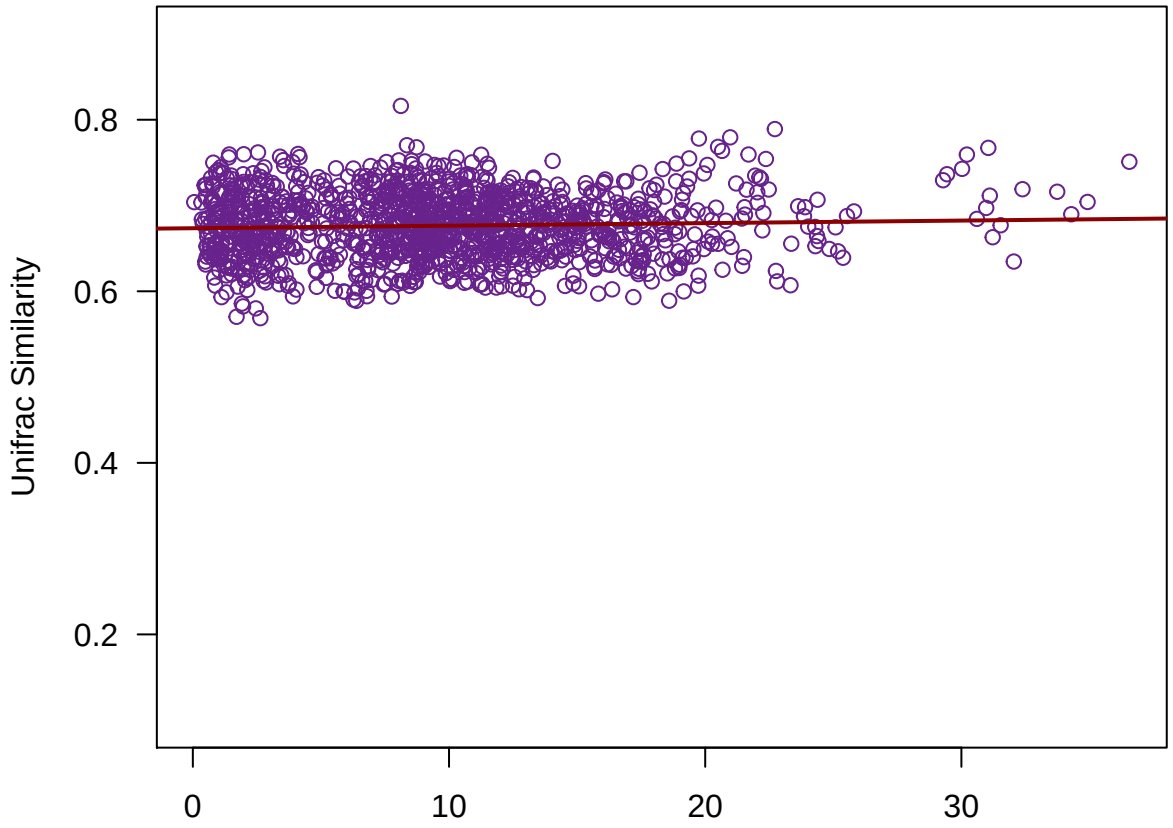
```r
abline(DD.reg.bc , col = "red4", lwd = 2)
```




```
# 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 =
# 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: 0.004
##
## Empirical upper confidence limits of r:
##      90%      95%    97.5%      99%
## 0.000738 0.000999 0.001156 0.001308
```

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

> ***Answer 7***: There is a marginal (0.001603) but significant difference (p = 0.001) in the slope values between the distance-decay relationships. The BC graph indicates a negative relationship

with slope of -0.0013 while the UniFrac indicates a slightly positive relationship with a slope of 0.00029. It is worthwhile to note that although the slopes itself is not very different from another, the variation around the slope for BC is greater than for UF distances. Hypothesis: Relatedness among individuals may not vary with increasing geographic distance, however there may be more variation in abundance than expected by chance between closer sites than far away owing to consequences of inter-specific competition.

B. Phylogenetic diversity-area relationship (PDAR)

i. Constructing the PDAR

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

```
PDAR <- function(comm, tree){
areas <- c()
diversity <- c()
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){
pond.sample <- sample(51, replace = FALSE, 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 = FALSE)
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))
}
```

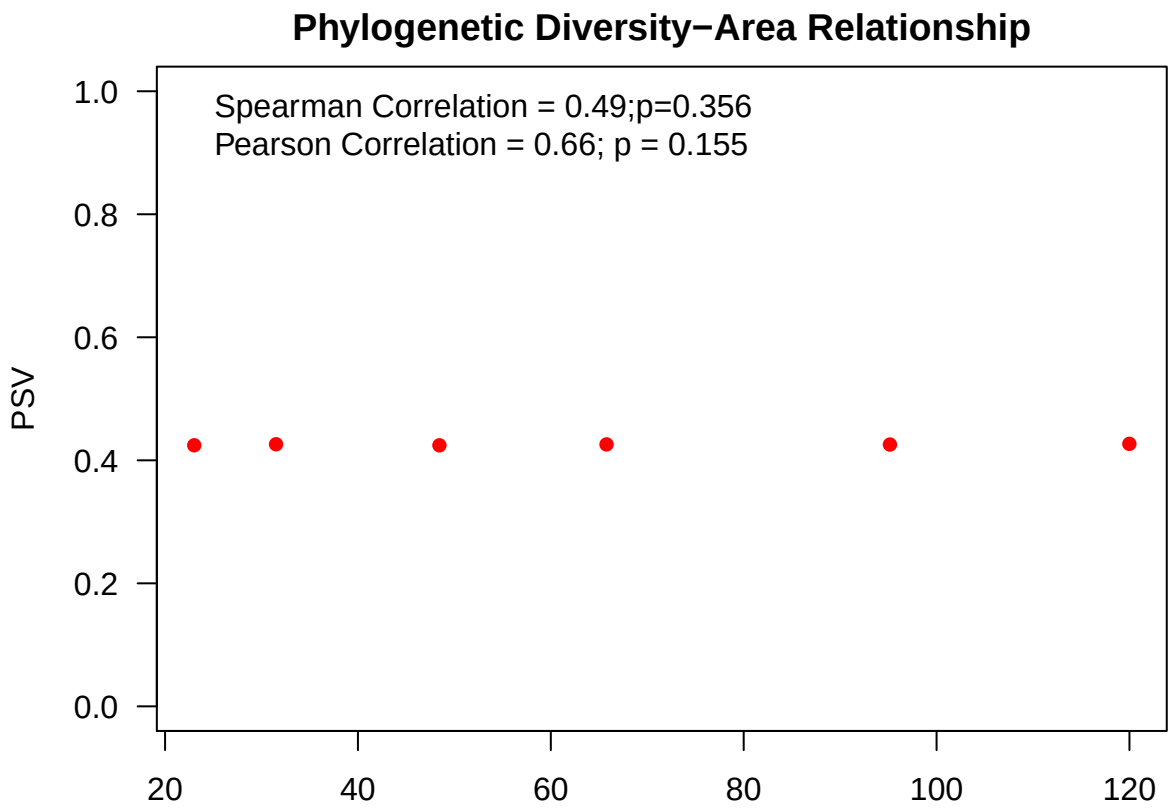
ii. Evaluating the PDAR

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   0.4244354 530.1586829
## [1]   4.0000000   0.4261009 992.8516635
## [1]    8.0000000    0.4243692 2347.3548339
## [1]   16.0000000    0.4258048 4326.7404180
## [1]   32.0000000    0.4256211 9056.0657244
## [1] 5.100000e+01 4.267902e-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-
legend("topleft", legend= c(paste("Spearman Correlation = ", rho, ";p=", rho.pval, sep=""),
paste("Pearson Correlation = ", P, "; p = ", P.pval, sep = "")),
bty = "n", col = "red")
```



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***:SAR indcates that with increasing area, one encounters greater species richness (slope ~ 0.14). In contrast, the PDAR analyisis indicates that there is no relationship between phylo diversity and are sampled. This is perhaps because incorporating phylogenetic relationship between species adjusts for species that are very cloesly related and accounting only for distant relatives (if any). The graph clearly indicates that even with increased sampled area, the probablity of findind species that are phylogeneticcally dispararte is still a rare event.

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.

> ***Answer*** : My current research interests lies in understanding the eco-evolutionary processes that govern host-parasite interactions and their consequences for disease spread.

-More specifically, I aim to understand to understand the impact that a particular parasitic species can have on a host due to its prior arrival in a host (i.e. priority effects), the evolutionary consequences for the host-parasite(s) system, and how this would impact the dynamics of disease spread on local and regional scales. Priority effects can manifest as either inhibitory or facilitative priority effects.

- During inhibitory priority effects, it has been widely regarded that the early colonizer can impede the

arrival of a later species by reducing the availability of space, or resource. In facilitative effects, early colonizers can alter the current environment to positively influence late colonizers.

- Previous data from our lab indicates that closely related species are more likely to cooperate with one another and engage in less spiteful interactions than those that are distantly related. If tis holds true, then it would be interesting to see how this phylogenetic relatedness would influence the outcome of priority effects. For instance, would an early colonizer that engages in inhibitory behaviour switch strategies if the late colonizer is a close relative versus a distant relative? What type of consequences would this result in with respect to abundance of each species within the community? Would this lead to greater cooperation in killing the host?
- For this I would require genomic data of all nematode species in a particular area to be able to calculate phylo distances and construct phylo trees. In addition, traits like growth, foraging and body-size could contribute to understanding more about the phylotraits in these species.