Phylogenetic Diversity - Communities

Erik Parker; Z620: Quantitative Biodiversity, Indiana University

26 February, 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] "/var/host/media/removable/USB Drive/GitHub/QB2017_Parker/Week7-PhyloCom"
```

```
setwd("./")

package.list <- c("picante", "ape", "seqinr", "vegan", "fossil", "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)
  }
}
```

```
## 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
```

```
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.

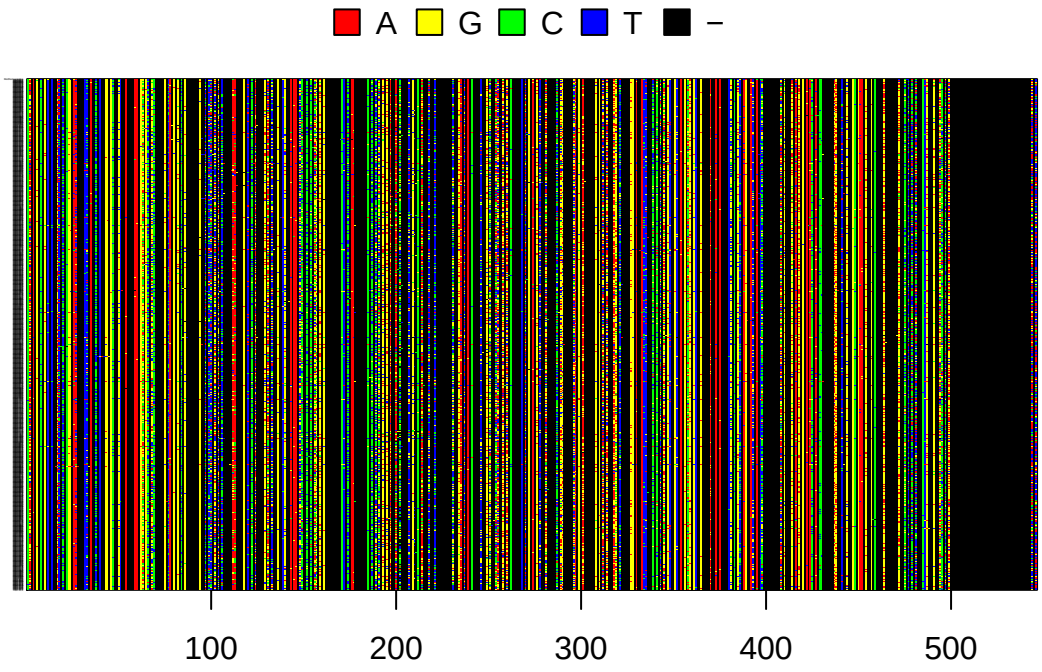
3) LOAD THE DATA

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.

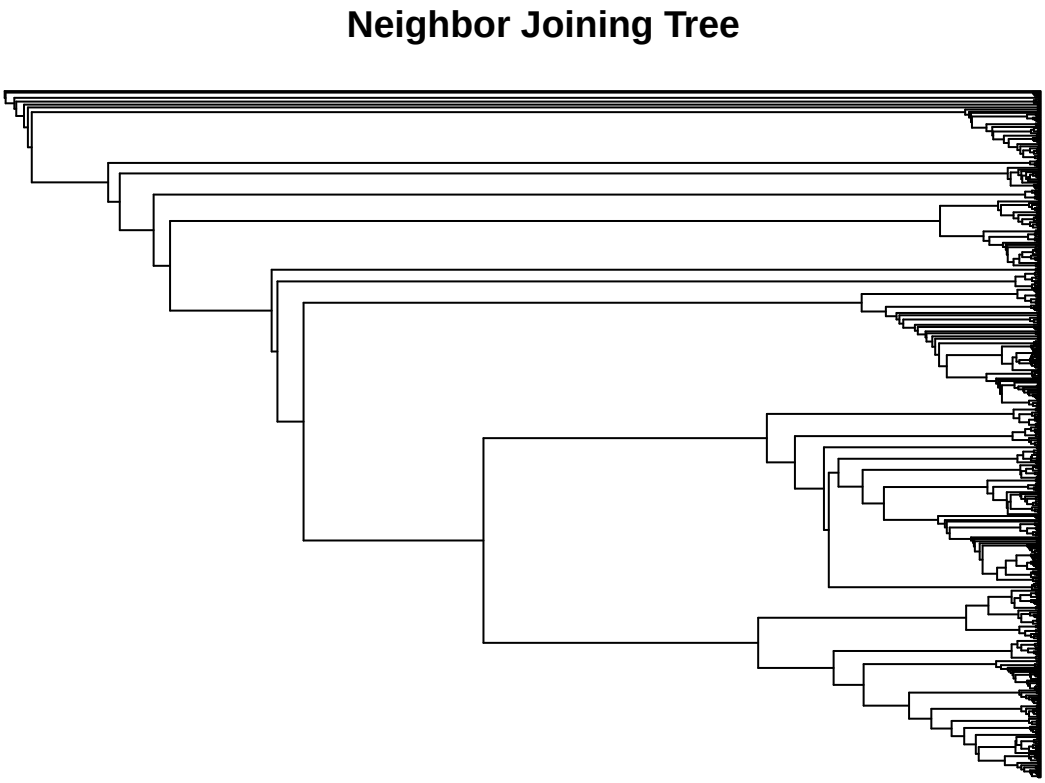
```
env <- read.table("data/20130801_PondDataMod.csv", sep = ",", header = TRUE)
env <- na.omit(env)

# read.otu used to read in OTU data
comm <- read.otu(shared = "./data/INPonds.final.rdp.shared", cutoff = "1")
# grep() used here to select only DNA data, while excluding cDNA data.
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]

# read.tax() used to load taxonomic data
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")

DNAbin <- rbind(as.DNAbin(outgroup), as.DNAbin(ponds.cons))

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




```
seq.dist.jc <- dist.dna(DNAbin, model = "JC", pairwise.deletion = FALSE)

phy.all <- bionj(seq.dist.jc)

# Removing tips with zero abundance
phy <- drop.tip(phy.all, phy.all$tip.label[!phy.all$tip.label %in% c(colnames(comm), "Methanosarcina")]

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

phy <- root(phy, outgroup, resolve.root = TRUE)

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



4) PHYLOGENETIC ALPHA DIVERSITY

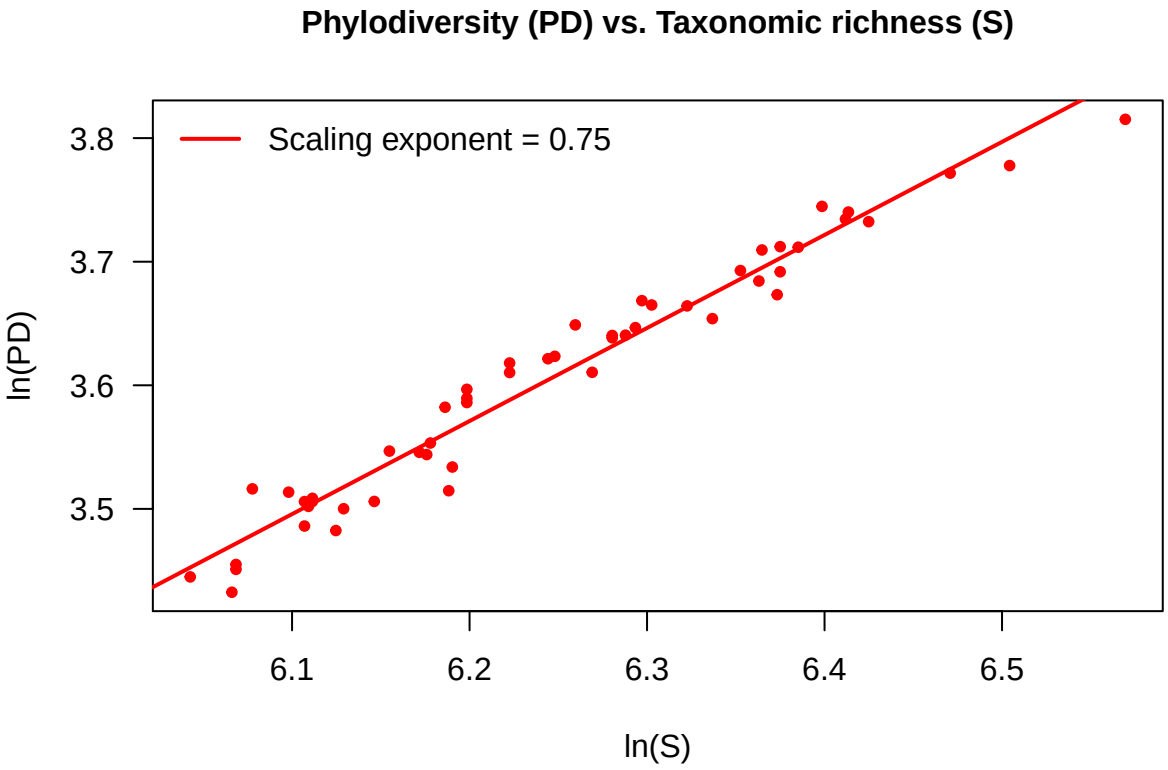
A. Faith's Phylogenetic Diversity (PD)

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

```
# Calculating Fait's phylogenetic diversity (PD) and species richness (S)
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.

```
# Transform PD and S with ln() to plot relationships as power-law exponents.
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)")

# Adding trend line to test for fit of data as 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 calculated based on the branch lengths of species (OTUs) present in the sample. So, it is basically just a measure of richness scaled by taxanomic relatedness.
>
> ***Answer 1b***:
> Phylodiversity seems like it is just a measure of taxonomic richness transformed with added information - the phylogenetic relatedness of the species found in the sample.
>
> ***Answer 1c***:
> Phylodiversity and richness should deviate from one another when samples consist of closely related species, so there are many shared branch lengths on the phylogenetic tree of the sample. When there are no closely related species, no shared branch lengths along the tree, phylodiverstiy and standard richness should be the same?
>
> ***Answer 1d***: The PD-S scaling exponent calculated above tells us that there is a non-significant amount of shared branches in tree calculated from the sample. When compared on a ln-ln scale, phylodiversity is 25% lower standard richness meaning that there are closely related species in our sample which add otherwise hidden structure to the data.

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)

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

```
ses.pd.freq <- ses.pd(comm[1:2,], phy, null.model = "frequency", 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     43.97865  0.6977168           9 -0.3719654
## BC002   587 40.94334     39.75017  1.0765422          22  1.1083372
##        pd.obs.p runs
## BC001 0.3461538   25
## BC002 0.8461538   25
```

```
ses.pd.ts
```

```
##       ntaxa   pd.obs pd.rand.mean pd.rand.sd pd.obs.rank   pd.obs.z
## BC001   668 43.71912     43.87869  0.3371381           7 -0.4733090
## BC002   587 40.94334     40.71827  0.3249985          22  0.6925203
##        pd.obs.p runs
## BC001 0.2692308   25
## BC002 0.8461538   25
```

```
ses.pd.freq
```

```
##       ntaxa   pd.obs pd.rand.mean pd.rand.sd pd.obs.rank  pd.obs.z
## BC001   668 43.71912     42.16642  0.6009204          26  2.583871
## BC002   587 40.94334     42.42327  0.5951193           1 -2.486787
##         pd.obs.p runs
## BC001 1.00000000   25
## BC002 0.03846154   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:
The null and alternative hypotheses being tested via randomizaion with ses.pd() are that the sample is as diverse as would be expected under the specified null model, and that it is more diverse than expected given the model respectively.

Answer 2b:
All three null models returned slightly different results for randomized means and standard deviations of phylodiversity for the two ponds. I guess that my model choices affected the output in this case because the three null models I chose (richness, frequency, and trialswap) all have different assumptions and maintain different characteristics of the dataset. That said, there is overlap between trialswap and the other two models (trialswap maintains both richness and frequency), so perhaps that could explain why my values were never really that different.

B. Phylogenetic Dispersion Within a Sample

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

i. Phylogenetic Resemblance Matrix

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

```
phydist <- cophenetic.phylo(phy)
```

ii. Net Relatedness Index (NRI)

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

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

# For NRI: negative values means taxa are less related to one another than expected under null. Positiv
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.mntd <- ses.mntd(comm, phydist, null.model = "taxa.labels",
abundance.weighted = TRUE, runs = 25)

# For NTI: negative values means taxa are less related to one another than expected under null. Positiv
NTI <- as.matrix(-1 * ((ses.mntd[,2] - ses.mntd[,3]) / ses.mntd[,4]))
rownames(NTI) <- row.names(ses.mntd)
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:
The NRI is calculated by finding the pairwise branch lengths between every unique taxa group in a sample (MPD), taking the average of that measure, and then subtracting from that the mean MPD value generated using null model randomization. The difference is then divided by the standard deviation of the randomply generated MPD values.

Answer 3b:
The NTI is calculated in much the same way as the NRI, with the only difference being the reliance on finding the mean nearest phylogenetic neighbor distance (MNND) instead of the MPD. The MNND is found by summing the minumum values for each taxon of the resemblance matrix (the distance between a taxa and it's closest related neighbor), which ends up weighting changes in terminal rather than root clustering as most important.

Answer 3c:
The overwhelming majority (all?) of the NRI values calculated were negative, indicating that taxa in the dataset are some level of less related to one another than would be expected under the null model of randomization. This is contrasted with the NTI calculation where we saw that a large number of taxa returned negative values, but a not insignificant number showed positive values, indicating that there was more clustering than would be expected under the null. This difference between the two indices seems to mean that there is a high level of terminal clustering going on that the NTI was able to pick up on which was missed by the NRI due to large branch length differences in the more root level branches.

Answer 3d:
Running the NRI and NTI calculations again while taking into account the abundance of the taxa in the sample lead to much higher (positive) values for both indices, meaning that there is more clustering found when abundance of taxa is taken into account.

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.

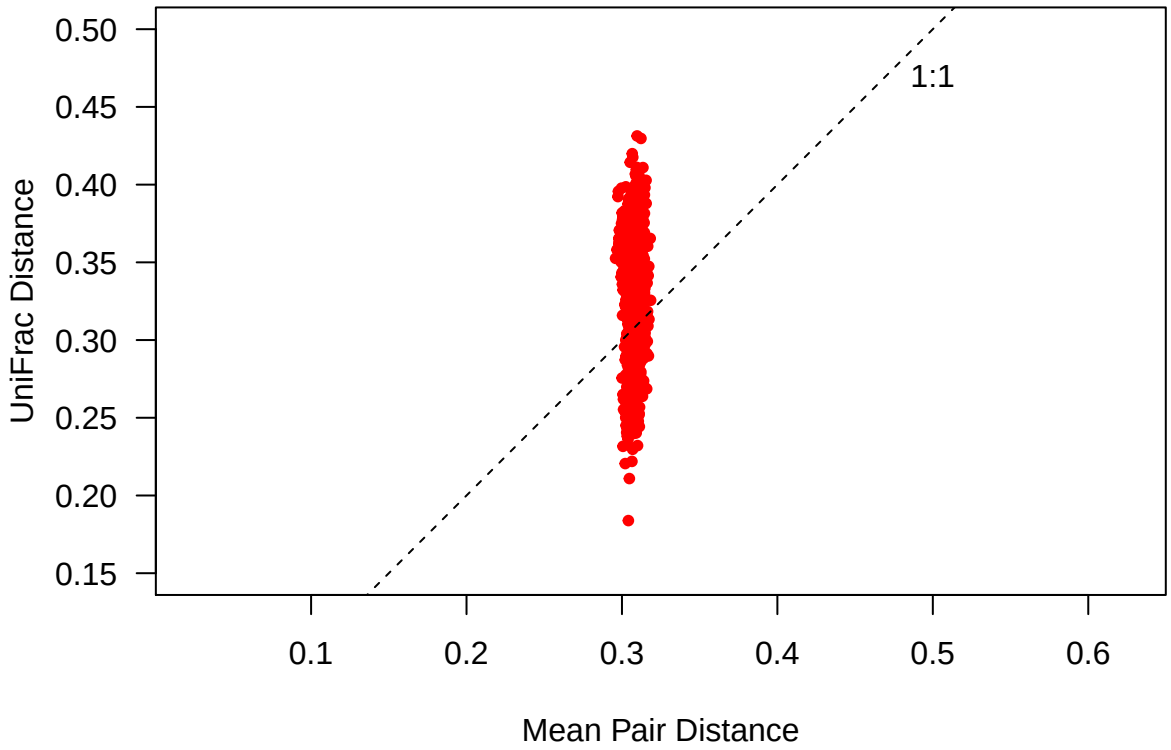
```
dist.mp <- comdist(comm, phydist)
```

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

```
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: Mean pairwise distance is calculated as the average length of all possible paths which connect all taxa pairs in a given tree for a given sample. UniFrac is calculated as the ratio of all unshared branch lengths to total (shared and unshared) branch lengths in a given tree (If finds the Unique Fraction of branches in a tree).

I can't find much detailed literature on what MPD is really doing, but it seems like UniFrac is doing more to seperate branch lengths into shared and unshared lengths, rather than just counting up the total length seperating species pairs as MPD is doing.

Answer 4b: Interestingly, mean pair distance seems overall to be a much less sensitive distance measure than UniFrac - all MPD values were clustered around 0.3, with movement of about 0.03 to each side, while the UniFrac values found ranged from 0.2 to 0.45.

Answer 4c: Again, I am not too confident about how MPD is actually calculated but that's never really stopped evolutionary biologists from reaching sweeping conclusions before, has it? From my understanding it seems that UniFrac is more directly taking into consideration the composition of the branch lengths used for the calculation. While MPD only looks at the length of these branches, UniFrac also looks at their identity - whether they are shared or not - when calculating a distance metric, making it a more sensitive distance metric.

B. Visualizing Phylogenetic Beta-Diversity

Now that we have our phylogenetically based community resemblance matrix, we can visualize phylogenetic diversity among samples using the same techniques that we used in the β-diversity module from earlier in the course.

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

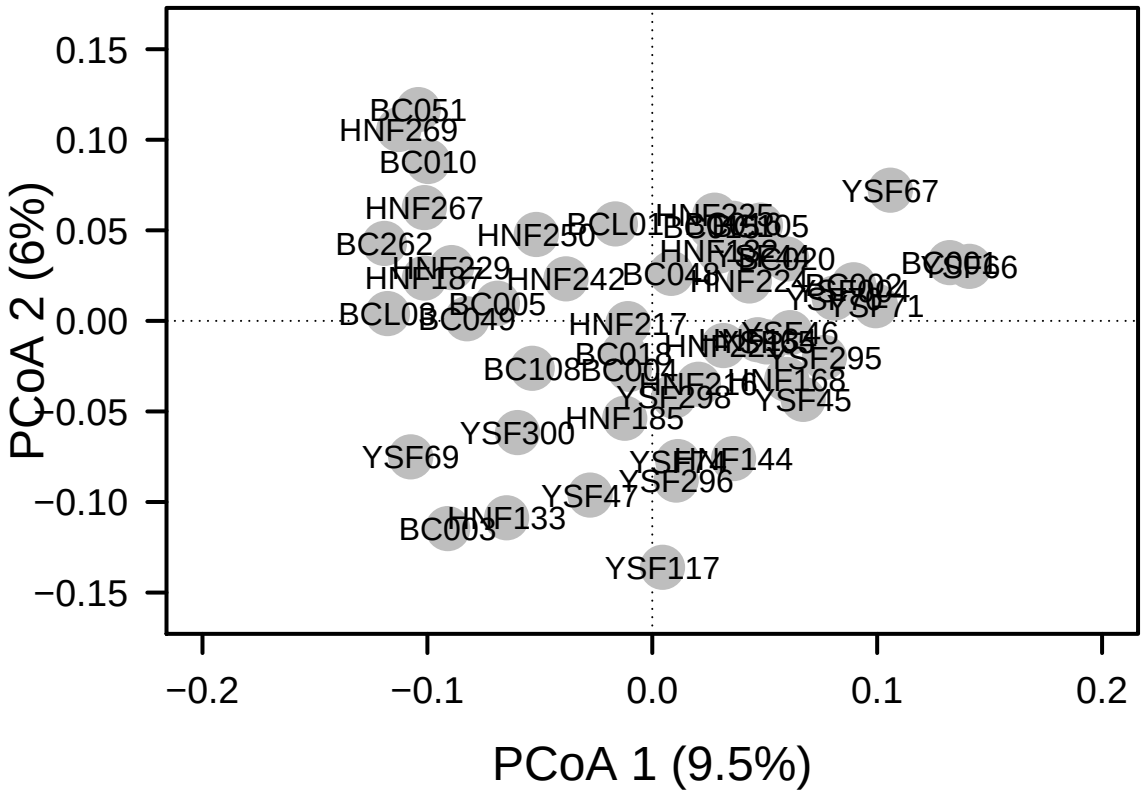
```
pond.pcoa <- cmdscale(dist.uf, eig = T, k = 3)
explainvar1 <- round(pond.pcoa$eig[1] / sum(pond.pcoa$eig), 3) * 100
explainvar2 <- round(pond.pcoa$eig[2] / sum(pond.pcoa$eig), 3) * 100
explainvar3 <- round(pond.pcoa$eig[3] / sum(pond.pcoa$eig), 3) * 100
sum.eig <- sum(explainvar1, explainvar2, explainvar3)
```

Now that we have calculated our PCoA, we can plot the results.

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

```
par(mar = c(5, 5, 1, 2) + 0.1)
plot(pond.pcoa$points[ ,1], pond.pcoa$points[ ,2],
xlim = c(-0.2, 0.2), ylim = c(-.16, 0.16),
xlab = paste("PCoA 1 (", explainvar1, "%)", sep = ""),
ylab = paste("PCoA 2 (", explainvar2, "%)", sep = ""),
pch = 16, cex = 2.0, type = "n", cex.lab = 1.5, cex.axis = 1.2, axes = FALSE)
axis(side = 1, labels = T, lwd.ticks = 2, cex.axis = 1.2, las = 1)
```

```
axis(side = 2, labels = T, lwd.ticks = 2, cex.axis = 1.2, las = 1)
abline(h = 0, v = 0, lty = 3)
box(lwd = 2)

points(pond.pcoa$points[ ,1], pond.pcoa$points[ ,2],
pch = 19, cex = 3, bg = "gray", col = "gray")
text(pond.pcoa$points[ ,1], pond.pcoa$points[ ,2],
labels = row.names(pond.pcoa$points))
```



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

```
pond.bray <- vegdist(comm, method = "bray")
pond.pcoa.b <- cmdscale(pond.bray, eig = T, k = 3)
explainvar1.b <- round(pond.pcoa.b$eig[1] / sum(pond.pcoa.b$eig), 3) * 100
explainvar2.b <- round(pond.pcoa.b$eig[2] / sum(pond.pcoa.b$eig), 3) * 100
explainvar3.b <- round(pond.pcoa.b$eig[3] / sum(pond.pcoa.b$eig), 3) * 100
sum.eig.b <- sum(explainvar1.b, explainvar2.b, explainvar3.b)
```

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 distance metric (UniFrac) explains much less variance on its first three axes than does the taxonomic ordination (Bray-Curtis). This means that using a phylogenetically based distance metric uncovers more variation in the data (when you have corresponding phylogentic information) when compared to a standard distance metric.

C. Hypothesis Testing

i. Categorical Approach

In the R code chunk below, do the following:
1. test the hypothesis that watershed has an effect on the phylogenetic diversity of bacterial communities.

```
watershed <- env$Location

# PERMANOVA
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  0.023 *
## 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
```

```
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  0.002 **
## 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()`).

```
envs <- env[, 5:19]

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

env.dist <- vegdist(scale(envs), method = "euclid")
```

In the R code chunk below, do the following:
1. conduct a Mantel test to evaluate whether or not UniFrac distance is correlated with environmental variation.

```
mantel(dist.uf, env.dist)
```

```
##
## Mantel statistic based on Pearson's product-moment correlation
##
## Call:
## mantel(xdis = dist.uf, ydis = env.dist)
##
## Mantel statistic r: 0.1604
##       Significance: 0.059
##
## Upper quantiles of permutations (null model):
##   90%   95% 97.5%   99%
## 0.130 0.167 0.194 0.258
## Permutation: free
## Number of permutations: 999
```

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

In the R code chunk below, do the following:
1. conduct a dbRDA to test the hypothesis that environmental variation effects the phylogenetic diversity of bacterial communities,
2. use a permutation test to determine significance, and 3. plot the dbRDA results

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

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

```
## Permutation test for dbrda under reduced model
## 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  0.001 ***
## dbRDA2    1  0.09258 1.7658  0.003 **
## dbRDA3    1  0.07555 1.4409  0.043 *
## dbRDA4    1  0.06677 1.2735  0.092 .
## dbRDA5    1  0.05666 1.0807  0.304
## dbRDA6    1  0.05293 1.0095  0.482
## dbRDA7    1  0.04750 0.9059  0.648
## dbRDA8    1  0.03941 0.7517  0.907
## dbRDA9    1  0.03775 0.7201  0.949
## dbRDA10   1  0.03280 0.6256  0.990
```

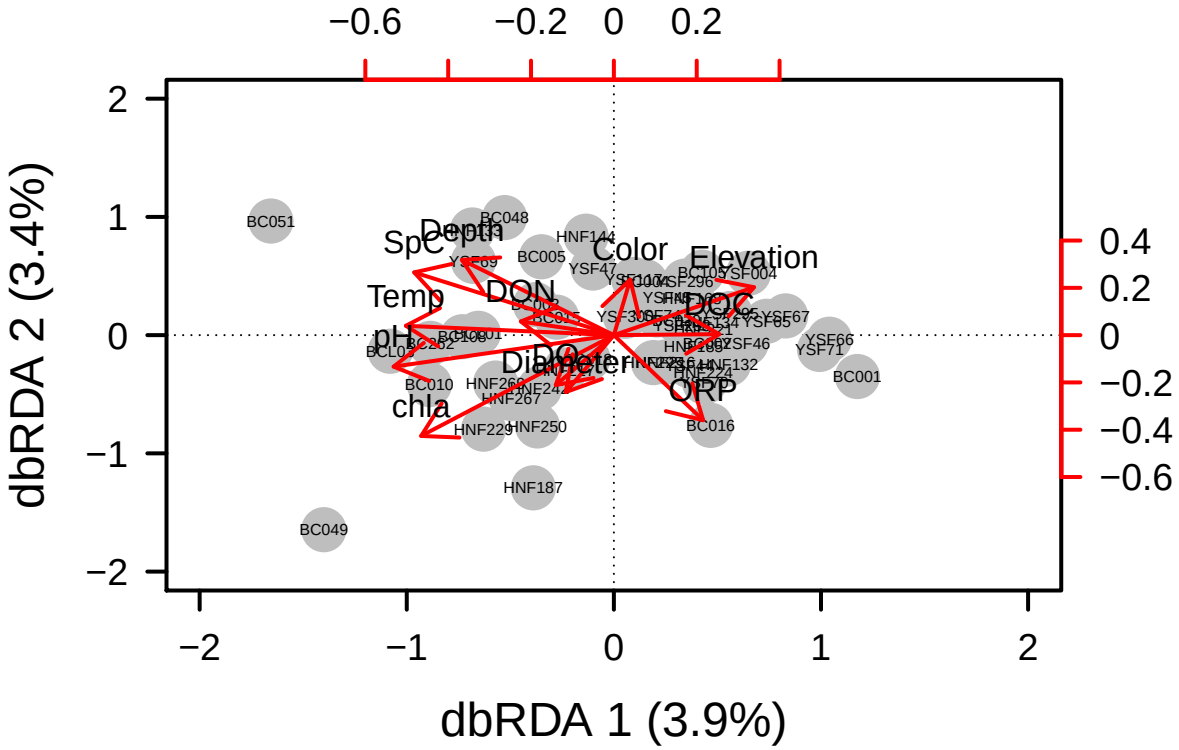
```
## dbRDA11   1  0.02876 0.5485  0.995
## dbRDA12   1  0.02501 0.4770  0.997
## Residual 39  2.04482
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

```
##
## ***VECTORS
##
##             dbRDA1   dbRDA2     r2 Pr(>r)
## Elevation  0.77670  0.62986 0.0959  0.088 .
## Diameter  -0.27972 -0.96008 0.0541  0.248
## Depth     -0.63137  0.77548 0.1756  0.016 *
## ORP        0.41879 -0.90808 0.1437  0.029 *
## Temp      -0.98250  0.18628 0.1523  0.013 *
## SpC       -0.77101  0.63682 0.2087  0.004 **
## DO        -0.39318 -0.91946 0.0464  0.341
## pH        -0.96210 -0.27270 0.1756  0.010 **
## Color      0.06353  0.99798 0.0464  0.319
## chla      -0.60392 -0.79704 0.2626  0.008 **
## DOC        0.99847 -0.05526 0.0382  0.387
## DON       -0.91633  0.40042 0.0339  0.430
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Permutation: free
## Number of permutations: 999
```

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

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

axis(side = 1, labels = T, lwd.ticks = 2, cex.axis = 1.2, las = 1)
axis(side = 2, labels = T, lwd.ticks = 2, cex.axis = 1.2, las = 1)
abline(h = 0, v = 0, lty = 3)
box(lwd = 2)

points(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)

vectors <- scores(ponds.dbrda, display = "bp")

arrows(0, 0, vectors[,1] * 2, vectors[, 2] * 2,
lwd = 2, lty = 1, length = 0.2, col = "red")
text(vectors[,1] * 2, vectors[, 2] * 2, pos = 3,
labels = 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:

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.

```
long.lat <- as.matrix(cbind(env$long, env$lat))
coord.dist <- earth.dist(long.lat, dist = TRUE)

bray.curtis.dist <- 1 - vegdist(comm)

unifrac.dist <- 1 - dist.uf
```

```
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")

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.

```
par(mfrow=c(2, 1), mar = c(1, 5, 2, 1) + 0.1, oma = c(2, 0, 0, 0))
plot(df$geo.dist, df$bray.curtis, xlab = "", xaxt = "n", las = 1, ylim = c(0.1, 0.9), ylab="Bray-Curtis
main = "Distance Decay", col = "SteelBlue")
DD.reg.bc <- lm(df$bray.curtis ~ df$geo.dist)
summary(DD.reg.bc)
```

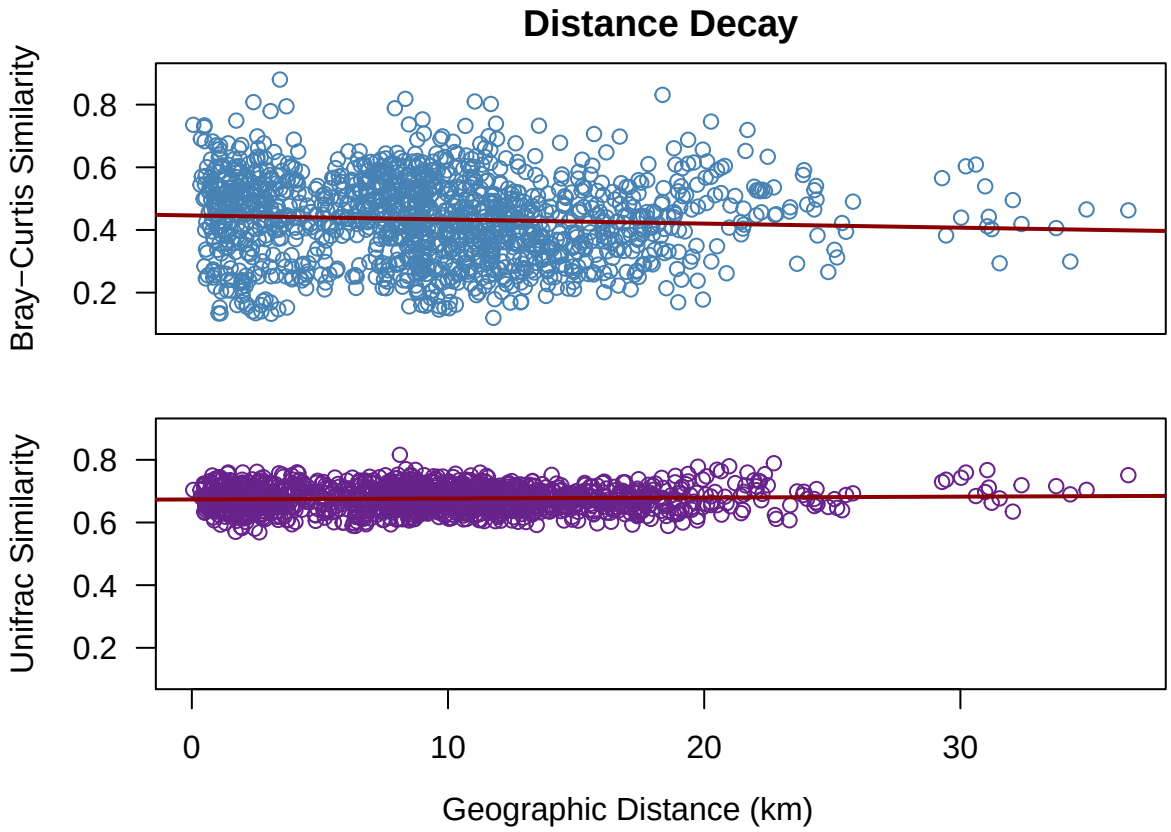
```
##
## Call:
## lm(formula = df$bray.curtis ~ df$geo.dist)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.31151 -0.08843  0.00315  0.09121  0.43817
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.4463453  0.0066883  66.735   <2e-16 ***
## 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)

par(mar = c(2, 5, 1, 1) + 0.1)
plot(df$geo.dist, df$unifrac, xlab = "", las = 1, ylim = c(0.1, 0.9),
     ylab = "Unifrac Similarity", col = "darkorchid4")
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)

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.

```
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%
## 0.00081 0.00102 0.00125 0.00145
```

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

> ***Answer 7***: The slopes of the UniFrac and Bray-Curtis distance decay plots are very (very) slightly different, by 0.0016. This doesn't really seem like a large or interesting difference, but when we look a little closer at the plots, it becomes clear that while this slope difference is small, it results in a change in our interpretation of the two relationships. While the Bray-Curtis plot shows a definite slight decrease in similarity over distance, the UniFrac plot seems to show a removal or even slight reversal of this trend with similarity ever so slightly increasing over distance.
>
> Perhaps we see this difference between the two measures, when phylogenetic relatedness is accounted for, due to range expansion ultimately leading to speciation at the edges of a species' distribution. That is, while species in close proximity to one another (small geographic distance) might be expected to show a wide range of relatedness, there may be instances in which far spaced species have a shared recent evolutionary history and are now far spaced due to recent dispersion to reduce intraspecies competition which lead to reduced interbreeding of subpopulations and eventual speciation by distance.

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){

areas.iter <- c()
diversity.iter <- c()

for (j in 1:10){
pond.sample <- sample(51, replace = FALSE, size = i)
area <- 0
sites <- c()
for (k in pond.sample) {
area <- area + pond.areas[k]
sites <- rbind(sites, comm[k, ])
}
areas.iter <- c(areas.iter, area)
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))
```

```
areas <- c(areas, mean(areas.iter))
print(c(i, mean(diversity.iter), mean(areas.iter)))
}
return(cbind(areas, diversity))
}
```

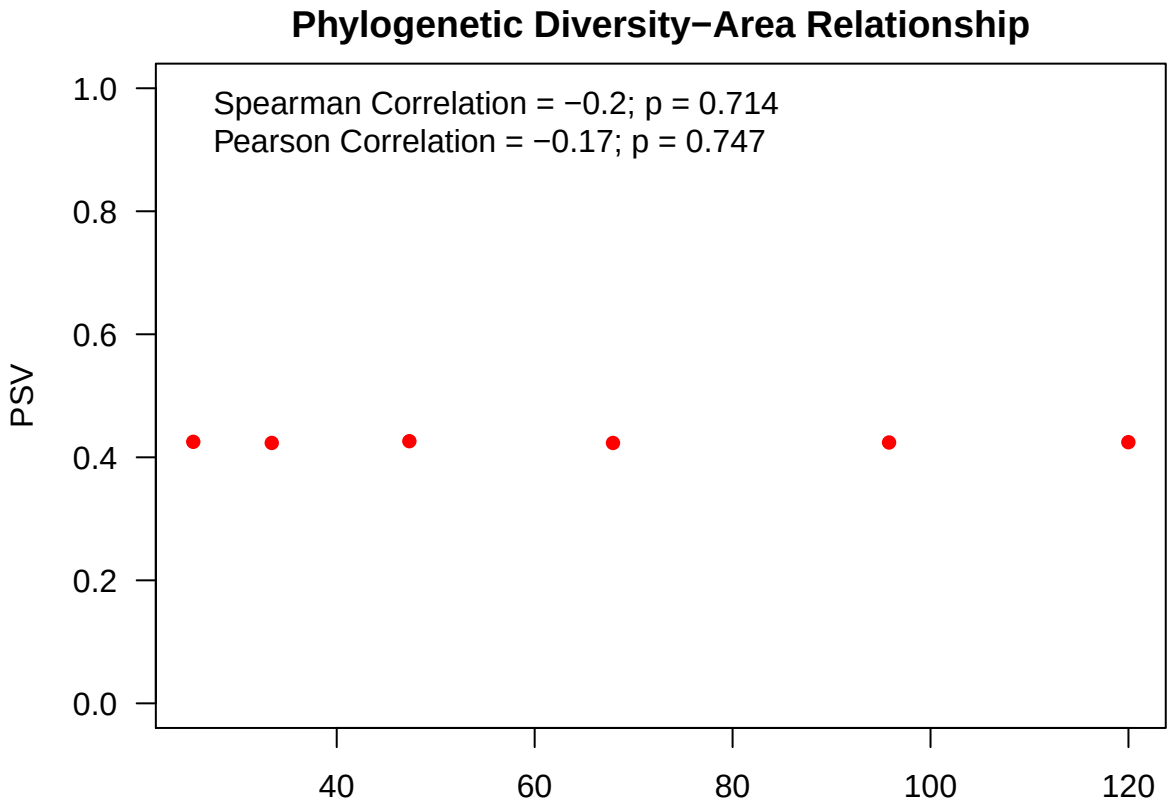
ii. Evaluating the PDAR

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.

```
pond.areas <- as.vector(pi * (env$Diameter/2)^2)

pdar <- PDAR(comm, phy)
```

```
## [1]   2.0000000   0.4251795 650.3638718
## [1]    4.0000000    0.4233742 1117.9608785
## [1]    8.0000000    0.4261951 2241.0213473
## [1]   16.00000    0.42329 4612.44314
## [1]   32.0000000    0.4240752 9179.9175170
## [1] 5.100000e+01 4.245479e-01 1.439763e+04
```

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

Pearson <- cor.test(pdar$areas, pdar$diversity, method = "pearson")
P <- round(Pearson$estimate, 2)
P.pval <- round(Pearson$p.value, 3)

Spearman <- cor.test(pdar$areas, pdar$diversity, method = "spearman")
rho <- round(Spearman$estimate, 2)
rho.pval <- round(Spearman$p.value, 3)

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("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***: The phylogenetic species variability (PSV) metric calculated is constant over geographic distance and hovers right around 0.4. Values of PSV near 1 indicate that species in a sample are unrelated, and values approach 0 as relatedness increases. Our finding then indicates that there is a moderate, constant level of relatedness between the species in our sample across space. This can be contrasted with the result of our standard, taxanomic based species-area relationship curve (SAR) which shows that as the sample area increases, the log of species richness increases at a slope of 0.144.
>
> The difference between our SAR and PDAR curves (increasing vs constant over space) seems to indicate that there is a signature of phylogenetic repulsion in our data. That is, as we expand our sampling area, we continue to find new species (leading to a positive SAR curve), but these new species found in distant locations are more closely related to species found in close proximity to one another (flattening out the PDAR curve). This relationship seen in the data fits with the prediction from question 7 above that closely related species may be found further apart than distantly related ones due to the influence of shared recent common ancestry.

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.

> As my currently planned research centers on an exploration of the predictions of the hologenome theory of evolution, and specifically phylosymbiosis (concordance between a host phylogeny and

associated microbial community assemblages): phylogenetic information will be integral to my research plans. Aside from the sequence data needed to assemble accurate host phylogenies, or at least validate existing assemblies, the techniques covered in the last two modules will be useful in constructing and comparing representitive microbiome community profiles (based on 16s sequence data) for each host I would like to compare. Though I currently don't plan on looking at the evolutionary relationships of the members of the microbiomes themselves per se, I will still need to apply the same types of phylogenetically informed distance metrics (UniFrac) and tree comparison methods covered this week and last in order to gather the kind of community composition data needed to answer the questions I have in mind.