Phylogenetic Diversity - Communities

RZ Moger-Reischer; BIOL-Z 620: Quantitative Biodiversity, Indiana University

25 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()
#setwd("~/GitHub/QB2017_Moger-Reischer/Week7-PhyloCom/")
#setwd("C:\\Users\\rmoge\\GitHub\\QB2017_Moger-Reischer\\Week7-PhyloCom")

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

```
##
## Attaching package: 'devtools'
```

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

```
## 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.

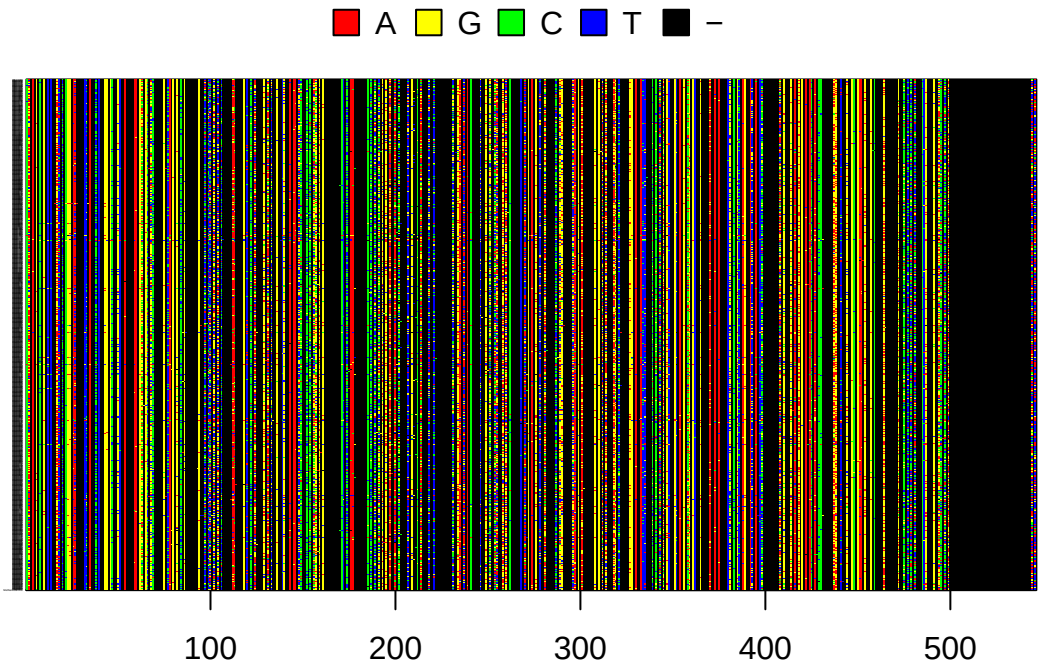
3) LOAD THE DATA

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.

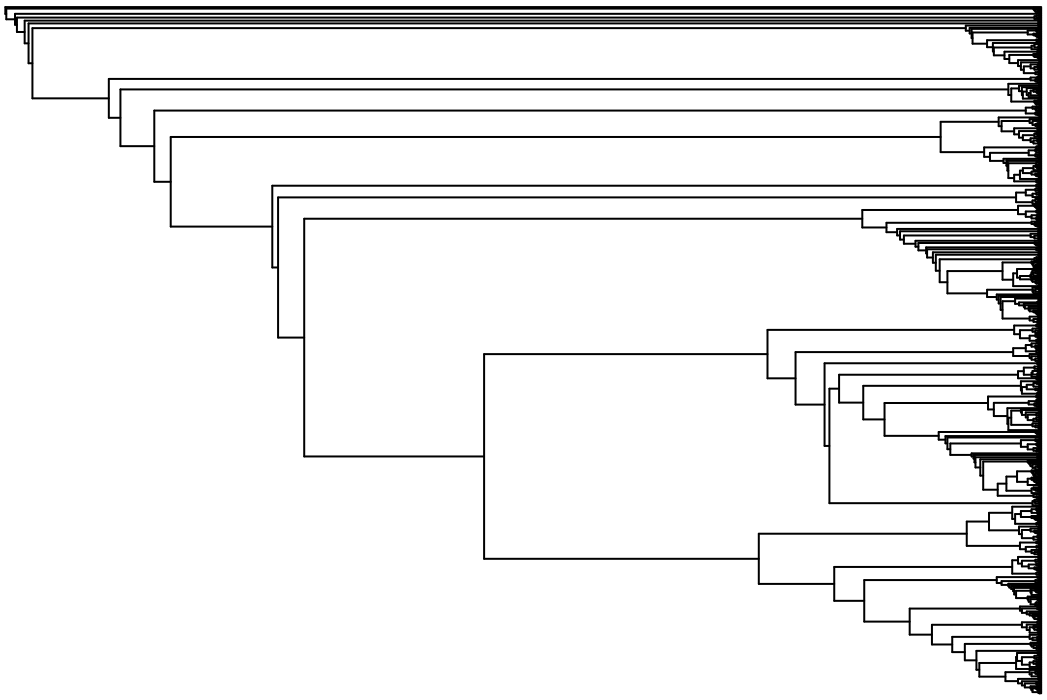
```
#1
env <- read.table("data/20130801_PondDataMod.csv", sep = ",", header = TRUE)
env <- na.omit(env)
#2
comm <- read.otu(shared = "./data/INPonds.final.rdp.shared", cutoff = "1")
#3
comm <- comm[grep("*-DNA", rownames(comm)), ]
#4
rownames(comm) <- gsub("\\-DNA", "", rownames(comm))
rownames(comm) <- gsub("\\_", "", rownames(comm))

comm <- comm[rownames(comm) %in% env$Sample_ID, ]
#5
comm <- comm[ , colSums(comm) > 0]

#6
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.

```
#1
ponds.cons <- read.alignment(file = "./data/INPonds.final.rdp.1.rep.fasta", format = "fasta")
#2
ponds.cons$nam <- gsub("\\|.*$", "", gsub("^.*?\t", "", ponds.cons$nam))
#3
outgroup <- read.alignment(file = "./data/methanosarcina.fasta", format = "fasta")
#4
DNAbin <- rbind(as.DNAbin(outgroup), as.DNAbin(ponds.cons))
#5
image.DNAbin(DNAbin, show.labels=T, cex.lab = 0.05, las = 1)
```




```
#6
seq.dist.jc <- dist.dna(DNAbin, model = "JC", pairwise.deletion = FALSE)
#7
phy.all <- bionj(seq.dist.jc)
#8
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)
#9
par(mar = c(1, 1, 2, 1) + 0.1)
plot.phylo(phy, main = "Neighbor Joining Tree", "phylogram", show.tip.label = FALSE, use.edge.length =
```



4) PHYLOGENETIC ALPHA DIVERSITY

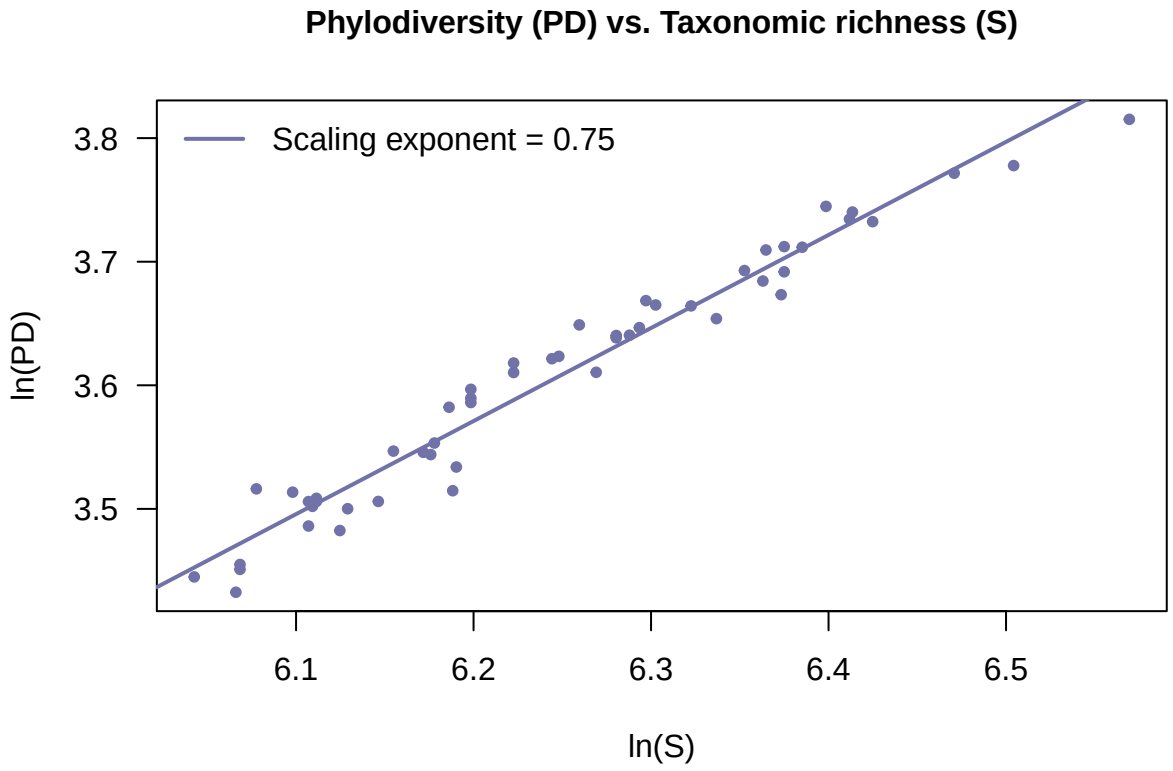
A. Faith's Phylogenetic Diversity (PD)

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)
#1
plot(log(pd$S), log(pd$PD),pch = 20, col = "#7072a8", las = 1,xlab = "ln(S)", ylab = "ln(PD)", cex.main
fit <- lm('log(pd$PD) ~ log(pd$S)')
#2
abline(fit, col = "#7072a8", lw = 2)
#3
exponent <- round(coefficients(fit)[2], 2)
legend("topleft", legend=paste("Scaling exponent = ", exponent, sep = ""), bty = "n", lw = 2, col = "#7
```



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 takes a sum of branch lengths. If there are more species present (higher richness), then there will be more branches, and the sum should be larger.

Answer 1b:

It looks like a powerlaw with exponent == 0.75 or 3/4. (Isn't that the same as metabolic rate scaling with body size?)

Answer 1c:

Perhaps if there were a recent radiation event on an island such that there are many, many taxa, but they all cluster into two or three very closely related groups which each descended from one small colonizer population. Then there might be high richness but low PD.

Answer 1d:

P <= 2x10^-16. We can be very sure that there is a real statistical relationship between (the logarithms of) these two variables.

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.pdr <- ses.pd(comm[1:2,], phy, null.model = "richness", runs = 25, include.root = FALSE)
ses.pdf <- ses.pd(comm[1:2,], phy, null.model = "frequency", runs = 25, include.root = FALSE)
ses.pdsp <- ses.pd(comm[1:2,], phy, null.model = "sample.pool", 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:

We are using simulation to ascertain whether our sample is more or less phylogentically diverse than some null expectation for a sample with a given X (where X depends on the specfic null model).

Answer 2b:

For the 'frequency' null model, the mean observed PD was 42.33. For the 'richness' null model, the mean observed PD was 42.33 also. The same was true for the 'sample.pool' null model. Perhaps the choice of null model did not affect the output because we ran too few randomizations to see subtle differences.

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 = FALSE, runs = 25)
NRI <- as.matrix(-1 * ((ses.mpd[,2] - ses.mpd[,3]) / ses.mpd[,4]))
rownames(NRI) <- row.names(ses.mpd)
colnames(NRI) <- "NRI"


ses.mpdabund <- ses.mpd(comm, phydist, null.model = "taxa.labels", abundance.weighted = T, runs = 25)
NRIabund <- as.matrix(-1 * ((ses.mpdabund[,2] - ses.mpdabund[,3]) / ses.mpdabund[,4]))
rownames(NRIabund) <- row.names(ses.mpdabund)
colnames(NRIabund) <- "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 = FALSE, runs = 25)
NTI <- as.matrix(-1 * ((ses.mntd[,2] - ses.mntd[,3]) / ses.mntd[,4]))
rownames(NTI) <- row.names(ses.mntd)
```

```
colnames(NTI) <- "NTI"

ses.mntdabund <- ses.mntd(comm, phydist, null.model = "taxa.labels", abundance.weighted = T, runs = 25)
NTIabund <- as.matrix(-1 * ((ses.mntdabund[,2] - ses.mntdabund[,3]) / ses.mntdabund[,4]))
rownames(NTIabund) <- row.names(ses.mntdabund)
colnames(NTIabund) <- "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:

NRI takes the average of pairwise distances along the phylogeny for the species in a site, and compares it to a null expectation.

Answer 3b:

NTI takes the average minimum pairwise distances along the phylogeny for the species in a site, and compares it to a null expectation.

Answer 3c:

NRI was negative for all sites. This indicates at the species are phylogenetically overdispersed.

NTI was closer to zero and in fact postive for some sites. Using this metric the samples appear to more closely match the null expectations.

Answer 3d:

NRI values were now close to zero and/or tending to be positive. Now the interpretation is that there is some phylogenetic clustering for the sites in our sample. NTI values were more-positive whilst accounting for the abundance data, implying more phylogenetic clustering. Overall this tells me that the most-abundant species tend to be phylogenetically clustered.

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.

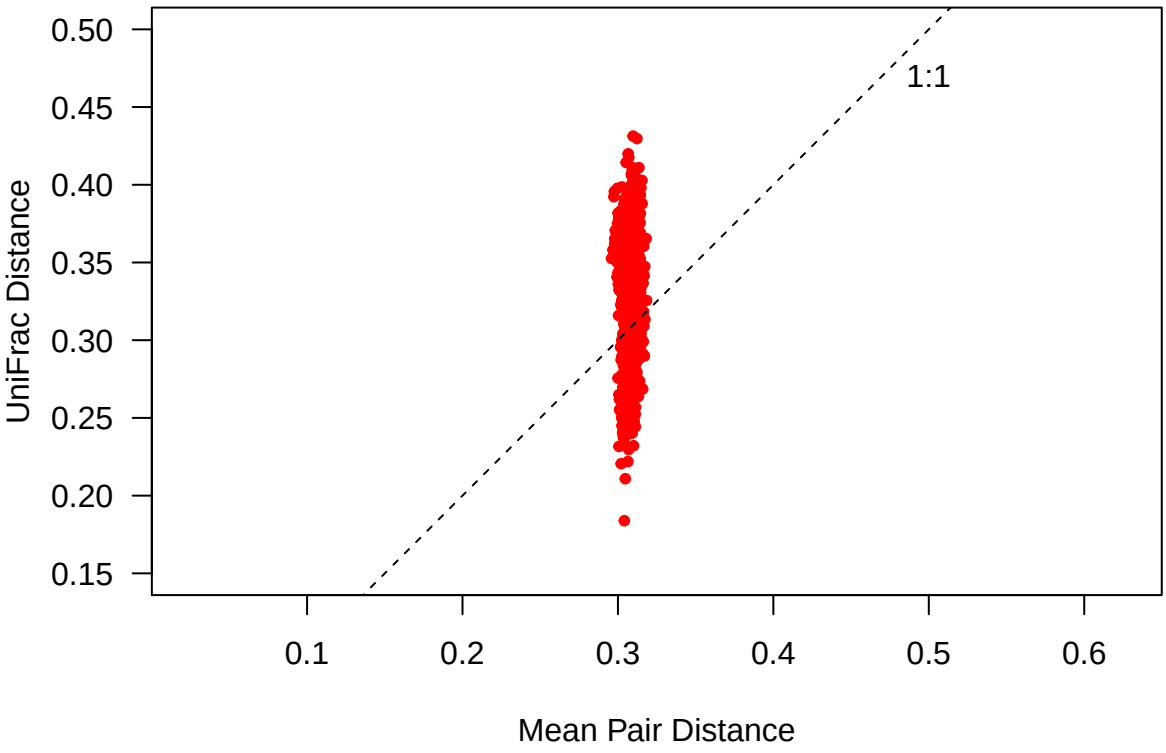
```
#1
dist.mp <- comdist(comm, phydist)
```

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

```
#2
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:

For all pairs of taxa in two samples, the mean pairwise distance looks at all those pairwise distances and takes average. On the other hand, UniFrac distance looks at the entire tree topology, and then calculates what proporortion of it comprises branches which are part of both samples and what proportion of its branch length is occupied by only one of those two samples.

Answer 4b:

The UniFrac distance can vary a lot, but the mean pair distance changes only a little.

Answer 4c:

The mean pairwise distance is an average of all pairs—an average of many calculations. Contrariwise, the UniFrac distance between two samples is a measurement of a single ratio. I would expect that a metric which is an average will tend to be less variable.

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.

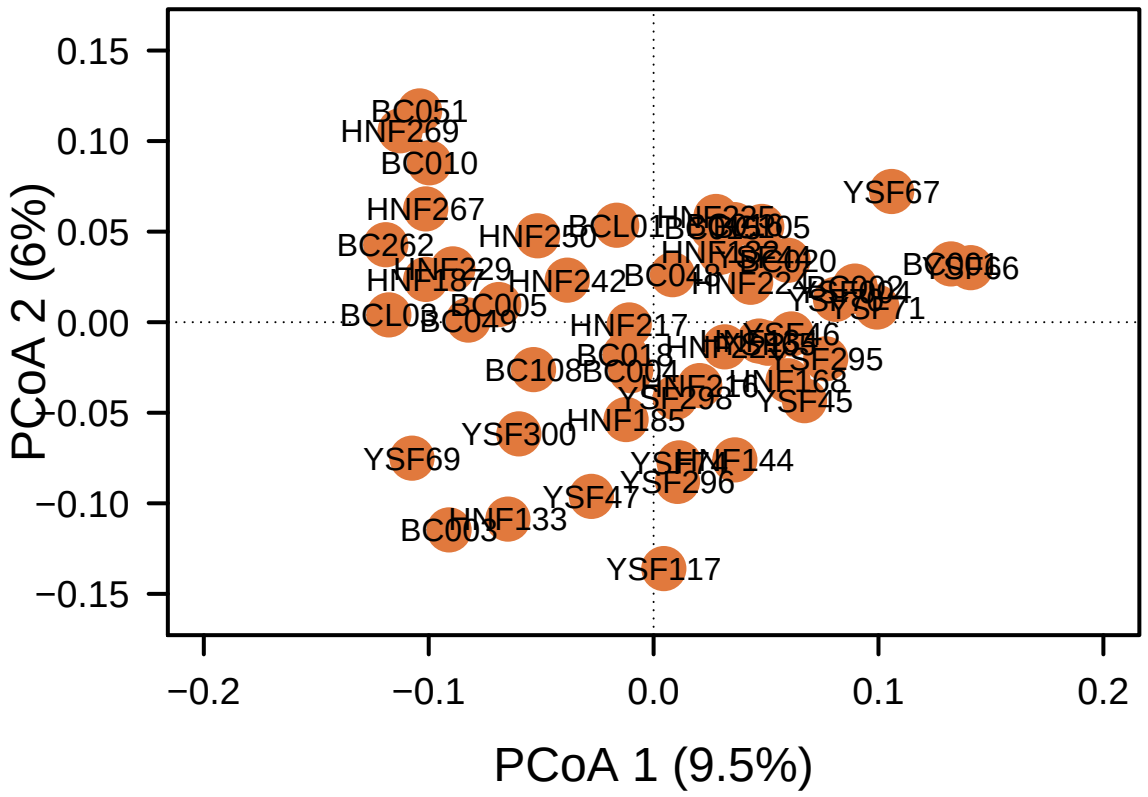
In the R code chunk below, do the following:
1. perform a PCoA based on the UniFrac distances, and
2. calculate the explained variation for the first three PCoA axes.

```
#1
pond.pcoa <- cmdscale(dist.uf, eig = T, k = 3)
#2
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.

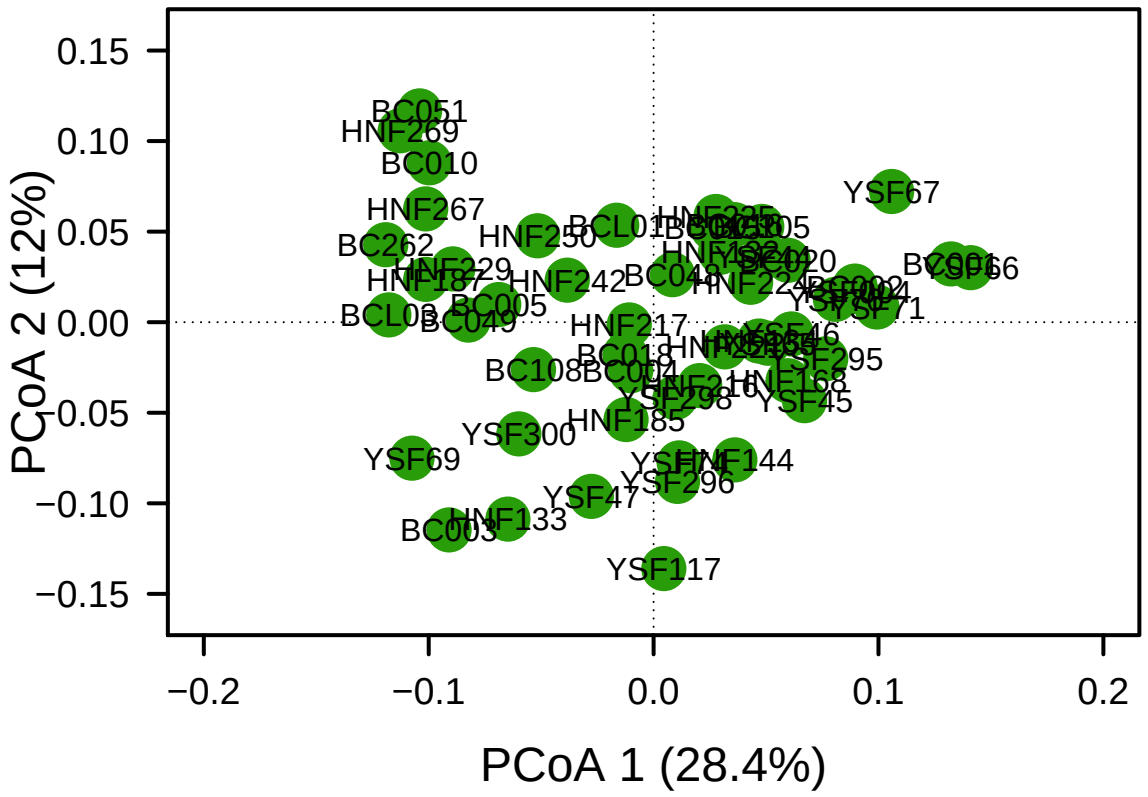
In the R code chunk below, do the following:
1. plot the PCoA results using either the R base package or the `ggplot` package,
2. include the appropriate axes,
3. add and label the points, and
4. customize the plot.

```
par(mar = c(5, 5, 1, 2) + 0.1)
#1, #2
plot(pond.pcoa$points[ ,1], pond.pcoa$points[ ,2],xlim = c(-0.2, 0.2), ylim = c(-.16, 0.16),xlab = past
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)
#3
points(pond.pcoa$points[ ,1], pond.pcoa$points[ ,2],
pch = 19, cex = 3, bg = "#000000", col = "#e1793d")
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.tax <- cmdscale(dist.db, eig = T, k = 3)
#2
explainvar1 <- round(pond.pcoa.tax$eig[1] / sum(pond.pcoa.tax$eig), 3) * 100
explainvar2 <- round(pond.pcoa.tax$eig[2] / sum(pond.pcoa.tax$eig), 3) * 100
explainvar3 <- round(pond.pcoa.tax$eig[3] / sum(pond.pcoa.tax$eig), 3) * 100
sum.eig.tax <- sum(explainvar1, explainvar2, explainvar3)

par(mar = c(5, 5, 1, 2) + 0.1)
#1, #2
plot(pond.pcoa$points[ ,1], pond.pcoa$points[ ,2],xlim = c(-0.2, 0.2), ylim = c(-.16, 0.16),xlab = paste
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)
#3
points(pond.pcoa$points[ ,1], pond.pcoa$points[ ,2],
pch = 19, cex = 3, bg = "#000000", col = "#299d09")
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 PCoA clusterings look mostly similar. However, for the taxonomic ordination, the first two PCoA axes explain a lot more (40.4%) of the variation than do the first two axes for the phylogenetic ordination (15.5%). This could mean that phylogenetic information is not very important for this system.

C. Hypothesis Testing

i. Categorical Approach

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

```
watershed <- env$Location
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  0.028 *
## 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
```

```
cata<-adonis(vegdist(decostand(comm, method = "log"),method = "bray") ~ watershed,permutations = 999)
```

ii. Continuous Approach

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

```
envs <- env[, 5:19]
envs <- envs[, -which(names(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.

```
my.mantel<-mantel(dist.uf, env.dist)
```

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

```
#1
ponds.dbrda <- vegan::dbrda(dist.uf ~ ., data = as.data.frame(scale(envs)))
#2
anova(ponds.dbrda, by = "axis")
```

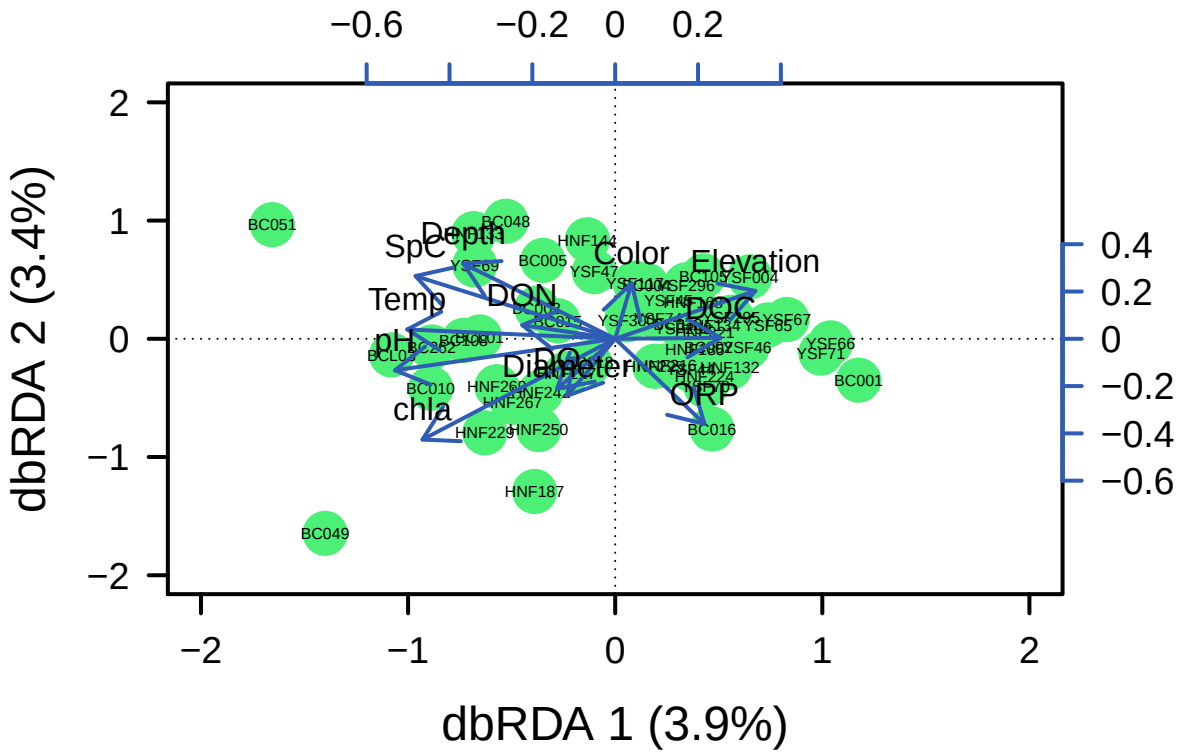
```
## Permutation test for dbrda under reduced model
## Marginal tests for axes
## Permutation: free
## Number of permutations: 999
##
## Model: vegan::dbrda(formula = dist.uf ~ Elevation + Diameter + Depth + ORP + Temp + SpC + DO + pH + C
##          Df SumOfSqs      F Pr(>F)
## dbRDA1    1  0.10566 2.0152  0.003 **
## dbRDA2    1  0.09258 1.7658  0.004 **
## dbRDA3    1  0.07555 1.4409  0.035 *
## dbRDA4    1  0.06677 1.2735  0.104
## dbRDA5    1  0.05666 1.0807  0.321
## dbRDA6    1  0.05293 1.0095  0.476
## dbRDA7    1  0.04750 0.9059  0.632
## dbRDA8    1  0.03941 0.7517  0.894
## dbRDA9    1  0.03775 0.7201  0.934
## dbRDA10   1  0.03280 0.6256  0.985
## 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
```

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

```
##
## ***VECTORS
##
##             dbRDA1   dbRDA2     r2 Pr(>r)
## Elevation  0.77670  0.62986 0.0959  0.091 .
## Diameter  -0.27972 -0.96008 0.0541  0.256
## Depth     -0.63137  0.77548 0.1756  0.008 **
## ORP        0.41879 -0.90808 0.1437  0.023 *
## Temp      -0.98250  0.18628 0.1523  0.018 *
## SpC       -0.77101  0.63682 0.2087  0.003 **
## DO        -0.39318 -0.91946 0.0464  0.332
## pH        -0.96210 -0.27270 0.1756  0.004 **
## Color      0.06353  0.99798 0.0464  0.305
## chla      -0.60392 -0.79704 0.2626  0.006 **
## DOC        0.99847 -0.05526 0.0382  0.393
## DON       -0.91633  0.40042 0.0339  0.419
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Permutation: free
## Number of permutations: 999
```

```
# Calculate explained variation
dbrda.explainvar1 <- round(ponds.dbrda$CCA$eig[1] / sum(c(ponds.dbrda$CCA$eig, ponds.dbrda$CA$eig)), 3)
dbrda.explainvar2 <- round(ponds.dbrda$CCA$eig[2] / sum(c(ponds.dbrda$CCA$eig, ponds.dbrda$CA$eig)), 3)
#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 * 100, "%)"), ylab = paste("dbRDA 2 (", dbrda.explainvar2 * 100, "%)"), pch = 16, cex = 2.0, type = "n", cex.lab = 1.5, cex.axis = 1.2, axes = FALSE)
# 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)
# labels, points
points(scores(ponds.dbrda, display = "wa"),
pch = 19, cex = 3, bg = "#4cf076", col = "#4cf076")
text(scores(ponds.dbrda, display = "wa"),
labels = row.names(scores(ponds.dbrda, display = "wa")), cex = 0.5)
#vectors
vectors <- scores(ponds.dbrda, display = "bp")
arrows(0, 0, vectors[,1] * 2, vectors[, 2] * 2,lwd = 2, lty = 1, length = 0.2, col = "#2f5cb5")
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 = "#2f5cb5", lwd = 2.2, at = pretty(range(vectors[,1] * 2)), labels = FALSE)
axis(side = 4, lwd.ticks = 2, cex.axis = 1.2, las = 1, col = "#2f5cb5", lwd = 2.2, at = pretty(range(vectors[,2] * 2)), labels = FALSE)
```



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***:
>
> Watershed of origin has a significant influence on phylogenetic distribution of species (P = 0.006). Among the environmental variables, it looks like depth; ORP; temperature; conductivity; pH; and chlorophyll a are significant predictors of phylogenetic composition.

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.

```
#1
long.lat <- as.matrix(cbind(env$long, env$lat))
coord.dist <- earth.dist(long.lat, dist = TRUE)
#2
bray.curtis.dist <- 1 - vegdist(comm)
```

```
#3
unifrac.dist <- 1 - dist.uf
# Transform all distances into list format:
unifrac.dist.ls <- liste(unifrac.dist, entry = "unifrac")
bray.curtis.dist.ls <- liste(bray.curtis.dist, entry = "bray.curtis")
coord.dist.ls <- liste(coord.dist, entry = "geo.dist")
env.dist.ls <- liste(env.dist, entry = "env.dist")
#4
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))
#1
plot(df$geo.dist, df$bray.curtis, xlab = "", xaxt = "n", las = 1, ylim = c(0.1, 0.9), ylab="Bray-Curtis
#3
DD.reg.bc <- lm(df$bray.curtis ~ df$geo.dist)
summary(DD.reg.bc)
```

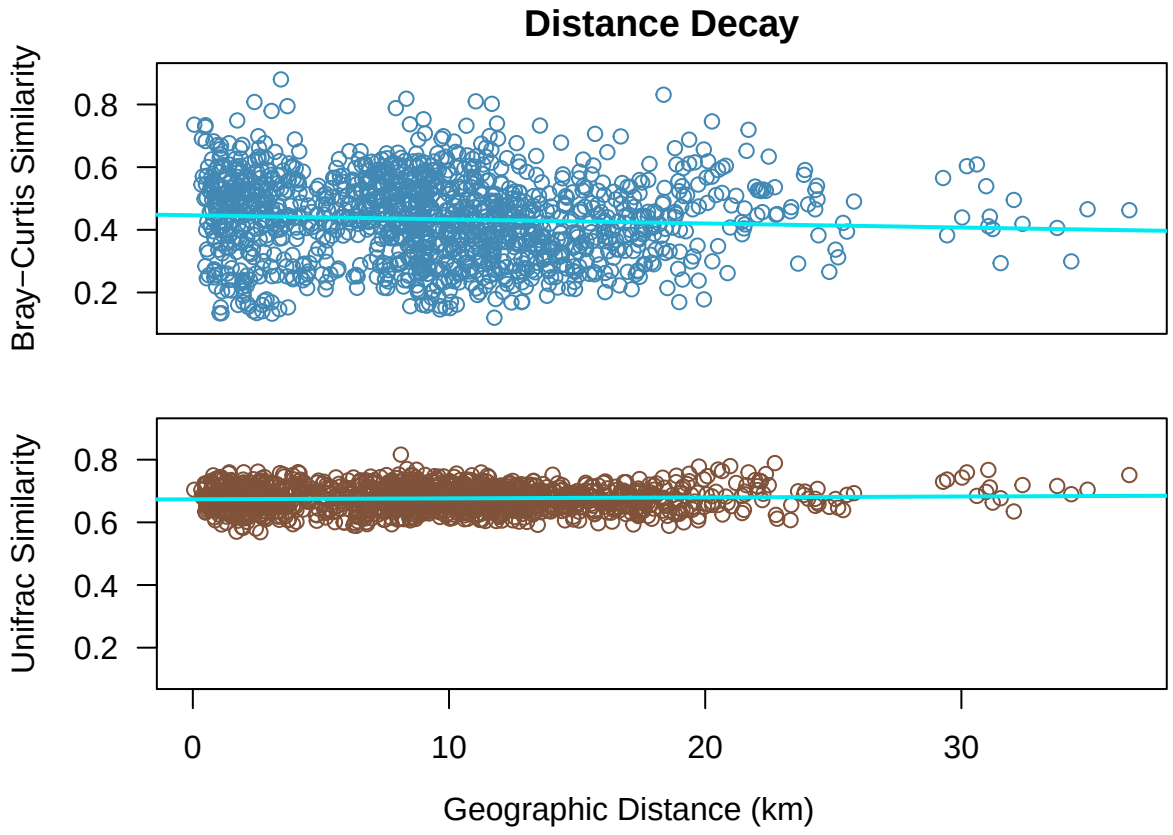
```
##
## Call:
## lm(formula = df$bray.curtis ~ df$geo.dist)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.31151 -0.08843  0.00315  0.09121  0.43817
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.4463453  0.0066883  66.735   <2e-16 ***
## 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 = "#00eaf3", lwd = 2)

par(mar = c(2, 5, 1, 1) + 0.1)
#2
plot(df$geo.dist, df$unifrac, xlab = "", las = 1, ylim = c(0.1, 0.9), ylab = "Unifrac Similarity", col =
#3
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 = "#00eaf3", 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: 0.005
##
## Empirical upper confidence limits of r:
##      90%      95%    97.5%      99%
## 0.000779 0.000996 0.001168 0.001446
```

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 are statistically significantly different. However, the difference is quite small in magnitude (0.001603). Therefore one might believe it would be wise to investigate the biology from which the statistical difference arises—it is possible that this statistically significant difference is not biologically meaningful.
>
> That said, the magnitude of the slopes is very small. And in fact the absolute value of the slope for the DD is an order of magnitude larger than that of the slope of the PDD. It appears that there is biological significance. In addition, for the PDD the slope is slightly positive (although the regression itself is not significant (P = 0.07738))—when accounting phylogeny there is no decrease in similarity over geographic distance. Whereas for the taxonomic DD there is a statistically signifcant (P = 0.0262) decrease in similarity across geographic 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))
  }
```

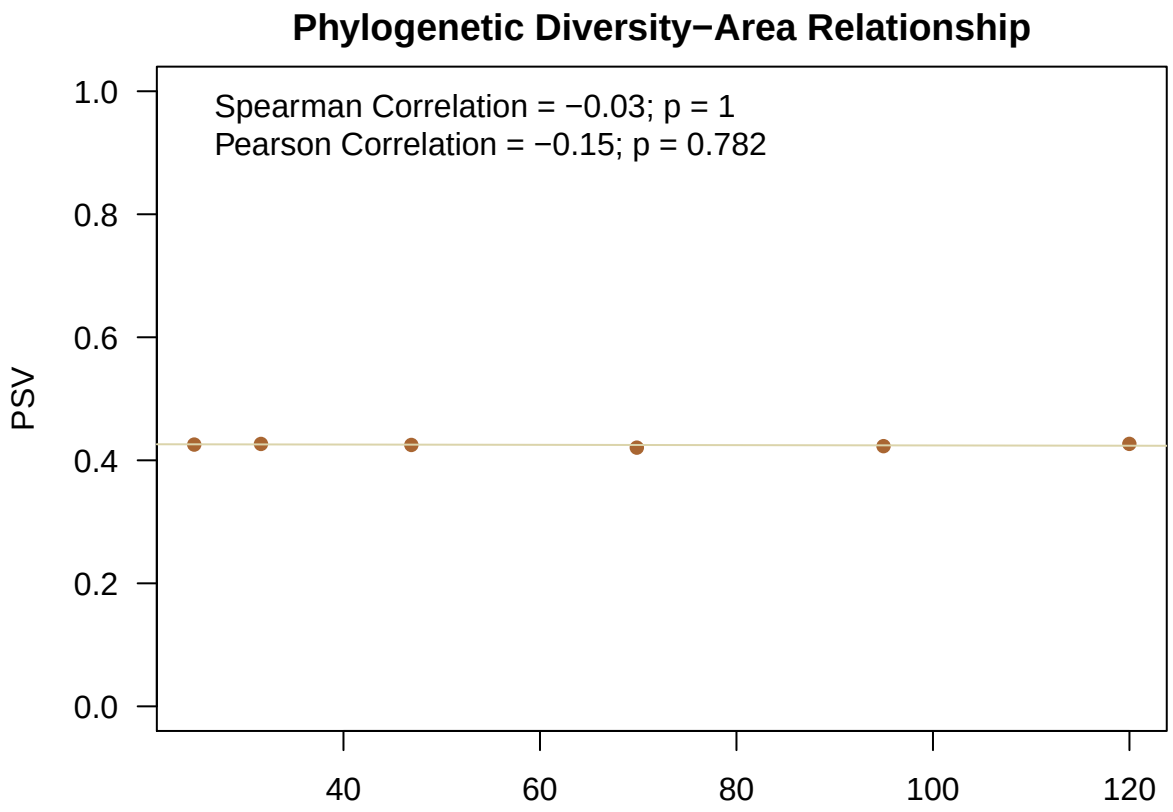
ii. Evaluating the PDAR

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.

```
#1
pond.areas <- as.vector(pi * (env$Diameter/2)^2)
#2
pdar <- PDAR(comm, phy)
```

```
## [1]   2.0000000   0.4255822 615.4874809
## [1]   4.000000   0.426611 998.673035
## [1]    8.0000000    0.4249637 2200.3275123
## [1]   16.0000000    0.4205598 4880.0133688
## [1]   32.000000    0.422996 9018.931314
## [1] 5.100000e+01 4.267004e-01 1.439763e+04
```

```
pdar <- as.data.frame(pdar)
pdar$areas <- sqrt(pdar$areas)
#3
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)
#4
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-A
legend("topleft", legend= c(paste("Spearman Correlation = ", rho, "; p = ", rho.pval, sep = ""),paste("
dufit<-lm(pdar$diversity~pdar$areas )
abline(.4266,-0.00002398, col="#dbd4ab")
```



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 z value for the SAR is 0.144. For the PDAR, however, the slope is essentially zero (very slightly negative, slope=-0.00002398). This makes sense. The SAR should always be positive, because you should not 'lose' species as area increases. Contrariwise, it is possible that as area increases the spatial extent will start to include new species that are very closely related to species which have already been sampled—perhaps these close relatives are occupying the same niche, but are located somewhere adjacent. In such a circumstance there is phylogenetic clustering of trait values, and local competitive exclusion among closely related species.

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.

Let us imagine that there are epigenetic mechanisms regulating a dormancy response to starvation in bacteria. There are a few ways I could use phylogenetic information. Firstly, I could construct a tree and assess the clustered/"overdispersed" pattern of the trait(i.e. up- or down-regulation of methylation when starved)'s appearance on the tree topology. Secondly, within and between methyltransferase families, I would like to BLAST flanking genes for each dormancy-related methylation motif in a custom library all-against-all search. By identifying reciprocal best hits, I

would be able to determine which SPECIFIC motifs are phylogenetically conserved. (Too, I might be able to learn which up- and down-stream GENEs could be involved in the dormancy response.)