Phylogenetic Diversity - Communities

Z620: Quantitative Biodiversity, Indiana University

February 27, 2015

OVERVIEW

Taxonomic measures of α- and β-diversity do not integrate evolutionary information. However, evolutionary information is important for a broad range of biodiversity issues including conservation, biogeography, and community assembly. In this handout, we introduce some commonly used methods in phylogenetic community ecology. These methods will allow us to account for phylogenetic diversity and provide insight into the mechansisms that give rise to the over- and under-dispersion (i.e., clustering) of biological communities.

After completing this exercise you will know how to:

1. measure phylogenetic α diversity
2. measure phylogenetic β diversity
3. evaluate the contribution of phylogeny to geographical patterns of biodiversity

1) SETUP

A. Retrieve and Set Your Working Directory

```
rm(list = ls())
getwd()
setwd("~/GitHub/QuantitativeBiodiversity/Assignments/PhyloCom")
```

B. Load Packages

We will be relying heavily on the R package `picante`. This package has many of the functions that are contained in the software Phylocom, which is used for the analysis of phylogenetic community structure and character evolution (http://phylodiversity.net/phylocom/). We will also use a few other packages packages, some of which were introduced in the Phylogenetic Traits module.

After the initial installation of these packages using the `install.packages()` fucntion, load the packages and their dependencies with the `require()` function:

```
require("picante")
require("ape")
require("seqinr")
require("vegan")
require("fossil")
require("simba")
```

C. Load Source Code

In addition to relying on contributed packages, we will also be using a source code file. A source code file has user-defined functions that are required for certain analyses. The benefit of source files is that they

contain "vetted" code that can be used across multiple projects. Here, we will be using a source code file that includes a function for reading in the output files from the popular community sequencing software `mothur` (http://www.mothur.org/).

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

2) DESCRIPTION OF DATA

We will revisit the data that was used in the Geographical Ecology module. As a reminder, in 2013 we sampled ~ 50 forested ponds located in Brown County State Park, Yellowood State Park, and Hoosier National Forest in southern Indiana. In addition to measuring a suite of geographic and environmental variables, we characterized the diversity of bacteria in the ponds using molecular-based approaches. Specifically, we amplified the 16S rRNA gene (i.e., "DNA") and 16S rRNA transcripts (i.e., "RNA") of bacteria using barcoded primers on the Illumina MiSeq platform. We then used a `mothur` pipeline to quality-trim our data set and assign sequences to operational taxonomic units (OTU). In this exercise, we will use the DNA sequence data for making community phylogenetic inference.

3) LOADING OF DATA

First, let's load the environmental data:

```
# Load Environmental Data
env <- read.table("data/20130801_PondDataMod.csv", sep = ",", header = TRUE)
```

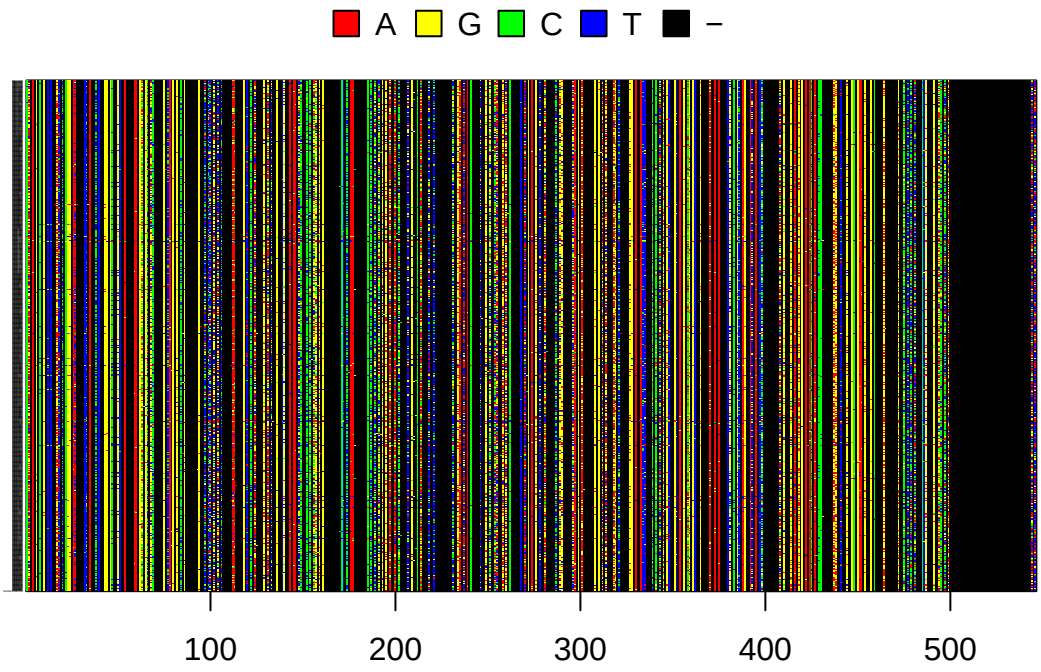
Next, let's load the bacterial OTU data (i.e., site-by-species matrix):

```
# Load Site-by-Species Matrix
comm <- read.otu(shared = "./data/INPonds.final.rdp.shared", cutoff = "1")

# Select DNA Data: Use the `grep()` Command and Rename with `gsub()`
comm <- comm[grep("*-DNA", rownames(comm)), ]
rownames(comm) <- gsub("\\-DNA", "", rownames(comm))
rownames(comm) <- gsub("\\_", "", rownames(comm))

# Remove Sites Not in the Environmental Data Set
comm <- comm[rownames(comm)  %in% env$Sample_ID, ]

# Remove Zero-Occurrence Taxa
comm <- comm[ , colSums(comm) > 0]

# Import Taxonomy Data Using `read.tax()` from Source Code
tax <- read.tax(taxonomy = "./data/INPonds.final.rdp.1.cons.taxonomy")
```

Now, let's load and process the phylogenetic data:

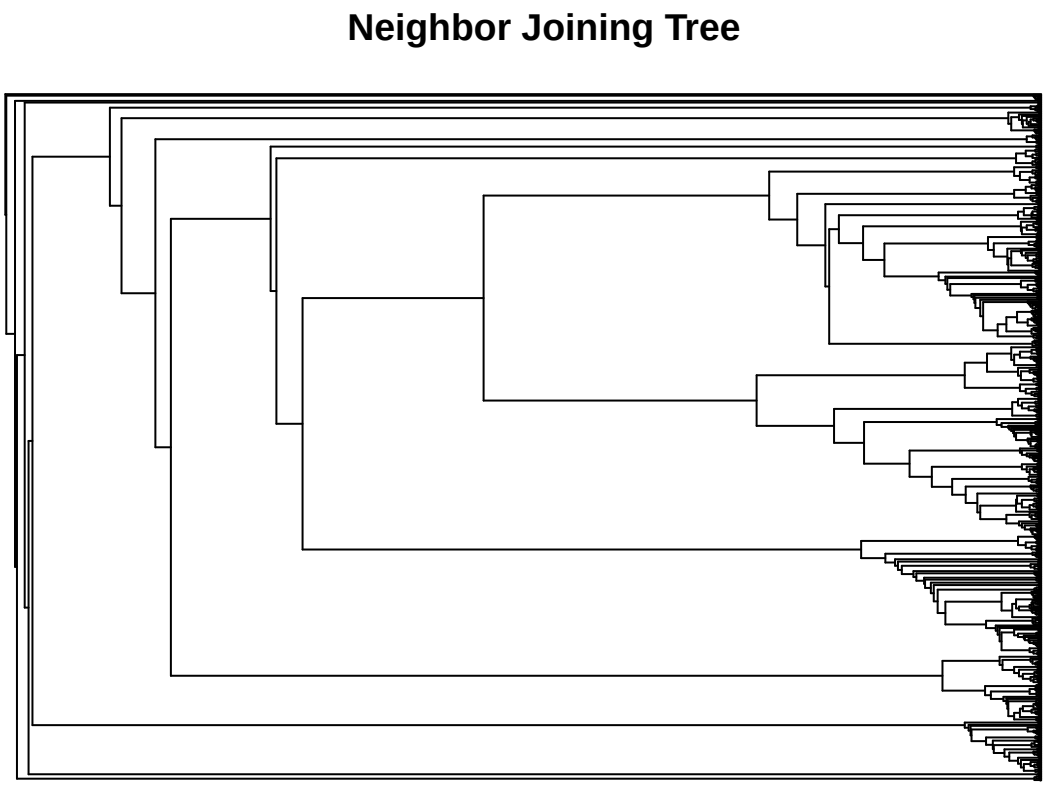
```
# Import the Alignment File {seqinr}
ponds.cons <- read.alignment(file = "./data/INPonds.final.rdp.1.rep.fasta",
                             format = "fasta")

# Rename OTUs in the FASTA File
ponds.cons$nam <- gsub("\\|.*$", "", gsub("^.*?\t", "", ponds.cons$nam))
```

```
# Import the Outgroup FASTA File {seqinr}
outgroup <- read.alignment(file = "./data/methanosarcina.fasta", format = "fasta")

# Convert Alignment File to DNAbin Object {ape}
DNAbin <- rbind(as.DNAbin(outgroup), as.DNAbin(ponds.cons))

# Visusalize Sequence Alignment {ape}
image.DNAbin(DNAbin, show.labels=T, cex.lab = 0.05, las = 1)
```




```
# Create Distance Matrix with the Jukes Cantor "JC" Model {ape}
seq.dist.jc <- dist.dna(DNAbin, model = "JC", pairwise.deletion = FALSE)

# Use Neighbor Joining Algorithm to Construct a Full Tree (DNA and RNA sequences) {ape}
phy.all <- bionj(seq.dist.jc)

# Drop Tips of Zero-Occurrence OTUs (Removes Taxa Only Found via RNA Sequencing) {ape}
phy <- drop.tip(phy.all, phy.all$tip.label[!phy.all$tip.label %in%
                                           c(colnames(comm), "Methanosarcina")])

# Identify Outgroup Sequence
outgroup <- match("Methanosarcina", phy$tip.label)

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

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



4) PHYLOGENETIC ALPHA DIVERSITY

A. Faith's Phylogenetic Diversity (PD)

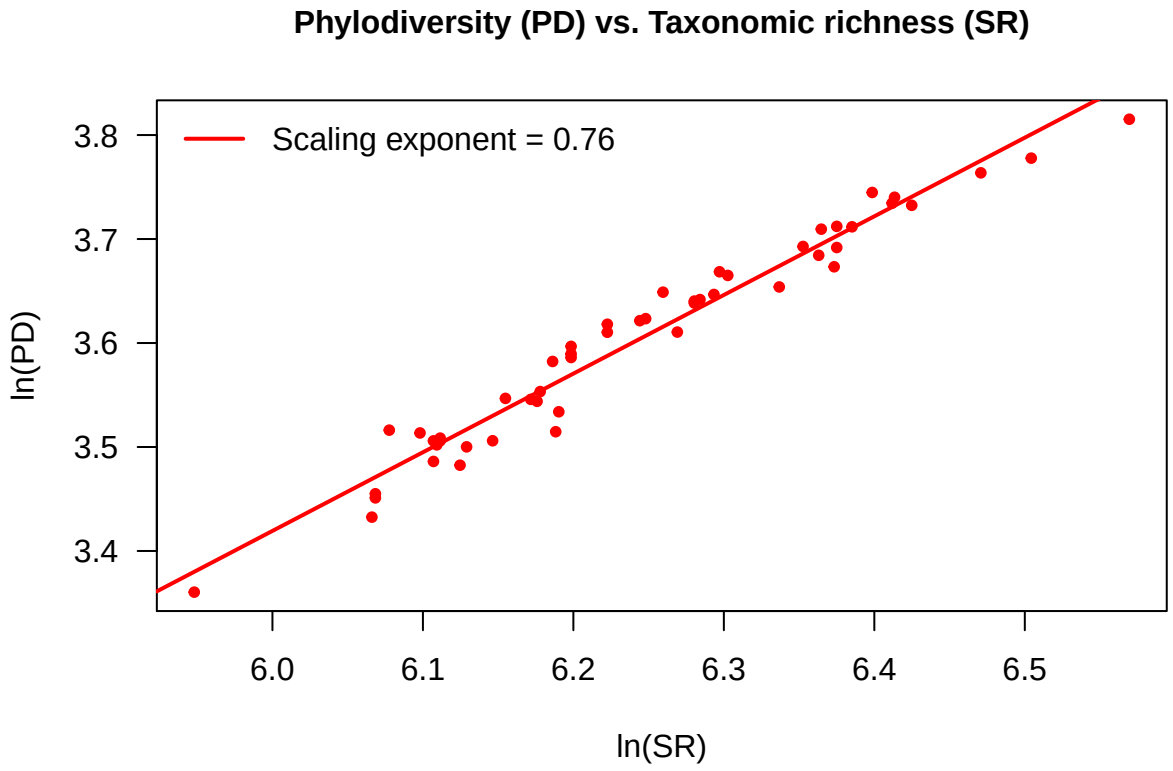
In 1992, Daniel Faith developed a diversity metric called Faith's PD (http://goo.gl/wM08Oy). The metric sums the branch lengths for each species found in a sample from the root to the tip of a the phylogenetic tree. The value of the metric captures the evolutionary history of an assemblage. Higher PD values indicate that an assemblage contains more evolutionarily divergent taxa, while lower PD values indicate that an assemblage contains taxa with a more restricted evolutionary history.

Faith's PD can be implemented in R using the `pd()` function in the `picante` package. A phylogenetic tree containing the species pool is required. In addition to returning Faith's PD, the `pd()` function also returns species richness (SR). SR is the same as observed richness ($S_{obs}), which we covered in the α diversity module.

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

Let's compare PD estimates with SR of our samples. We'll natural-log-transform our data so that the slope of the relationship gives us a power-law exponent, which describes how PD scales with SR.

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

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

fit <- lm('log(pd$PD) ~ log(pd$SR)')
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-SR pattern.
a. Describe the relationship between taxonomic richness and phylodiversity?
b. Mathematically, why should they be correlated?
c. Under what conditions would you expect these two estimates of diversity to deviate from one another?

Answer 1a:
Answer 1b:
Answer 1c:

i. Randomizations and Null Models Randomizations are a way of resampling data to assess whether or not observed patterns are different from a null expectation. A number of the functions in `picante` allow

us to specify different null models as an argument. These null models can control for features such as species richness, species occurrence frequencey, and the diversity of the regional species pool. We will use some of these models for assesing the degree to which phylogenetic measures of α diversity deviate from null expecations. The following table describes some of the null models that are available to us when using `picante`:

Null Model	Description
taxa.labels	Shuffles taxa labels across tips of phylogeny (across all taxa included in phylogeny)
richness	Randomizes community data matrix abundances within samples (maintains sample species richness)
frequency	Randomizes community data matrix abundances within species (maintains species occurence frequency)
sample.pool	Randomizes community data matrix by drawing species from pool of species occurring in at least one community (sample pool) with equal probability
phylogeny.pool	Randomize community data matrix by drawing species from pool of species occurring in at least one community (sample pool) with equal probability
independentswap	Randomizes community data matrix with the independent swap algorithm (Gotelli 2000) maintaining species occurrence frequency and sample species richness
trialswap	Randomizes community data matrix with the trial-swap algorithm (Miklos & Podani 2004) maintaining species occurrence frequency and sample species richness

Now, we are going to use the `ses.pd()` function in `picante`. This function estimates the standardized effect size ("ses") using the following equation: `ses.pd` = (`pd.obs` - `pd.rand.mean`) / `pd.rand.sd`, where `pd.obs` is the observed PD, `pd.rand.mean` is the mean of the PD values generated via randomization under a null model, and `pd.rand.sd` is the standard deviation of the PD values generated via randomization under a null model (see table above). Given the size of both our site-by-species matrix and the phylogenetic tree, the randomizaiton process is computationally intensive. Therfore, we are only going to run the `ses.pd` function for two ponds with a limited number of randomizations (i.e., "runs" argument).

```
# Estimate Standardized Effect Size of PD via Randomization {picante}
ses.pd <- ses.pd(comm[1:2,], phy, null.model = "richness", runs = 25,
                 include.root = FALSE)
```

Question 2: Using `help()` and the table above, identify two null models that can be used with the `ses.pd()` function. Run `ses.pd()` with these null models and answer the following questions:

a. What hypotheses are being tested with the p-values associated with `ses.pd`?
b. What features might affect the intepretation of the `ses.pd` output?

Answer 2a:
Answer 2b:

B. Phylogenetic Dispersion Within a Sample

Another way to assess phylogenetic α-diversity is to look at dispersion within a sample. In the following section we will introduce two commonly used metrics — the Net Relatedness Index (NRI) and the Nearest Taxon Index — to quantify the degree to which closely related taxa co-occur. We will use randomization procedures to test whether species are phylogenetically clustered or overdispersed.

i. Phylogenetic Resemblance Matrix Before estimating dispersion metrics, we need to create a phylogenetic resemblance matrix. This type of matrix is nearly identical to the resemblance matrix introduced in the β-diversity module. The only difference is that the phylogenetic resemblance matrix contains distances between taxa in a tree, whereas the community resemblance matrix contains distances among sites. The elements in phylogenetic resemblance matrix are calculated as the pairwise branch-length distances between tips (i.e., taxa) on a phylogenetic tree. The phylogenetic resemblance matrix is sometimes referred to as the phylogenetic variance-covariance matrix. We will use the `cophenetic.phylo()` function in `picante` to calculate the phylogenetic resemblance matrix.

```
# Create a Phylogenetic Distance Matrix {picante}
phydist <- cophenetic.phylo(phy)
```

ii. Net Relatedness Index (NRI) One common way to test for phylogenetic clustering and overdispersion is to use the Net Relatedness Index (NRI). NRI is based on the mean phylogenetic distance (MPD). MPD is the mean phylogenetic distance from pairwise branch lengths in a sample. With this information in hand, NRI is expressed as: - (`mpd.obs` - `mpd.rand.mean`) / `mpd.rand.sd` where `mpd.obs` is the observed MPD, `mpd.rand.mean` is the mean of the MPD values generated via randomization under a null model, and `mpd.rand.sd` is the standard deviation of the MPD values generated via randomization under a null model.

Negative NRI values indicate that a sample is phylogenetically overdispersed; that is, taxa are less related to one another than expected under the null model. Positive NRI values indicate that a sample is phylogentically underdispersed, or clustered, such that taxa are more closely related to one another than expected under the null model.

As with Faith's PD, the radomization procedures are computationally intensive, so we are only going to peform a relatively small number of "runs".

```
# 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)

Another way to test for phylogenetic clustering and oversdispersion in a sample is to use the Nearest Taxon Index (NTI). This index is mathemically similar to NRI, but uses the mean nearest phylogenetic neighbor distance (MNND) instead of MPD. MNND is the mean phylogenetic distance between all taxa in a sample and their phylogenetically closest neighbor. As a result, NTI tends to emphasize terminal clustering, independent of deep level clustering (Webb et al. 2002; http://goo.gl/WikgWE). Just like NRI, we peform randomizations and use this information to estimate the standardized effect size. Negative NTI values indicate phylogenetic overdispersion and positive NTI values indicate phylogenetic clustering.

```r
# Estimate Standardized Effect Size of NRI via Randomization {picante}
ses.mntd <- ses.mntd(comm, phydist, null.model = "taxa.labels",
                     abundance.weighted = FALSE, runs = 25)

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

Question 3: 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 caculated using abundance data. How does this affect the interpretation of NRI and NTI?

Answer 3:

C) Mini-Exercise On Phylogenetic Alpha-Diversity

Conduct an exploratory analysis of the phylogenetic α-diversity for the Indiana refuge ponds. Using techniques from the *Week1_Handout.Rmd* and any other statistical methods at your disposal, identify enviornmental variables that may influence the taxonomic richness (SR), phylodiveristy (PD), and dispersion (NRI and NTI) of the bacterial communities in the sampled ponds. Generate some hypotheses as to what ecological and/or evolutionary processes are giving rise to these patterns of phylogenetic α-diversity.

Mini-Exercise: Code, Output, and Discussion:

5) PHYLOGENETIC BETA DIVERSITY

A. Phylogenetically Based Community Resemblance Matrix

As you may recall, in order to quantify β-diversity for more than two samples, one needs to create a resemblance matrix. When quantifying taxonomic β-diversity, we calculated the pairwise **similarity** or **dissimiliarity** for all samples in a site-by-species matrix using metrics such as the Sørensen index or the Bray-Curtis index. More or less, we need to go through the same process for quantifying phylogenetic β-diversity. Instead of making the resemblance matrix based on incidence or abundance of taxa, we are going to incorporate information about the phylogenetic relationships among taxa. Similar to other measures of β-diversity, there are numerous ways to calculate the phylogenetic distances in the community resemblance matrix. Here, we will explore two: Mean Pairwise Distance and UniFrac distance.

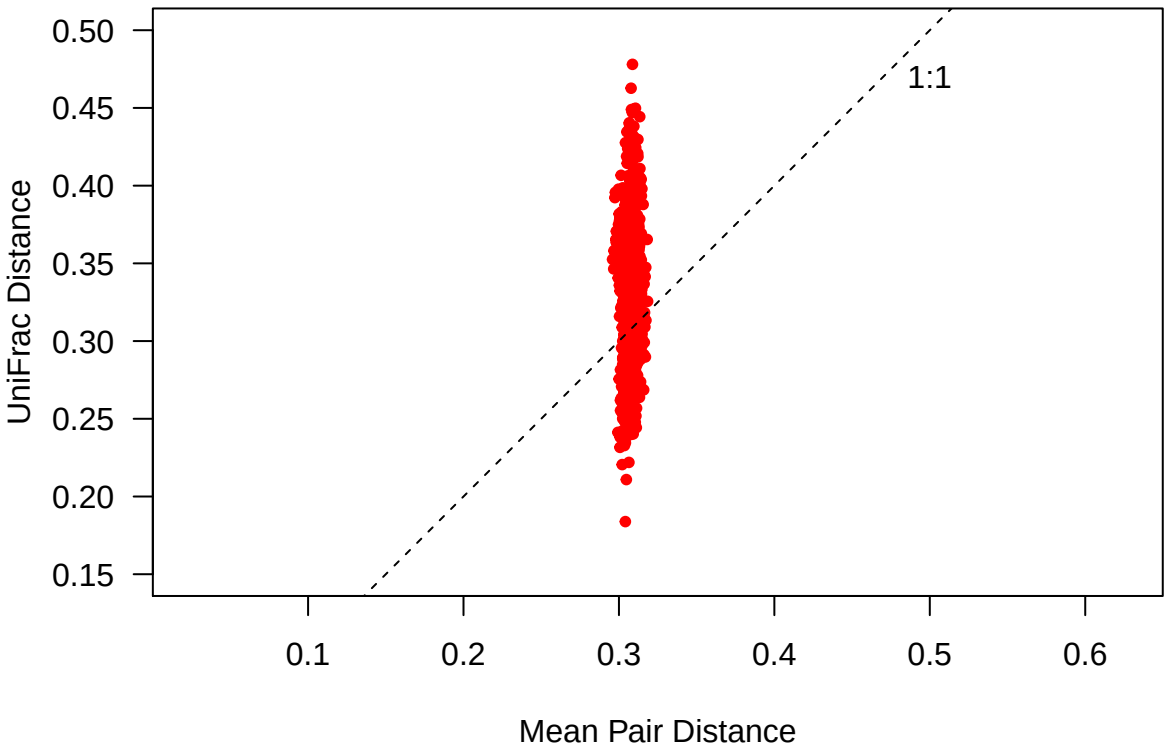
Index	Description
Mean Pairwise Distance	Distance between two samples calculated as the mean phylogenetic distance between pairs of taxa
UniFrac	Distance between two samples calculated as $\Sigma_{unshared}$ / Σ_{total}, where $\Sigma_{unshared}$ is the sum of unshared branch lengths between samples and Σ_{total} is the total (shared and unshared) branch lengths in a rooted tree

Let's calculate the phylogenetically based community resemblance matrices for our pond data set.

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

# UniFrac Distance (Note: This Takes a Few Minutes; Be Patient)
dist.uf <- unifrac(comm, phy)
```

Now, let's compare the Mean Pair Distance and UniFrac distance matrices.

```
par(mar = c(5, 5, 2, 1) + 0.1)
plot(dist.mp, dist.uf,
     pch = 20, col = "red", las = 1, asp = 1, xlim = c(0.15, 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: Using the plot above, describe the relationship between Mean Pair Distance and UniFrac distance. Note: we are calcuating unweighted phylogenetic distances (similar to incidence based measures). That means that we are not taking into account the abundance of each taxa in each site.

Answer 4:

B. Visualizing Phylogenetic Beta Diversity

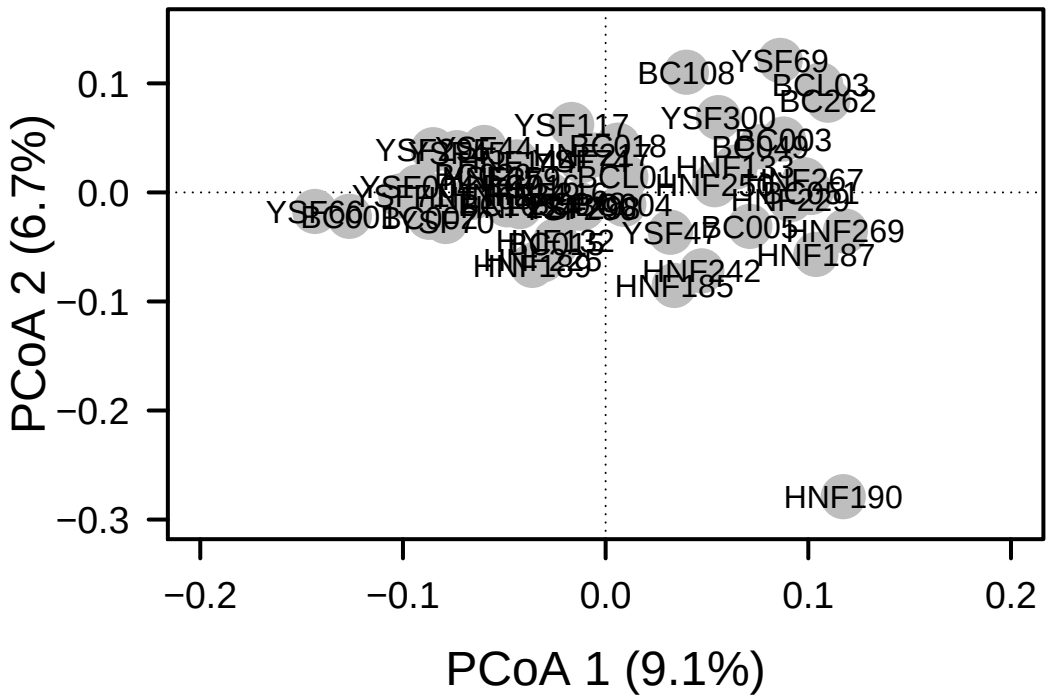
Now that we have our phylogenetically based community resemblance matrix, we can visualize phyogenetic diversity among samples using the same techniques that we used in the β-diversity module. As an example, we

will use ordination, but any of the other β-diversity viualization techniques would also work (e.g., heatmaps and cluster analysis). Specifically, we will use the `cmdscale()` function to conduct a Principal Coordinates Analysis (PCoA) using the UniFrac distance matrix. Additionally, we will calculate the amount of explained variation for each phylogenetically informed PCoA axis.

```
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. Remember, you should check the eigenvalues to determine your confidence in the data reduction approach.

```
# Define Plot Parameters
par(mar = c(5, 5, 1, 2) + 0.1)

# Initiate Plot
plot(pond.pcoa$points[ ,1], pond.pcoa$points[ ,2],
     xlim = c(-0.2, 0.2), ylim = c(-0.3, 0.15),
     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))
```



Question 5: How much variation is explained in the first three PCoA axes if Mean Pair Distance is used as the resemblance matrix instead of UniFrac?

Answer 5:

C. Hypothesis Testing

i. Categorical Approach: Watershed Effect The ponds that we sampled in southern Indiana were located in three distinct watersheds (BCSP, HNF, and YSF). For many organisms, watershed boundaries represent major dispersal barriers, which may influence the phylogenetic distribution of species. Here, we will test for this watershed affect using the Permutational Multivariate Analysis of Variance (PERMANOVA) test that we learned about in the β-diversity module.

```
# Define Environmental Category
water.shed <- env$Location

# Run PERMANOVA with `adonis()` Function {vegan}
adonis(dist.uf ~ water.shed, permutations = 999)
```

ii. Continuous Approach: Environmental Gradients In the Geographic Ecology module, we demonstrated that there was substantial variation in enviornmental variables that are known to influence the structure and function of microbial communities. In the following section, we will revist two methods that are used to test for relationships among multivariate environmental and biological data.

First, let's define the environmental data and then create an enviornmental distance matrix:

```r
# Define Environmental Variables
envs <- env[, 5:19]

# Distance Matrix for Environmental Variables
env.dist <- vegdist(scale(envs), method = "euclid")
```

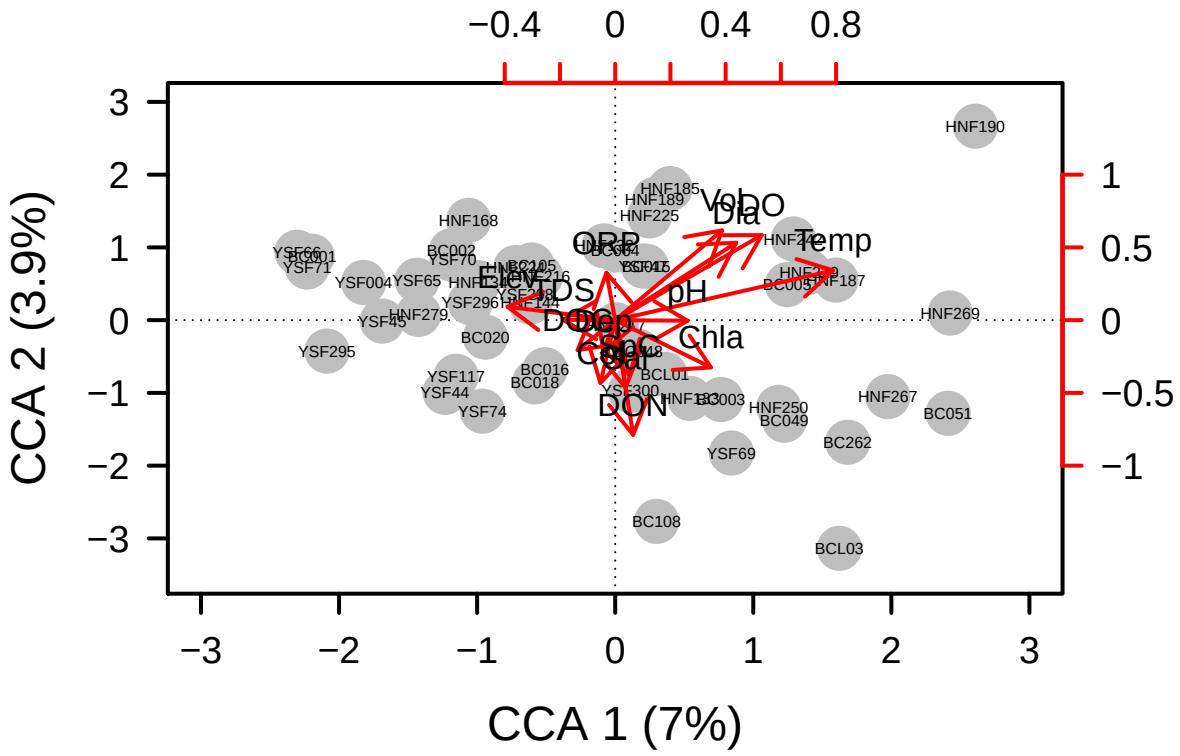
Second, let's conduct a Mantel's test to see whether UniFrac distance is correlated with environmental variation:

```r
# Conduct Mantel Test {vegan}
mantel(dist.uf, env.dist)
```

Last, let's conduct a Canonical Correspondence Analysis (CCA). You will recall that this constrained ordination technique allows one to test for the effects of an explanatory matrix (e.g., environmental data) on a response matrix (e.g., phylogenetic distance matrix).

```r
# Conduct CCA {vegan}
ponds.cca <- vegan::cca(dist.uf ~ scale(envs))

# Permutation Tests: Axes and Env Variables
anova(ponds.cca, by = "axis")
ponds.fit <- envfit(ponds.cca, envs, perm = 999)
ponds.fit

# Calculate Explained Variation
cca.explainvar1 <- round(ponds.cca$CCA$eig[1] /
                         sum(c(ponds.cca$CCA$eig, ponds.cca$CA$eig)), 3) * 100
cca.explainvar2 <- round(ponds.cca$CCA$eig[2] /
                         sum(c(ponds.cca$CCA$eig, ponds.cca$CA$eig)), 3) * 100

# Define Plot Parameters
par(mar = c(5, 5, 4, 4) + 0.1)

# Initiate Plot
plot(scores(ponds.cca, display = "wa"), xlim = c(-3, 3), ylim = c(-3.5, 3),
     xlab = paste("CCA 1 (", cca.explainvar1, "%)", sep = ""),
     ylab = paste("CCA 2 (", cca.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.cca, display = "wa"),
       pch = 19, cex = 3, bg = "gray", col = "gray")
text(scores(ponds.cca, display = "wa"),
    labels = row.names(scores(ponds.cca, display = "wa")), cex = 0.5)

# Add Environmental Vectors
vectors <- scores(ponds.cca, display = "bp")
```

```
row.names(vectors) <- c("Elev", "Dia", "Dep", "Vol", "ORP", "Temp", "SpC", "DO",
                        "TDS", "Sal", "pH", "Col", "Chla", "DOC", "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 procecures conducted above, describe some of the phylogenetic patterns of β-diversity for bacterial communities in the Indiana refuge ponds.

Answer 6:

D) Mini-Exercise On Phylogenetic Beta-Diversity

Generate and test hypotheses about the phylogenetic β-diversity of the Indiana refuge ponds. Using the environemtanal variables identified in the α-diversity mini-exercise, redo either the PERMANOVA or CCA analysis above using just these variables. Feel free to add any additional approaches we used in *Beta_Handout.Rmd*. Compare your results to the findings from above and discuss any differences.

Mini-Exercise Discussion:

6) GEOGRAPHICAL PHYLOGENETIC COMMUNITY ECOLOGY

A. Phylogenetic Distance-Decay (PDD)

You will recall from the Geographical Ecology module, that the distance decay (DD) relationship reflects the spatial autocorrelation of community similarity. That is, geographically near communities should be more similar than geographically distant communities. Here, we will test to what degree spatial autocorrelation can also affect phylogenetic DD.

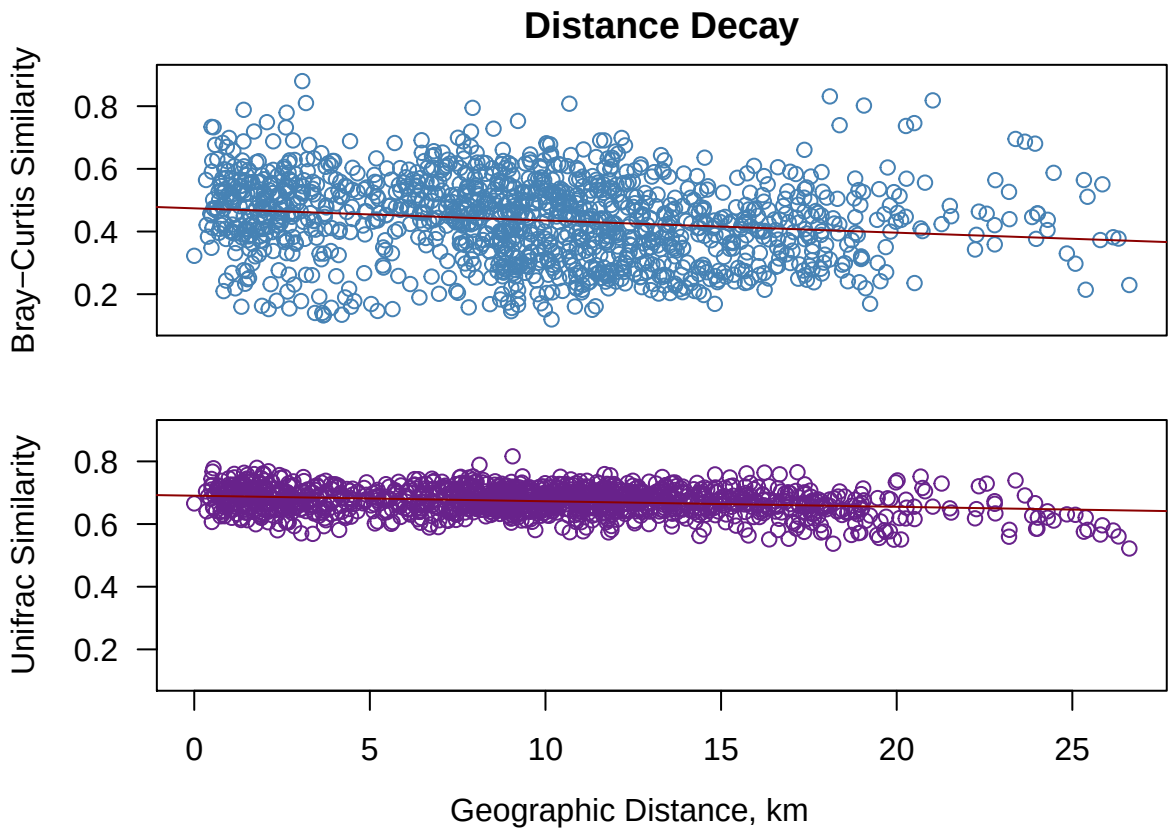
First, we will calcualte distances for geographic data, taxonomic data, and phylogenetic data among all unique pair-wise combinations of ponds.

```
# Geographic Distances (Kilometers) Among Ponds
long.lat <- as.matrix(cbind(env$long, env$lat))
coord.dist <- earth.dist(long.lat, dist = TRUE)

# Taxonomic Distances Among Ponds (Bray-Curits)
bray.curtis.dist <- 1 - vegdist(comm)

# Phylogenetic Distances Among Ponds
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")

# Create a Data Frame from the Lists of Distances
df <- data.frame(coord.dist.ls, bray.curtis.dist.ls[, 3], unifrac.dist.ls[, 3])
names(df)[4:5] <- c("bray.curtis", "unifrac")
attach(df)
```

Now, let's plot the DD relationships:

```
# Set Initial Plot Parameters
par(mfrow=c(2, 1), mar = c(1, 5, 2, 1) + 0.1, oma = c(2, 0, 0, 0))

# Make Plot for Taxonomic DD
plot(coord.dist, bray.curtis, xlab = "", xaxt = "n", las = 1, ylim = c(0.1, 0.9),
     ylab="Bray-Curtis Similarity",
     main = "Distance Decay", col = "SteelBlue")

# Regression for Taxonomic DD
DD.reg.bc <- lm(bray.curtis ~ geo.dist)
summary(DD.reg.bc)
abline(DD.reg.bc , col = "red4")

# New Plot Parameters
par(mar = c(2, 5, 1, 1) + 0.1)

# Make Plot for Phylogenetic DD
plot(coord.dist, unifrac, xlab = "", las = 1, ylim = c(0.1, 0.9),
     ylab = "Unifrac Similarity", col = "darkorchid4")
```

```r
# Regression for Phylogenetic DD
DD.reg.uni <- lm(unifrac ~ coord.dist)
summary(DD.reg.uni)
abline(DD.reg.uni, col = "red4")

# Add X-Axis Label to Plot
mtext("Geographic Distance, km", side = 1, adj = 0.55,
      line = 0.5, outer = TRUE)
```



Finally, let's test whether the slopes for taxonomic and phylogenetic DD are significantly different from one another.

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

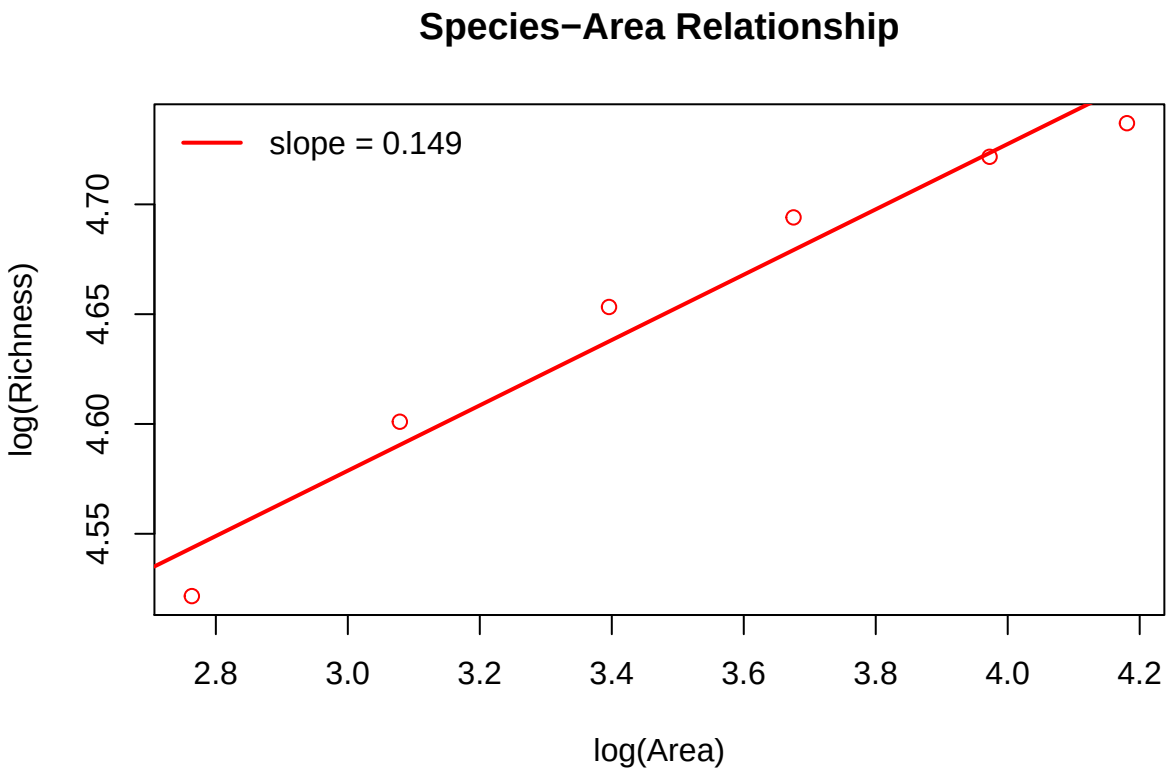
Question 7: Interpret the slopes from the taxonomic and phylogenetic DD relationships. If there are differences, hypothesize why this might be.

Answer 7:

A. Phlogenetic diversity-area relationship (PDAR)

Helmus and Ives (2012) developed methods to study how phylogenetic diversity changes with increasing area. This phylogenetic diversity-area relationship (PDAR) is analogous to the species-area relationship (SAR) that we learned about in the Geographical Ecology module. You will recall that the SAR is a cumulative

relationship (S = cA^z) and so, cannot be negative. In fact, the shallowest the SAR can get is to have a slope of 0.0, meaning that all species are found in all samples. Below is a plot of the taxonomically based SAR for the pond data.



In contrast to SAR, phylogenetic divesity-area relationships can increase or decrease with area. This provoked Helmus and Ives (2012) to develop theoretical expectations for the effects that ecological and evolutionary processes might have on the shape of the PDAR.

i. Constructing the PDAR Helmus and Ives (2012) used the phylogenetic species variability (PSV) metric to quantify phylogenetic diversity. PSV quantifies how phylogenetic relatedness decreases the variance of a hypothetical neutral trait shared by all species in a community. Type 'help(psv)' to learn more.

In this exercise, we will construct PDARs using the random aggregation approach of Helmus and Ives (2012). This approach is similar to how we constructed SARs in the Geographical Ecology module.

Let's begin by writing a function to generate the PDAR.

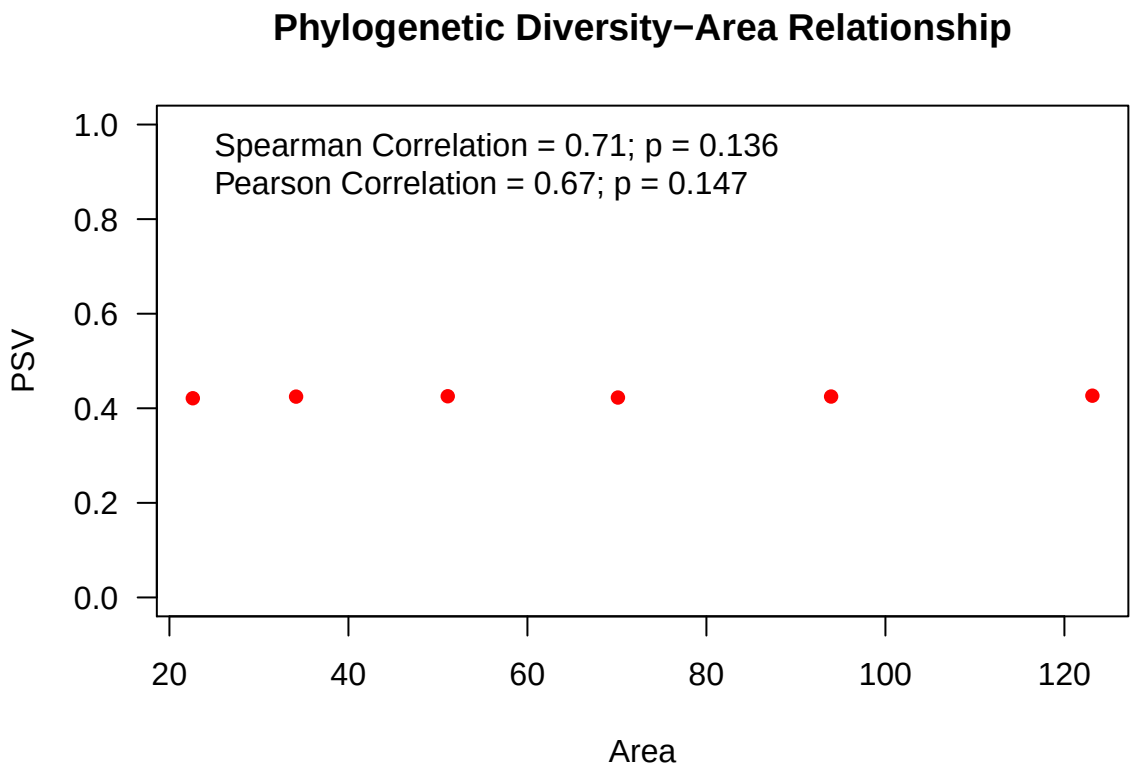
```
PDAR <- function(comm, tree){

  # Create Objects to Hold Areas and Diversity
  areas <- c()
  diversity <- c()

  # Create Vector Increasing Number of Plots by 2x
  num.plots <- c(2, 4, 8, 16, 32, 51)

  for (i in num.plots){
```

```
    # Create vectors to hold areas and diversity form iterations, used for means
    areas.iter <- c()
    diversity.iter <- c()

    # Iterate 10 Times Per Sample Size
    for (j in 1:10){
      # Sample w/o replacement
      pond.sample <- sample(51, replace = 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 rea 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 We will examine the relationship between phylogenetic diversity and area using both Spearman's correlation coefficient (S) and Pearson's correlation coefficient (P). It is informative to use both because while S is computed on ranks and depicts monotonic relationships (the degree to which the relationship is continually increasing or decreasing), P is computed on the observed values and therefore depicts linear relationships.

```
# Calculate Areas for Ponds: Find Areas of All 51 ponds
pond.areas <- as.vector(pi * (env$Diameter/2)^2)

# Compute the PDAR
pdar <- PDAR(comm, phy)
pdar <- as.data.frame(pdar)
pdar$areas <- sqrt(pdar$areas)

# Calculate Perason's Correlation Coefficient
Pearson <- cor.test(pdar$areas, pdar$diversity, method = "pearson")
P <- round(Pearson$estimate, 2)
Pp <- round(Pearson$p.value, 3)
```

```r
# Calculate Spearman's Correlation Coefficient
Spearman <- cor.test(pdar$areas, pdar$diversity, method = "spearman")
S <- round(Spearman$estimate, 2)
Sp <- round(Spearman$p.value, 3)

# Plot the PDAR
par(mar = c(5, 5, 4, 2) + 0.1)
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 = ", S, "; p = ", Sp, sep = ""),
                            paste("Pearson Correlation = ", P, "; p = ", Pp, sep = "")),
                            bty = "n", col = "red")
```



Question 8: For the bacteria in the Indiana ponds, the slope (z) of the SAR was 0.14. This is slightly lower than z-values observed for many macroscopic organisms (e.g., fish, birds, plants), but is higher than what has been reported for other microbial systems. However, what did we observe for the microbial PDAR in the Indiana pond? How might we explain the differences between the taxonomic (SAR) and phylogenetic (PDAR)?

Answer 8:

7) HOMEWORK

1. In their study of phylogenetic diversity-area relationships (PDARs), Helmus and Ives, increased area by aggregating plots in two ways. First, they aggregated plots with respect to whether the plots were adjacent; what they called 'spatial'. Second, they aggregated plots at random ('non-spatial'). While both spatial and non-spatial sampling methods capture the effect of increasing area, sampling with respect to location (i.e., spatial) also captures the effect of increasing distance or spatial autocorrelation. Explain the general difference that Helmus and Ives (2012) observed between spatial and non-spatial PDARs, and why this difference should be expected in both the species-area relationship (SAR) and the phylogenetic diversity-area relationship.

2. Use Knitr to create a pdf of your completed PhyloTraits_handout.Rmd document, push it to GitHub, and create a pull request. The due date for this assignment is March 4, 2015 at 12:00 PM (noon).