12.Phylogenetic Diversity - Communities

Mark Hibbins; Z620: Quantitative Biodiversity, Indiana University

25 February, 2019

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 worksheet, 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. In the Markdown version of this document in your cloned repo, change "Student Name" on line 3 (above) with your name.
2. Complete as much of the worksheet as possible during class.
3. Use the handout as a guide; it contains a more complete description of data sets along with examples of proper scripting needed to carry out the exercises.
4. Answer questions in the worksheet. Space for your answers is provided in this document and is indicated by the ">" character. If you need a second paragraph be sure to start the first line with ">". You should notice that the answer is highlighted in green by RStudio (color may vary if you changed the editor theme).
5. Before you leave the classroom today, it is *imperative* that you **push** this file to your GitHub repo, at whatever stage you are. This will enable you to pull your work onto your own computer.
6. When you have completed the worksheet, **Knit** the text and code into a single PDF file by pressing the `Knit` button in the RStudio scripting panel. This will save the PDF output in your '8.BetaDiversity' folder.
7. After Knitting, please submit the worksheet by making a **push** to your GitHub repo and then create a **pull request** via GitHub. Your pull request should include this file *12.PhyloCom_Worksheet.Rmd* and the PDF output of `Knitr` (*12.PhyloCom_Worksheet.pdf*).

1) SETUP

Typically, the first thing you will do in either an R script or an RMarkdown file is setup your environment. This includes things such as setting the working directory and loading any packages that you will need.

In the R code chunk below, provide the code to:
1. clear your R environment,
2. print your current working directory,
3. set your working directory to your `/Week7-PhyloCom` folder,
4. load all of the required R packages (be sure to install if needed), and
5. load the required R source file.

```
rm(list = ls())
getwd()
```

```
## [1] "/Users/mark/Box Sync/Courses/Quantitative Biodiversity/QB2019_Hibbins/2.Worksheets/12.PhyloCom"
```

```
setwd('/Users/mark/Box Sync/Courses/Quantitative Biodiversity/QB2019_Hibbins/2.Worksheets/12.PhyloCom')

package.list <- c('picante', 'ape', 'seqinr', 'vegan',
                  'fossil', 'reshape', '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 2.5-3
```

```
##
## 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')
```

2) DESCRIPTION OF DATA

need to discuss data set from spatial ecology!

In 2013 we sampled > 50 forested ponds in Brown County State Park, Yellowwood 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., the DNA sequence) and 16S rRNA transcripts (i.e., the RNA transcript of the gene) of bacteria. We used a program called `mothur` to quality-trim our data set and assign sequences to operational taxonomic units (OTUs), which resulted in a site-by-OTU matrix.
In this module we will focus on taxa that were present (i.e., DNA), but there will be a few steps where we need to parse out the transcript (i.e., RNA) samples. See the handout for a further description of this week's dataset.

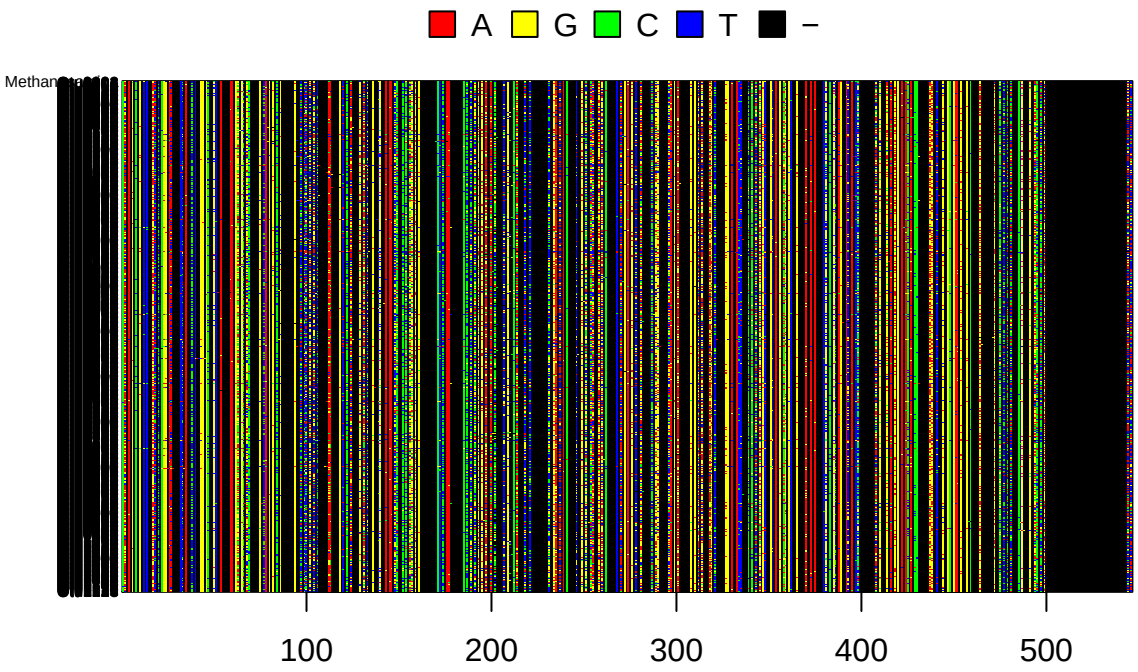
3) LOAD THE DATA

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.

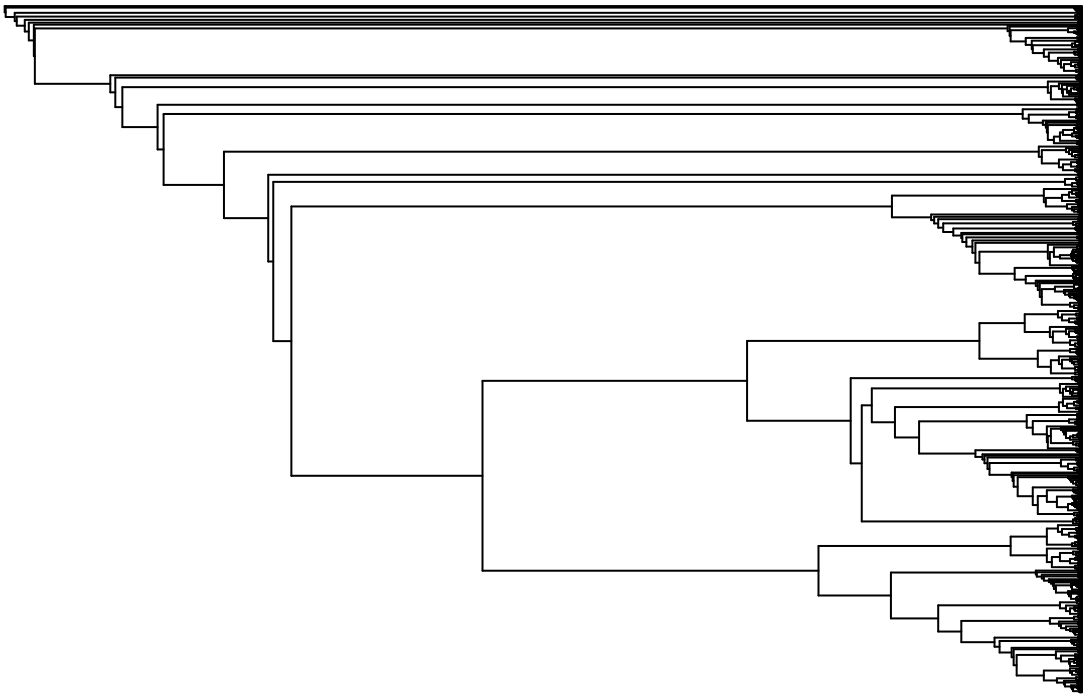
```
env <- na.omit(read.table('data/20130801_PondDataMod.csv',
                          sep = ',', header = TRUE))

comm <- read.otu(shared = './data/INPonds.final.rdp.shared',
                 cutoff = '1')
comm <- comm[grep('*-DNA', rownames(comm)), ]
rownames(comm) <- gsub('\\-DNA', '', rownames(comm))
rownames(comm) <- gsub('\\_', '', rownames(comm))
comm <- comm[rownames(comm) %in% env$Sample_ID, ]
comm <- comm[ , colSums(comm) > 0]

tax <- read.tax(taxonomy = './data/INPonds.final.rdp.1.cons.taxonomy')
```

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

```
ponds.cons <- read.alignment(file = './data/INPonds.final.rdp.1.rep.fasta',
                             format = 'fasta')
ponds.cons$nam <- gsub('\\|.*$', '', gsub('^.*?\t', '', ponds.cons$nam))

outgroup <- read.alignment(file = './data/methanosarcina.fasta',
                           format = 'fasta')
DNAbin <- rbind(as.DNAbin(outgroup), as.DNAbin(ponds.cons))

image.DNAbin(DNAbin, show.labels = T, cex.lab = 0.5, las = 1)
```




```
seq.dist.t92 <- dist.dna(DNAbin, model = 'T92', pairwise.deletion = FALSE)

phy.all <- bionj(seq.dist.t92)

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 = 'Neighbour Joining Tree', 'phylogram', show.tip.label = FALSE,
           use.edge.length = FALSE, direction = 'right', cex = 0.6, label.offset = 1)
```



4) PHYLOGENETIC ALPHA DIVERSITY

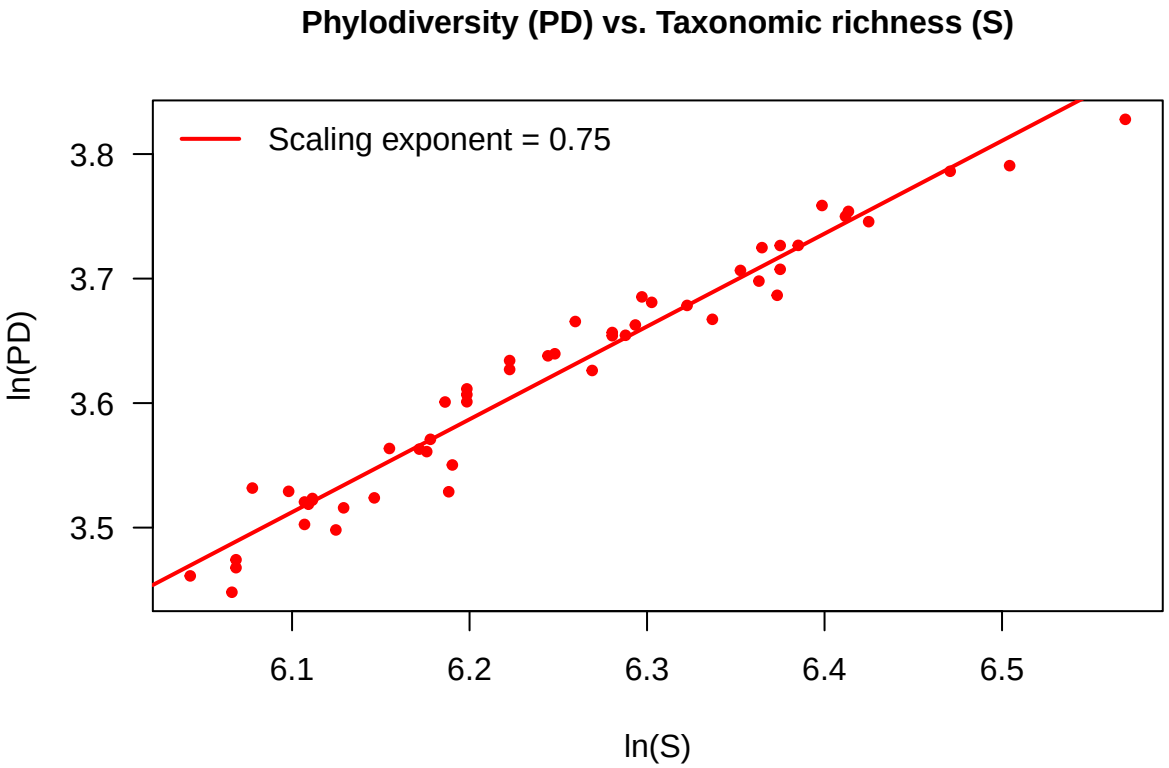
A. Faith's Phylogenetic Diversity (PD)

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

```
pd <- pd(comm, phy, include.root = FALSE)
```

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

```
par(mar = c(5, 5, 4, 1) + 0.1)

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

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***: Faith's PD sums the branch lengths leading to each species in the community from root to tip. Since sampling new taxa naturally adds more tips to the tree, it also increases the number of branches and therefore the sum of branch lengths. Additionally, as you sample more species, it is increasingly likely that you will sample taxa more distantly related to the ones already in your data set. ***Answer 1b***: Taxonomic richness and phylodiversity are tightly correlated, and appear to follow a 3/4 power law (translating to a slope of 3/4 on the log scale). ***Answer 1c***: They would deviate if there isn't a tight correspondence between taxonomic classification and evolutionary distance. For example, two differently labelled taxa could actually be very closely related, or two individuals labelled as the same taxon could actually be very distant. ***Answer 1d***: This scaling coefficient implies that there is not a 1:1 relationship between species richness and phylodiversity. This is because while different species are given equal weight in terms of taxonomic classification, there may be a lot of variation in how diverged they are from each other in a phylogeny.

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_rich <- ses.pd(comm[1:2,], phy, null.model = 'richness',
                  runs = 25, include.root = FALSE)
ses_pd_taxa <- ses.pd(comm[1:2,], phy, null.model = 'taxa.labels',
                  runs = 25, include.root = FALSE)
ses_pd_swap <- ses.pd(comm[1:2,], phy, null.model = 'independentswap',
                  runs = 25, include.root = FALSE)
```

```
ses_pd_rich
```

```
##       ntaxa   pd.obs pd.rand.mean pd.rand.sd pd.obs.rank   pd.obs.z
## BC001   668 44.28691     44.49636  0.9155984          14 -0.2287617
## BC002   587 41.53580     40.42588  0.8080104          24  1.3736525
##        pd.obs.p runs
## BC001 0.5384615   25
## BC002 0.9230769   25
```

```
ses_pd_taxa
```

```
##       ntaxa   pd.obs pd.rand.mean pd.rand.sd pd.obs.rank   pd.obs.z
## BC001   668 44.28691     44.64924  0.7975138          10 -0.4543260
## BC002   587 41.53580     40.60794  1.0203216          21  0.9093767
##        pd.obs.p runs
## BC001 0.3846154   25
## BC002 0.8076923   25
```

```
ses_pd_swap
```

```
##       ntaxa   pd.obs pd.rand.mean pd.rand.sd pd.obs.rank  pd.obs.z
## BC001   668 44.28691     45.07812  0.5287625           4 -1.496352
## BC002   587 41.53580     40.63179  0.5712078          23  1.582634
##        pd.obs.p runs
## BC001 0.1538462   25
## BC002 0.8846154   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 randomization approach generates a null distribution of phylodiversity values. If our data is consistent with the specified null model, then we should expect to see no significant difference between the randomized phylodiversity value and the one that was observed. ***Answer 2b***: The p-value for observed phylodiversity is not significant for any of the null models, so practically speaking, there isn't much of an effect. The mean randomized phylodiversity values are almost identical for all three models, but there are some interesting differences in the standard deviation, being lower for the independentswap null model than the other two.

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.

```r
ses_mpd <- ses.mpd(comm, phydist, null.model = 'taxa.labels',
                   abundance.weighted = TRUE, runs = 25)
NRI <- as.matrix(-1 * ((ses_mpd[,2] - ses_mpd[,3]) / ses_mpd[,4]))
rownames(NRI) <- row.names(ses.mpd)
colnames(NRI) <- 'NRI'
NRI
```

```
##                NRI
##  [1,]  0.320090248
##  [2,]  0.472890533
##  [3,]  0.628561524
##  [4,] -0.052864356
##  [5,]  0.605410473
##  [6,]  0.398661558
##  [7,]  0.174237493
##  [8,]  0.574344536
##  [9,]  0.326621709
## [10,]  0.345447379
## [11,] -0.006226352
## [12,]  0.366983182
## [13,] -0.487652519
## [14,] -0.364967016
## [15,] -0.042399131
## [16,] -0.384648125
## [17,] -0.310197470
## [18,] -0.301921824
## [19,] -0.026969864
## [20,]  0.160480405
## [21,]  0.396153590
## [22,] -0.122595199
## [23,]  0.078041491
## [24,]  0.329047396
## [25,]  0.545366970
## [26,]  0.514063786
## [27,]  0.108316242
## [28,] -0.338993584
## [29,]  0.253091739
## [30,]  0.485840494
## [31,] -0.354587530
## [32,] -0.022383081
## [33,] -0.337566489
## [34,] -0.458084176
## [35,] -0.339194388
## [36,] -0.498188147
## [37,]  0.609818957
## [38,] -0.671124646
## [39,]  1.010196684
## [40,]  0.673013862
## [41,]  0.361438998
## [42,]  0.499359395
## [43,]  0.644617574
## [44,]  1.335929722
## [45,]  0.356382916
## [46,]  0.581722878
```

```
## [47,]  0.138272826
## [48,]  0.135911178
## [49,]  0.177483653
## [50,] -0.283289857
## [51,]  0.527874672
## [52,]  0.812227674
```

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

```
##                NTI
## BC001   1.28516155
## BC002   1.94473366
## BC003   1.52603600
## BC004   1.09819559
## BC005   1.93028595
## BC010   0.38457464
## BC015   1.12296187
## BC016   1.76461694
## BC018   1.49473051
## BC020   1.48381946
## BC048   1.38617333
## BC049   1.81536613
## BC051   1.73299414
## BC105   1.72406528
## BC108   1.31645321
## BC262   1.00028342
## BCL01   1.38858807
## BCL03   0.84822938
## HNF132  1.26002642
## HNF133  1.75337693
## HNF134  1.41143147
## HNF144  0.95775617
## HNF168  0.71973886
## HNF185  1.34699432
## HNF187  0.44579222
## HNF216  0.09974943
## HNF217  0.10522870
## HNF221  0.70539748
## HNF224  1.13961207
## HNF225  0.14672692
## HNF229  1.40153657
## HNF242  1.64308203
## HNF250  1.20291709
## HNF267  0.64635513
## HNF269  0.77971418
## YSF004  0.48168758
## YSF117  1.81511194
```

```
## YSF295 -0.66766358
## YSF296  1.48865308
## YSF298  2.27447170
## YSF300  2.27471756
## YSF44   1.41861845
## YSF45   1.73966723
## YSF46   2.14610483
## YSF47   1.09321286
## YSF65   1.99885973
## YSF66   1.35481779
## YSF67   1.15254525
## YSF69   1.26029953
## YSF70   0.81414530
## YSF71   1.70933912
## YSF74   1.92923801
```

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 uses the mean phylogenetic distance (MPD), which is simply the mean of all the pairwise phylogenetic distances. First, a null distribution of the MPD is obtained using randomization. Then, the NRI is calculated simply as the difference between the observed and randomized values, relative to the standard deviation. ***Answer 3b***: The NTI is calculated in the same way, except that it uses the mean nearest phylogenetic neighbour distance instead of MPD. This value is the mean of the distances between each taxon and their nearest neighbour, rather than the overall distance between each pair of taxa. ***Answer 3c***: The results for NRI are all negative, indicating that every site in the Indiana ponds dataset is phylogenetically overdispersed. For NTI, most of the values are negative, but there are more positive values, suggesting there is less overall overdispersion when the tips of the tree are emphasized. ***Answer 3d***: Using the abundances as weights appears to have shifted the results more towards clustering. For NRI, there no longer appears to be a consistent signal, with many negative and positive values. For NTI, there is now a stronger signal towards clustering, as almost all the values are positive.

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.

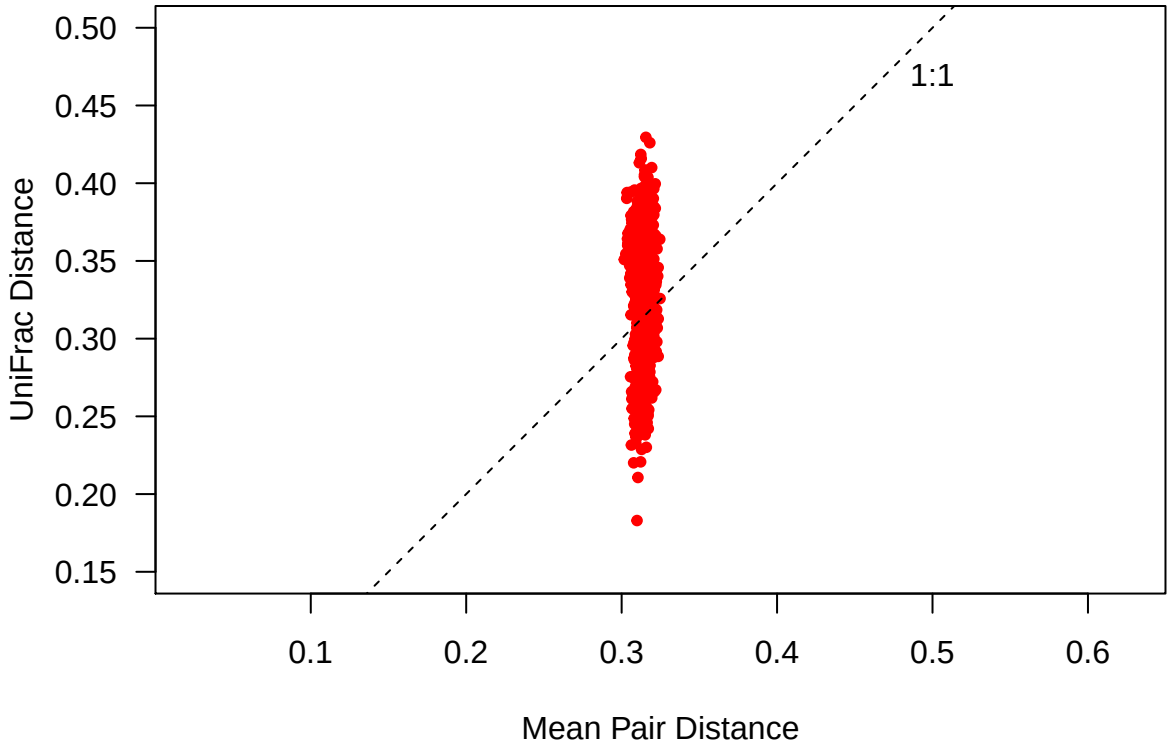
```
dist.mp <- comdist(comm, phydist)
```

```
## [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 Pair Distance is the expected distance between two taxa that are drawn from different communities. UniFrac distance, on the other hand, is the mean proportion of phylogenetic divergence contained in branches that are not shared by pairs of sites. Mean pair distance essentially measures how variable the branch lengths are between sites, whereas UniFrac distance measures how similar the tree topologies are. ***Answer 4b***: Mean pair distance has very little variance, being centered around a value of approximately 0.31. Therefore there is not much of a relationship between it and UniFrac distance. ***Answer 4c***: While both measures are fundamentally incidence-based, UniFrac is more sensitive to the presence or absence of specific taxa, because that affects the number of branches that are shared between sites. Mean pair distance doesn't really "care" about the exact topology of the tree, because it just sums across the branches that are there.

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.

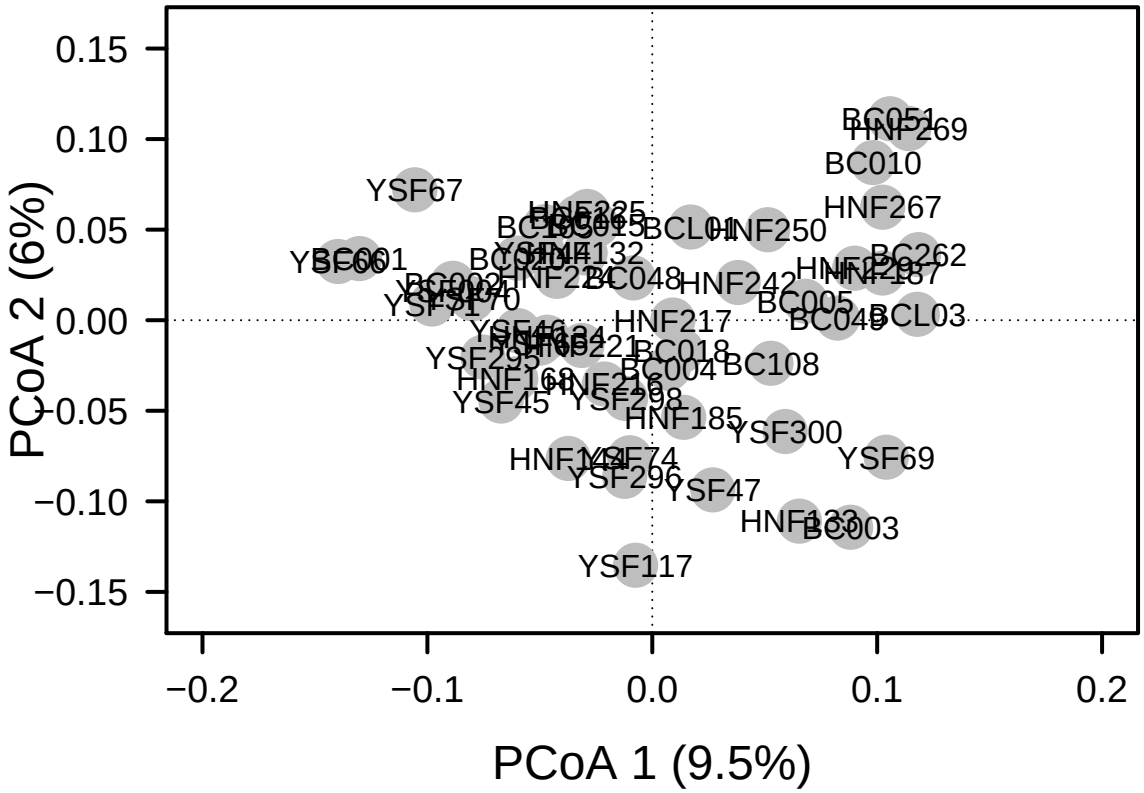
In the R code chunk below, do the following:
1. perform a PCoA based on the UniFrac distances, and
2. calculate the explained variation for the first three PCoA axes.

```
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.

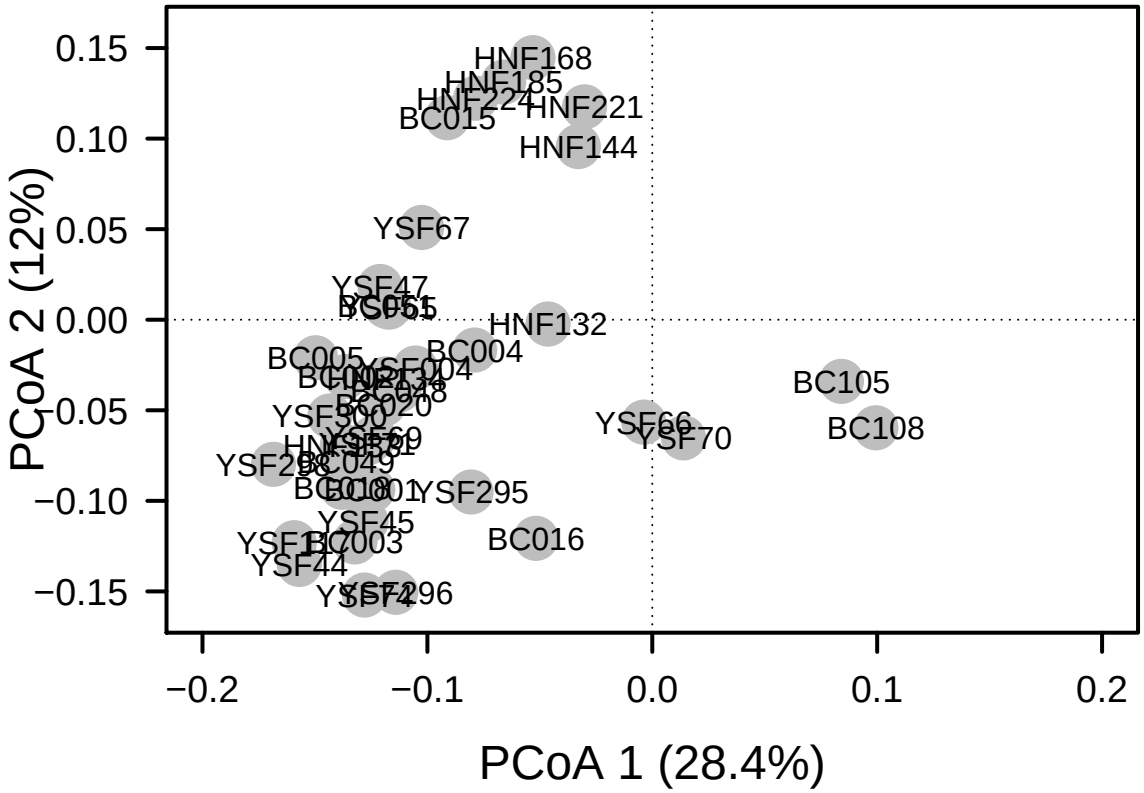
```
par(mar = c(5, 5, 1, 2) + 0.1)

plot(pond.pcoa$points[ ,1], pond.pcoa$points[ ,2],
     xlim = c(-0.2, 0.2), ylim = c(-0.16, 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.db <- vegdist(comm, method = 'bray')
pond.pcoa_tax <- cmdscale(pond.db, eig = T, k = 3)

explainvar1_tax <- round(pond.pcoa_tax$eig[1] / sum(pond.pcoa_tax$eig), 3) * 100
explainvar2_tax <- round(pond.pcoa_tax$eig[2] / sum(pond.pcoa_tax$eig), 3) * 100
explainvar3_tax <- round(pond.pcoa_tax$eig[3] / sum(pond.pcoa_tax$eig), 3) * 100

sum.eig_tax <- sum(explainvar1_tax, explainvar2_tax, explainvar3_tax)

par(mar = c(5, 5, 1, 2) + 0.1)

plot(pond.pcoa_tax$points[ ,1], pond.pcoa_tax$points[ ,2],
     xlim = c(-0.2, 0.2), ylim = c(-0.16, 0.16),
     xlab = paste('PCoA 1 (', explainvar1_tax, '%)', sep = ''),
     ylab = paste('PCoA 2 (', explainvar2_tax, '%)', 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_tax$points[ ,1], pond.pcoa_tax$points[ ,2],
       pch = 19, cex = 3, bg = 'gray', col = 'gray')
text(pond.pcoa_tax$points[ ,1], pond.pcoa_tax$points[ ,2],
```

```
    labels = row.names(pond.pcoa_tax$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***: When a phylogenetic distance matrix is used, the first two axes explain less overall variation and there is far less visual clustering of the data. This suggests that phylogeny is not as important for driving community composition in this particular 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   0.13163 0.065813  1.2676 0.04919  0.026 *
## Residuals 49   2.54398 0.051918         0.95081
## Total     51   2.67560                 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]
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: 0.1577
##       Significance: 0.069
##
## Upper quantiles of permutations (null model):
##   90%   95% 97.5%   99%
## 0.137 0.176 0.204 0.239
## 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
## Forward 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  0.10503 2.0260  0.449
```

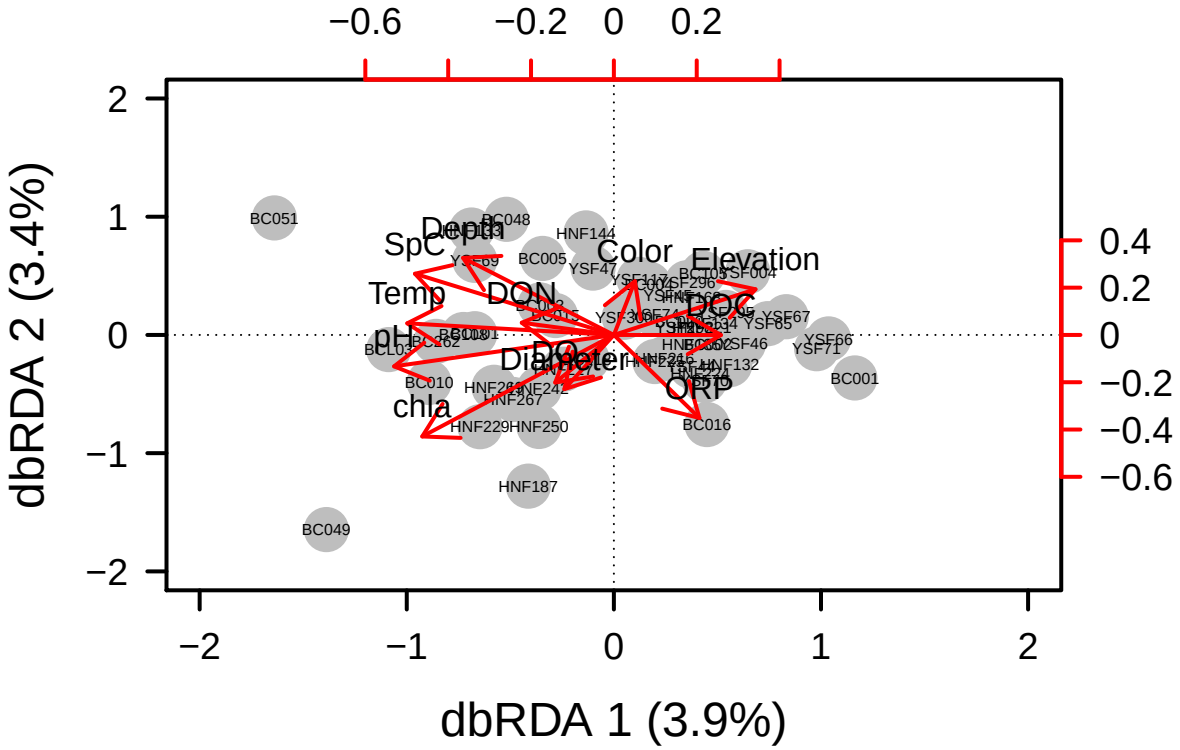
```
## dbRDA2    1  0.09197 1.7741  0.639
## dbRDA3    1  0.07402 1.4280  0.967
## dbRDA4    1  0.06687 1.2899  0.996
## dbRDA5    1  0.05674 1.0945  1.000
## dbRDA6    1  0.05171 0.9976  1.000
## dbRDA7    1  0.04671 0.9010  1.000
## dbRDA8    1  0.03867 0.7459  1.000
## dbRDA9    1  0.03717 0.7170  1.000
## dbRDA10   1  0.03204 0.6180  1.000
## dbRDA11   1  0.02842 0.5482  1.000
## dbRDA12   1  0.02454 0.4734  1.000
## Residual 39  2.02173
```

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

```
##
## ***VECTORS
##
##             dbRDA1   dbRDA2     r2 Pr(>r)
## Elevation  0.79872  0.60171 0.0946  0.094 .
## Diameter  -0.30762 -0.95151 0.0500  0.282
## Depth     -0.62323  0.78204 0.1787  0.005 **
## ORP        0.41540 -0.90964 0.1356  0.024 *
## Temp      -0.97762  0.21040 0.1520  0.020 *
## SpC       -0.77882  0.62725 0.2046  0.003 **
## DO        -0.41860 -0.90817 0.0433  0.314
## pH        -0.96265 -0.27076 0.1754  0.009 **
## Color      0.10366  0.99461 0.0437  0.341
## chla      -0.60121 -0.79909 0.2634  0.010 **
## DOC        0.99682 -0.07973 0.0397  0.404
## DON       -0.92894  0.37022 0.0327  0.445
## ---
## 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) * 100
dbrda.explainvar2 <- round(ponds.dbrda$CCA$eig[2] /
                           sum(c(ponds.dbrda$CCA$eig, ponds.dbrda$CA$eig)), 3) * 100

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***: The above results suggest that phylogenetic diversity for these communities is at least somewhat affected by environmental variables. There is no significant clustering by environment, but this is probably because there are many continuously varying variables across sites.

6) SPATIAL PHYLOGENETIC COMMUNITY ECOLOGY

A. Phylogenetic Distance-Decay (PDD)

A distance decay (DD) relationship reflects the spatial autocorrelation of community similarity. That is, communities located near one another should be more similar to one another in taxonomic composition than distant communities. (This is analagous to the isolation by distance (IBD) pattern that is commonly found

when examining genetic similarity of a populations as a function of space.) Historically, the two most common explanations for the taxonomic DD are that it reflects spatially autocorrelated environmental variables and the influence of dispersal limitation. However, if phylogenetic diversity is also spatially autocorrelated, then evolutionary history may also explain some of the taxonomic DD pattern. Here, we will construct the phylogenetic distance-decay (PDD) relationship

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 Similarity",
     main = "Distance Decay", col = "SteelBlue")

DD.reg.bc <- lm(df$bray.curtis ~ df$geo.dist)
summary(DD.reg.bc)
```

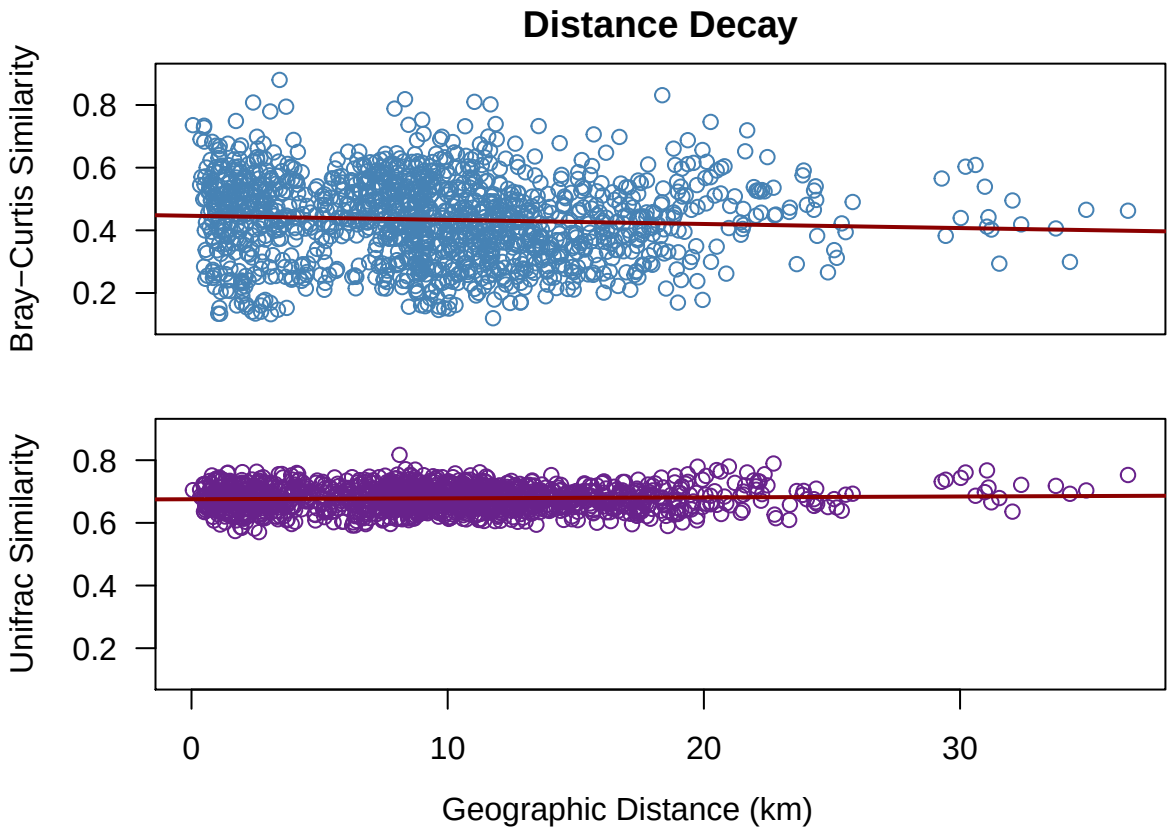
```
##
## Call:
## lm(formula = df$bray.curtis ~ df$geo.dist)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.31151 -0.08843  0.00315  0.09121  0.43817
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (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
## -0.105716 -0.027135  0.000119  0.026781  0.139246
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.6754253  0.0019087 353.857   <2e-16 ***
## df$geo.dist 0.0002899  0.0001673   1.733   0.0834 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03718 on 1324 degrees of freedom
## Multiple R-squared:  0.002262,   Adjusted R-squared:  0.001509
## F-statistic: 3.002 on 1 and 1324 DF,  p-value: 0.0834
```

```
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: 0.001595
## Significance: 0.005
##
## Empirical upper confidence limits of r:
##     90%     95%   97.5%     99%
## 0.00076 0.00101 0.00117 0.00139
```

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

> ***Answer 7***: Both slopes are only weakly significant and explain less than 1% of the total variance in similarity, so distance is not that strong of a driver. The slope for Bray-Curtis similarity is slightly negative, whereas for Unifrac similarity it is slightly positive. These slopes are significantly different. The negative slope for Bray-Curtis similarity makes sense; more distant communities should be less similar. The positive slope makes less sense, but it is marginally significant at best ($p = 0.08$), so it is probably best interpreted as a negative result. This suggests that phylogenetic diversity in these sites is probably not spatially autocorrelated, ie. there are not significant

differences in evolutionary history over space.

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.

Most of my research up to this point has focused on theoretical aspects of phylogenetics / population genetics, but much of the empirical work that is done in our lab is phylogenetically informed due to the nature of what we study. For example, I have been working on a project to detect positive selection in the coding regions of Nancy Ma's night monkey, a nocturnal monkey native to South America. For this project, I have to "polarize" amino acid subsitutions, by identifying them as specific to the lineage leading to owl monkeys. This requires that I have an alignment between owl monkey, a closely related species, and an outgroup, for each ortholog. In terms of future research, I have an interest in introgression analyses from whole-genome sequence, which necessitates a phylogenetic framework.