Phylogenetic Diversity - Communities

Matt Gibson; Z620: Quantitative Biodiversity, Indiana University

01 March, 2017

OVERVIEW

Complementing taxonomic measures of α- and β-diversity with evolutionary information yields insight into a broad range of biodiversity issues including conservation, biogeography, and community assembly. In this assignment, you will be introduced to some commonly used methods in phylogenetic community ecology.

After completing this assignment you will know how to:

1. incorporate an evolutionary perspective into your understanding of community ecology
2. quantify and interpret phylogenetic α- and β-diversity
3. evaluate the contribution of phylogeny to spatial patterns of biodiversity

Directions:

1. Change "Student Name" on line 3 (above) with your name.
2. Complete as much of the assignment as possible during class; what you do not complete in class will need to be done outside of class.
3. Use the handout as a guide; it contains a more complete description of data sets along with the proper scripting needed to carry out the exercise.
4. Be sure to **answer the questions** in this exercise document; they also correspond to the handout. Space for your answer is provided in this document and indicated by the ">" character. If you need a second paragraph be sure to start the first line with ">".
5. Before you leave the classroom, **push** this file to your GitHub repo.
6. When you are done, **Knit** the text and code into a PDF file.
7. After Knitting, please submit the completed assignment by creating a **pull request** via GitHub. Your pull request should include this file *PhyloCom_assignment.Rmd* and the PDF output of `Knitr` (*PhyloCom_assignment.pdf*).

1) SETUP

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

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

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

```
## [1] "C:/Users/matth/Documents/bin/QB2017_Gibson/Week7-PhyloCom"
```

```
setwd("c:/Users/matth/Documents/bin/QB2017_Gibson/Week7-PhyloCom/")

package.list <- c('picante', 'ape', 'seqinr', 'vegan', 'fossil', 'simba')
for (package in package.list){
  if (!require(package, character.only = T, quietly=T)){
    install.packages(package, repos='http://cran.us.r-project.org')
    library(package, character.only=T)
  }
}
```

```
## Warning: package 'picante' was built under R version 3.2.5

## Warning: package 'ape' was built under R version 3.2.5

## Warning: package 'vegan' was built under R version 3.2.5

## Warning: package 'permute' was built under R version 3.2.5

## This is vegan 2.4-2

## Warning: package 'nlme' was built under R version 3.2.4

## Warning: package 'seqinr' was built under R version 3.2.5

##
## 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

## Warning: package 'fossil' was built under R version 3.2.5

## Warning: package 'sp' was built under R version 3.2.5

## Warning: package 'maps' was built under R version 3.2.4

##
##  # maps v3.1: updated 'world': all lakes moved to separate new #
##  # 'lakes' database. Type '?world' or 'news(package="maps")'.  #
```

```
## Warning: package 'shapefiles' was built under R version 3.2.5

##
## Attaching package: 'shapefiles'

## The following objects are masked from 'package:foreign':
##
##     read.dbf, write.dbf

## Warning: package 'simba' was built under R version 3.2.5

## 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

## Warning: package 'reshape' was built under R version 3.2.5
```

2) DESCRIPTION OF DATA

We will revisit the data that was used in the Spatial Diversity module. As a reminder, in 2013 we sampled ~ 50 forested ponds located in Brown County State Park, Yellowwood State Park, and Hoosier National Forest in southern Indiana. See the handout for a further description of this week's dataset.

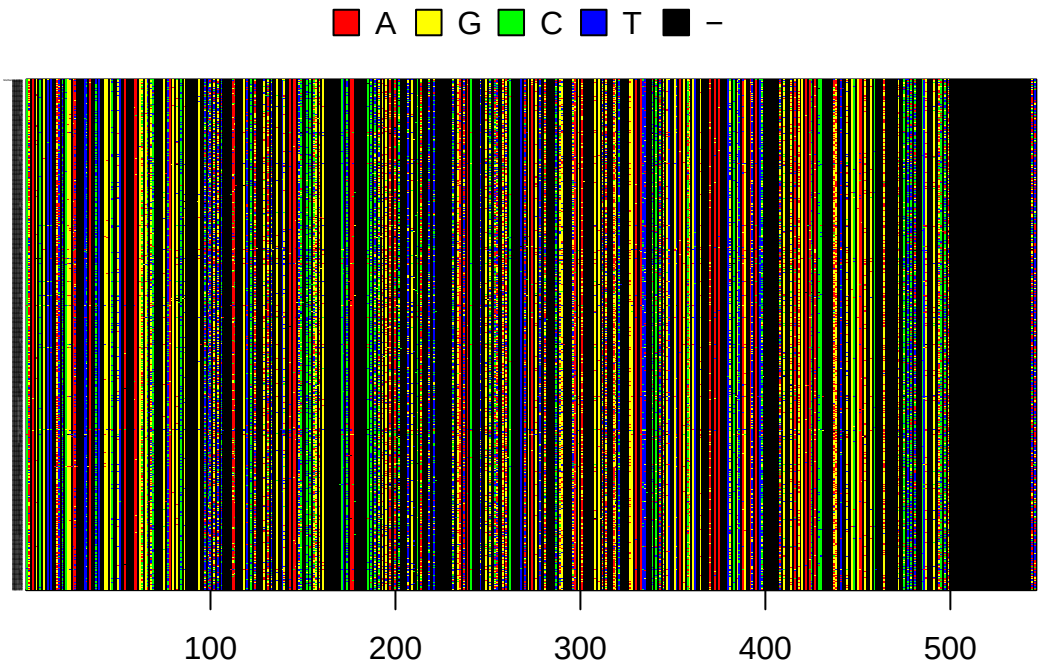
3) LOAD THE DATA

In the R code chunk below, do the following:
1. load the environmental data for the Brown County ponds (*20130801_PondDataMod.csv*),
2. load the site-by-species matrix using the `read.otu()` function,
3. subset the data to include only DNA-based identifications of bacteria,
4. rename the sites by removing extra characters,
5. remove unnecessary OTUs in the site-by-species, and
6. load the taxonomic data using the `read.tax()` function from the source-code file.

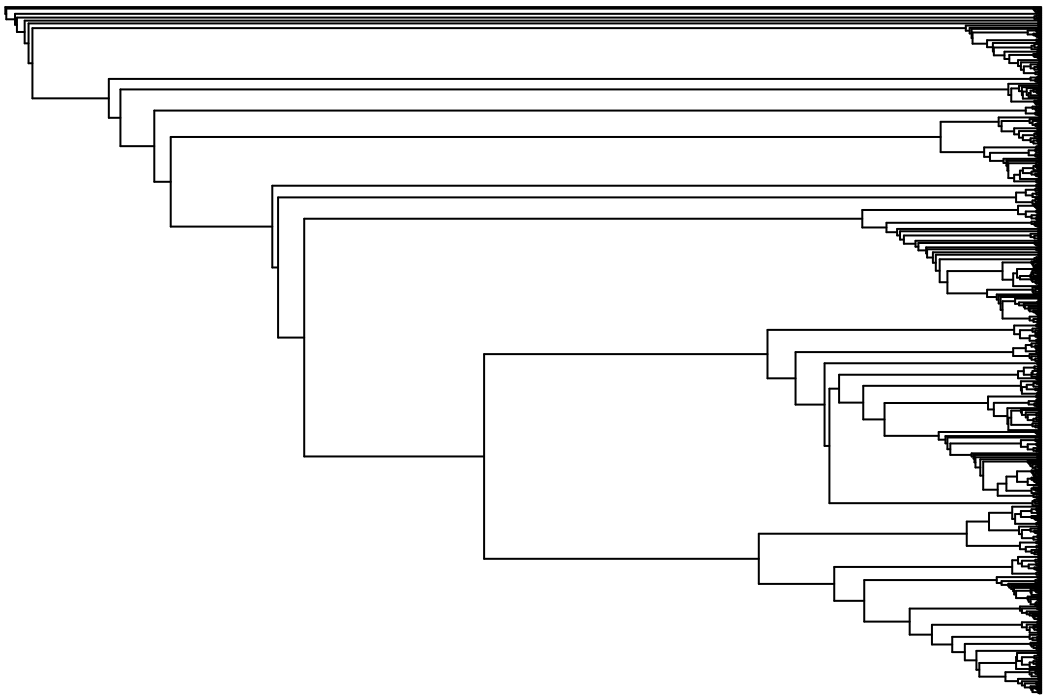
```
env <- read.table("data/20130801_PondDataMod.csv", sep = ",", header=T)
env <- na.omit(env)
comm <- read.otu(shared = "./data/INPonds.final.rdp.shared", cutoff="1")
comm <- comm[grep("*-DNA", rownames(comm)), ]
rownames(comm) <- gsub("\\-DNA", "", rownames(comm))
rownames(comm) <- gsub("\\_", "",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

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

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

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

```
ponds.cons <- read.alignment(file = "./data/INPonds.final.rdp.1.rep.fasta", format = "fasta")
ponds.cons$nam <- gsub("\\|.*$", "", gsub("^.*?\t", "", ponds.cons$nam))
outgroup<- read.alignment(file = "./data/methanosarcina.fasta", format = "fasta")

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

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




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

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

phy <- drop.tip(phy.all, phy.all$tip.label[!phy.all$tip.label %in%
                                           c(colnames(comm), "Methanosarcina")])

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

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

par(mar = c(1,1,2,1) + 0.1)
plot.phylo(phy, main = "NJ TREE", "phylogram", show.tip.label = F,
           use.edge.length = F, 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 = F)

pd
```

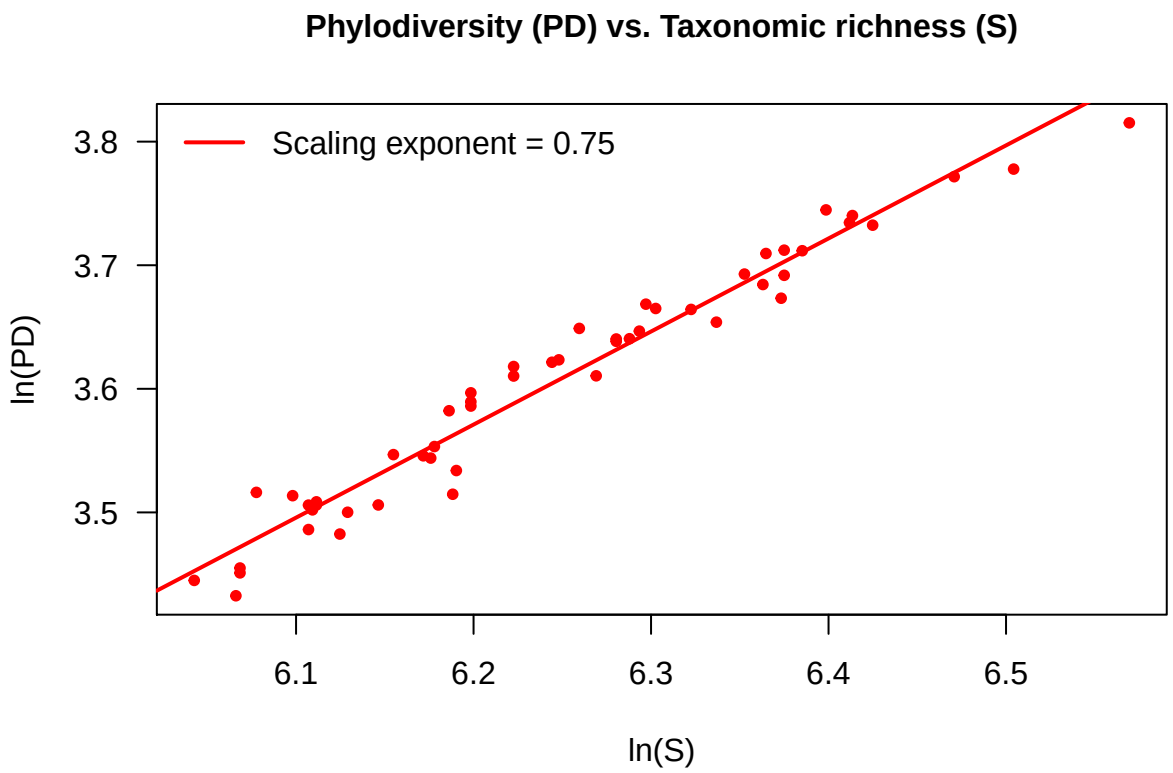
```
##              PD  SR
## BC001  43.71912 668
## BC002  40.94334 587
## BC003  31.53402 432
## BC004  35.95465 486
## BC005  33.65632 436
## BC010  31.34254 421
## BC015  40.15954 574
## BC016  38.62593 565
## BC018  36.98545 528
## BC020  40.92505 593
## BC048  37.39332 515
## BC049  32.65870 449
```

```
## BC051  33.56599 445
## BC105  41.86524 609
## BC108  37.46606 517
## BC262  33.31506 467
## BCL01  38.43171 523
## BCL03  33.11983 459
## HNF132 38.03250 534
## HNF133 33.31136 449
## HNF134 38.34699 541
## HNF144 37.26483 504
## HNF168 39.19321 543
## HNF185 33.39952 451
## HNF187 30.95549 431
## HNF216 36.47943 492
## HNF217 36.97870 504
## HNF221 38.11234 538
## HNF224 39.05581 546
## HNF225 38.10391 534
## HNF229 34.25624 488
## HNF242 34.92780 482
## HNF250 34.66701 479
## HNF267 33.60526 487
## HNF269 32.54033 457
## YSF004 42.29978 601
## YSF117 34.70052 471
## YSF295 42.10615 610
## YSF296 34.60291 481
## YSF298 36.21693 492
## YSF300 33.31466 451
## YSF44  40.11835 587
## YSF45  40.83293 581
## YSF46  39.02660 557
## YSF47  33.18340 450
## YSF65  39.38228 586
## YSF66  45.38554 713
## YSF67  43.45101 646
## YSF69  31.65875 432
## YSF70  39.82013 580
## YSF71  41.77992 617
## YSF74  36.09142 492
```

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) + .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")
```




```
print(fit)
```

```
##
## Call:
## lm(formula = "log(pd$PD) ~ log(pd$S)")
##
## Coefficients:
## (Intercept)    log(pd$S)
##      -1.098        0.753
```

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***: Because the presence of more species (higher richness) will necessarily lead to a larger sum of branch lengths. Even if the new species at a higher richness are closely related, there will have to be a positive (though small) increase in PD. ***Answer 1b***: It is linear and positive in the log scale. Slope = 0.75. ***Answer 1c***: When there are very few highly diverged species at a site. Or when there are many very closesly related species at a site. ***Answer 1d***: It is not a

1:1 relationship. A unit increase in `S` is not met with a unit increase in `PD`. In the log scale, a 4 unit increase in `S` is met with a 3 unit increase in `PD`. One wouldn't expect it to be 1:1 since `PD` is affected by how closely related species at increasing richness are.

i. Randomizations and Null Models

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

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

ses.pd
```

```
##       ntaxa   pd.obs pd.rand.mean pd.rand.sd pd.obs.rank   pd.obs.z
## BC001   668 43.71912     44.24037  0.9842151           7 -0.5296079
## BC002   587 40.94334     39.67874  0.8161399          25  1.5494817
##        pd.obs.p runs
## BC001 0.2692308   25
## BC002 0.9615385   25
```

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

```
##       ntaxa   pd.obs pd.rand.mean pd.rand.sd pd.obs.rank  pd.obs.z
## BC001   668 43.71912     42.28261  0.7852100          26  1.829460
## BC002   587 40.94334     42.29003  0.7543007           1 -1.785358
##         pd.obs.p runs
## BC001 1.00000000   25
## BC002 0.03846154   25
```

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

ses.pd.p
```

```
##       ntaxa   pd.obs pd.rand.mean pd.rand.sd pd.obs.rank   pd.obs.z
## BC001   668 43.71912     43.95428  0.7945398          12 -0.2959709
## BC002   587 40.94334     40.01670  0.6665476          22  1.3901993
##        pd.obs.p runs
## BC001 0.4615385   25
## BC002 0.8461538   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: Null: Observed `PD` is equal to `PD` obtained from a community dataset with randomized abundances within samples (but maintaining total site species richness). Alternative: Observed `PD` is not equal to the `PD` obtained through randomizations. /// The richness null model allows us to control for species richness in our tests. ***Answer 2b***: The choice of model (as expected) did effect our estimates of standard effect size. Choosing a different randomization null model effected the `pd.rand.mean` and `pd.rand.sd` terms in the ses formula. (note that these are random, so my discussion here may not reflect the actual output of the functions once I re-knit this document for submission). The `frequency` null model resulted in observed standardized effect sizes of higher magnitude in both site commpared to the two other null models `richness` and `sample.pool`. `Richness` and `sample.pool` null models had similar effect sizes. No choice of null model effected our conclusions (at leat at alpha = 0.05). P-values for both sites using all null models were greater than .05. The choice of null model affected the output because in chosing different models, we are controlling for different things (richness with the `richness` model, species occurence frequency with the `frequency` model, and so on).

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 = F, 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
## BC001  -3.5514293
## BC002  -3.8734989
## BC003  -1.9927861
## BC004  -3.9843752
## BC005  -4.4332790
## BC010  -3.1307356
## BC015  -2.7848617
## BC016  -2.4879478
## BC018  -3.4071661
## BC020  -2.1128837
## BC048  -2.7535885
## BC049  -1.0507282
## BC051  -3.0727673
## BC105  -3.1340842
```

```
## BC108  -2.5851856
## BC262  -2.4719106
## BCL01  -4.6618146
## BCL03  -1.8793340
## HNF132 -3.1951109
## HNF133 -3.2729746
## HNF134 -2.5342165
## HNF144 -5.1501436
## HNF168 -2.3254928
## HNF185 -4.2470603
## HNF187  0.1774809
## HNF216 -1.9232022
## HNF217 -2.5871250
## HNF221 -2.1670675
## HNF224 -3.4625727
## HNF225 -2.4557934
## HNF229 -1.5992215
## HNF242 -1.9471665
## HNF250 -2.5922799
## HNF267 -1.5881561
## HNF269 -2.3673693
## YSF004 -3.3084632
## YSF117 -2.7770000
## YSF295 -3.4182762
## YSF296 -2.9265812
## YSF298 -3.6246425
## YSF300 -2.8019626
## YSF44  -2.1970127
## YSF45  -3.1433941
## YSF46  -1.8995517
## YSF47  -3.6120729
## YSF65  -0.8417328
## YSF66  -1.2546914
## YSF67  -3.2620878
## YSF69  -2.0978509
## YSF70  -2.1845613
## YSF71  -1.7756832
## YSF74  -3.1941658
```

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 = F, 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   0.59216665
## BC002  -1.21051894
## BC003  -0.26313751
```

```
## BC004  -1.56140848
## BC005  -1.95791927
## BC010  -0.82420137
## BC015  -0.65531037
## BC016   0.09242089
## BC018  -0.40782601
## BC020  -0.86361178
## BC048  -2.05638612
## BC049  -0.02255801
## BC051  -1.08950875
## BC105  -1.29655520
## BC108  -0.94699500
## BC262   0.11127385
## BCL01  -1.51915810
## BCL03   0.01623603
## HNF132 -0.60652147
## HNF133 -0.80024968
## HNF134 -0.76989428
## HNF144 -1.86274747
## HNF168 -1.34275068
## HNF185 -0.68615073
## HNF187  0.37211528
## HNF216 -2.09937450
## HNF217 -1.25837836
## HNF221 -1.03298889
## HNF224 -2.63656165
## HNF225 -1.89211383
## HNF229  0.16330801
## HNF242 -1.25233014
## HNF250 -0.57573436
## HNF267  0.87651243
## HNF269  0.30863422
## YSF004 -1.98015851
## YSF117 -1.16755307
## YSF295 -0.95730309
## YSF296  0.11864890
## YSF298 -0.91584452
## YSF300 -0.78134737
## YSF44  -1.14007398
## YSF45  -1.28008267
## YSF46  -0.83833994
## YSF47  -0.71636301
## YSF65   0.21913075
## YSF66   0.89088856
## YSF67  -0.64308811
## YSF69   0.03079696
## YSF70  -0.28408231
## YSF71  -0.54147092
## YSF74  -1.23166222
```

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

NRI <- as.matrix(-1*((ses.mpd[,2] - ses.mpd[,3]) / ses.mpd[,4]))
```

```r
rownames(NRI) <- row.names(ses.mpd)
colnames(NRI) <- "NRI"
NRI
```

```
##                NRI
## BC001   0.20172201
## BC002   0.76734683
## BC003   0.90249760
## BC004  -0.01111229
## BC005   0.83949781
## BC010   0.42358451
## BC015   0.37267575
## BC016   0.51859854
## BC018   0.38470267
## BC020   0.57619509
## BC048   0.28818937
## BC049   0.49910261
## BC051  -0.23346281
## BC105  -0.09244506
## BC108  -0.25458816
## BC262  -0.62279211
## BCL01  -0.48945229
## BCL03  -0.67225571
## HNF132  0.08664766
## HNF133  0.39560865
## HNF134  0.47735486
## HNF144 -0.10341686
## HNF168 -0.28211370
## HNF185  0.27448223
## HNF187  0.86168227
## HNF216  0.74886107
## HNF217 -0.03764532
## HNF221 -0.19996212
## HNF224  0.30127513
## HNF225  0.79284063
## HNF229 -0.71718110
## HNF242 -0.12779027
## HNF250 -0.65557308
## HNF267 -0.88379985
## HNF269 -0.80075260
## YSF004 -0.32491929
## YSF117  0.90075293
## YSF295 -0.57987005
## YSF296  1.34255235
## YSF298  1.28908088
## YSF300  0.55504318
## YSF44   0.73863309
## YSF45   0.75160483
## YSF46   1.60702782
## YSF47   0.55935027
## YSF65   0.85619781
## YSF66  -0.25004437
## YSF67   0.21929471
```

```
## YSF69   0.01461245
## YSF70  -0.01439505
## YSF71   0.86704835
## YSF74   1.11482025
```

```
ses.mntd <- ses.mntd(comm, phydist, null.model = "taxa.labels",
                     abundance.weighted = T, 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   0.9024107
## BC002   1.6660594
## BC003   1.3004652
## BC004   1.0481697
## BC005   1.8065587
## BC010   0.5103753
## BC015   1.1287443
## BC016   1.7391368
## BC018   1.0806681
## BC020   1.1733140
## BC048   1.1107971
## BC049   1.3213836
## BC051   1.7848007
## BC105   1.3198585
## BC108   1.7064726
## BC262   1.1048825
## BCL01   1.5134535
## BCL03   1.4781592
## HNF132  1.2424093
## HNF133  1.1896427
## HNF134  1.3271685
## HNF144  0.8847214
## HNF168  0.5908694
## HNF185  1.0991264
## HNF187  0.5556366
## HNF216  0.3219022
## HNF217  0.3567630
## HNF221  0.6104066
## HNF224  1.1525511
## HNF225  0.4090754
## HNF229  1.3233682
## HNF242  1.6394086
## HNF250  1.2458413
## HNF267  0.8882484
## HNF269  1.0761726
## YSF004  0.3101564
## YSF117  1.3978772
## YSF295 -1.0306161
## YSF296  1.0875288
## YSF298  1.9592040
## YSF300  1.6630732
```

```
## YSF44   1.0432957
## YSF45   0.8858712
## YSF46   1.7453179
## YSF47   0.9215342
## YSF65   1.2202491
## YSF66   1.3095077
## YSF67   1.2001774
## YSF69   1.0614427
## YSF70   0.9933571
## YSF71   1.2763118
## YSF74   1.4979612
```

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: We are finding the difference between the observed mean phylogenetic distance (MPD; average pairwise branch length between taxa in the sample) and the mean MPD generted from randomization of the data under a particular null model. This difference is divided by the standard deviation of the randomized data so that it is standardized and comparable. We can use the NRI to assess overdispersion vs. underdispersion (i.e. not clustered vs. clustered) of taxa at the site in question. ***Answer 3b***: NTI is similar to NRI, but uses a different metric of distance (mean nearest phylogenetic neighbor distance (MNND) instead of MPD). MNND is the the mean distance separating each species in the sample from the species it is most closely related to. ***Answer 3c***: Negative NRI values tell us that species in the site/sample are overdispersed (highly unrelated). For all but 1 sample (HNF187), values for NRI are negative. This indicates that sites contain more phylogenetically distant taxa than would be expected by chance. Values given by the NTI are interpreted the same way as NRI and, in general, show the same pattern. Most sites have negative values indicating overdispersion. Though importantly, multiple more sites have positive values for NTI (BC001, BCLO3, HNF267, YSF65, YSF66, and HNF269). Interestingly, the only site that showed a positive NRI value (HNF187), did not have a positive NTI value. This could be due to the fact that NTI emphasizes the clustering of taxa at the tips independent of how they cluster at deeper nodes. ***Answer 3d***: If we weight nearest taxon distances by abundance, we are placing more emphasis on species with higher abundance. There could be a site dominated by a single species (that may otherwise have a very low pairwise or nearest taxon distance) be given an inappropriately high distance value and therefore overestimate NTI or NRI. This is probably why the default is FALSE. If one were to use this option, it would probably be a good idea to evaluate the data itself before drawing conclusions. Indeed, when calculating NRI and NTI weighting by abundance, the values are much higher (and positive) at all sites. In most cases, this flips our conclusions about the overdispersion of taxa at sites.

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, abundance.weighted = F)
```

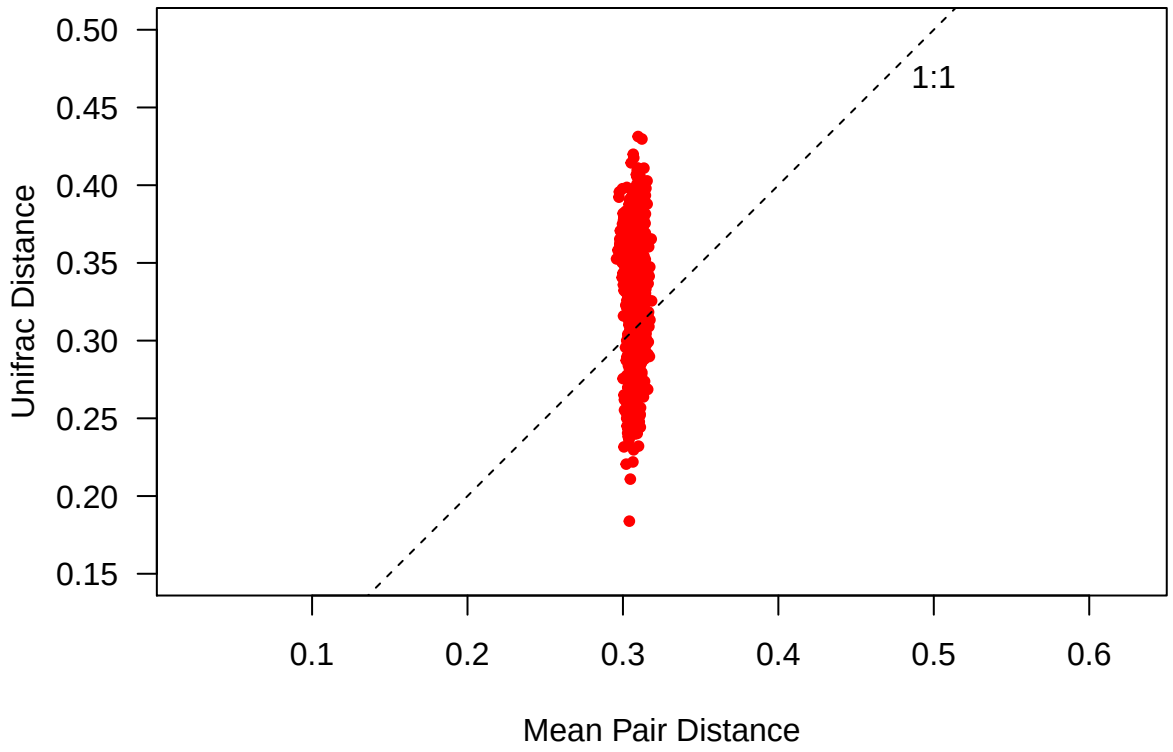
```
## [1] "Dropping taxa from the distance matrix because they are not present in the community data:"
## [1] "Methanosarcina"
```

```
dist.uf <- unifrac(comm, phy)
mean(dist.uf)
```

```
## [1] 0.3236131
```

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) + .1)
plot(dist.mp, dist.uf,
     pch=20, col="red", las = 1, asp=1, xlim=c(0.15,0.5), ylim = c(0.15,0.5),
     xlab = "Mean Pair Distance", ylab = "Unifrac Distance")
abline(b = 1, a = 0, lty = 2)
text(0.5, 0.47, "1:1")
```



Question 4:

a. In your own words describe Mean Pair Distance, UniFrac distance, and the difference between them.

b. Using the plot above, describe the relationship between Mean Pair Distance and UniFrac distance. Note: we are calculating unweighted phylogenetic distances (similar to incidence based measures). That means that we are not taking into account the abundance of each taxon in each site.
c. Why might MPD show less variation than UniFrac?

Answer 4a: Mean pairwise distance is the average phylogenetic distance between species in site 1 and species in site 2. That is, we average over all pairwise distances between the two sites. With UniFrac, we calcuate two quantities: the sum of all unshared branches between samples and the sum of off all branch lengths in our tree. The Unifrac distance is then calculated as sum(unshared)/sum(total). If two sites contain no unshared branches, UniFrac equals 0 and the communities are identical. ***Answer 4b***: They are not related. The mean values for the two metrics are roughly the same. MPD is much less variable than UniFrac and the two metrics do not have a strong relationship, positive or otherwise. ***Answer 4c***: Because with UniFrac, two sites could have anywhere from 0 to all but 2 unshared branches. Compared to the MPD which is an average and more or else equalizes, UniFrac will have a higher variance just based on how it is calculated.

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)

explainvar1
```

```
## [1] 9.5
```

```
explainvar2
```

```
## [1] 6
```

```
explainvar3
```

```
## [1] 5.4
```

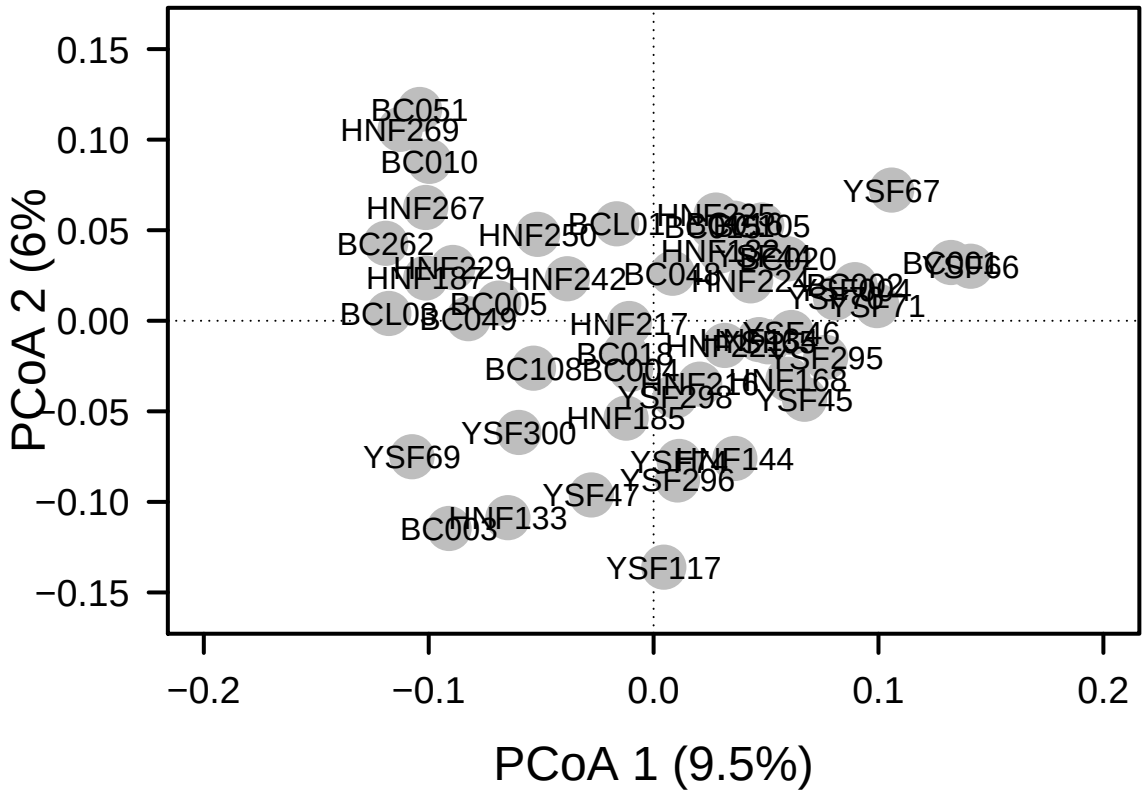
```
sum.eig
```

```
## [1] 20.9
```

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

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

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

```
library(BiodiversityR)
```

```
## Warning: package 'BiodiversityR' was built under R version 3.2.5
```

```
## Loading required package: tcltk
```

```
## Warning in FUN(X[[i]], ...): failed to assign NativeSymbolInfo for env
## since env is already defined in the 'lazyeval' namespace
```

```
## BiodiversityR 2.8-0: Use command BiodiversityRGUI() to launch the Graphical User Interface and to lea
```

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

#Variance explained by first axis
explainvar1
```

```
## [1] 28.4
```

```
#variance explained by second axis
explainvar2
```

```
## [1] 12
```

```
#variance explained by third axis
explainvar3
```

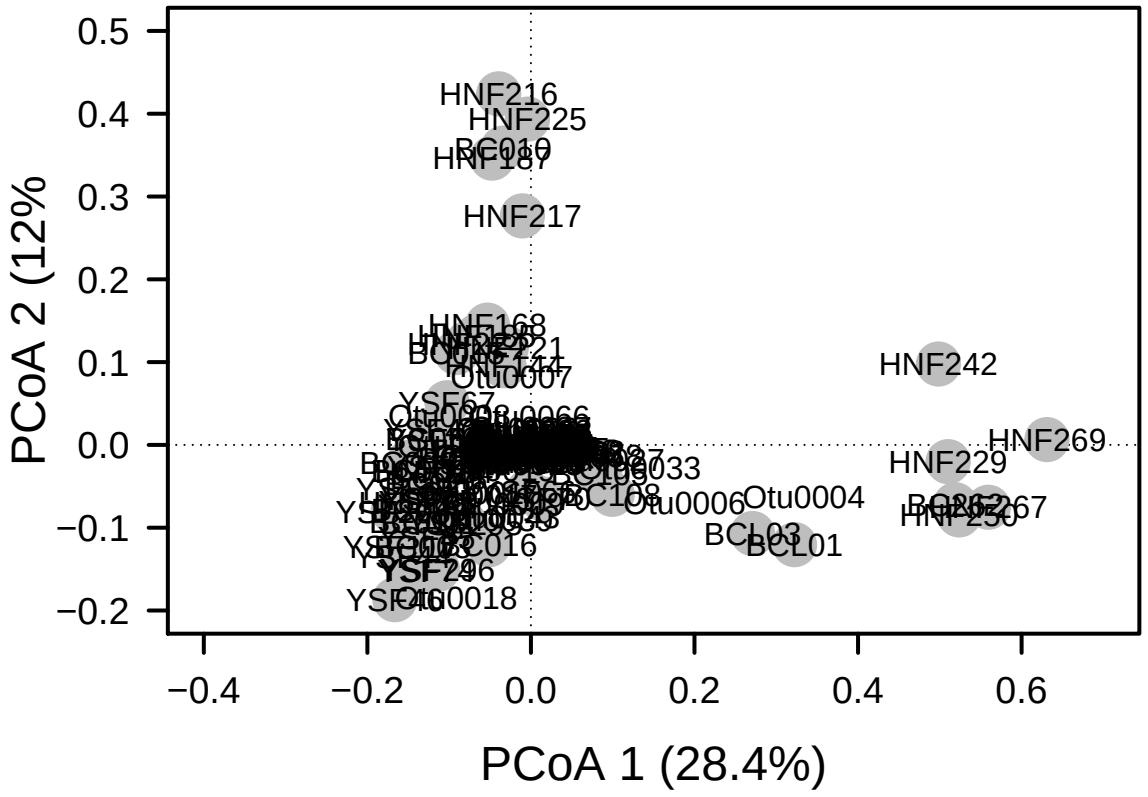
```
## [1] 8.6
```

```
sum.eig
```

```
## [1] 49
```

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

plot(comm.pcoa$points[ ,1], comm.pcoa$points[ ,2], ylim = c(-0.2, 0.5), xlim = c(-.4, .7),
     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 = F)

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


commREL <- comm
  for(i in 1:nrow(comm)){
    commREL[i, ] = comm[i, ]/ sum(comm[i, ])
  }

comm.pcoa <- add.spec.scores(comm.pcoa, commREL, method = "pcoa.scores")
text(comm.pcoa$cproj[ ,1], comm.pcoa$cproj[ ,2],
     labels = row.names(comm.pcoa$cproj), col = "black")
```



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***: More variation was explained in the taxonomic ordination (axis 1: 28.4%, axis 2: 12%, axis 3: 8.6%) than in the phylogenetic ordination (axis 1: 9.5%, axis 2: 6%, axis 3: 5.4%). There is clear clustering of sites in the taxonomic ordination that is missing in the phylogenetic ordination. So, it would seem that, phylogenetically, the sites are all fairly similar. Although overall less variation was explained in the phylogenetic ordination, I would argue that phylogenetic information is very important in this system. Sites may differ taxonomically, but these differences are in taxa that are closely related.

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  0.033 *
## Residuals 49   2.57305 0.052511         0.9508
## Total     51   2.70621                  1.0000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

```
##
## Call:
## adonis(formula = vegdist(decostand(comm, method = "log"), method = "bray") ~      watershed, permuta
##
## Permutation: free
## Number of permutations: 999
##
## Terms added sequentially (first to last)
##
##           Df SumsOfSqs  MeanSqs F.Model      R2 Pr(>F)
## watershed  2   0.16601 0.083003  1.5689 0.06018  0.005 **
## Residuals 49   2.59229 0.052904         0.93982
## Total     51   2.75829                  1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ii. Continuous Approach

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

```
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: 0.1604
##       Significance: 0.058
##
## Upper quantiles of permutations (null model):
##   90%   95% 97.5%   99%
## 0.125 0.170 0.218 0.273
## Permutation: free
## Number of permutations: 999
```

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

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

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

```
## Permutation test for dbrda under reduced model
## Marginal tests for axes
## Permutation: free
## Number of permutations: 999
##
## Model: vegan::dbrda(formula = dist.uf ~ Elevation + Diameter + Depth + ORP + Temp + SpC + DO + pH + (
##          Df SumOfSqs      F Pr(>F)
## dbRDA1    1  0.10566 2.0152  0.004 **
## dbRDA2    1  0.09258 1.7658  0.003 **
## dbRDA3    1  0.07555 1.4409  0.032 *
## dbRDA4    1  0.06677 1.2735  0.099 .
## dbRDA5    1  0.05666 1.0807  0.344
## dbRDA6    1  0.05293 1.0095  0.420
## dbRDA7    1  0.04750 0.9059  0.645
## dbRDA8    1  0.03941 0.7517  0.899
## dbRDA9    1  0.03775 0.7201  0.916
## dbRDA10   1  0.03280 0.6256  0.987
```

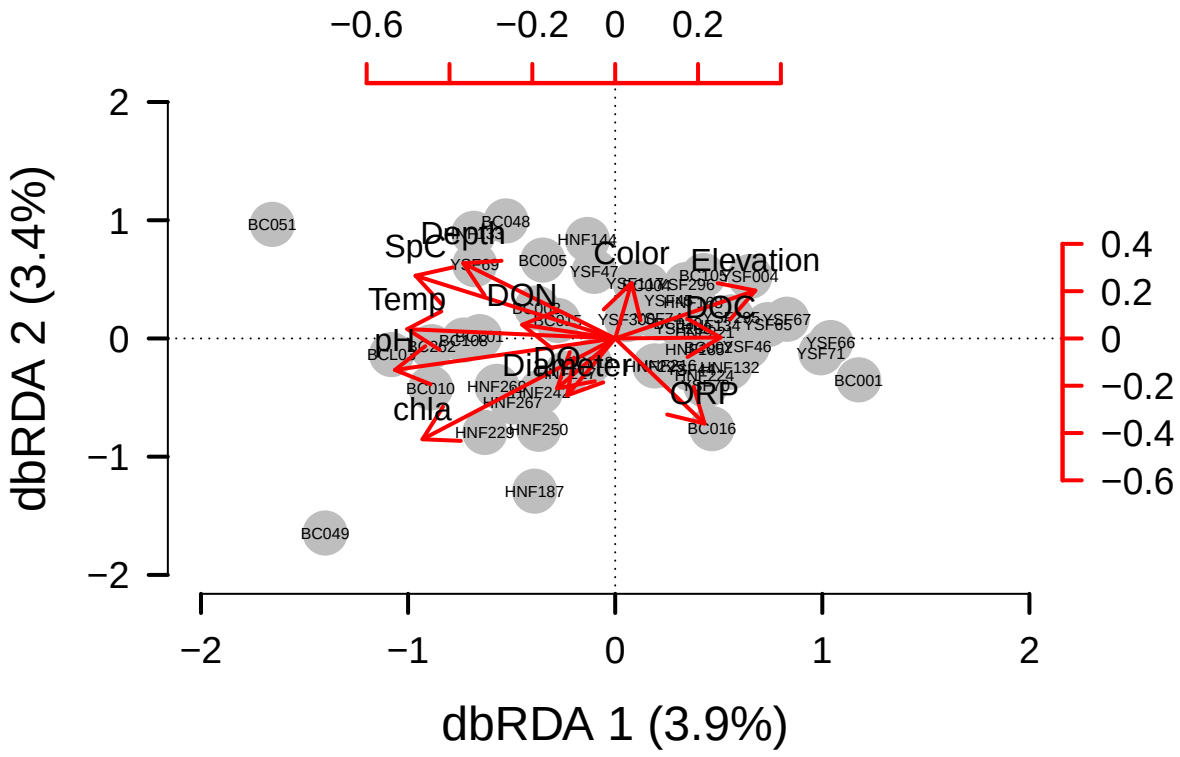
```
## dbRDA11   1  0.02876 0.5485  0.994
## dbRDA12   1  0.02501 0.4770  1.000
## Residual 39  2.04482
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

```
##
## ***VECTORS
##
##             dbRDA1   dbRDA2     r2 Pr(>r)
## Elevation  0.77670  0.62986 0.0959  0.088 .
## Diameter  -0.27972 -0.96008 0.0541  0.255
## Depth     -0.63137  0.77548 0.1756  0.013 *
## ORP        0.41879 -0.90808 0.1437  0.020 *
## Temp      -0.98250  0.18628 0.1523  0.010 **
## SpC       -0.77101  0.63682 0.2087  0.004 **
## DO        -0.39318 -0.91946 0.0464  0.294
## pH        -0.96210 -0.27270 0.1756  0.009 **
## Color      0.06353  0.99798 0.0464  0.329
## chla      -0.60392 -0.79704 0.2626  0.008 **
## DOC        0.99847 -0.05526 0.0382  0.398
## DON       -0.91633  0.40042 0.0339  0.434
## ---
## 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 = F)
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)

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***: Based on a PERMANOVA, `watershed` has a significant effect on phylogenetic diversity (P = 0.026) of the bacterial communities as measured by UniFrac distance. The Mantel statistic for the correlation of UniFrac phylogenetic distance with environmental distance was 0.1604 which is significant (0.044) based on permutation tests. Based on the redundancy analysis, environmental variation effects phylogenetic distance between sites and using permutation tests, the environmental variables driving these effects are `Depth of pond`, `Oxidation-red potential`, `Temp`, `specific water conductivity`, `pH`, `chlorophyll a concentration`.

6) SPATIAL PHYLOGENETIC COMMUNITY ECOLOGY

A. Phylogenetic Distance-Decay (PDD)

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

In the R code chunk below, do the following:
1. calculate the geographic distances among ponds,
2. calculate the taxonomic similarity among ponds,
3. calculate the phylogenetic similarity among ponds, and
4. create a dataframe that includes all of the above information.

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

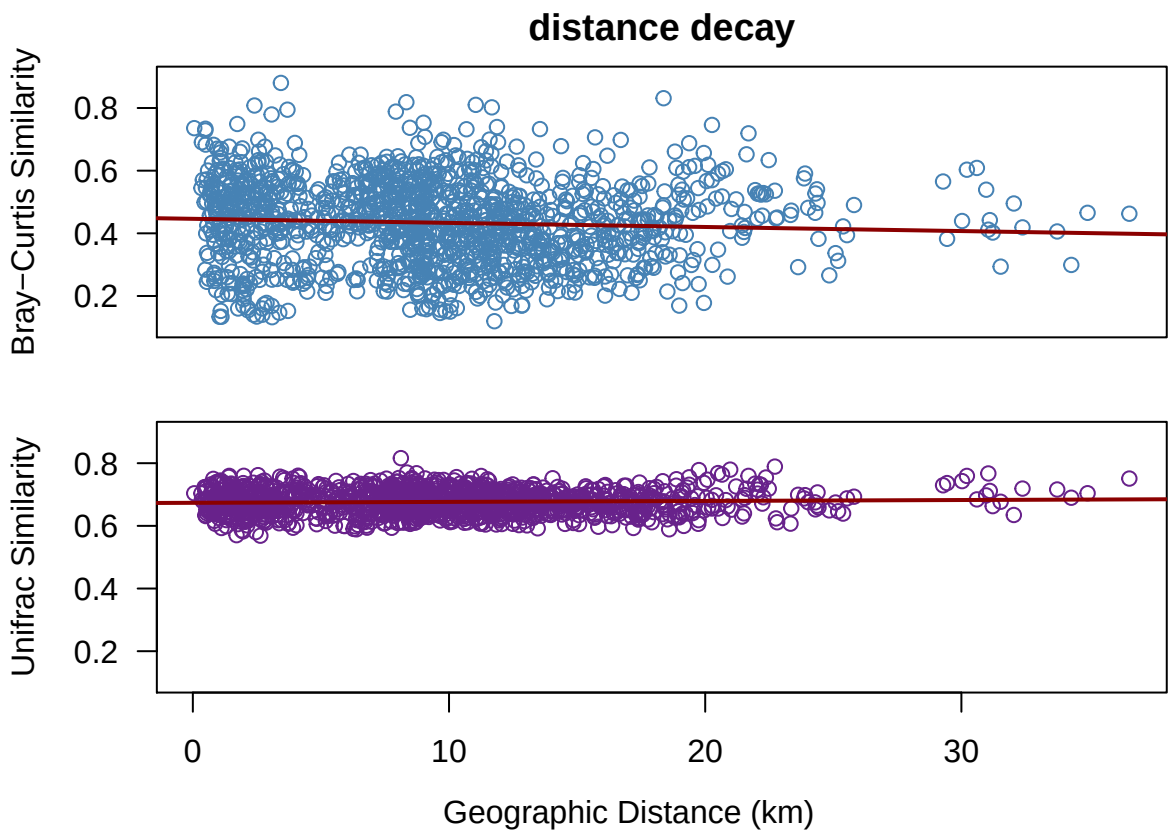
```
##
## Call:
## lm(formula = df$bray.curtis ~ df$geo.dist)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.31151 -0.08843  0.00315  0.09121  0.43817
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (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) + .1)

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

DD.reg.uni <- lm(df$unifrac ~ df$geo.dist)
summary(DD.reg.uni)
```

```
##
## Call:
## lm(formula = df$unifrac ~ df$geo.dist)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.105629 -0.027107 -0.000077  0.026761  0.140215
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.6735186  0.0019206 350.677   <2e-16 ***
## df$geo.dist 0.0002976  0.0001684   1.767   0.0774 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03741 on 1324 degrees of freedom
## Multiple R-squared:  0.002354,   Adjusted R-squared:  0.0016
## F-statistic: 3.124 on 1 and 1324 DF,  p-value: 0.07738
```

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

mtext("Geographic Distance (km)", side = 1, adj = 0.55,
      line = 0.5, outer = T)
```



In the R code chunk below, test if the trend lines in the above distance decay relationships are different from one another.

```
diffslope(df$geo.dist, df$unifrac, df$geo.dist, df$bray.curtis)
```

```
##
## Is difference in slope significant?
## Significance is based on 1000 permutations
##
## Call:
## diffslope(x1 = df$geo.dist, y1 = df$unifrac, x2 = df$geo.dist,      y2 = df$bray.curtis)
##
## Difference in Slope: 0.001603
## Significance: 0.002
##
## Empirical upper confidence limits of r:
##      90%      95%    97.5%      99%
## 0.000708 0.000924 0.001150 0.001390
```

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 slightly different (0.001603) and this difference is significant (0.002). The distance decay curve slope based on Bray-Curtis similarity is -.001 (P = 0.026) while the distance decay based on phylogenetic distance is 0.00029 (P = 0.07). Based on these analyses, there

does not appear to be a phylogenetic signal in variation in diversity across space. Consequently, one might infer that evolutionary events such as adaptive radiations or ecological speciation are not likely to have occured. Rather, species at sites separated by a large amount of space are roughly equally related to those nearby. This could mean that the radiation of these populations occured very recently and they have not have time to diverge or the process by which they were dispersed resulted in the placement of related populations across the landscape.

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 = F, 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 = F)
      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))
}
```

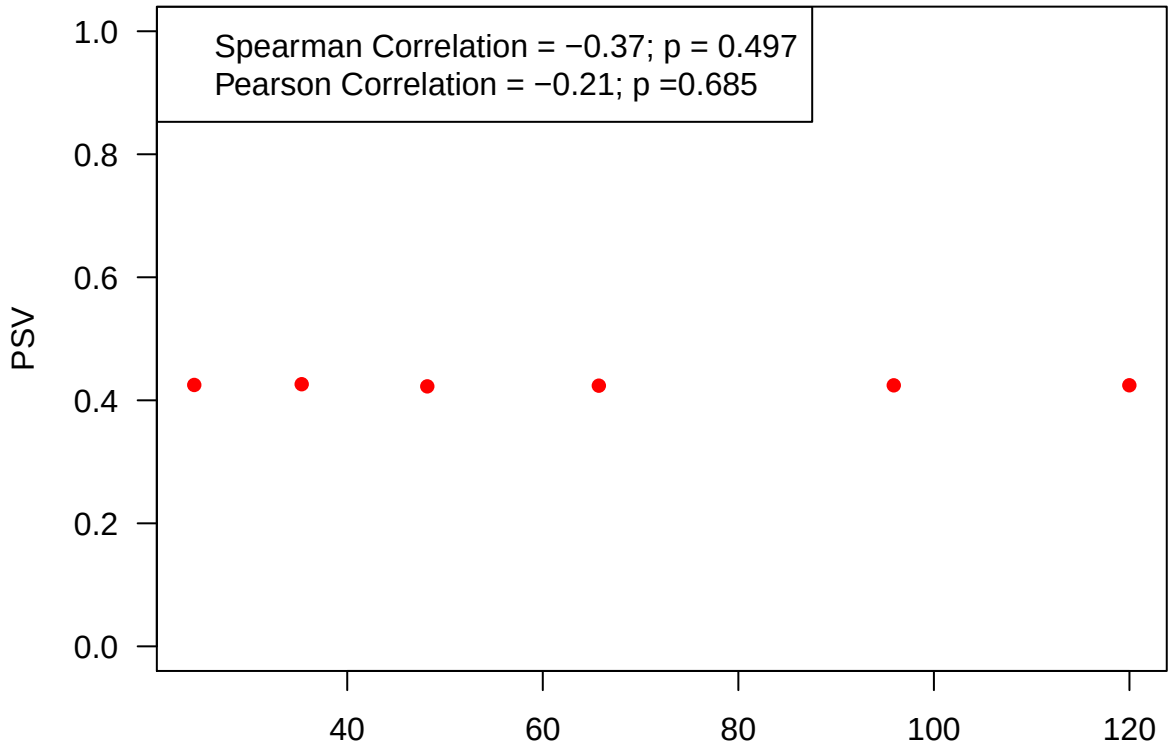
ii. Evaluating the PDAR

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.

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

#Could not recreate plot in handout......

#S.obs <- function(x = ""){
#  rowSums(x > 0) * 1
#    }

#richness <- S.obs(comm)#

#results <- lm(log(richness) ~ pond.areas)
#plot(pond.areas, log(richness), col="dark red", pch=20, cex=2,
#     main="Species-area relationship",
#     xlab='ln(Area)', ylab='ln(Richness)')
#abline(results)


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

```
## [1]   2.0000000   0.4249366 593.0298058
## [1]    4.0000000    0.4261737 1249.0241970
## [1]    8.0000000    0.4226763 2321.9369931
## [1]   16.0000000    0.4237997 4320.6928521
## [1]   32.0000000    0.4243516 9196.2781461
## [1] 5.100000e+01 4.244303e-01 1.439763e+04
```

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

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

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

plot.new()
par(mfrow=c(1,1), mar = c(1,5,2,1) + 0.1, oma = c(2,0,0,0))
plot(pdar[,1], pdar[,2], xlab = "Area", ylab = "PSV", ylim = c(0,1),
     col = "red", pch = 16, las = 1)

legend("topleft", legend = c(paste("Spearman Correlation = ", rho, "; p = ", rho.pval, sep = ""),
                             paste("Pearson Correlation = ", P, "; p =", P.pval, sep = "")))
```



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***: There is no correlation between phylogenetic diversity (as measured by PSV) and area (Spearman P = 0.242, Pearson P = 0.663). This is in contrast to the species-area relationship of richness vs. area which showed a positive relationship (slope = 0.144). As we sample larger areas, we will likely encounter more taxa (SAR) but these taxa need not mean we increase in evolutionary distance as the new species encountered could be very closely related to the current set of species! So as pond area increases, species richness will increase. But if newly encountered species are closely related to the existing set, we will not see a corresponding pattern in the PDAR.

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.

> ***Synthesis***: Most species of tomato are very closely related but nonetheless display tremendous functional diversity. Phylogenetics has played a key role in studying the evolution of self-compatibility (the ability to self pollinate) in the tomato section of Solanum. I am interested in studying the distribution of adaptive traits across climatic and environmental gradients and in identifying the loci responsible for local adaptation, adaptive divergence, and ultimately speciation.

Phylogenetics could play a key roll in understanding the evolution of adaptive traits such as salt tolerance, drought tolerance, and herbivore response. If multiple taxa display strong salt tolerance, for example, did this independently evolve multiple times (convergent evolution) or is it likely that this trait evolved in a common ancestor of all the salt-tolerant taxa? To do this, independent contrasts needs to be made. Similarly, I could ask how salt tolerance is related to local salt concentrations by performing a phylogenetic regression to account for shared evolutionary history. To answer these questions, I would need sequence data from the taxa of interests, phenotype data (degree of salt tolerance), and local salt concentration data from the field.