12.Phylogenetic Diversity - Communities

Student Name; Z620: Quantitative Biodiversity, Indiana University

07 May, 2021

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 '12.PhyloCom' 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*).

The completed exercise is due on **Monday, May 10th, 2021 before 09:00 AM**.

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 `/12.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())
setwd("~/Classwork/Current/QB/")
```

2) DESCRIPTION OF DATA

need to discuss data set from spatial ecology!

In 2013 we sampled > 50 forested ponds 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., 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.

```
for (p in c('picante', 'ape', 'seqinr', 'vegan', 'fossil', 'reshape', 'simba')) {
  require(p, character.only=TRUE)
}
```

```
## Loading required package: picante
## Loading required package: ape
## Loading required package: vegan
## Loading required package: permute
## Loading required package: lattice
## This is vegan 2.5-7
## Loading required package: nlme
## Loading required package: seqinr
##
## 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

## Loading required package: fossil

## Loading required package: sp

## Loading required package: maps

## Loading required package: shapefiles

## Loading required package: foreign

##
## Attaching package: 'shapefiles'

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

## Loading required package: reshape

## Loading required package: simba

## 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
```

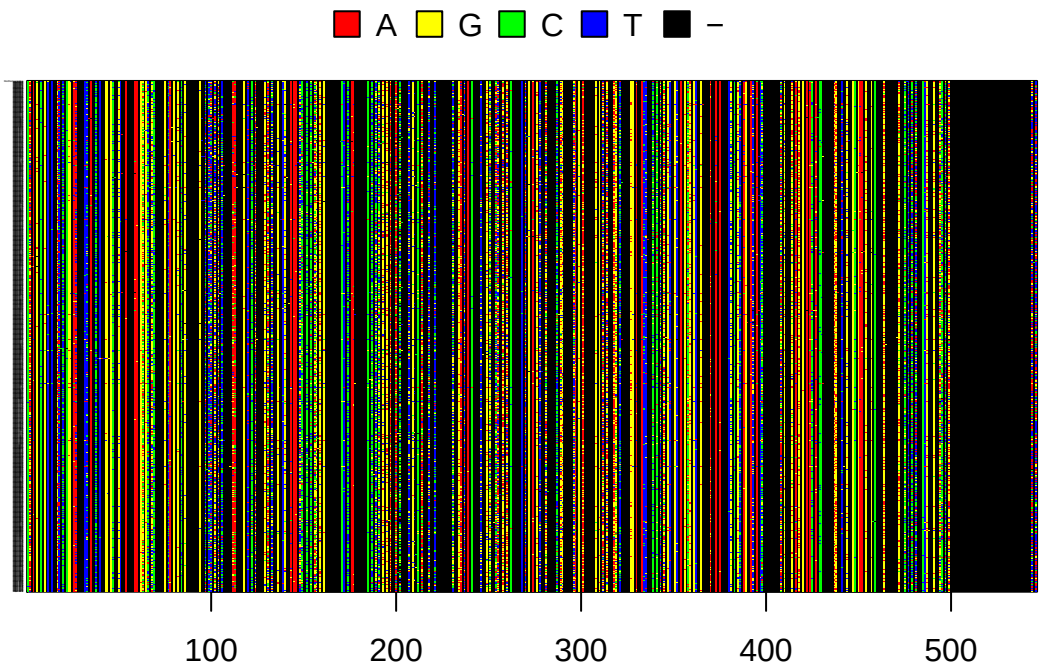
```
source("./bin/MothurTools.R")
```

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.

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

# load a select data
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))

# remove missing sites and zeros
comm <- comm[rownames(comm) %in% env$Sample_ID,]
comm <- comm[, colSums(comm) > 0]
```

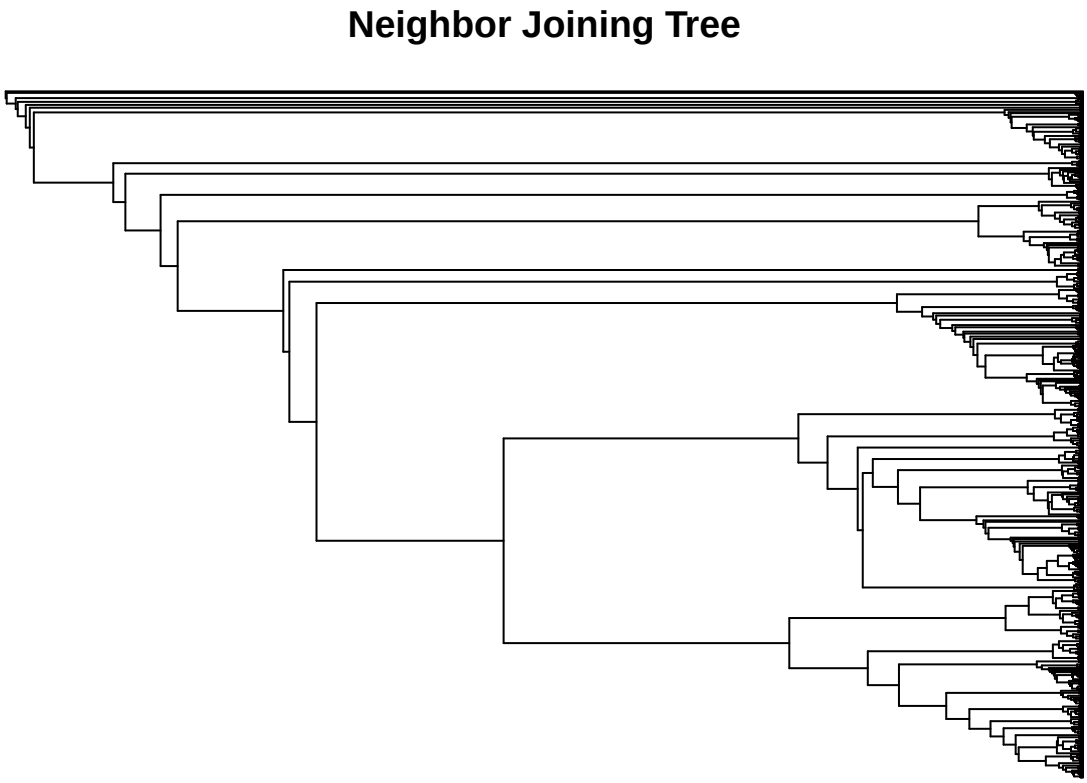
```
tax <- read.tax(taxonomy = "./data/INPonds.final.rdp.1.cons.taxonomy")

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

# import outgroup sequence
outgroup <- read.alignment(file = "./data/methanosarcina.fasta", format = "fasta")
DNAbin <- rbind(as.DNAbin(outgroup), as.DNAbin(ponds.cons))
image.DNAbin(DNAbin, shot.labels = T, cex.lab = 0.05, las = 1)
```

```
## Warning in plot.window(...): "shot.labels" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "shot.labels" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "shot.labels" is
## not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "shot.labels" is
## not a graphical parameter

## Warning in box(...): "shot.labels" is not a graphical parameter

## Warning in title(...): "shot.labels" is not a graphical parameter
```




```
# distence matrix
seq.dist.jc <- dist.dna(DNAbin, model = "JC", pairwise.deletion = FALSE)
#tree
phy.all <- bionj(seq.dist.jc)
phy <- drop.tip(phy.all, phy.all$tip.label[!phy.all$tip.label %in% c(colnames(comm), "Methanosarcina")])

# pull out the outgroup w.r.t the tree and root it
outgroup <- match("Methanosarcina", phy$tip.label)
phy <- root(phy, outgroup, resolve.root = TRUE)
```

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



4) PHYLOGENETIC ALPHA DIVERSITY

A. Faith's Phylogenetic Diversity (PD)

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

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
faith_pd <- pd(comm, phy, include.root = FALSE)
line <- lm(log(PD) ~ log(SR), faith_pd)
faith_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
```

```
anova(line)
```

```
## Analysis of Variance Table
##
## Response: log(PD)
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## log(SR)    1 0.46940 0.46940  1189.6 < 2.2e-16 ***
## Residuals 50 0.01973 0.00039
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

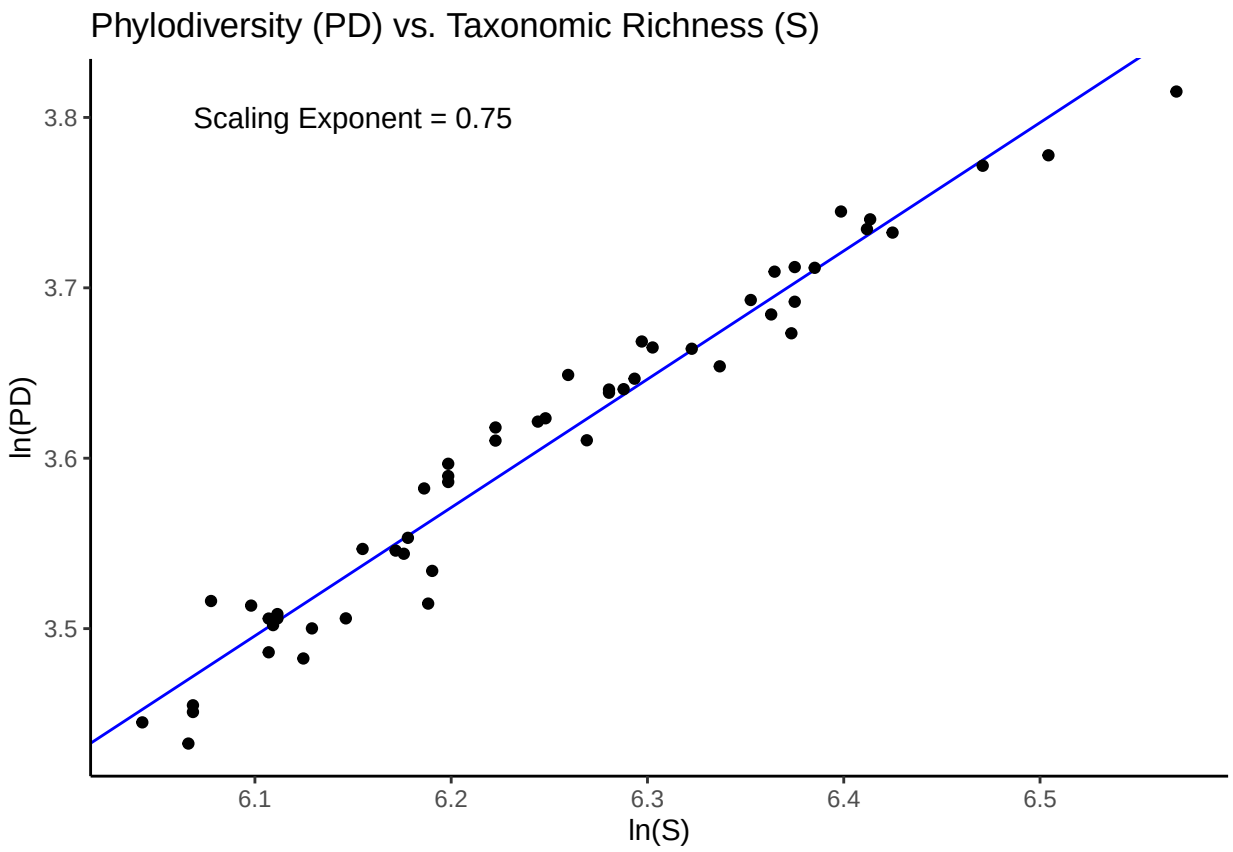
In the R code chunk below, do the following:

1. plot species richness (S) versus phylogenetic diversity (PD),
2. add the trend line, and
3. calculate the scaling exponent.

```
print(line)
```

```
##
## Call:
## lm(formula = log(PD) ~ log(SR), data = faith_pd)
##
## Coefficients:
## (Intercept)      log(SR)
##      -1.098        0.753
```

```
g <- ggplot(faith_pd, aes(x = log(SR), y = log(PD))) +
  geom_abline(slope = line$coefficients[2], intercept = line$coefficients[1], color = 'blue') +
  geom_point() +
  annotate("text", x = 6.15, y = 3.8, label = paste("Scaling Exponent = ", round(line$coefficients[2], 2
  xlab("ln(S)") +
  ylab("ln(PD)") +
  labs(title = "Phylodiversity (PD) vs. Taxonomic Richness (S)") +
  theme_classic()
g
```



Question 1: Answer the following questions about the PD-S pattern.
a. Based on how PD is calculated, why should this metric be related to taxonmic richness? b. Describe the relationship between taxonomic richness and phylodiversity. c. When would you expect these two estimates of diversity to deviate from one another? d. Interpret the significance of the scaling PD-S scaling exponent.

Answer 1a: PD requires different taxa to provide diversity but not all additional taxa will provide the same amount of phylodiversity. Initial samples are more likely to represent taxa less related to those already sampled thus we expect a saturating relationship for a given site. ***Answer 1b***: We see a sublinear power law relationship meaning that each additional taxon sampled provides less overall phylodiversity. ***Answer 1c***: These measures should diverge anytime there are a lot of taxa to sample when PD will saturate. ***Answer 1d***: The scaling exponent of 0.75 is highly significant giving us confidence that it is accurate.

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.

```
rich_ses <- ses.pd(comm[1:2,], phy, null.model = "richness", runs = 25,
                   include.root = FALSE)
labl_ses <- ses.pd(comm[1:2,], phy, null.model = "taxa.labels", runs = 25,
                   include.root = FALSE)
ppool_ses <- ses.pd(comm[1:2,], phy, null.model = "phylogeny.pool", runs = 25,
                   include.root = FALSE)

rich_ses
```

```
##       ntaxa   pd.obs pd.rand.mean pd.rand.sd pd.obs.rank   pd.obs.z  pd.obs.p
## BC001   668 43.71912     44.03658  0.8855939           9 -0.3584691 0.3461538
## BC002   587 40.94334     40.03402  0.7947700          21  1.1441312 0.8076923
##       runs
## BC001   25
## BC002   25
```

```
labl_ses
```

```
##       ntaxa   pd.obs pd.rand.mean pd.rand.sd pd.obs.rank   pd.obs.z  pd.obs.p
## BC001   668 43.71912     43.87270  0.7560382          13 -0.2031344 0.5000000
## BC002   587 40.94334     39.83877  0.7571344          25  1.4588818 0.9615385
##       runs
## BC001   25
## BC002   25
```

```
ppool_ses
```

```
##       ntaxa   pd.obs pd.rand.mean pd.rand.sd pd.obs.rank   pd.obs.z  pd.obs.p
## BC001   668 43.71912     44.22870  0.5837828           5 -0.8728875 0.1923077
## BC002   587 40.94334     39.98721  0.9172474          23  1.0423873 0.8846154
##       runs
## BC001   25
## BC002   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: I am testing three null hypotheses: that the identity of the species in the local site do not effect Faith's PD (`richness`); that the position of taxa on the phylogeny in our sites has no effect on the value of Faith's PD (`taxa.labels`); and that the local site is not different than the region in terms of Faith's PD (`phylogeny.pool`). The alternative hypotheses are that

there are effects on Faith's PD on these features. ***Answer 2b***: Under these null models, with the stochastic run I am looking at now, the interpretation is the same for all three. This is that at the second included pond, we almost see an effect and in the first we definitely don't. In the second pond PD is about 1 standard deviation off of the mean under each of the null models which is not a very powerful effect. This kind of makes sense in that each of these null models only allow changes to the site species identities but don't constrain anything about the species themselves. [I dont quite understand some of the null models so I've chosen ones that I do understand. In particular I dont get what maintaining species occurance frequency means? Is it the number of sites in which a species occurs?? `independentswap` is a null model I talked about in my quals today if so!]

B. Phylogenetic Dispersion Within a Sample

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

i. Phylogenetic Resemblance Matrix

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

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

ii. Net Relatedness Index (NRI)

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

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

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   0.001164092
## BC002   0.407149078
## BC003   0.825480708
## BC004  -0.312406100
## BC005   0.554963649
## BC010   0.357469482
## BC015  -0.145613145
## BC016   0.265827649
## BC018   0.096834474
## BC020   0.295282993
## BC048  -0.360129025
## BC049   0.054832439
## BC051  -0.991623817
## BC105  -0.697162860
## BC108  -0.512781447
## BC262  -0.343842367
## BCL01  -0.432353567
## BCL03  -0.783748873
## HNF132 -0.420769946
## HNF133  0.083353090
## HNF134  0.237560029
```

```
## HNF144 -0.600303624
## HNF168 -0.474429780
## HNF185  0.339676856
## HNF187  0.813627055
## HNF216  0.647786349
## HNF217 -0.154942178
## HNF221 -0.908075333
## HNF224 -0.009031772
## HNF225  0.570607040
## HNF229 -0.398153651
## HNF242 -0.114753536
## HNF250 -0.342255799
## HNF267 -0.452658899
## HNF269 -0.347779989
## YSF004 -0.454665502
## YSF117  0.737912302
## YSF295 -1.142902605
## YSF296  0.696066458
## YSF298  0.957493590
## YSF300  0.271985394
## YSF44   0.694809918
## YSF45   0.729630201
## YSF46   1.431920661
## YSF47   0.132383059
## YSF65   0.506981778
## YSF66  -0.297902105
## YSF67  -0.298247288
## YSF69  -0.096404796
## YSF70  -0.301639565
## YSF71   0.859190711
## YSF74   0.976864880
```

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   0.8770754
## BC002   1.8241387
## BC003   1.5387459
## BC004   1.4598034
## BC005   2.2916237
## BC010   0.7555187
## BC015   1.1281394
## BC016   1.8789328
## BC018   1.7272443
## BC020   1.2326310
## BC048   1.6248967
```

```
## BC049   2.4130552
## BC051   2.4106166
## BC105   1.6099244
## BC108   1.6890340
## BC262   1.2197982
## BCL01   1.4000997
## BCL03   1.0964019
## HNF132  1.5716123
## HNF133  1.6212925
## HNF134  1.5317678
## HNF144  1.5985305
## HNF168  0.7579868
## HNF185  1.9347786
## HNF187  0.5685072
## HNF216  0.2724065
## HNF217  0.3413237
## HNF221  0.6066441
## HNF224  1.4247972
## HNF225  0.4714239
## HNF229  1.5083953
## HNF242  1.9194966
## HNF250  1.2509613
## HNF267  0.9931219
## HNF269  1.1646262
## YSF004  0.5632992
## YSF117  2.0512585
## YSF295 -0.6288545
## YSF296  2.0368423
## YSF298  1.6361359
## YSF300  1.6752431
## YSF44   1.4151545
## YSF45   1.3064526
## YSF46   2.1159824
## YSF47   1.0305314
## YSF65   1.5723223
## YSF66   1.3326370
## YSF67   1.4970988
## YSF69   1.2568173
## YSF70   1.4031876
## YSF71   2.1267231
## YSF74   2.1070572
```

Question 3:

a. In your own words describe what you are doing when you calculate the NRI.
b. In your own words describe what you are doing when you calculate the NTI.
c. Interpret the NRI and NTI values you observed for this dataset.
d. In the NRI and NTI examples above, the arguments "abundance.weighted = FALSE" means that the indices were calculated using presence-absence data. Modify and rerun the code so that NRI and NTI are calculated using abundance data. How does this affect the interpretation of NRI and NTI?

Answer 3a: NRI is the offset of the average branch lengths within a site from a null distribution of the same measure but expressed in terms of null distribution standard deviations. In other words it is the negative Z-score for the average branch length for the focal site against a null distribution. ***Answer 3b***: NTI is the same idea—negative Z-score—except that the measure is

the phylogenetic distance from each taxon to its nearest neighbor averaged over taxa. ***Answer 3c***: Both of these measures, particularly NRI, indicate that sites are overdispersed with respect to the phylogeny. This is shown in the negative values and indicates that present taxa are less related than expected by chance. ***Answer 3d***: This change completely changes the results and instead reports that most sites are underdispersed. Maybe a small number of closely related and highly abundant species are present in each of these 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.

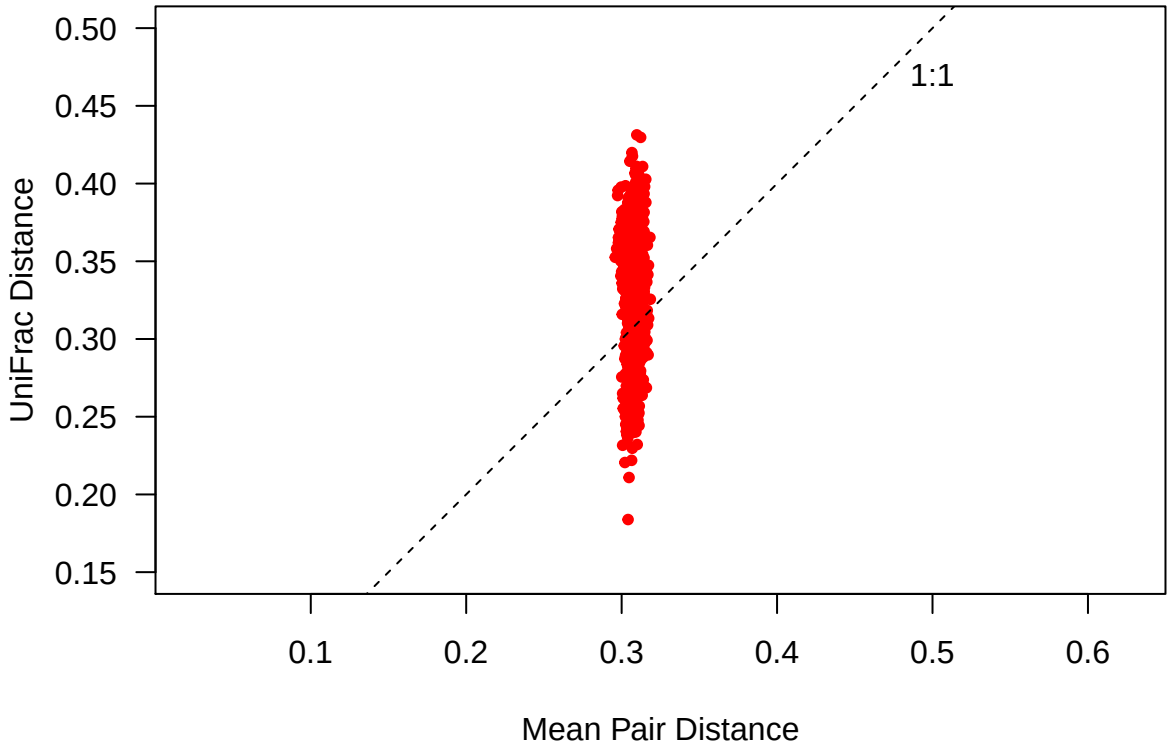
```
dist_mp <- comdist(comm, phydist)
```

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

```
dist_uni <- 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_uni, 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 what the name says: the mean pairwise phylogenetic distance between taxa. UniFrac is the sum of branch lengths for branches that are not in the intersection of both sites' set of branches normalized by the sum of the total branch lengths. UniFrac is not based on pairwise measures. ***Answer 4b***: These measures provide very different results. Mean pairwise distance shows much less variability than UniFrac. ***Answer 4c***: Thos is probably because of the fac that mean pairwise difference is always iusing the same tree without weights according to abundance so the differences between sites may just be short branches near te tips.

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.

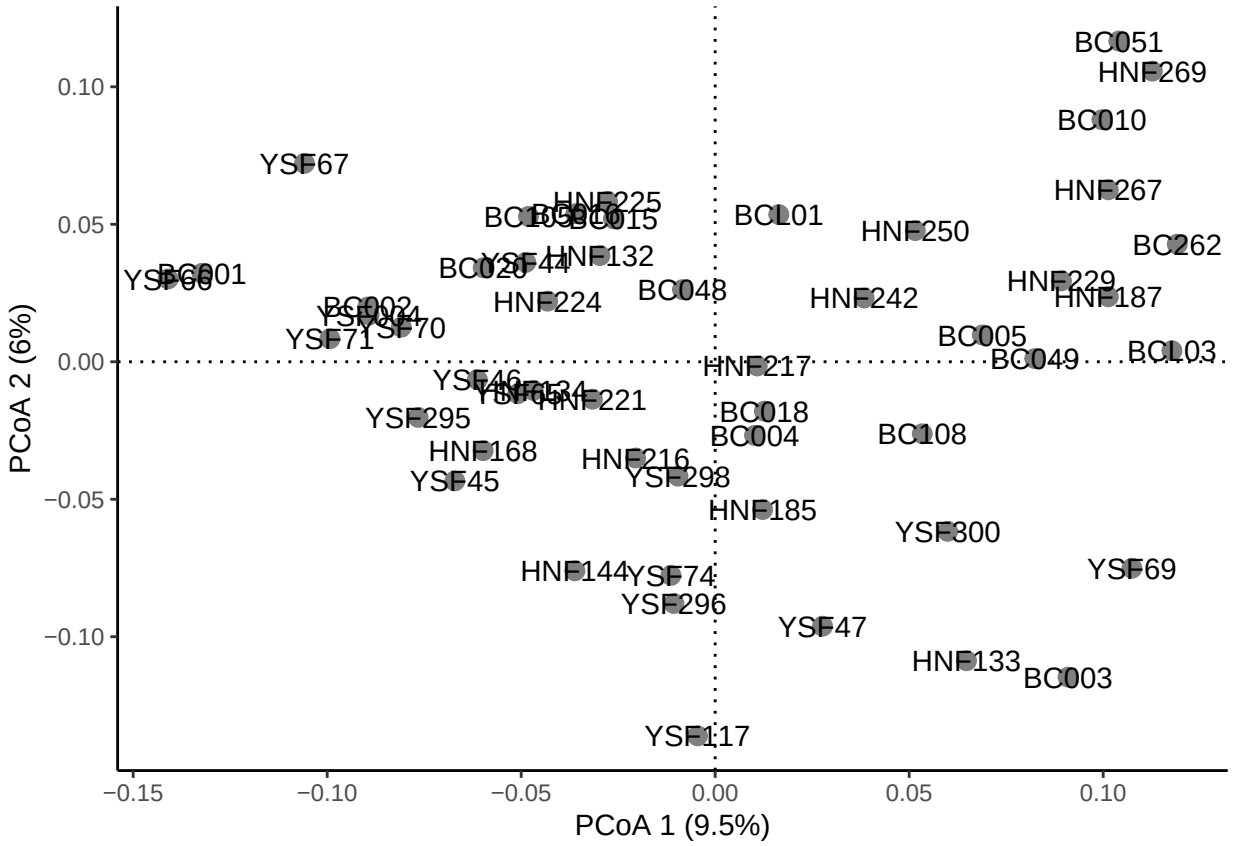
```
# calculate PCoA, explain variance, plot it
pond_pcoa <- cmdscale(dist_uni, eig = TRUE, k = 3)
expvar <- round(sapply(pond_pcoa$eig, function(x) x / sum(pond_pcoa$eig)), 3) * 100
plot_pcoa <- tibble::as_tibble(pond_pcoa$points)
```

```
## Warning: The `x` argument of `as_tibble.matrix()` must have unique column names if `.name_repair` is
## Using compatibility `.name_repair`.
```

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.

```
g <- ggplot(plot_pcoa, aes(x = V1, y = V2)) +
  geom_vline(xintercept = 0, linetype=3) +
  geom_hline(yintercept = 0, linetype=3) +
  geom_point(size = 3, color="grey50") +
  geom_text(label = row.names(pond_pcoa$points)) +
  xlab(paste("PCoA 1 (", expvar[1], "%)", sep = "")) +
  ylab(paste("PCoA 2 (", expvar[2], "%)", sep = "")) +
  theme_classic()
g
```



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.

```
# calculate PCoA, explain variance, plot it
comm_pcoa <- cmdscale(vegdist(comm, binary = TRUE), eig = TRUE, k = 3)
expvar <- round(sapply(comm_pcoa$eig, function(x) x / sum(comm_pcoa$eig)), 3) * 100
print(expvar)
```

```
##  [1] 10.4  6.4  5.3  4.6  3.9  3.7  3.5  3.1  3.0  2.9  2.7  2.6  2.5  2.5  2.3
## [16]  2.3  2.2  2.1  2.0  1.9  1.9  1.8  1.7  1.7  1.6  1.6  1.4  1.4  1.3  1.3
## [31]  1.2  1.2  1.1  1.1  1.0  1.0  0.9  0.8  0.8  0.8  0.7  0.7  0.7  0.6  0.5
## [46]  0.5  0.4  0.4  0.2  0.2  0.2  0.0
```

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***: We actually see less variance explained in the first few axes of the phylogenetically informed diatance matrix actually explains less variance than Sorenson dissimilarity. Maybe phylogenetic information is not as important as abundance information in this system.

C. Hypothesis Testing

i. Categorical Approach

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

```
watershed <- env$Location
adonis(dist_uni ~ watershed, permutations = 999)
```

```
##
## Call:
## adonis(formula = dist_uni ~ 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  0.022 *
## 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
```

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_uni, env_dist)
```

```
##
## Mantel statistic based on Pearson's product-moment correlation
##
## Call:
## mantel(xdis = dist_uni, ydis = env_dist)
##
## Mantel statistic r: 0.1604
##       Significance: 0.057
##
## Upper quantiles of permutations (null model):
##   90%   95% 97.5%   99%
## 0.119 0.168 0.207 0.252
## 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

```
require(dplyr)
```

```
## Loading required package: dplyr
```

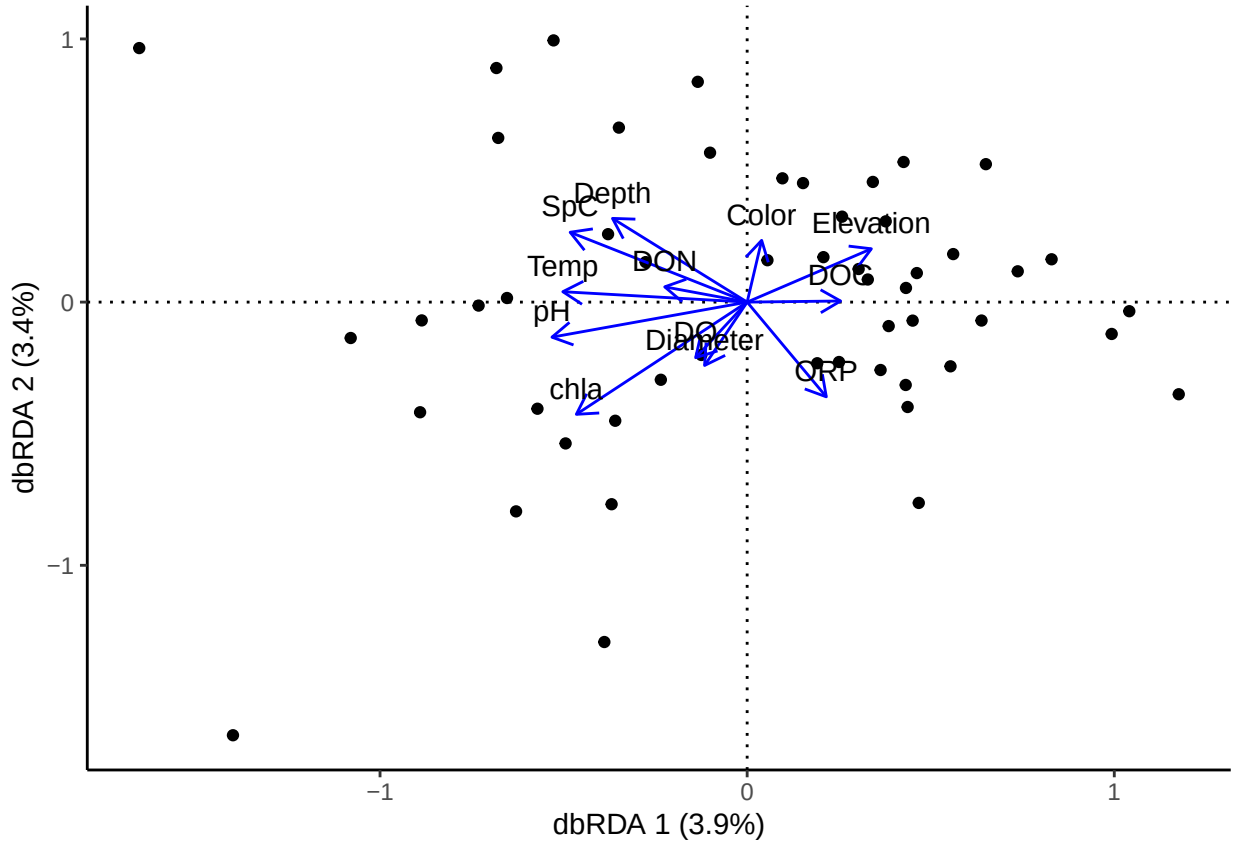
```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:reshape':
##
##     rename

## The following object is masked from 'package:seqinr':
##
##     count

## The following object is masked from 'package:nlme':
##
##     collapse

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# hypothesis test
ponds_dbrda <- vegan::dbrda(dist_uni ~ ., as.data.frame(scale(envs)))
#anova(ponds_dbrda, by = "axis")
#ponds_fit <- envfit(ponds_dbrda, envs, perm = 999)
#ponds_fit

# visualize
expvar <- round(sapply(ponds_dbrda$CCA$eig, function(x) x / sum(c(ponds_dbrda$CCA$eig, ponds_dbrda$CA$ei
plot_dbrda <- tibble::as_tibble(scores(ponds_dbrda)$sites)
env_vecs <- tibble::as_tibble(scores(ponds_dbrda, display = "bp")) %>% mutate(name = names(envs))
plot_dbrda
```

```
## # A tibble: 52 x 2
##    dbRDA1  dbRDA2
##     <dbl>   <dbl>
##  1  1.18  -0.350
##  2  0.451 -0.0704
##  3 -0.379  0.258
##  4  0.152  0.452
##  5 -0.349  0.663
##  6 -0.891 -0.419
##  7 -0.276  0.152
##  8  0.467 -0.763
##  9 -0.125 -0.201
## 10  0.304  0.126
## # ... with 42 more rows
```

```
g <- ggplot() +
  geom_hline(yintercept = 0, linetype = 3) +
  geom_vline(xintercept = 0, linetype = 3) +
  geom_point(data = plot_dbrda, aes(x = dbRDA1, y = dbRDA2)) +
  geom_segment(data = env_vecs, aes(x = 0, y = 0, xend = dbRDA1, yend = dbRDA2), color='blue', arrow = a
  geom_text(data = env_vecs, aes(x = dbRDA1, y = dbRDA2 + 0.1, label = name)) +
  xlab(paste("dbRDA 1 (", expvar[1], "%)", sep = "")) +
  ylab(paste("dbRDA 2 (", expvar[2], "%)", sep = "")) +
```

```
  theme_classic()
g
```



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***: Something was going wrong with my PERANOVA but the constrained ordination doesn't seem to give significant results at tha $\alpha = 0.05$ level (if I'm interpreting that correctly). The sites differ from one another in terms of manty environmental variables but none seem to really be powerful discriminators.

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.

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.

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

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

> ***Answer 7***:

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.

> In my own research, I almost require a phylogeny. I use concatenated ribosomal genes that *in principle* will make a nice well resolved phylogeny of bacterial species with and without sporulation genes. This will allow me to look at evolutionary patterns of sporulation loss and potentially even help me Identify different sets of sporulation genes corresponding to different shared evolutionary histories. I want to use my sporulation predictor to look at the kinds of communities in which sporulation may be more likely to be lost and I could use the phylogeny to look at dispersion in those communities.