12.Phylogenetic Diversity - Communities

Herbert Sizek; 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())
getwd()
```

```
## [1] "D:/GitHub/QB2021_Sizek/2.Worksheets/12.PhyloCom"
```

```
setwd("D:/GitHub/QB2021_Sizek/2.Worksheets/12.PhyloCom")
```

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.

```
package.list <- c('ape', 'seqinr',  'picante','vegan','fossil','reshape','simba','dplyr')
for (package in package.list) {
  if (!require(package, character.only=TRUE, quietly=TRUE)) {
    install.packages(package)
    library(package, character.only=TRUE)
  }
}
```

```
##
## Attaching package: 'seqinr'
```

```
## The following objects are masked from 'package:ape':
##
##     as.alignment, consensus
```

```
##
## Attaching package: 'permute'
```

```
## The following object is masked from 'package:seqinr':
##
##     getType
```

```
## This is vegan 2.5-7
```

```
##
## Attaching package: 'nlme'

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

##
## 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

##
## Attaching package: 'dplyr'

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

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

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

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

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

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

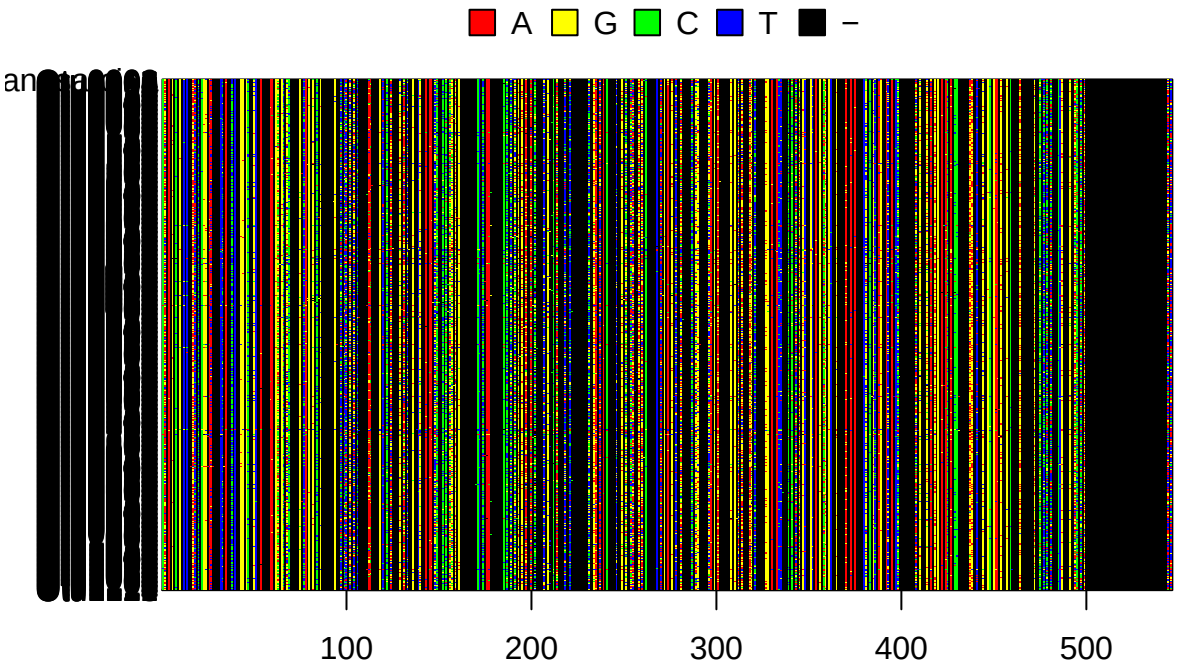
3) LOAD THE DATA

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=TRUE)
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,

```
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 = TRUE)
```



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.

```
seq.dist.jc <- dist.dna(DNAbin,model="JC",pairwise.deletion = FALSE)
phy.all <-bionj(seq.dist.jc)
phy <- drop.tip(phy.all,phy.all$tip.label[!phy.all$tim.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 = "NJ tree","phylogram",show.tip.label = FALSE,use.edge.length = FALSE,direction="
```

4) PHYLOGENETIC ALPHA DIVERSITY

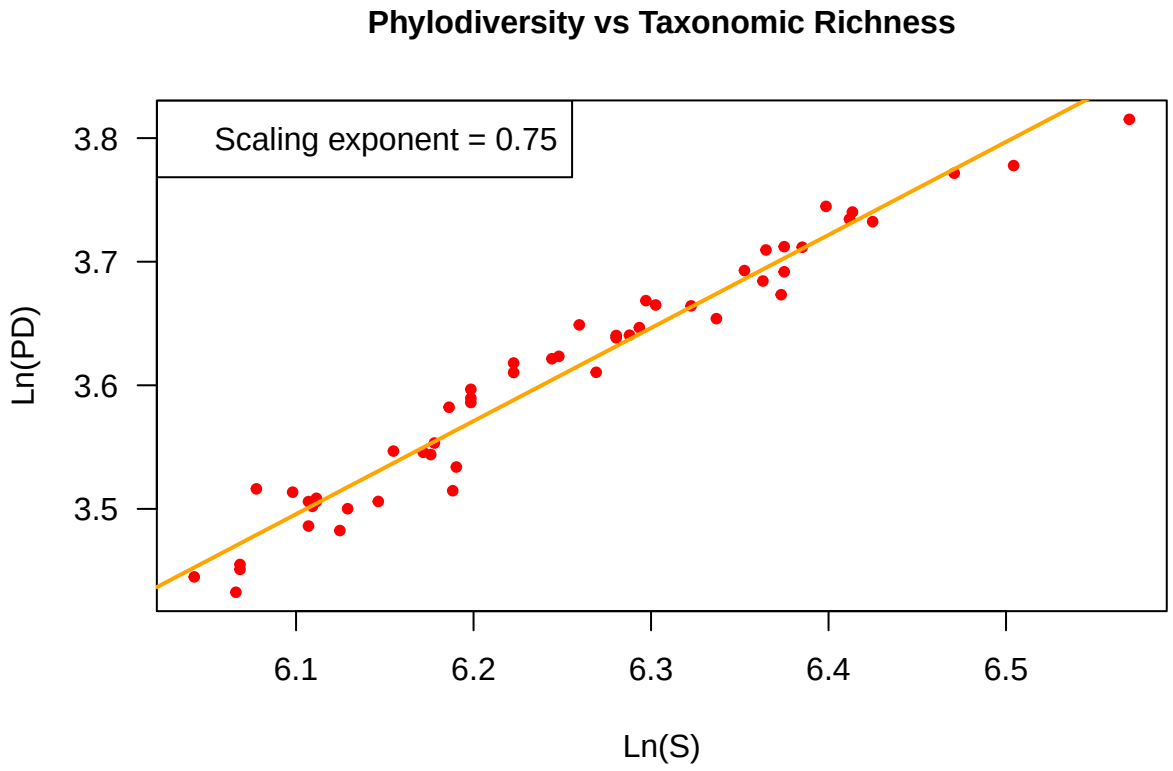
A. Faith's Phylogenetic Diversity (PD)

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="Phylod
fit <-lm('log(pd$PD) ~log(pd$S)')
abline(fit,col="orange",lw=2)
exponent <-round(coefficients(fit)[2],2)
legend("topleft",legend=paste("Scaling exponent =",exponent))
```



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***: The number of branches in the tree is 2(N-1). So as N increases, the possible value of Faith PD increases. Faith PD is the sum of all pairwise paths of a tree network. This means that as the average distance to the central node increases, Faiths PD increases. ***Answer 1b***: Taxonomic richness == species richness. I've describe in A the relations between Faith PD and the number of species and with the increase in distance from the central node of the tree
> ***Answer 1c***: As the number of species increases/diverge, Fiath's PD would have a higher scaling exponent, if the newer species diverge instead of older species. They shouldn't deviate from one another as long as the divergance pattern among all sites isn't different (one site that has older species diverging and another only having new species diverging) ***Answer 1d***: The scaling exponent falls out of the structure of the graph and the calculations done on the graph.

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(comm[1:2,],phy,null.model="richness",runs=25,include.root=FALSE)
```

```
##       ntaxa   pd.obs pd.rand.mean pd.rand.sd pd.obs.rank   pd.obs.z  pd.obs.p
## BC001   668 43.71912     44.23384  0.7337099           6 -0.7015274 0.2307692
## BC002   587 40.94334     40.37370  0.7975435          19  0.7142401 0.7307692
```

```
##       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***: What happens if the species at a site (richness) has a different evolutionary history.
> ***Answer 2b***:In the case of the frequency model, the difference is that species get shuffled between sites, so if there are species that are highly localized that get shuffled around sites, then it would increase ses.pd, along with the inverse (recently migrated species would decrease ses.pd or keep it the same). Under the richness there is not significant differences between BC001 and 002 and the random model.

```
ses.pd(comm[1:2,],phy,null.model="independentswap",runs=25,include.root=FALSE)
```

```
##       ntaxa   pd.obs pd.rand.mean pd.rand.sd pd.obs.rank  pd.obs.z   pd.obs.p
## BC001   668 43.71912     44.28892  0.4325345           2 -1.317347 0.07692308
## BC002   587 40.94334     40.26132  0.4544664          25  1.500693 0.96153846
##       runs
## BC001   25
## BC002   25
```

The independent swap model produces a more significant relationship as the distribution of the random model is narrower, because it constrains the possible null models more than the richness model.

B. Phylogenetic Dispersion Within a Sample

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

i. Phylogenetic Resemblance Matrix

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

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

ii. Net Relatedness Index (NRI)

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

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

```
##               NRI
## BC001  -2.0457568
## BC002  -3.8571010
## BC003  -1.5304121
## BC004  -3.6460632
## BC005  -3.3246214
## BC010  -2.9861745
## BC015  -2.5901020
## BC016  -1.8074829
## BC018  -2.2472351
## BC020  -1.5307874
## BC048  -1.8081896
## BC049  -0.7552927
## BC051  -2.7912703
## BC105  -2.6767777
## BC108  -2.4599868
## BC262  -2.2807607
## BCL01  -4.1320472
## BCL03  -2.2653639
## HNF132 -4.0313408
## HNF133 -2.3848881
## HNF134 -3.0116297
## HNF144 -3.5775813
## HNF168 -2.5946456
## HNF185 -3.4439243
## HNF187  0.2783119
## HNF216 -1.7849621
## HNF217 -2.8525620
## HNF221 -2.0375865
## HNF224 -3.1662965
## HNF225 -2.1100773
## HNF229 -1.1661935
## HNF242 -2.1511827
## HNF250 -2.4983251
## HNF267 -1.6950395
## HNF269 -2.8213461
## YSF004 -3.1758977
## YSF117 -2.1815612
## YSF295 -2.0496666
## YSF296 -2.0982130
## YSF298 -3.9746951
## YSF300 -1.7483486
## YSF44  -1.4652214
## YSF45  -1.8542848
## YSF46  -1.2928028
## YSF47  -2.1708560
## YSF65  -0.8330100
## YSF66  -1.3722900
## YSF67  -3.3239771
## YSF69  -1.5548861
## YSF70  -1.8645182
## YSF71  -1.3903267
## YSF74  -2.7262008
```

iii. Nearest Taxon Index (NTI)

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

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

```
##                NTI
## BC001   0.47355679
## BC002  -1.38126054
## BC003  -0.38459852
## BC004  -1.69418815
## BC005  -1.89456996
## BC010  -1.11637472
## BC015  -1.02118797
## BC016  -0.03721187
## BC018  -0.48777420
## BC020  -0.55518920
## BC048  -1.63090624
## BC049  -0.05172011
## BC051  -0.72007845
## BC105  -0.96109570
## BC108  -0.82817049
## BC262   0.15419845
## BCL01  -1.50613552
## BCL03  -0.08401903
## HNF132 -0.46579614
## HNF133 -1.23167914
## HNF134 -1.36365642
## HNF144 -1.58760824
## HNF168 -1.87062035
## HNF185 -0.94936541
## HNF187  0.25407855
## HNF216 -2.02309419
## HNF217 -1.60869430
## HNF221 -1.30722582
## HNF224 -2.11208429
## HNF225 -1.57524224
## HNF229 -0.04297678
## HNF242 -1.35931106
## HNF250 -0.56374177
## HNF267  1.12220130
## HNF269  0.49586550
## YSF004 -1.68760573
## YSF117 -1.40721396
## YSF295 -0.91225937
## YSF296 -0.05582595
## YSF298 -1.05445818
## YSF300 -1.28200266
## YSF44  -0.72240330
```

```
## YSF45  -1.16670568
## YSF46  -0.65361663
## YSF47  -0.80802951
## YSF65   0.46794314
## YSF66   0.70919380
## YSF67  -0.70139930
## YSF69   0.06720784
## YSF70  -0.16523330
## YSF71  -0.01850998
## YSF74  -1.75025481
```

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: It is the negative z-score of the mean phylogenetic distance.
Answer 3b: It is the negative z-score of the nearest neighbor on the phylogenetic tree. ***Answer 3c***: The negative values indicate that there is overdispersion. ***Answer 3d***:

```
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   0.206470281
## BC002   0.544189941
## BC003   0.719299296
## BC004   0.002246866
## BC005   0.459424721
## BC010   0.462670151
## BC015   0.220007174
## BC016   0.479736045
## BC018   0.410905562
## BC020   0.189363551
## BC048   0.011742477
## BC049   0.394773645
## BC051  -0.453617665
## BC105  -0.505958278
## BC108  -0.008164137
## BC262  -0.040901823
## BCL01  -0.146239526
## BCL03  -0.328458989
## HNF132 -0.130229881
## HNF133  0.165185696
## HNF134  0.371834894
## HNF144 -0.191900399
```

```
## HNF168 -0.087052896
## HNF185  0.541001271
## HNF187  0.771630046
## HNF216  0.724872880
## HNF217 -0.060901779
## HNF221 -0.292659251
## HNF224  0.233938008
## HNF225  0.733758965
## HNF229 -0.015609177
## HNF242  0.417022436
## HNF250 -0.039608244
## HNF267 -0.279381796
## HNF269 -0.044363873
## YSF004 -0.298436769
## YSF117  0.754533406
## YSF295 -0.527303971
## YSF296  1.111376864
## YSF298  0.882063836
## YSF300  0.346780335
## YSF44   0.545337451
## YSF45   0.696254804
## YSF46   1.450478478
## YSF47   0.288953107
## YSF65   0.444042134
## YSF66  -0.412026893
## YSF67  -0.097885225
## YSF69   0.100390031
## YSF70  -0.416696747
## YSF71   0.524723113
## YSF74   0.844780898
```

```
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   1.31548728
## BC002   1.82930878
## BC003   1.99461594
## BC004   1.48101878
## BC005   2.18797015
## BC010   0.36247320
## BC015   1.03614998
## BC016   1.84297740
## BC018   1.46845348
## BC020   1.17136750
## BC048   1.13740997
## BC049   1.74768864
## BC051   2.26867367
## BC105   1.35138292
## BC108   1.58935426
## BC262   1.15819088
```

```
## BCL01   1.51765159
## BCL03   0.93427706
## HNF132  1.53175761
## HNF133  2.04928567
## HNF134  1.50784913
## HNF144  0.74499414
## HNF168  0.59976662
## HNF185  1.47740013
## HNF187  0.36409192
## HNF216  0.08931756
## HNF217  0.18937237
## HNF221  0.73067642
## HNF224  1.22748983
## HNF225  0.21765512
## HNF229  1.64939733
## HNF242  1.35955335
## HNF250  1.49537776
## HNF267  1.14138546
## HNF269  1.19298251
## YSF004  0.17441460
## YSF117  1.75186980
## YSF295 -0.89237613
## YSF296  1.92702512
## YSF298  2.14776729
## YSF300  2.17598122
## YSF44   1.80578794
## YSF45   1.57045269
## YSF46   2.99125027
## YSF47   1.15262092
## YSF65   1.42251638
## YSF66   1.12732224
## YSF67   1.09826435
## YSF69   1.46093395
## YSF70   1.46203452
## YSF71   2.13880425
## YSF74   2.80371920
```

> This results in a more heterogeneous result where there is some overdispersion and underdispersion/clustering in the phylogonies. I'm not sure what this means practically. It suggests that the samples are have more close relations than the entire set of relations but I don't know how to interpret that physically.

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)
```

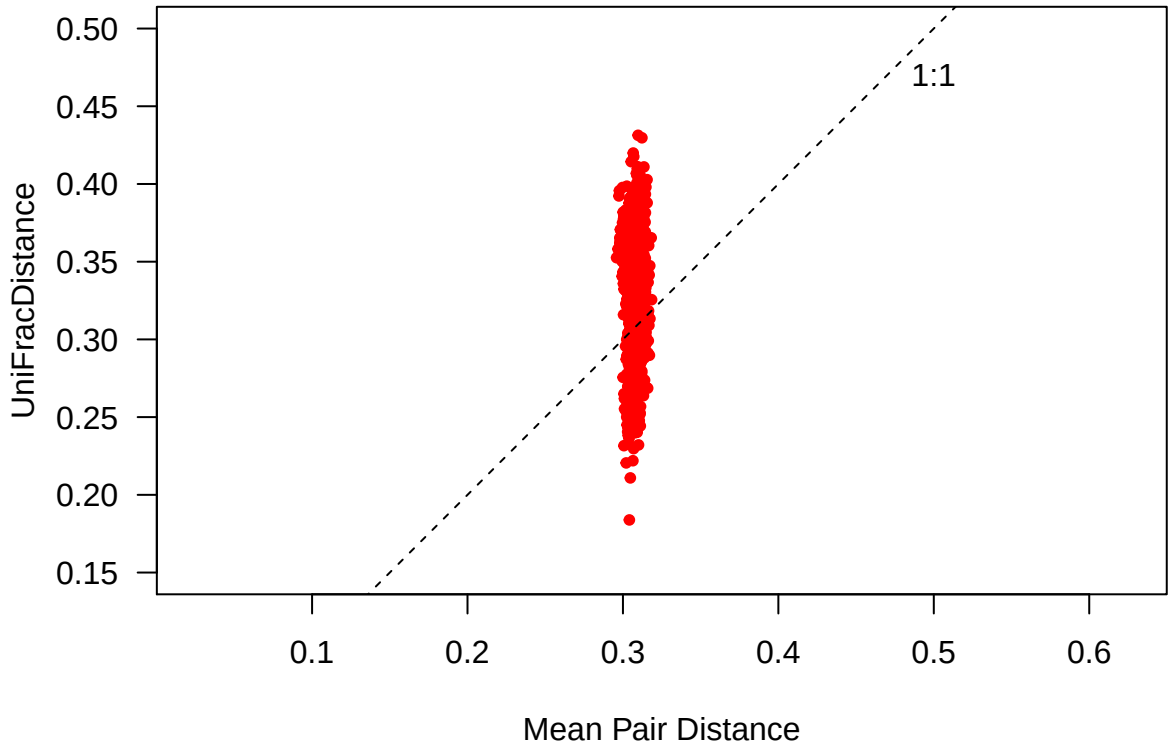
```
## [1] "Dropping taxa from the distance matrix because they are not present in the community data:"
```

```
##  [1] "Methanosarcina" "Otu0881"        "Otu0963"        "Otu0969"
##  [5] "Otu0984"        "Otu0990"        "Otu0991"        "Otu0997"
##  [9] "Otu0998"        "Otu1002"        "Otu1004"        "Otu1007"
## [13] "Otu1011"        "Otu1013"        "Otu1019"        "Otu1022"
## [17] "Otu1023"        "Otu1025"        "Otu1029"        "Otu1030"
## [21] "Otu1034"        "Otu1039"        "Otu1049"        "Otu1050"
## [25] "Otu1052"        "Otu1057"        "Otu1058"        "Otu1059"
## [29] "Otu1060"        "Otu1061"        "Otu1062"        "Otu1069"
## [33] "Otu1072"        "Otu1073"        "Otu1074"        "Otu1079"
## [37] "Otu1083"        "Otu1084"        "Otu1085"        "Otu1089"
## [41] "Otu1090"        "Otu1091"        "Otu1093"        "Otu1094"
## [45] "Otu1096"        "Otu1097"        "Otu1098"        "Otu1112"
## [49] "Otu1113"        "Otu1115"        "Otu1116"        "Otu1119"
## [53] "Otu1120"        "Otu1123"
```

```
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 Dis
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 takes all of the pairs of samples and basically calcualtes the correlation between them, while unifrac determines the shared tree components versus the unshared tree components, this provides a measure that will be less standardized across the entire tree. ***Answer 4b***: The mean pair distance is consistent across all communities, while Unifrac is not. ***Answer 4c***: Pairwise distances will not express the structure of the tree, just the distances on the tree, while unifrac is comparing the distances between species on the tree and then also comparing that to distances to the rooted specie.

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.

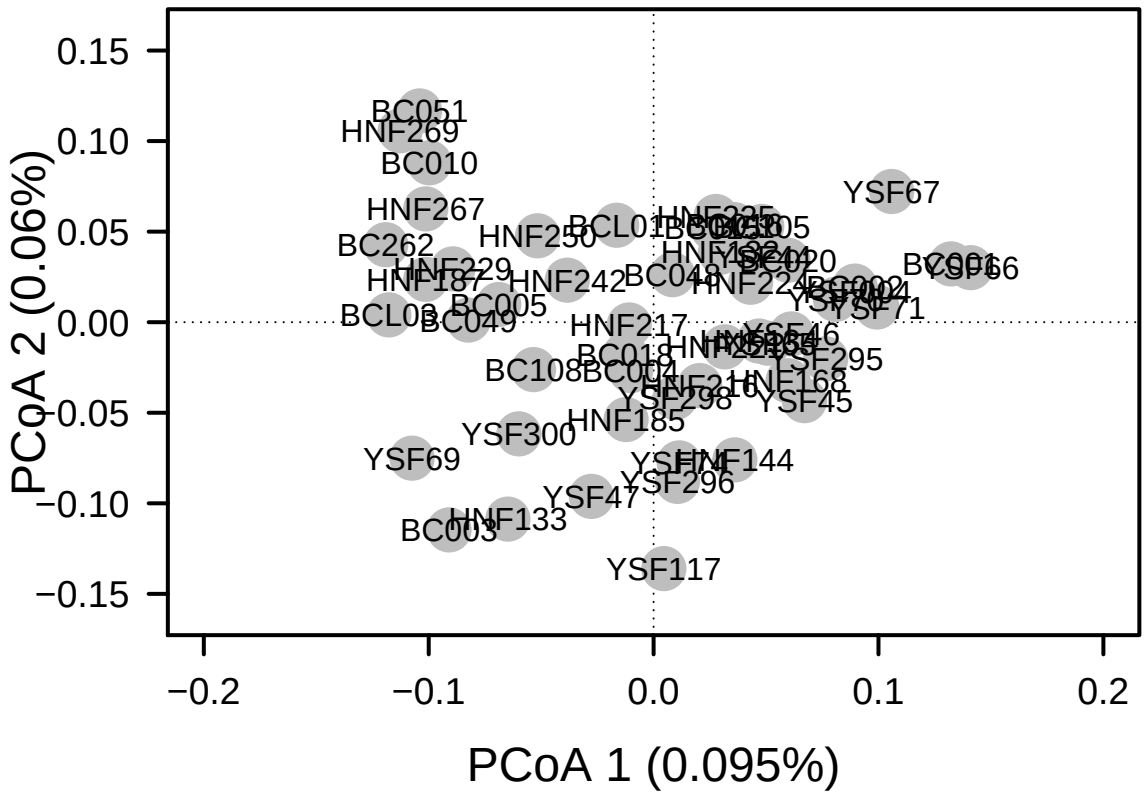
In the R code chunk below, do the following:
1. perform a PCoA based on the UniFrac distances, and
2. calculate the explained variation for the first three PCoA axes.

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

explainvar1 <- round(pond.pcoa$eig[1]/sum(pond.pcoa$eig),3)
explainvar2 <- round(pond.pcoa$eig[2]/sum(pond.pcoa$eig),3)
explainvar3 <- round(pond.pcoa$eig[3]/sum(pond.pcoa$eig),3)
sum.eig <- sum(explainvar1,explainvar2,explainvar3)
```

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

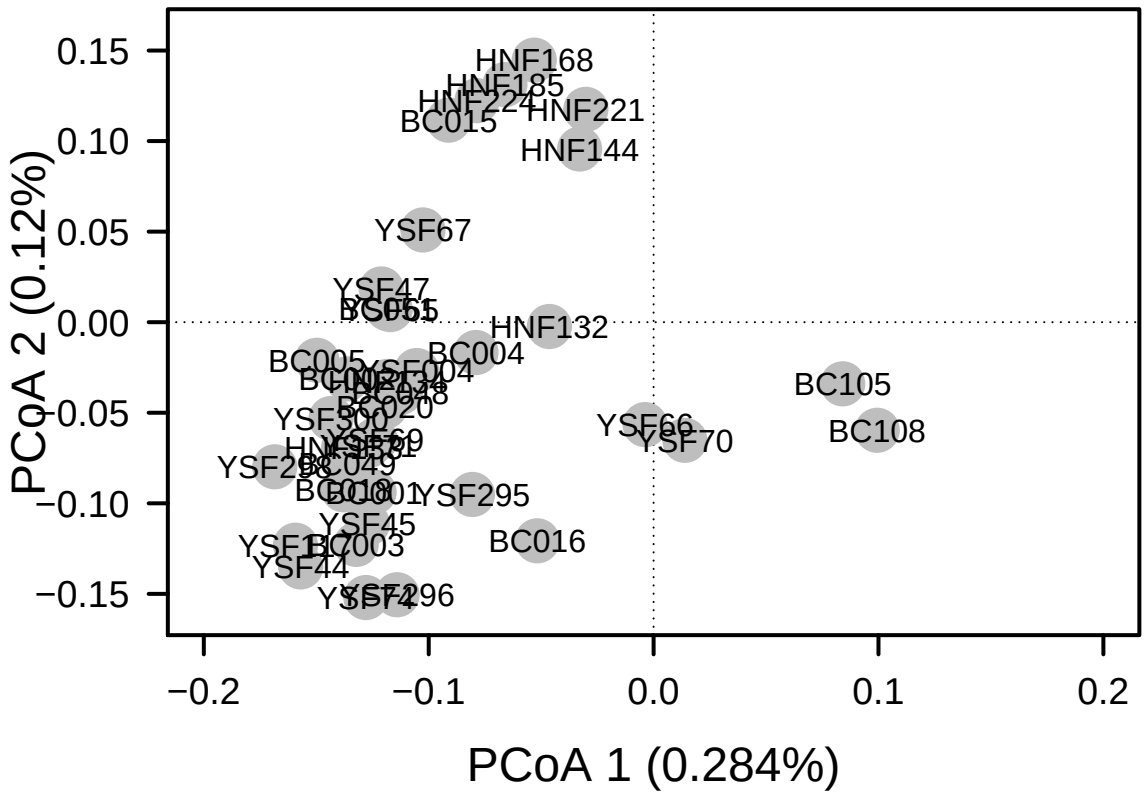
In the R code chunk below, do the following:
1. plot the PCoA results using either the R base package or the `ggplot` package,
2. include the appropriate axes,
3. add and label the points, and
4. customize the plot.

```
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,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.db.pcoa <- cmdscale(pond.db,eig=T,k=3)

explainvar1 <- round(pond.db.pcoa$eig[1]/sum(pond.db.pcoa$eig),3)
explainvar2 <- round(pond.db.pcoa$eig[2]/sum(pond.db.pcoa$eig),3)
explainvar3 <- round(pond.db.pcoa$eig[3]/sum(pond.db.pcoa$eig),3)
sum.eig <- sum(explainvar1,explainvar2,explainvar3)

par(mar = c(5,5,1,2)+0.1)
plot(pond.db.pcoa$points[,1],pond.db.pcoa$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,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.db.pcoa$points[,1],pond.db.pcoa$points[,2],
       pch=19,cex=3,bg="gray",col="gray")
text(pond.db.pcoa$points[,1],pond.db.pcoa$points[,2],
     labels=row.names(pond.db.pcoa$points))
```



Question 5: Using a combination of visualization tools and percent variation explained, how does the phylogenetically based ordination compare or contrast with the taxonomic ordination? What does this tell you about the importance of phylogenetic information in this system?

> ***Answer 5***: In the environmental PCoA, there is more variaion accounted for by the first component and it separates out particular sites, suggesting that there is some environmental variation that is fairly strong. The second PCoA similarly separates out a few envirnmental variables. The separation in the phylogonies does not suggest such separations, suggesting that the phylogonies are reducing the amount of information in the system, this means that the environmental variations are correlated with t he phylogienies.

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  0.026 *
## 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  0.004 **
## 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: 0.044
##
## Upper quantiles of permutations (null model):
##   90%   95% 97.5%   99%
## 0.117 0.154 0.178 0.217
## Permutation: free
## Number of permutations: 999
```

Matricies are possitively correlated most of the time.

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")
```

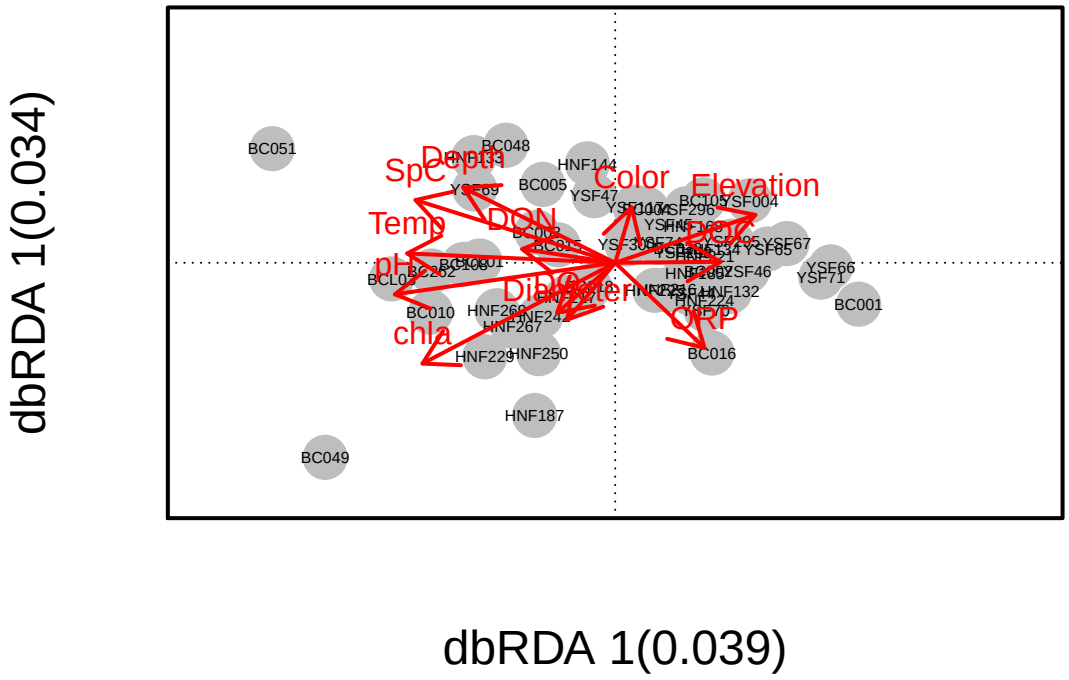
```
## Permutation test for dbrda under reduced model
## Forward tests for axes
## Permutation: free
## Number of permutations: 999
##
## Model: vegan::dbrda(formula = dist.uf ~ Elevation + Diameter + Depth + ORP + Temp + SpC + DO + pH + (
##          Df SumOfSqs      F Pr(>F)
## dbRDA1    1  0.10566 2.0152  0.431
## dbRDA2    1  0.09258 1.7658  0.634
## dbRDA3    1  0.07555 1.4409  0.970
## dbRDA4    1  0.06677 1.2735  0.997
## dbRDA5    1  0.05666 1.0807  1.000
## dbRDA6    1  0.05293 1.0095  0.999
## dbRDA7    1  0.04750 0.9059  1.000
## dbRDA8    1  0.03941 0.7517  1.000
## dbRDA9    1  0.03775 0.7201  1.000
## dbRDA10   1  0.03280 0.6256  1.000
## dbRDA11   1  0.02876 0.5485  1.000
## dbRDA12   1  0.02501 0.4770  0.999
## Residual 39  2.04482
```

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

```
##
## ***VECTORS
##
##             dbRDA1   dbRDA2     r2 Pr(>r)
## Elevation  0.77670  0.62986 0.0959  0.083 .
## Diameter  -0.27972 -0.96008 0.0541  0.259
## Depth     -0.63137  0.77548 0.1756  0.010 **
```

```
## ORP        0.41879 -0.90808 0.1437  0.023 *
## Temp      -0.98250  0.18628 0.1523  0.013 *
## SpC       -0.77101  0.63682 0.2087  0.004 **
## DO        -0.39318 -0.91946 0.0464  0.299
## pH        -0.96210 -0.27270 0.1756  0.002 **
## Color      0.06353  0.99798 0.0464  0.323
## chla      -0.60392 -0.79704 0.2626  0.008 **
## DOC        0.99847 -0.05526 0.0382  0.388
## DON       -0.91633  0.40042 0.0339  0.422
## ---
## 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)
dbrda.explainvar2 <- round(ponds.dbrda$CCA$eig[2]/sum(c(ponds.dbrda$CCA$eig,ponds.dbrda$CA$eig)),3)

par(mar=c(5,5,4,4)+0.1)
plot(scores(ponds.dbrda,display ="wa"),xlim = c(-2,2),ylim = c(-2,2),
            xlab=paste("dbRDA 1(",dbrda.explainvar1,")",sep=""),
            ylab=paste("dbRDA 1(",dbrda.explainvar2,")",sep=""),
     pch=16,cex=2,type="n",cex.lab=1.5,cex.axis=1.2,axes=FALSE)
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),col="red")
```



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 environmental variables do a poor job of expalining the distances in the phylogonies computed by unifrac among 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.

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.

> I don't get these methods, it seems to me that we are just continously manipulating different correlation matrices and seeing if we can find structures in them. I don't know how generated data would be structured within these systems and influence the outcome of the calculations, i.e. that the measures we're computing using the correlation matrices are not skewed by the calculations themselves and actually represent something real in the system. Here, we are calculating a correlation matrix for species, constructing a phylogeny for that matrix, then constructing another correlation matrix on top of that which we are comparing in structure to the initial correlation matrix (mantel test). I would think that the matrices would have structures based off of the calculations done on them, and we would have to correct for that bias. With the agricultural datasets I work with I don't have any phylogenetic data. In the crop world right now most crops varieties are designed to operate with low variation across environmental varients and practices aim to homogonize environmental conditions. Sometimes the vareity history is listed in the catalogue, but the public data is not there yet in detecting varieties from satallite imaging, just crop types. In terms of species relations work that I am starting with agregation of plant communities, the phylogonies of the plant communities could be useful later on, but I would want to first find relations from the data and then compare if species are more likely to associate with those related to them or those who are farther away. I would have to understand these methods better to apply them, I think I would want a more careful test than what was presented here and really go into the null model constructructions to make sure that the associations are not spurious and that we can replicate the distributions of measures in reality with careful construction of our null models.