12.Phylogenetic Diversity - Communities

Caroline Edwards; Z620: Quantitative Biodiversity, Indiana University

10 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("~/quant_bio/GitHub/QB2021_Edwards/2.Worksheets/12.PhyloCom/")

package.list <- c(’ape’, ’seqinr’, ’vegan’, ’fossil’, ’reshape’,’simba’, ’picante’)
for (package in package.list) {
  if (!require(package, character.only=TRUE, quietly=TRUE)) {
    install.packages(package, repos = ’https://cran.us.r-project.org’)
    library(package, character.only=TRUE)
  }
}
```

```
## Warning: package ’seqinr’ was built under R version 3.6.2
```

```
##
## 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 2.5-6
```

```
##
## 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:stats’:
##
##     mad
```

```
## Warning: package ’picante’ was built under R version 3.6.2
```

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

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

##
## Attaching package: 'picante'

## The following object is masked from 'package:simba':
##
##     mpd
```

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

2) DESCRIPTION OF DATA

need to discuss data set from spatial ecology!

In 2013 we sampled > 50 forested ponds in Brown County State Park, 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.

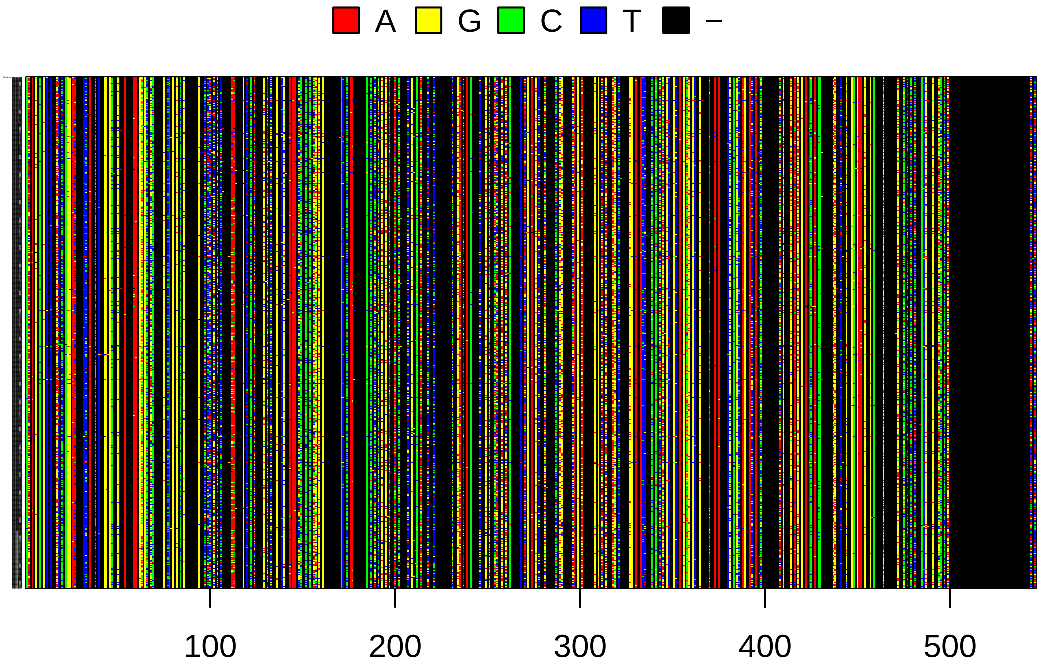
3) LOAD THE DATA

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.

```
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, ] #comment out this line if need be
comm<-comm[ , colSums(comm)>0]

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

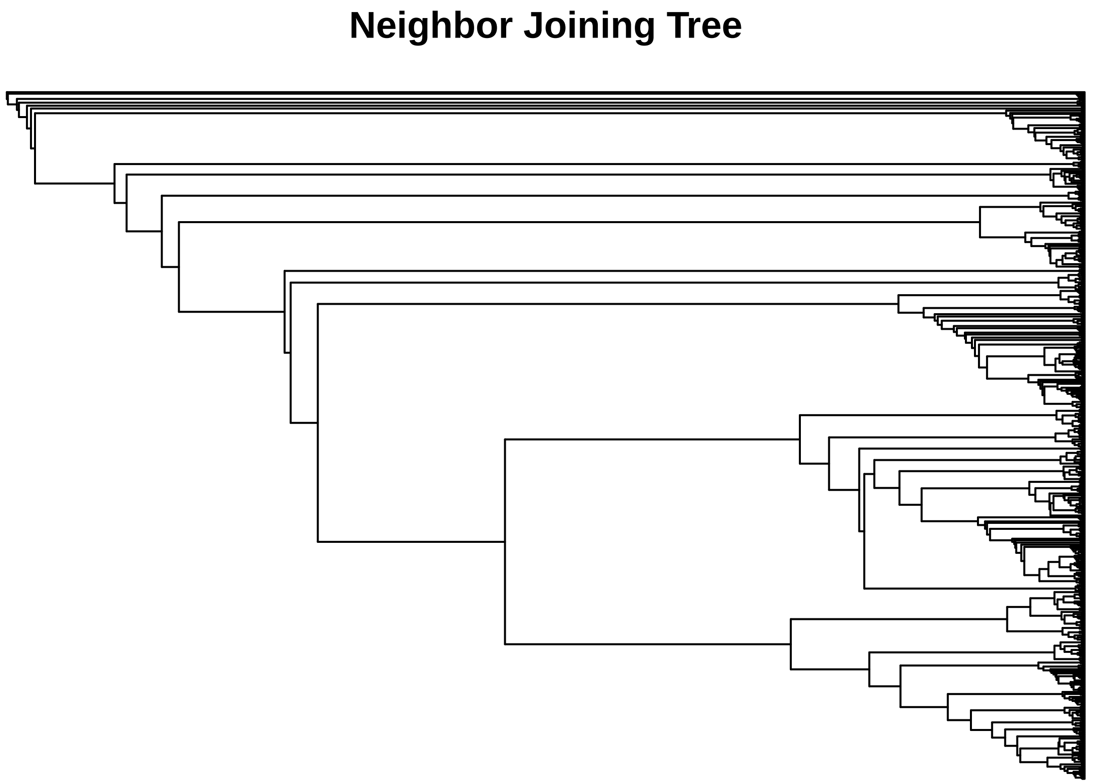
Next, in the R code chunk below, do the following:
1. load the FASTA alignment for the bacterial operational taxonomic units (OTUs),
2. rename the OTUs by removing everything before the tab (\t) and after the bar (|),
3. import the *Methanosarcina* outgroup FASTA file,
4. convert both FASTA files into 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 = FALSE)
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=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)

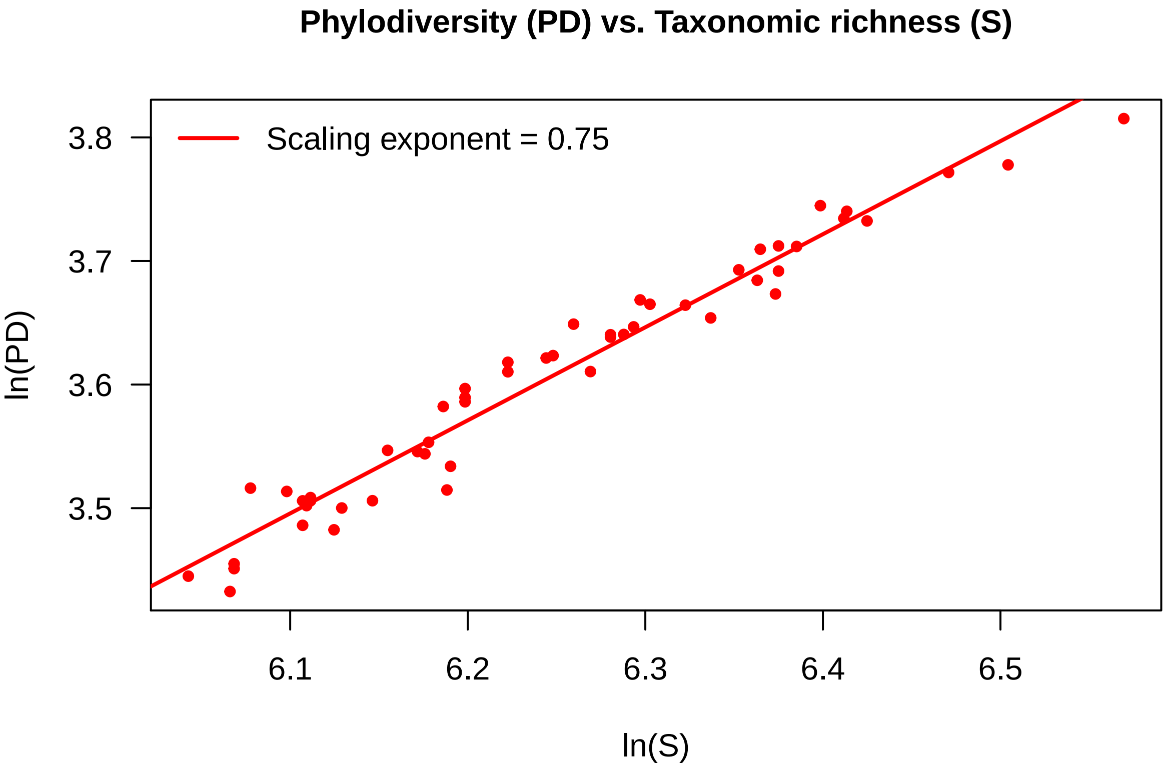
In the R code chunk below, do the following:
1. calculate Faith's D using the `pd()` function.

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

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

```
par(mar=c(5,5,4,1)+0.1)
plot(log(pd$S), log(pd$PD),
     pch=20, col="red", las=1,
     xlab="ln(S)", ylab="ln(PD)", cex.main=1,
     main="Phylodiversity (PD) vs. Taxonomic richness (S)")
```

```r
fit<-lm(’log(pd$PD)~log(pd$S)’)
abline(fit, col="red",lw=2)
exponent<-round(coefficients(fit)[2],2)
legend("topleft", legend=paste("Scaling exponent = ", exponent, sep=""),
       bty="n", lw=2, col="red")
```



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

> ***Answer 1a***: Phylodiversity is calculated by adding together all the branch lengths for each species in a sample, so it should be related to taxonomic richness because the higher the species richness (the more species exist in a community), the more likely the community is to have greater phylodiversity because with more species, you would expect to get more species that are related to varying degrees. ***Answer 1b***: The relationship between taxonomic richness and phylodiversity is of the type y=ax^k, since the relationship is linear with both axis log transformed. ***Answer 1c***: I would expect these two estimated of diversity to deviate from one another if your community has only very closely related species or very distantly related species, because the species richness will be the same, but the PD will be at either extreme. ***Answer 1d***: The scaling exponent is what defines how PD and S relate to each other, and since 0.75 is less than 1, we know that PD scales with S in less than a one to one relationship. PD gets bigger as S gets bigger, but at a slower rate than S.

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

Question 2: Using `help()` and the table above, run the `ses.pd()` function using two other null models and answer the following questions:

a. What are the null and alternative hypotheses you are testing via randomization when calculating `ses.pd`?
b. How did your choice of null model influence your observed ses.pd values? Explain why this choice affected or did not affect the output.

> ***Answer 2a***: The null model is what we expect our phylogenetic diversity to be when the data is resampled randomly with respect to certain things such as species richness or species occurrence frequency, and is used as a comparison. I tested null models "richness", "frequency", and "samples.pool", which randomized community data matrix abundances within samples, within species, and by drawing species from a regional pool. The alternative hypothesis is that the phylogenetic diversity is that the phylogenetic diversity is greater than expected by chance, which is supported when the observed PD is greater than the null PD (when ses.pd>0). ***Answer 2b***: The choice of null model didn't affect the observed ses.pd values, because those values are inherent to the dataset and won't change when the null model changes, just how those observed values compare to the new null distributions will change.

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"
```

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"

#rerun using abundance data
ses.mpd.a<-ses.mpd(comm, phydist, null.model="taxa.labels",
                 abundance.weighted = TRUE, runs = 25)
NRI_abund<-as.matrix(-1*((ses.mpd.a[,2]-ses.mpd.a[,3])/ses.mpd.a[,4]))
rownames(NRI_abund)<-row.names(ses.mpd)
colnames(NRI_abund)<-"NRI"

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

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: When you are calculating the NRI, you are calculating a metric that will help you determine whether your taxa are clustered or overdispersed. The NRI metric is calculated using the phylogenetic distances between pairwise species in your dataset and taking the average of those distances. You then take the observed mean distance and compare it to a null mean phylogenetic distance and standard deviation (generated through randomization). ***Answer 3b***: The NTI is another way to test for phylogenetic clustering or overdispersion. This metric is calculated in the exact same was as the NRI, except instead of using a mean phylogenetic distance, it used a mean nearest neighbor distance, which is calculated using the phylogenetic distance of all taxa and their closest related taxa. So instead of all pairwis phylogenetic distances, this metric only used the mean distance found between taxa and their nearest neighbor. ***Answer 3c***: The NRI values were almost all pretty negative, whereas the NTI values were mostly negative, but less so, and also had some positive values. Negative NRI values mean a sample is phylogenetically overdispersed, so I think these data are showing mostly patterns of overdispersion according to the NRI values, and since NTI emphasizes tip-level clustering, so maybe at the tips, these taxa are less overdispersed than when looking at all depths of the tree. ***Answer 3d***: When using abundance data instead of presence-absence data, the NRI values and NTI values all become more positive, with the NRI values being both positive and negative, and the NTI values switching to to all postive values. This changes the interpretation of the data, because it no longer seems like there is a strong signal for overdispersion, but there might even be some phylogenetic clustering.

5) PHYLOGENETIC BETA DIVERSITY

A. Phylogenetically Based Community Resemblance Matrix

In the R code chunk below, do the following:
1. calculate the phylogenetically based community resemblance matrix using Mean Pair Distance, and
2. calculate the phylogenetically based community resemblance matrix using UniFrac distance.

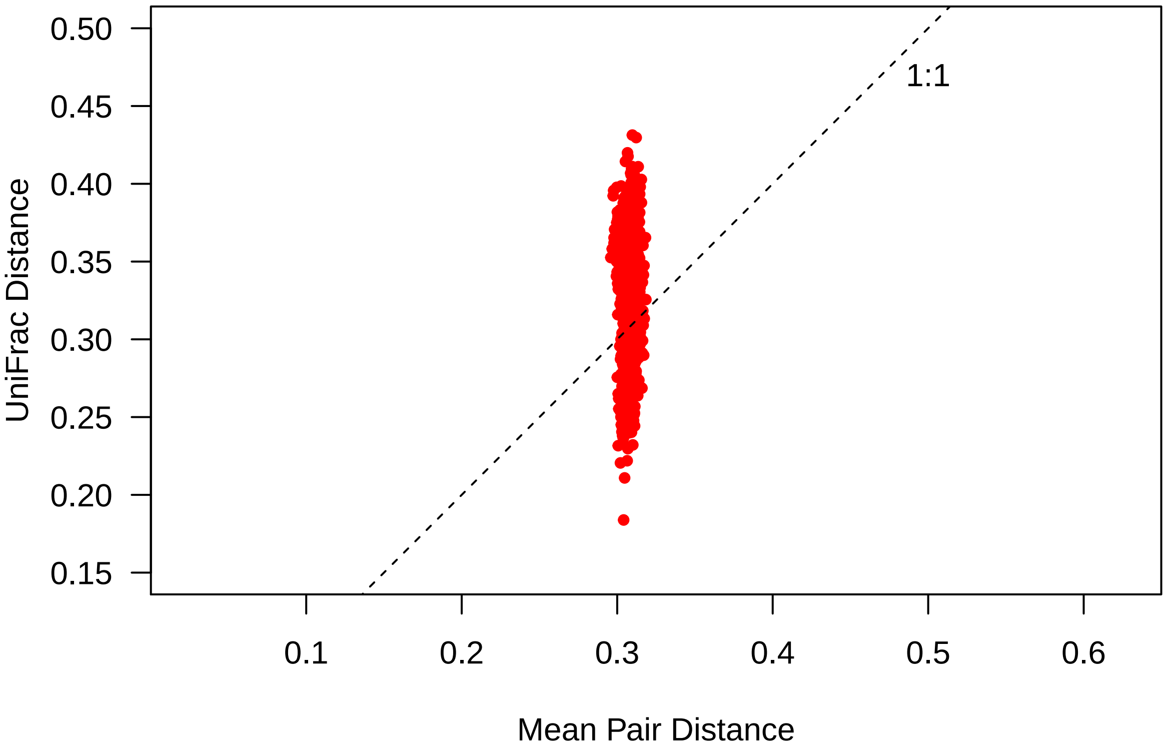
```
dist.mp<-comdist(comm, phydist)
```

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

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

In the R code chunk below, do the following:
1. plot Mean Pair Distance versus UniFrac distance and compare.

```
par(mar=c(5,5,2,1)+0.1)
plot(dist.mp, dist.uf,
     pch=20, col="red", las=1, asp=1, xlim=c(0.15,0.5), ylim=c(0.15,0.5),
     xlab="Mean Pair Distance", ylab="UniFrac Distance")
abline(b=1, a=0, lty=2)
text(0.5, 0.47, "1:1")
```



Question 4:

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

> ***Answer 4a***: Both of these metrics are used for making community resemblance matrices, but the mean pair distance method uses branch lengths to calculate how closely related two species are through this phylogenetic distance, whereas UniFrac distance is calculated more based on shared evolutionary history, using shared and unshared branches between two species to calculate their relatedness. ***Answer 4b***: There doesn't seem to be a correlation between mean pair distance and UniFrac distance, because all the mean pair distances around 0.3, but the UniFrac distances vary from ~0.20-0.45. There is far less variation in mean pair distance than in UniFrac distance. ***Answer 4c***: The MPD might show less bariation than UniFrac because MPD is a mean value.

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.

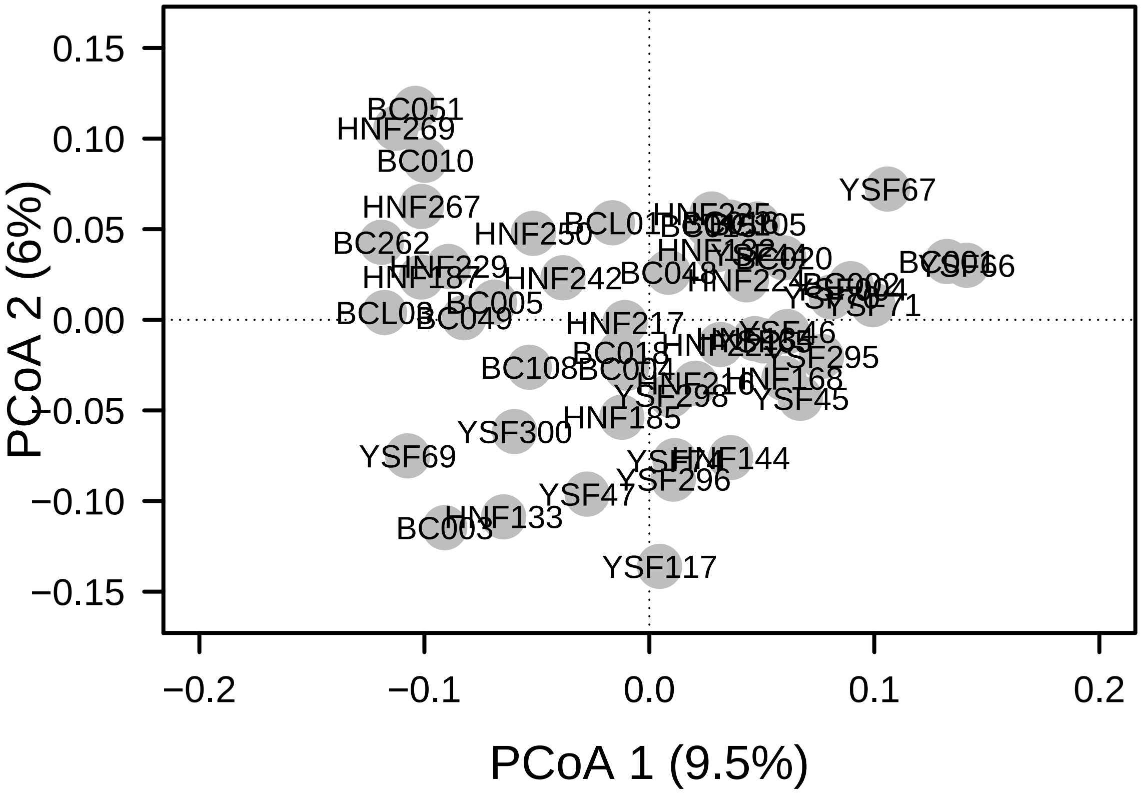
```r
pond.pcoa<-cmdscale(dist.uf, eig=T, k=3)
explainvar1<-round(pond.pcoa$eig[1]/sum(pond.pcoa$eig), 3)*100
explainvar2<-round(pond.pcoa$eig[2]/sum(pond.pcoa$eig), 3)*100
explainvar3<-round(pond.pcoa$eig[3]/sum(pond.pcoa$eig), 3)*100
sum.eig<-sum(explainvar1, explainvar2, explainvar3)
```

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

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

```r
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=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")
```

```r
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.

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***: This phylogenetically based ordination seems to have less variation explained by the two PCoA axis than previous taxonomic ordinations.

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.

```r
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  0.025 *
## 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  0.008 **
## 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: 0.068
##
## Upper quantiles of permutations (null model):
##   90%   95% 97.5%   99%
## 0.127 0.172 0.196 0.260
## Permutation: free
## Number of permutations: 999
```

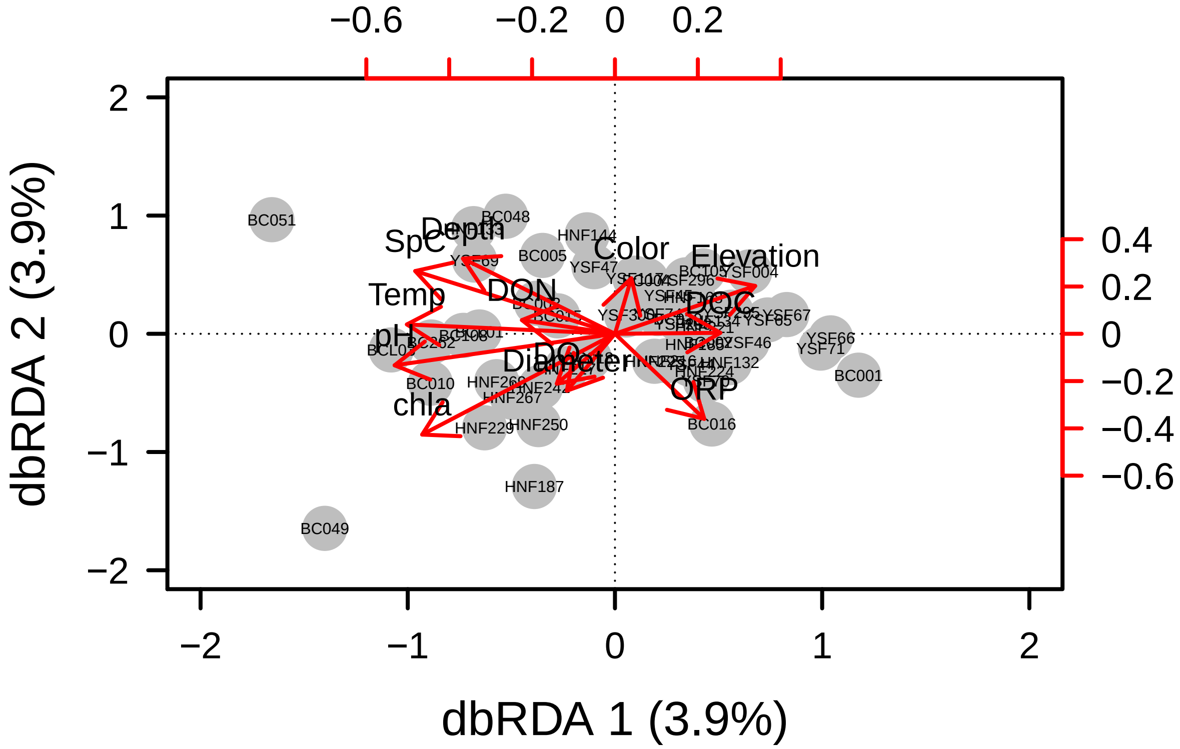
Last, conduct a distance-based Redundancy Analysis (dbRDA).

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

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

```
## Permutation test for dbrda under reduced model
## Forward tests for axes
## Permutation: free
## Number of permutations: 999
##
## Model: vegan::dbrda(formula = dist.uf ~ Elevation + Diameter + Depth + ORP + Temp + SpC + DO + pH + C
##          Df SumOfSqs      F Pr(>F)
## dbRDA1    1  0.10566 2.0152  0.462
## dbRDA2    1  0.09258 1.7658  0.632
## dbRDA3    1  0.07555 1.4409  0.980
## dbRDA4    1  0.06677 1.2735  0.995
## dbRDA5    1  0.05666 1.0807  1.000
## dbRDA6    1  0.05293 1.0095  1.000
## 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  0.998
## Residual 39  2.04482
```

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

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.explainvar1, "%)", sep=""),
     pch=16, cex=2.0, type="n", cex.lab=1.5, cex.axis=1.2, axes=FALSE)

axis(side=1, labels=T, lwd.ticks = 2, cex.axis=1.2, las=1)
axis(side=2, labels=T, lwd.ticks = 2, cex.axis=1.2, las=1)
abline(h=0, v=0, lty=3)
box(lwd=2)

points(scores(ponds.dbrda, display = "wa"),
       pch=19, cex=3, bg="gray", col="gray")
text(scores(ponds.dbrda, display = "wa"),
     labels = row.names(scores(ponds.dbrda, display = "wa")), cex = 0.5)

vectors<-scores(ponds.dbrda, display = "bp")
arrows(0, 0, vectors[,1]*2, vectors[,2]*2,
       lwd=2, lty=1, length=0.2, col="red")
text(vectors[,1]*2, vectors[,2]*2, pos=3,
     labels = row.names(vectors))
axis(side=3, lwd.ticks = 2, cex.axis=1.2, las=1, col="red", lwd=2.2,
     at=pretty(range(vectors[,1]))*2, labels= pretty(range(vectors[,1])))
axis(side=4, lwd.ticks = 2, cex.axis=1.2, las=1, col="red", lwd=2.2,
     at=pretty(range(vectors[,2]))*2, labels= pretty(range(vectors[,2])))
```



Question 6: Based on the multivariate procedures conducted above, describe the phylogenetic patterns of β-diversity for bacterial communities in the Indiana ponds.

> ***Answer 6***: From the PERMANOVA test, there is a significant effect of watershed on the phylogenetic diversity of bacterial communities, but it isn't a large effect (r2=0.06). The mantel test also tells us that the phylogenetic diversity of bacteria are significantly correlated with the environmental variables, with an r statistic of 0.16. The dbRDA shows the same type of result, that the environmental variables significantly effect the bacteria communities phylogenetic diversity, but the first two axes each only explains 3.9% of the variation, which is pretty low.

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.

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

bray.curtis.dist<- 1 - vegdist(comm)
unifrac.dist<- 1 - dist.uf

unifrac.dist.ls<-liste(unifrac.dist, entry="unifrac")
bray.curtis.dist.ls<-liste(bray.curtis.dist, entry="bray.curtis")
coord.dist.ls<-liste(coord.dist, entry="geo.dist")
env.dist.ls<-liste(env.dist, entry="env.dist")

df<-data.frame(coord.dist.ls, bray.curtis.dist.ls[,3], unifrac.dist.ls[,3],
               env.dist.ls[,3])
names(df)[4:6]<-c("bray.curtis", "unifrac", "env.dist")
```

Now, let’s plot the DD relationships:
In the R code chunk below, do the following:
1. plot the taxonomic distance decay relationship,
2. plot the phylogenetic distance decay relationship, and
3. add trend lines to each.

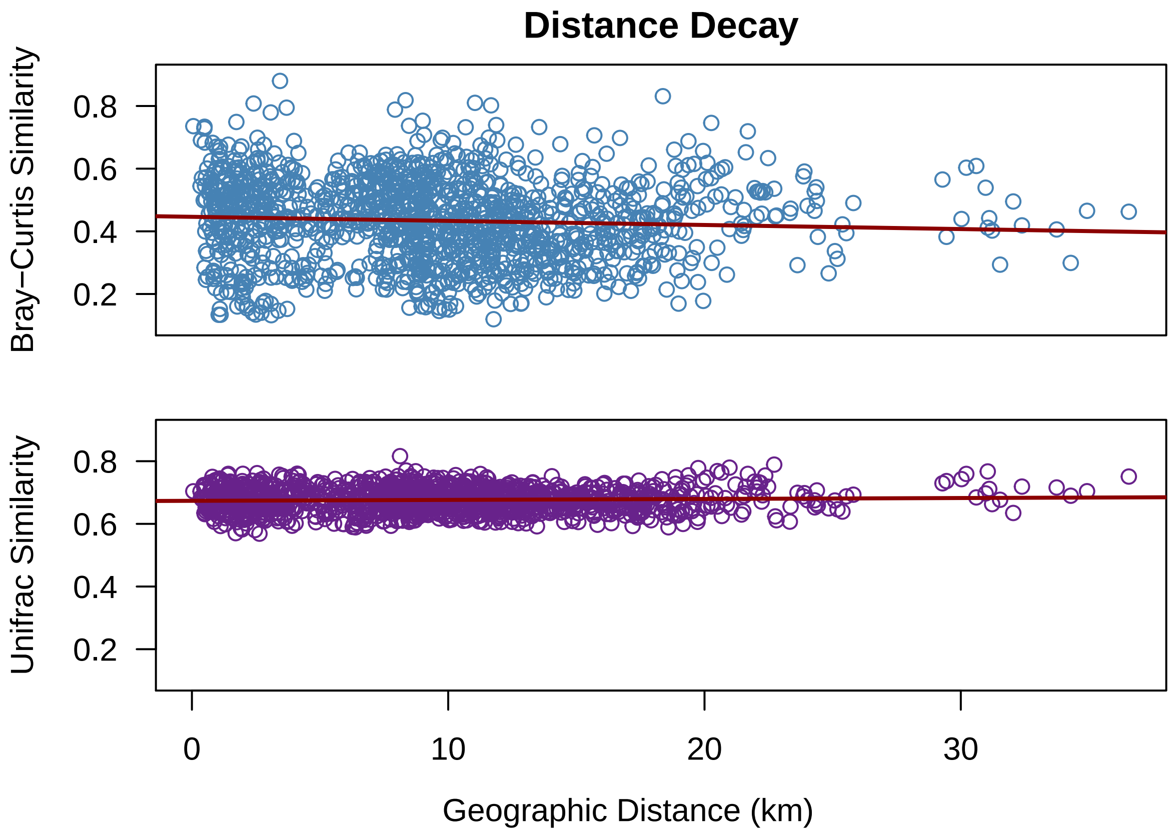
```r
par(mfrow=c(2,1), mar=c(1,5,2,1)+0.1, oma=c(2,0,0,0))
plot(df$geo.dist, df$bray.curtis, xlab="", xaxt="n", las=1, ylim=c(0.1,0.9),
     ylab="Bray-Curtis Similarity",
     main="Distance Decay", col="SteelBlue")
DD.reg.bc<- lm(df$bray.curtis ~ df$geo.dist)
summary(DD.reg.bc)
```

```
##
## 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
```

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

par(mar=c(2,5,1,1)+0.1)
plot(df$geo.dist, df$unifrac, xlab="", las=1, ylim=c(0.1, 0.9),
     ylab="Unifrac Similarity", col="darkorchid4")
DD.reg.uni<-lm(df$unifrac~df$geo.dist)
summary(DD.reg.uni)
```

```
##
## Call:
## lm(formula = df$unifrac ~ df$geo.dist)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.105629 -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
```

```r
abline(DD.reg.uni, col="red4", lwd=2)
mtext("Geographic Distance (km)", side=1, adj=0.55,
      line=0.5, outer=TRUE)
```



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

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

```
##
## Is difference in slope significant?
## Significance is based on 1000 permutations
##
## Call:
## diffslope(x1 = df$geo.dist, y1 = df$unifrac, x2 = df$geo.dist,      y2 = df$bray.curtis)
##
## Difference in Slope: 0.001603
## Significance: 0.005
##
## Empirical upper confidence limits of r:
##      90%      95%    97.5%      99%
## 0.000794 0.000945 0.001127 0.001421
```

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

> ***Answer 7***: These slopes seem pretty flat, with community similarities only decreasing slighlty with increased geographic distance. There is a significant different between the slopes of the two different distance metrics, which could be because bray-curtis uses abundance data and unifrac doesn't, however, the magnitude of the different seems pretty small.

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.

> Phylogenetic information would be useful in my own research for doing comparative studies to look at how strong different reproductive isolating barriers are for species pairs that are at different time since divergence. This is a common study setup, and used a phylogenetic distance matrix and also often environmental dissimilarity matrix because ecological divergence can be a premating isolating barrier, especially in plants. Spatial distance matrices are also used to show whether species pairs are more fully reproductively isolated based on geographic distance.