7. Worksheet: Diversity Synthesis

Thomas Zambiasi; Z620: Quantitative Biodiversity, Indiana University

15 February, 2023

OVERVIEW

In this worksheet, you will conduct exercises that reinforce fundamental concepts of biodiversity. Specifically, you will construct a a site-by-species matrix by sampling confectionery taxa. With this primary data structure, you will then answer questions and generate figures using tools from previous weeks, along with wrangling techniques that we learned about in class.

Directions:

1. In the Markdown version of this document in your cloned repo, change "Student Name" on line 3 (above) to your name.
2. Complete as much of the worksheet as possible during class.
3. Refer to previous handouts to help with developing of questions and writing of code.
4. Answer questions in the worksheet. Space for your answer is provided in this document and indicated by the ">" character. If you need a second paragraph be sure to start the first line with ">". 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, **push** this file to your GitHub repo.
6. For the assignment portion of the worksheet, follow the directions at the bottom of this file.
7. When you are done, **Knit** the text and code into a PDF file.
8. After Knitting, submit the completed exercise by creating a **pull request** via GitHub. Your pull request should include this file `7.DiversitySynthesis_Worskheet.Rmd` and the PDF output of `Knitr` (`DiversitySynthesis_Worskheet.pdf`).

CONFECTIONARY EXERCISE GOALS

We will construct a site-by-species matrix using confectionery taxa (i.e, gummies). The instructors have created distinct **sources communities** that vary in the composition of gummy taxa with even and uneven communities. It might be fun to consider them as distinct geographical regions experiencing different environmental regimes, or different experimental units under different treatments. Each student will sample a source community and then use a taxonomic key to identify gummies and their abundances.

In the end, students will use the site-by-species matrix to:

1) explore their sampling efforts and their effects on species richness using **coverage** and **rarefaction** concept,

2) measure **alpha diversity** for each sub-sample collated from data with their peers from the same source community,

3) examine **beta diversity** between each source community using the data generated across each source community, and

4) use **data wrangling** tools they have learned during the class to accomplish the above goals.

SAMPLING PROTOCOL TO CONSTRUCT A SITE-BY-SPECIES MATRIX

1. Instructors will assign you to sample confectionery taxa from one of the two designated source community bucket (A and B).
2. After randomly sampling one unit (imagine as an equal biomass) from the source community, each student will count the total number of individuals (N), identify the taxa using the species key and quantify the abundance of each taxon.
3. Work with other students in your group to assemble data into a site-by-species matrix on the white board. One person needs to create a .csv or .txt file and share your group's site-by-species matrix with the class using GitHub. Make sure that you include a sample identifier (student name) and what community you sampled from.

GROUP BRAINSTORM

In smaller groups, take 15 minutes to brainstorm questions, code, statistical tests, and "fantasy figures" using the site-by-species matrix the class generated.

1. Using this data, explore how well your sampling effort was. You can use rarefaction and coverage tools you have learned earlier.
2. Investigate alpha diversity based on the methods you have learned in the rest of the handout and accompanying worksheet. For example, you can measure richness, Shannon diversity and Simpson index. You can also convert them to effective number of species using the Hill numbers concept.
3. Measure beta diversity using ordination and multivariate statistical methods. For example, you can create a PCoA plot, based on Bray-Curtis dissimilarity, of sites and communities using different shape and color codes. Use Permanova to test if there are differences between communities.

```
#Workspace setup
rm(list=ls())
getwd()
```

```
## [1] "C:/Users/tmzam/GitHub/QB2023_Zambiasi/2.Worksheets/7.DiversitySynthesis"
```

```
setwd("C:/Users/tmzam/GitHub/QB2023_Zambiasi/2.Worksheets/7.DiversitySynthesis")

#loading packages
package.list<-c("vegan","tidyverse","ggplot2","dplyr","broom")
for (package in package.list){
  if(!require(package,character.only=TRUE,quietly=TRUE)){
    install.packages(package)
  }
  library(c(package),character.only=TRUE)
}
```

```
## Warning: package 'vegan' was built under R version 4.2.2
```

```
## This is vegan 2.6-4

## Warning: package 'tidyverse' was built under R version 4.2.2

## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --

## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.3     v forcats 0.5.2

## Warning: package 'readr' was built under R version 4.2.2

## Warning: package 'forcats' was built under R version 4.2.2

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
#loading data
gummy.data<-read.csv("QB23_gummydiv.csv")
```

DATA ANALYSIS

1) Sampling coverage and rarefaction curves

Question 1: Using this data, explore how well your sampling effort was. Compare your sampling efforts with other groups. Do you think that your samples cover the actual diversity found in each source community? You can use rarefaction and coverage tools you have learned earlier.

Answer 1: Use the space below to generate a rarefaction curve/sample coverage based on the data we collected in class for each community. Make sure to annotate your code using # symbols so others (including instructors) understand what you have done and why you have done it.

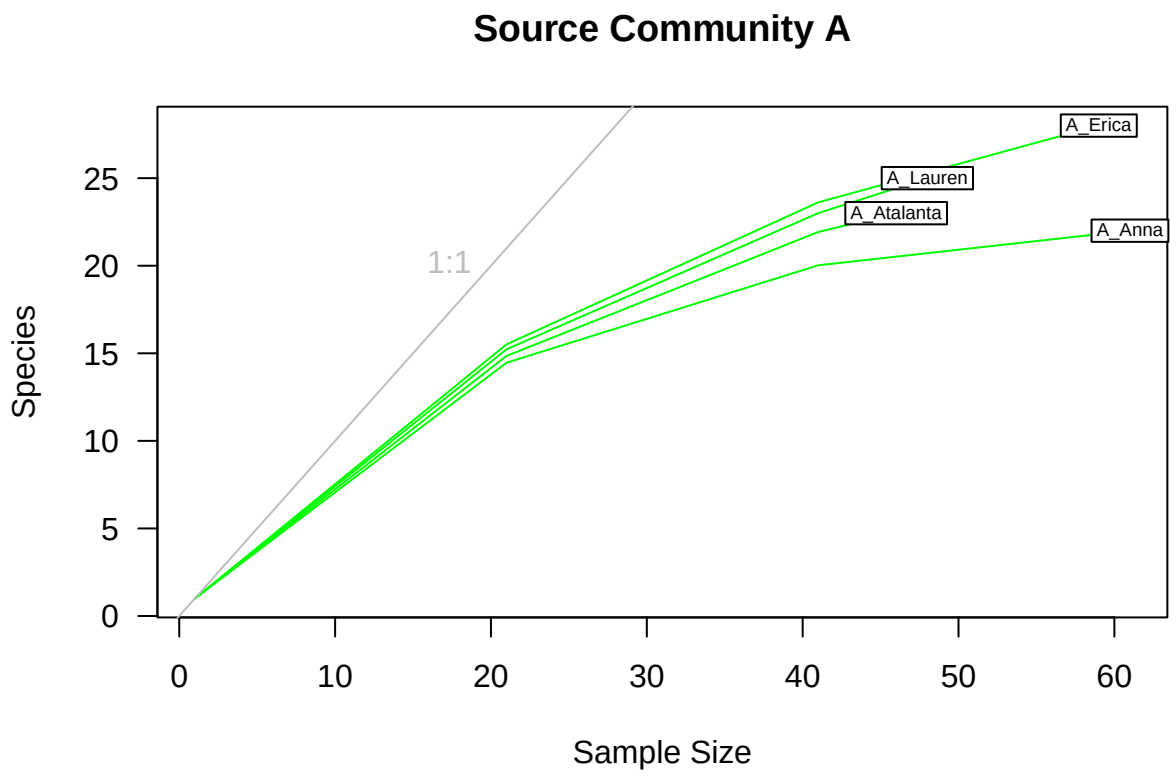
```
#first subsetting out charater columns
gummy.nums<-gummy.data[,3:32]
#changing data from integer to numeric (R was unhappy with integer data)
gummy.num2<-sapply(gummy.nums,FUN=function(x){as.numeric(x)})
#subsetting source community A (rows 1-4) and adding rownames
commA<-gummy.num2[1:4,]
rownames(commA)<-c("A_Erica","A_Lauren","A_Atalanta","A_Anna")
#subsetting source community B (rows 5-8) and adding rownames
commB<-gummy.num2[5:8,]
rownames(commB)<-c("B_Joy","B_Thomas","B_Jonathan","B_Madison")

#estimating coverage using Good's Coverage (C)
#function for good's coverage
C<-function(x=""){
  1-(rowSums(x==1)/rowSums(x))
}
#Good's Coverage for source community A
C(commA)
```

```
##    A_Erica   A_Lauren A_Atalanta     A_Anna
##  0.8135593  0.7291667  0.8043478  0.9508197
```

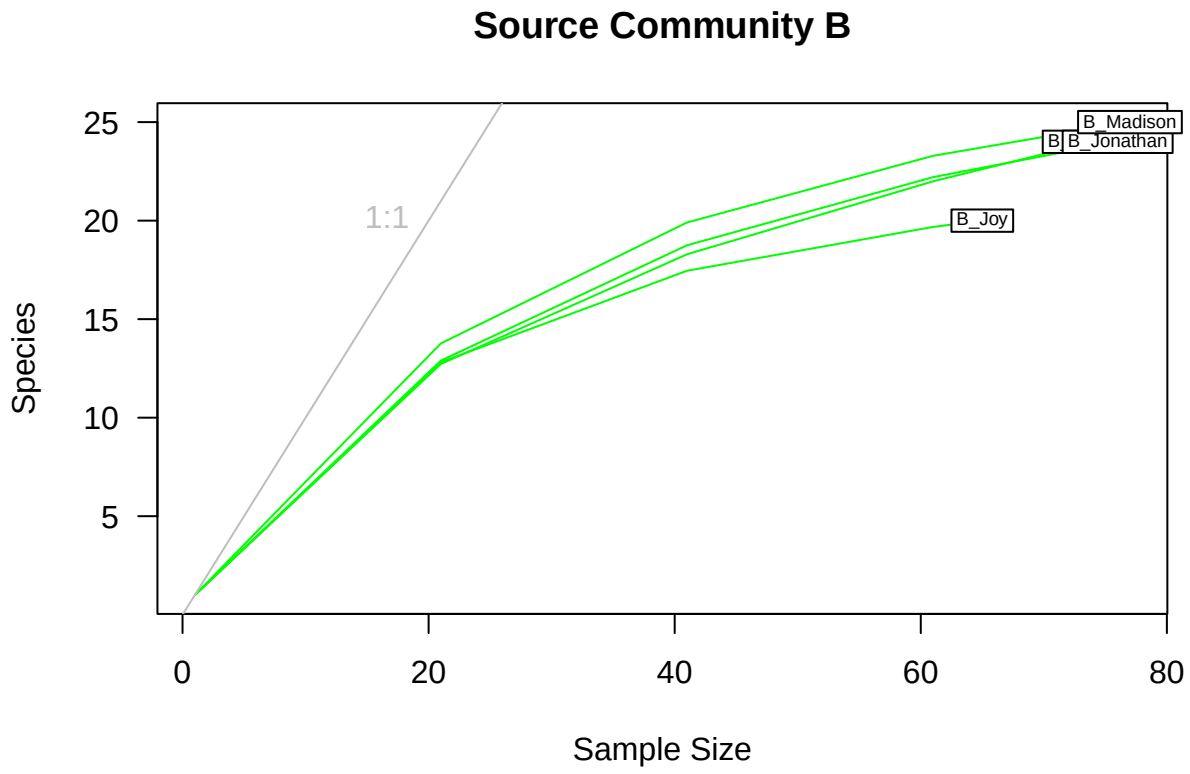
```
#Good's Coverage for source community B
C(commB)
```

```
##      B_Joy   B_Thomas B_Jonathan  B_Madison
##  0.9230769  0.8513514  0.8947368  0.9090909
```

```
#Rarefaction curves
#first need to make observed richness function
S.obs<-function(x=""){
  rowSums(x>0)*1
}
#community A rarefaction curve
commA.S<-S.obs(commA) #finding observed richness at sites in comm. A
min.N.A<-min(rowSums(commA)) #finding lowest sampling effort
#seeing how many spp. at each site w/ lowest sampling effort
S.rarefy.A<-rarefy(x=commA,sample=min.N.A,se=TRUE)
#plotting rarefaction curve
rarecurve(x=commA,step=20,col="green",cex=0.6,las=1,main="Source Community A")
abline(0,1,col="gray")
text(20,20,"1:1",pos=2,col="gray")
```



```
#community B
commB.S<-S.obs(commB) #finding observed richness at sites in comm. B
min.N.B<-min(rowSums(commB)) #finding lowest sampling effort
#seeing how many spp. at each site w/ lowest sampling effort
S.rarefy.B<-rarefy(x=commB,sample=min.N.B,se=TRUE)
#plotting rarefaction curve
rarecurve(x=commB,step=20,col="green",cex=0.6,las=1,main="Source Community B")
abline(0,1,col="gray")
text(20,20,"1:1",pos=2,col="gray")
```



> Based on the Good's Coverage metrics, I would say that the sampling coverage was acceptable foor this application. It's very unlikely that there would be many singletons in the gummy communities (given that most don't come in single-serving packages) and the proportion of non-singleton species is similar across sites, which would indicate to me that sampling efforts were relatively consistent across the whole class. The rarefaction curves also seem to indicate that sampling in communities A and B was relatively consistent with a few exceptions. This means that for the most part, apparent differences in observed richness could be explained by variation in sample sizes. For the outliers in this curve, I would wonder how the order in which each person sampled from their community bowl affected their site's richness (late sampling may come across fewer species due to the nature of the exercise).

2) Alpha diversity

Question 2: Compare alpha diversity measures within sites and among communities. You can calculate and plot richness, Shannon diversity, and Simpson index. You can also convert these indices to effective number of species using the Hill numbers concept by generating a diversity profile, which will make comparisons easier across sites.

What is the variation among the samples in your group and between the communities of other groups for the alpha diversity indices? Generate a hypothesis around the diversity metrics you chose and test your hypothesis. Interpret your findings.

> Hypothesis: Per the Good's Coverage metrics calculted in the previous section, source community A seems to contain a higher proportion of rare species (lower C values = higher proportion of singleton spp. in community). Because it seems that source community A has a greater proportion of rare species, I hypothesize that both the sites and community A as a whole will have a higher diversity (via Shannon diversity index) than those in community B.

Answer 2a - Analysis: Use the space below for code that is being used to analyze your data and test your hypotheses on your chosen alpha diversity tool. Make sure to annotate your code using # symbols so others (including instructors) understand what you have done and why you have done it.

```
#Calculating Shannon diversity values for both source communities and indv. sites
#starting with source community A
commA.H<-diversity(commA,index="shannon")
commA.H
```

```
##    A_Erica   A_Lauren A_Atalanta     A_Anna
##   3.166063   3.053443   2.975329   2.998721
```

```
#source community B
commB.H<-diversity(commB,index="shannon")
commB.H
```

```
##      B_Joy   B_Thomas B_Jonathan  B_Madison
##   2.787795   2.834947   2.836906   2.978193
```

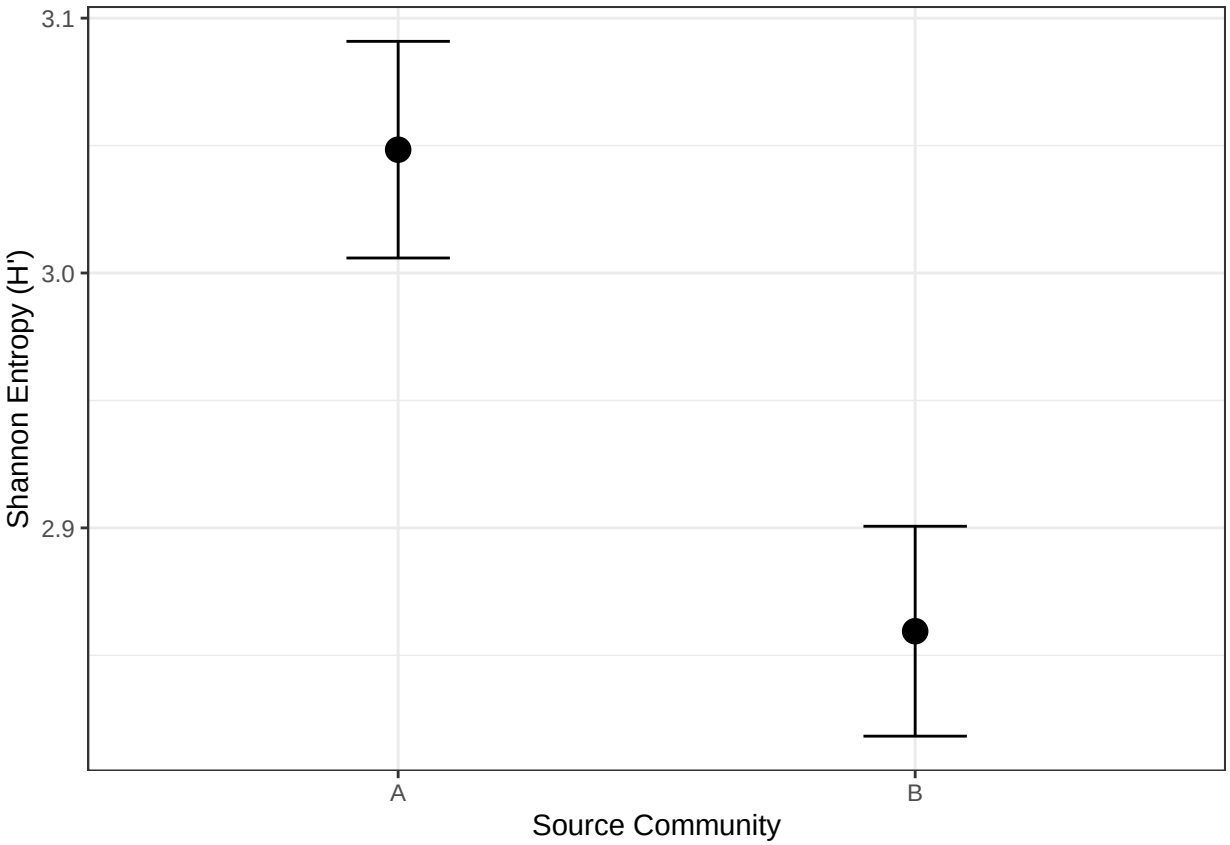
```
#need to summarize data within each community to compare communities
#summarizing spp. counts for community A
Asumm<-as.vector(apply(commA,2,sum))
#summarizing spp. counts for community B
Bsumm<-as.vector(apply(commB,2,sum))
#combining community spp. counts into matrix
combosum<-matrix(c(Asumm,Bsumm),nrow=2,ncol=30,byrow=TRUE)
#changing column names so they make more sense
colnames(combosum)<-colnames(commA)
#changing row names so they make sense
rownames(combosum)<-c("commA","commB")
#calculating shannon diversity for source communities
combo.H<-diversity(combosum,index="shannon")
combo.H
```

```
##    commA    commB
## 3.312211 3.095398
```

> As we can see from the Shannon index results, source community A generally had higher values for diversity. This can be seen both in the individual sites and between the communities as a whole.

Answer 2b - Plot: With your analysis, create one (and only one, although it can have multiple panels) *publication-quality* figure.

```
#creating a graph to compare values of Shannon diversity between sites and comms.
#first need to make standard error function (to get error bars)
sem<-function(x){
  return(sd(x)/sqrt(length(x)))
}
#combining site-specific Shannon values into one dataframe
commcombo<-t(cbind.data.frame(commA.H,commB.H))
#changing column names
colnames(commcombo)<-c("site1","site2","site3","site4")
#changing row names
rownames(commcombo)<-c("A","B")
#making new dataframe for summary statistics
commsum<-as.data.frame(matrix(ncol=2,nrow=2))
colnames(commsum)<-c("mean","SE")
#filling in df with mean and SE for each community
for(i in 1:2){
  x=as.numeric(commcombo[i,])
  commsum[i,1]<-mean(x) #makes col 1 means for each community
  commsum[i,2]<-sem(x) #makes col 2 SE for each community
}
#adding column specifying community to summary table
commsum$community<-c("A","B")

#using ggplot to plot differences in shannon diversity between two communities
ggplot(data=commsum,aes(x=community,y=mean))+
  geom_point(size=4)+
  geom_errorbar(aes(ymin=mean-SE,ymax=mean+SE),width=0.2)+
  ylab("Shannon Entropy (H')")+
  xlab("Source Community")+theme_bw()
```



Answer 2c - Interpret results: Write an informative yet succinct (~5 sentences) caption that creates a "stand-alone" figure. Take a peek at figures and figure captions in a paper published in your favorite journal for inspiration.

> **Figure 1.** Mean diversity indices of sites within gummy source communities A and B. Sites within Community A had a diversity value of 3.048. Sites at Community B had a diversity value of 2.859. Diversity values for each of the four sites within each community were calculated with the Shannon Diversity Index and were then averaged to find the mean site diversity values. Error bars reflect standard error of the mean.

3) Beta diversity

Question 3: Measure beta diversity using ordination and multivariate statistics methods. You can create a PCoA plot, based on Bray-Curtis dissimilarity, of sites and communities using different shape and color codes. Then, you can use a Permanova to test if there are differences between communities. Generate a hypothesis around your chosen analysis and test your hypothesis. Interpret your findings.

> Hypothesis: I found in the previous section that sites within source community A had higher gummy diversity than those within source community B. Due to these results, I hypothesize that the higher site-specific diversity across community A gives a greater likelihood of differences in community composition between sites, so source community A will have higher turnover (beta diversity) than source community B.

Can you detect compositional differences between each source community sampled?

Answer 3a - Analysis: Use the space below for code that is being used to analyze your data and test your hypotheses on your chosen beta diversity tool. Make sure to annotate your code using # symbols so others (including instructors) understand what you have done and why you have done it.

```
#I'll use PERMANOVA to look for differences in community composition
#between source communities
#first, need to make source community vector
source_comm<-c(rep("A",4),rep("B",4))
#now will conduct PERMANOVA for difference between comm. A and B
adonis2(gummy.num2~source_comm,method="bray",permutations=999)
```

```
## Permutation test for adonis under reduced model
## Terms added sequentially (first to last)
## Permutation: free
## Number of permutations: 999
##
## adonis2(formula = gummy.num2 ~ source_comm, permutations = 999, method = "bray")
##             Df SumOfSqs      R2      F Pr(>F)
## source_comm  1  0.31895 0.39151 3.8605  0.031 *
## Residual     6  0.49571 0.60849
## Total        7  0.81466 1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#I'll also calculate indicator values to see if any species are particularly
#associated with either of the source communities
#loading indicspecies package
library(indicspecies)
```

```
## Warning: package 'indicspecies' was built under R version 4.2.2
```

```
#calculating indicator values
gummy.indval<-multipatt(gummy.num2,cluster=source_comm,func="IndVal.g",
                        control=how(nperm=999))
summary(gummy.indval)
```

```
##
##  Multilevel pattern analysis
##  ---------------------------
##
##  Association function: IndVal.g
##  Significance level (alpha): 0.05
##
##  Total number of species: 30
##  Selected number of species: 2
##  Number of species associated to 1 group: 2
##
##  List of species associated to each combination:
##
##  Group A  #sps.  1
##       stat p.value
## SP28 0.913   0.025 *
```

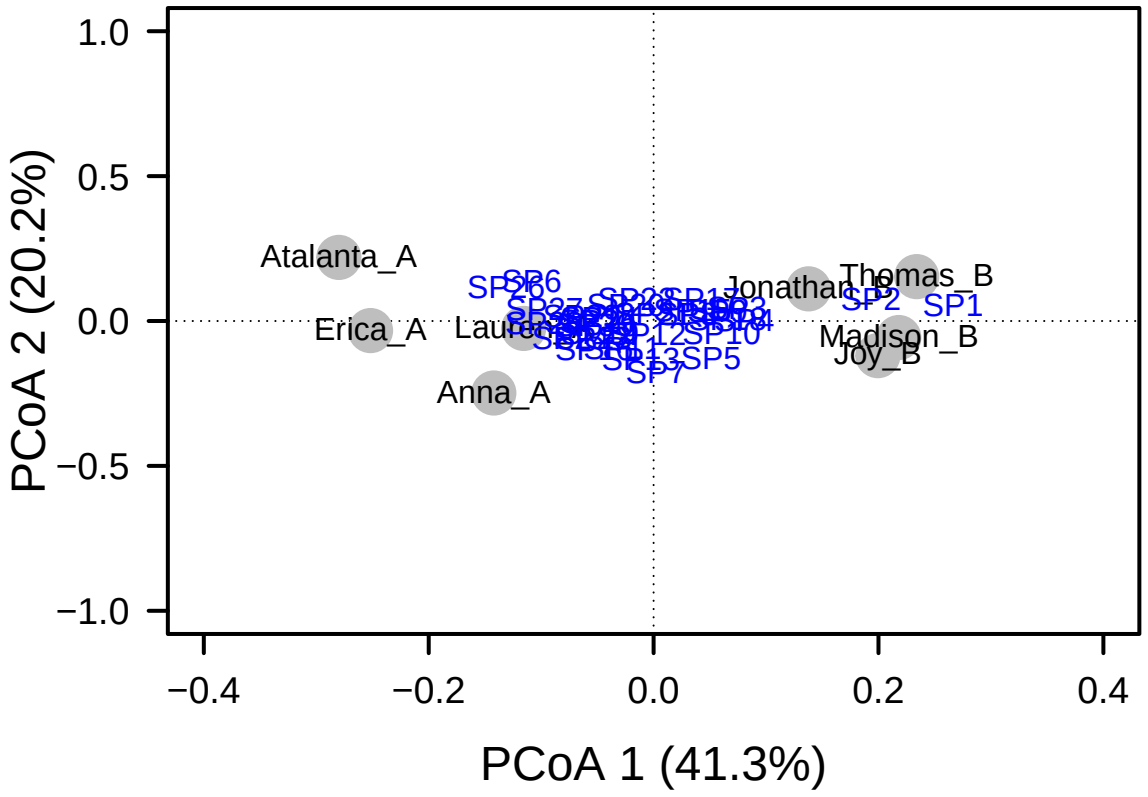
```
##
##  Group B  #sps.  1
##      stat p.value
## SP1 0.963   0.025 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

I first performed a PERMANOVA to assess whether there was a notable difference in the composition of the two gummy source communities, A and B. This analysis found that there was a difference in community composition between A and B, and that about 39.15% of the variance in composition was explained by the different communities ($F_{1,6} = 3.8605$, $p = 0.033$). Each community only had one gummy "species" significantly associated with it; the blackberry gummies were primarily associated with source community A (0.913, $p = 0.03$) and the jelly beans were primarily associated with source community B (0.963, $p = 0.03$). *Note: I wanted to find beta-diversity values within each source community but couldn't quite figure out how to make that part work.*

Answer 3b - Plot: With your analysis, create one (and only one, although it can have multiple panels) *publication-quality* figure.

```
#Creating PCoA ordination for different gummy communities
#changing rownames really quickly to sites
row.names(gummy.num2)<-c("Erica_A","Lauren_A","Atalanta_A","Anna_A","Joy_B",
                         "Thomas_B","Jonathan_B","Madison_B")
#first need to build resemblance matrix (using Bray-Curtis %diff.)
gummy.db<-vegdist(gummy.num2,method="bray")
#doing pcoa test
gummy.pcoa<-cmdscale(gummy.db,eig=TRUE,k=3)
#finding variance explained by pcoa
explvar1<-round(gummy.pcoa$eig[1]/sum(gummy.pcoa$eig),3)*100
explvar2<-round(gummy.pcoa$eig[2]/sum(gummy.pcoa$eig),3)*100
explvar3<-round(gummy.pcoa$eig[3]/sum(gummy.pcoa$eig),3)*100
sum.eig<-sum(explvar1,explvar2,explvar3)
#creating pcoa plot
#plot parameters
par(mar=c(5,5,1,2)+0.1)
#starting plot
plot(gummy.pcoa$points[,1],gummy.pcoa$points[,2],ylim=c(-1.0,1.0),
     xlim=c(-0.4,0.4),xlab=paste("PCoA 1 (", explvar1, "%)",sep=""),
     ylab=paste("PCoA 2 (", explvar2, "%)",sep=""),
     pch=16,cex=2.0,type="n",cex.lab=1.5,cex.axis=1.2,axes=FALSE)
#adding axes to plot
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)
#adding points and labels
points(gummy.pcoa$points[,1],gummy.pcoa$points[,2],pch=19,cex=3,
       bg="gray",col="gray")
text(gummy.pcoa$points[,1],gummy.pcoa$points[,2],
     labels=row.names(gummy.pcoa$points))
#now finding gummy species scores so they can be included in ordination
#calculating relative abundances of spp. at each site
gummyREL<-gummy.num2
```

```r
for(i in 1:nrow(gummy.num2)){
  gummyREL[i,]=gummy.num2[i,]/sum(gummy.num2[i,])
}
#using relative abundance to calculate and add spp. scores to figure
#reading in new spec.score function
source("C:/Users/tmzam/OneDrive/Documents/R/Functions_SourceCodes/spec.scores.function.R")
gummy.pcoa<-add.spec.scores.class(gummy.pcoa,gummyREL,method="pcoa.scores")
text(gummy.pcoa$cproj[,1],gummy.pcoa$cproj[,2],
     labels=row.names(gummy.pcoa$cproj),col="blue")
```



Answer 3c - Interpret results: Write an informative yet succinct (~5 sentences) caption that creates a "stand-alone" figure. Take a peek at figures and figure captions in a paper published in your favorite journal for inspiration.

Figure 2. Principal coordinates analysis generated from a site-by-species matrix for gummies ("species") collected by class members ("sites"). Each site is labeled by class member and which source community they collected from (black text); Atalanta, Erica, Lauren, and Anna were in source community A, while Thomas, Madison, Jonathan, and Joy were in source community B. Species codes are given in blue and are representative of species scores (association with particular sites). Species 1 and 2 (jelly beans and sour watermelons) are clearly associated with community B. The majority of the remaining species seem to be fairly evenly associated with both source communities (some associate more with one than the other); species 6 and 26 (teal bear and yellow-red dinosaur) appear to be the most associated with community A of this cluster. Source communities A and B are clearly separated based on their compositions.

SUBMITTING YOUR ASSIGNMENT

Use Knitr to create a PDF of your completed 7.DiversitySynthesis_Worksheet.Rmd document, push it to GitHub, and create a pull request. Please make sure your updated repo includes both the pdf and RMarkdown files.

Unless otherwise noted, this assignment is due on **Wednesday, February 15th, 2023 at 12:00 PM (noon)**.