8. Worksheet: Among Site (Beta) Diversity – Part 1

Mackenzie Caple; Z620: Quantitative Biodiversity, Indiana University

05 February, 2019

OVERVIEW

In this worksheet, we move beyond the investigation of within-site α-diversity. We will explore β-diversity, which is defined as the diversity that occurs among sites. This requires that we examine the compositional similarity of assemblages that vary in space or time.

After completing this exercise you will know how to:

1. formally quantify β-diversity
2. visualize β-diversity with heatmaps, cluster analysis, and ordination
3. test hypotheses about β-diversity using multivariate statistics

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. Ths 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 '8.BetaDiversity' 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 (**8.BetaDiversity_1_Worksheet.Rmd**) with all code blocks filled out and questions answered) and the PDF output of `Knitr` (**8.BetaDiversity_1_Worksheet.pdf**).

The completed exercise is due on **Wednesday, February 6th, 2019 before 12:00 PM (noon)**.

1) R 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 "*/8.BetaDiversity*" folder, and
4. load the `vegan` R package (be sure to install if needed).

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

```
## [1] "/Users/mcaple/GitHub/QB2019_Caple/2.Worksheets/8.BetaDiversity"
```

```r
setwd("~/GitHub/QB2019_Caple/2.Worksheets/8.BetaDiversity")

package.list <- c('vegan', 'ade4', 'viridis', 'gplots', 'BiodiversityR', 'indicspecies')
for(package in package.list){
  if(!require(package, character.only = TRUE, quietly = TRUE)){
    install.packages(package)
    library(package, character.only = TRUE)
  }
}
```

```
## This is vegan 2.5-3
```

```
## Warning: package 'gplots' was built under R version 3.5.2
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
## Warning: package 'BiodiversityR' was built under R version 3.5.2
```

```
## BiodiversityR 2.11-1: Use command BiodiversityRGUI() to launch the Graphical User Interface;
## to see changes use BiodiversityRGUI(changeLog=TRUE, backward.compatibility.messages=TRUE)
```

2) LOADING DATA

Load dataset

In the R code chunk below, do the following:

1. load the doubs dataset from the ade4 package, and
2. explore the structure of the dataset.

```r
# note, pleae do not print the dataset when submitting
data(doubs)
#doubs

str(doubs, max.level = 1)
```

```
## List of 4
##  $ env    :'data.frame': 30 obs. of  11 variables:
##  $ fish   :'data.frame': 30 obs. of  27 variables:
##  $ xy     :'data.frame': 30 obs. of  2 variables:
##  $ species:'data.frame': 27 obs. of  4 variables:
```

```r
#head(doubs$env)
```

Question 1: Describe some of the attributes of the doubs dataset.

a. How many objects are in doubs?
b. How many fish species are there in the doubs dataset?
c. How many sites are in the doubs dataset?

Answer 1a: 4 (4 different data frames) ***Answer 1b***: 27 ***Answer 1c***: 30

Visualizing the Doubs River Dataset

Question 2: Answer the following questions based on the spatial patterns of richness (i.e., α-diversity) and Brown Trout (*Salmo trutta*) abundance in the Doubs River.

a. How does fish richness vary along the sampled reach of the Doubs River?
b. How does Brown Trout (*Salmo trutta*) abundance vary along the sampled reach of the Doubs River?
c. What do these patterns say about the limitations of using richness when examining patterns of biodiversity?

> ***Answer 2a***: Fish richness tends to increase further downstream, although there is a patch of lower richness roughly two thirds of the way through the sites ***Answer 2b***: Abundance is patchy, with a high abundance patch near the most upstream group of sites, and another just above halfway downstream; abundance is low in every site outside these two patches ***Answer 2c***: The richness plot omits information about abundances, so we are missing a lot of the important information about the biodiversity of the stream

3) QUANTIFYING BETA-DIVERSITY

In the R code chunk below, do the following:

1. write a function (`beta.w()`) to calculate Whittaker's β-diversity (i.e., β_w) that accepts a site-by-species matrix with optional arguments to specify pairwise turnover between two sites, and
2. use this function to analyze various aspects of β-diversity in the Doubs River.

```r
beta.w <- function(site.by.species = "", sitenum1 = "", sitenum2 = "", pairwise = FALSE){
  # ONLY if we specify pairwise as TRUE, do this:
  if(pairwise == TRUE){

    # As a check, print an error if we do not provide needed arguments
    if(sitenum1 == "" | sitenum2 == ""){
    print("Error: please specify sites to compare")
    return(NA)
    }

    # If our function made it this far, calculate pairwise beta diversity
    site1 = site.by.species[sitenum1, ]                   # select site 1
    site2 = site.by.species[sitenum2, ]                   # select site 2
    site1 = subset(site1, select = site1 > 0)             # removes absences
    site2 = subset(site2, select = site2 > 0)             # removes absences
    gamma = union(colnames(site1), colnames(site2))       # gamma species pool
    s = length(gamma)                                     # gamma richness
    a.bar = mean(c(specnumber(site1), specnumber(site2))) # mean sample richness
    b.w = round(s/a.bar - 1, 3)
    return(b.w)
  }

  # OTHERWISE pairwise defaults to FALSE, so do this, like before:
  else{
    SbyS.pa <- decostand(site.by.species, method = "pa")  # convert to presence-absence
    S <- ncol(SbyS.pa[ , which(colSums(SbyS.pa) > 0)])    # number of species in region
    a.bar <- mean(specnumber(SbyS.pa))                    # average richness at each site
    b.w <- round(S/a.bar, 3)
    return(b.w)
  }
}

# remember: for pairwise beta.w for turnover:
# 0 is min diversity (complete similarity, sites are identical), 1 is max diversity (complete disimilar
beta.w(doubs$fish, "1", "2", pairwise = TRUE)
```

```
## [1] 0.5
```

```
beta.w(doubs$fish, "1", "10", pairwise = TRUE)
```

```
## [1] 0.714
```

Question 3: Using your `beta.w()` function above, answer the following questions:

a. Describe how local richness (α) and turnover (β) contribute to regional (γ) fish diversity in the Doubs.
b. Is the fish assemblage at site 1 more similar to the one at site 2 or site 10?
c. Using your understanding of the equation $\beta_w = \gamma/\alpha$, how would your interpretation of β change if we instead defined beta additively (i.e., $\beta = \gamma - \alpha$)?

> ***Answer 3a***: Regional diversity is a multiplicative combination of local richness and turnover; increasing either alpha or beta diversity will also increase gamma diversity. ***Answer 3b***: Site 1 is more similar to site 2 than to site 10 ***Answer 3c***: Instead of interpreting beta as the amount of turnover, I would interpret it as the additive difference between local and regional diversity; instead of being a measure from 0 to 1 of how the average site differs from the total region, it would be a numeric value describing how many fewer species the site has than the region (from 0 to the number of species in the region)

The Resemblance Matrix

In order to quantify β-diversity for more than two samples, we need to introduce a new primary ecological data structure: the **Resemblance Matrix**.

Question 4: How do incidence- and abundance-based metrics differ in their treatment of rare species?

> ***Answer 4***: Incidence-based metrics give equal weight to rare and abundant species, whereas abundance-based metrics give more weight to species based on how abundant they are in the community. Both approaches have pros and cons, depending on what questions are being asked.

In the R code chunk below, do the following:

1. make a new object, `fish`, containing the fish abundance data for the Doubs River,
2. remove any sites where no fish were observed (i.e., rows with sum of zero),
3. construct a resemblance matrix based on Sørensen's Similarity ("fish.ds"), and
4. construct a resemblance matrix based on Bray-Curtis Distance ("fish.db").

```
fish <- doubs$fish
fish <- fish[-8, ] # remove site 8 from the data (it has no observations)

# calculate Jaccard
fish.dj <- vegdist(fish, method = "jaccard", binary = TRUE)

# calculate Bray-Curtis
fish.db <- vegdist(fish, method = "bray")

# calculate Sørensen
```

```
fish.ds <- vegdist(fish, method = "bray", binary = TRUE)

fish.db.full <- vegdist(fish, method = "bray", upper = TRUE, diag = TRUE) # creates a rectangular matri
```

Question 5: Using the distance matrices from above, answer the following questions:

a. Does the resemblance matrix (`fish.db`) represent similarity or dissimilarity? What information in the resemblance matrix led you to arrive at your answer?
b. Compare the resemblance matrices (`fish.db` or `fish.ds`) you just created. How does the choice of the Sørensen or Bray-Curtis distance influence your interpretation of site (dis)similarity?

Answer 5a: This matrix measures dissimilarity; I concluded this because the diagonals are 0. Since comparing a site to itself yields a value of 0, that means that higher values mean higher levels of dissimilarity. ***Answer 5b***: Since Sørensen uses presence-absence data and Bray-Curtis takes abundance into account, Sørensen gives more information about how similar the species composition at sites are; in other words, rare species are weighted more heavily, which is either useful or not depending on the question. The Bray-Curtis distance says more about the structure of the community, since it takes species abundance into account.

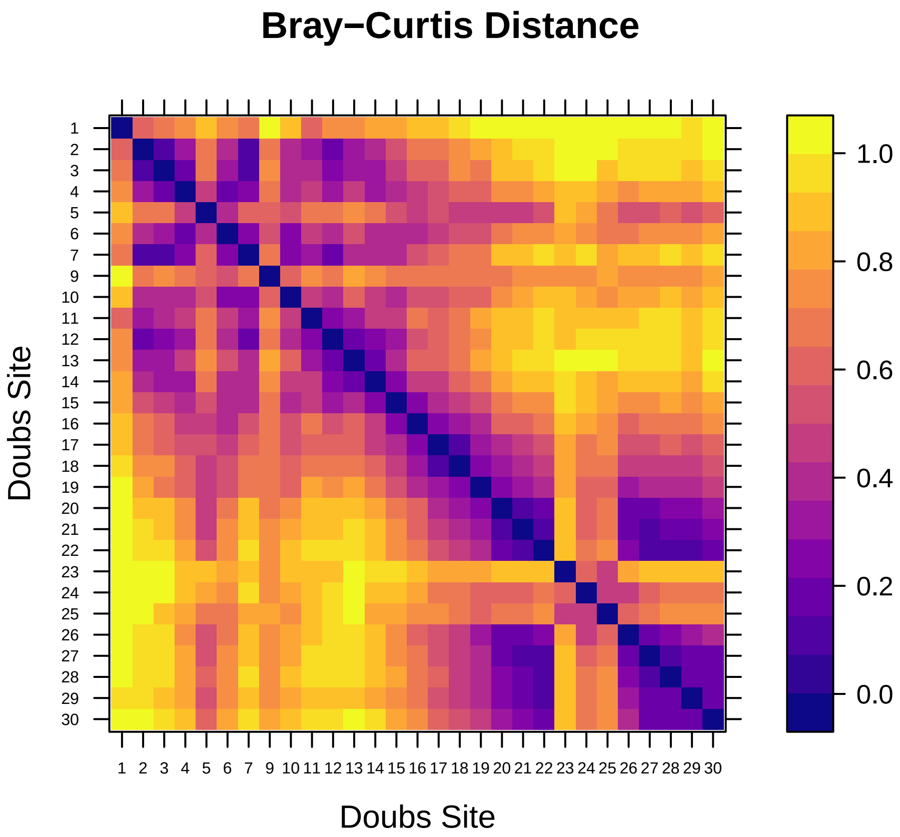
4) VISUALIZING BETA-DIVERSITY

A. Heatmaps

In the R code chunk below, do the following:

1. define a color palette,
2. define the order of sites in the Doubs River, and
3. use the `levelplot()` function to create a heatmap of fish abundances in the Doubs River.

```
# define order of sites
order <- rev(attr(fish.db, "Labels"))

# plot heatmap
levelplot(as.matrix(fish.db)[, order], aspect = "iso", col.regions = plasma,
          xlab = "Doubs Site", ylab = "Doubs Site", scales = list(cex = 0.5),
          main = "Bray-Curtis Distance")
```

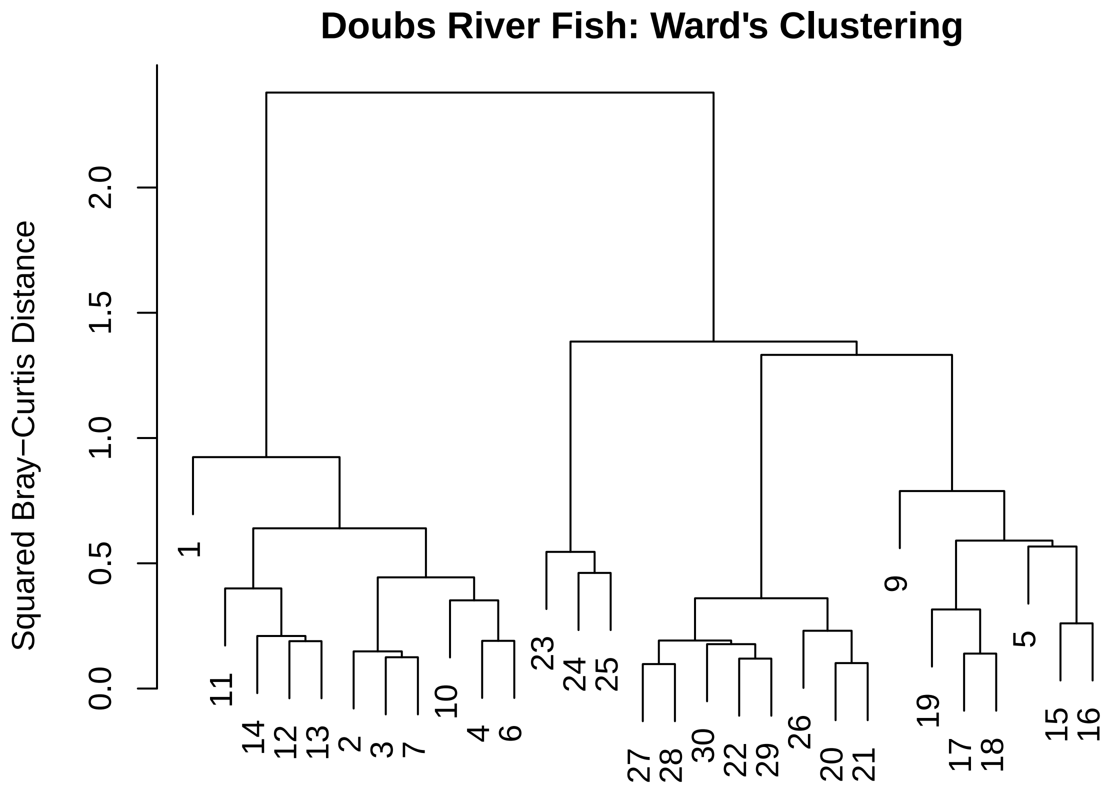


B. Cluster Analysis

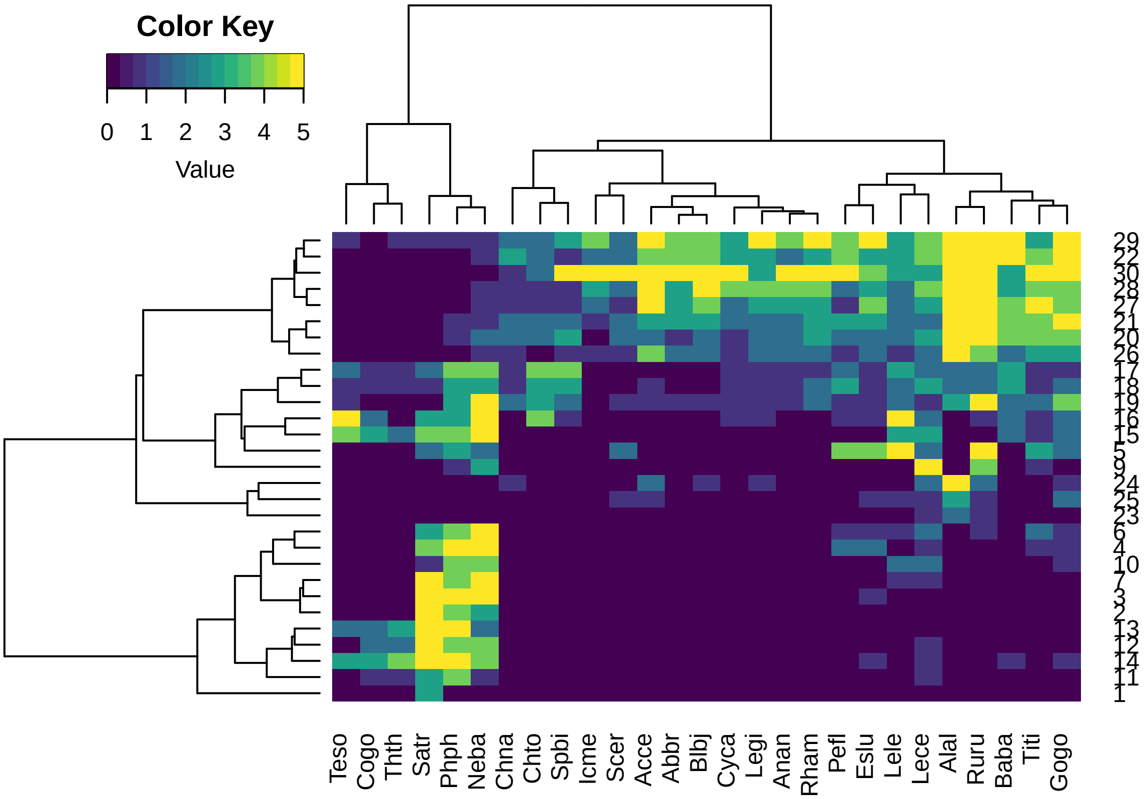
In the R code chunk below, do the following:

1. perform a cluster analysis using Ward's Clustering, and
2. plot your cluster analysis (use either `hclust` or `heatmap.2`).

```
# perform cluster analysis
fish.ward <- hclust(fish.db, method = "ward.D2")

# plot cluster
par(mar = c(1, 5, 2, 2) + 0.1)
plot(fish.ward, main = "Doubs River Fish: Ward's Clustering",
     ylab = "Squared Bray-Curtis Distance")
```



```r
# plot cluster heatmap
gplots::heatmap.2(as.matrix(fish), distfun = function(x) vegdist(x, method = "bray"),
                  hclustfun = function(x) hclust(x, method = "ward.D2"),
                  col = viridis, trace = "none", density.info = "none")
```



Question 6: Based on cluster analyses and the introductory plots that we generated after loading the data, develop an ecological hypothesis for fish diversity the `doubs` data set?

> ***Answer 6***: Sites are generally not primarily clustered by distance, so there must be something else happening in the stream to structure the fish community. This could be water quality, depth, flow rate, etc. One testable hypothesis is that water quality, not site number (subbing for distance downstream) is the primary driver of fish community composition.

C. Ordination

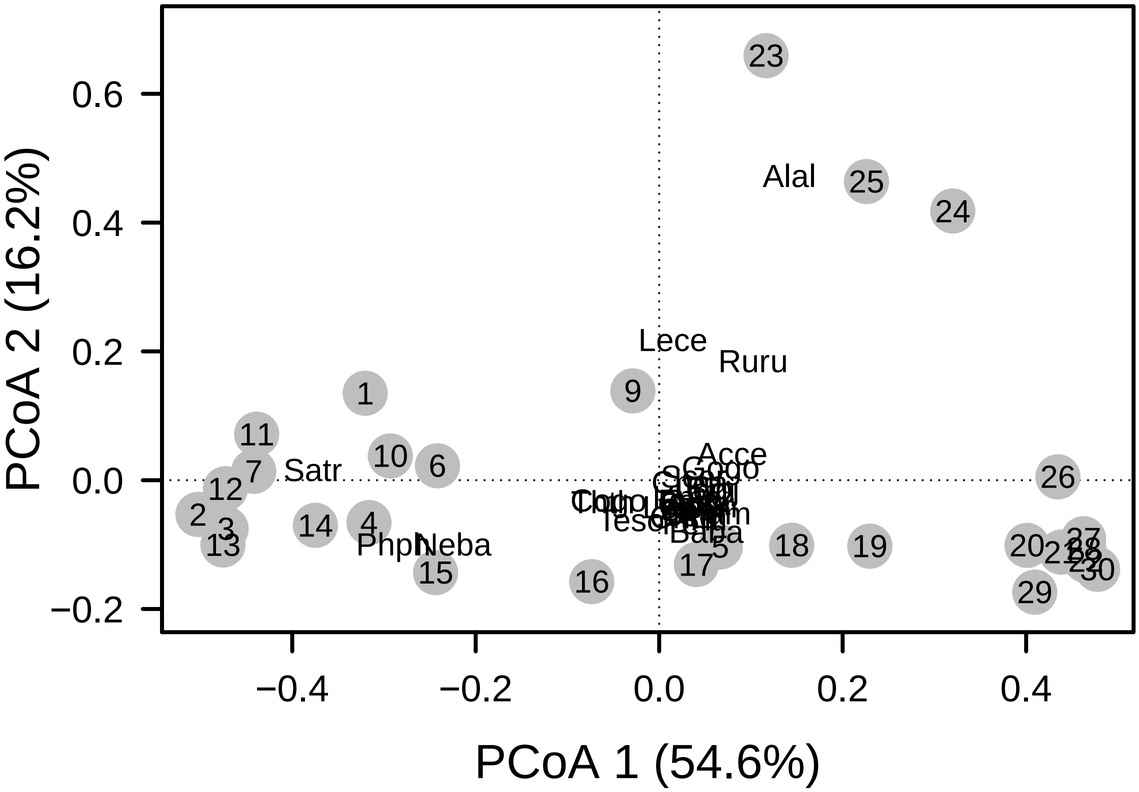
Principal Coordinates Analysis (PCoA)

In the R code chunk below, do the following:

1. perform a Principal Coordinates Analysis to visualize beta-diversity
2. calculate the variation explained by the first three axes in your ordination
3. plot the PCoA ordination,
4. label the sites as points using the Doubs River site number, and
5. identify influential species and add species coordinates to PCoA plot.

```
# conduct PCoA
fish.pcoa <- cmdscale(fish.db, eig = TRUE, k = 3)

# examine eigenvalues
explainvar1 <- round(fish.pcoa$eig[1] / sum(fish.pcoa$eig), 3) *100
explainvar2 <- round(fish.pcoa$eig[2] / sum(fish.pcoa$eig), 3) *100
explainvar3 <- round(fish.pcoa$eig[3] / sum(fish.pcoa$eig), 3) *100
sum.eig <- sum(explainvar1, explainvar2, explainvar3)
```

```
# create PCoA ordination plot

# define plot parameters
par(mar = c(5, 5, 1, 2) + 0.1)

# initiate plot
plot(fish.pcoa$points[ , 1], fish.pcoa$points[ , 2], ylim = c(-0.2, 0.7),
     xlab = paste("PCoA 1 (", explainvar1, "%)", sep = ""),
     ylab = paste("PCoA 2 (", explainvar2, "%)", sep = ""),
     pch = 16, cex = 2.0, type = "n", cex.lab = 1.5, cex.axis = 1.2, axes = FALSE)

# add axes
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)

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

# calculate the relative abundances of each species at each site
fishREL <- fish
  for(i in 1:nrow(fish)){
    fishREL[i, ] = fish[i, ] / sum(fish[i, ])
  }

# use above info to calculate and add species scores
fish.pcoa <- add.spec.scores(fish.pcoa, fishREL, method = "pcoa.scores")
text(fish.pcoa$cproj[ , 1], fish.pcoa$cproj[ , 2],
     labels = row.names(fish.pcoa$cproj), col = "black")
```



In the R code chunk below, do the following:

1. identify influential species based on correlations along each PCoA axis (use a cutoff of 0.70), and
2. use a permutation test (999 permutations) to test the correlations of each species along each axis.

```
# more quantitative way
spe.corr <- add.spec.scores(fish.pcoa, fishREL, method = "cor.scores")$cproj
corrcut <- 0.7 # user defined cutoff
imp.spp <- spe.corr[abs(spe.corr[ , 1]) >= corrcut | abs(spe.corr[ , 2]) >= corrcut, ]

# permutation test for species abundances across axes
fit <- envfit(fish.pcoa, fishREL, perm = 999)
```

Question 7: Address the following questions about the ordination results of the `doubs` data set:

a. Describe the grouping of sites in the Doubs River based on fish community composition.
b. Generate a hypothesis about which fish species are potential indicators of river quality.

> ***Answer 7a***: Sites are primarily separated along the first axis, differences in which seem to be driven primarily by Salmo trutta fario (Satr), Phoxinus phoxinus (Phph), and Nemacheilus barbatulus (Neba). Along the second axis there is less separation, and it is driven primarily by Alburnus alburnus (Alal), Leuciscus cephalus cephalus (Lece), and Rutilus rutilus (Ruru). ***Answer 7b***: Nearly all the fish species are highly significant, and all but Teso, Lele, and Eslu are significant. This makes sense, since the river mostly had fairly low species richness except for a couple patches of sites with high richness.

SYNTHESIS

Using the jelly bean data from class (i.e., JellyBeans.Source.txt and JellyBeans.txt):

1) Compare the average pairwise similarity among subsamples in group A to the average pairswise similarity among subsamples in group B. Use a t-test to determine whether compositional similarity was affected by the "vicariance" event. Finally, compare the compositional similarity of jelly beans in group A and group B to the source community?

```
setwd("~/GitHub/QB2019_Caple/2.Worksheets/6.DiversitySampling")

jellysource <- read.table("JellyBeans.Source.txt", sep = "\t", header = TRUE, row.names = 1)

jellysource.num <- Filter(is.numeric, jellysource)


jellysample <- read.table("JellyBeans.std.txt", sep = "\t", header = TRUE)

jellysample.a <- subset(jellysample, Group == "A")
jellya <- Filter(is.numeric, jellysample.a)
jellysample.b <- subset(jellysample, Group == "B")
jellyb <- Filter(is.numeric, jellysample.b)

jellya.db <- vegdist(jellya, method = "bray")
jellyb.db <- vegdist(jellyb, method = "bray")


mean(jellya.db)
```

```
## [1] 0.2649123
```

```
mean(jellyb.db)
```

```
## [1] 0.3302977
```

```
t.test(jellya.db, jellyb.db)
```

```
##
##  Welch Two Sample t-test
##
## data:  jellya.db and jellyb.db
## t = -2.5912, df = 7.5291, p-value = 0.03372
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.124214114 -0.006556611
## sample estimates:
## mean of x mean of y
## 0.2649123 0.3302977
```

```
jellysource.db <- vegdist(jellysource.num, method = "bray")

t.test(jellysource.db, jellya.db)
```

```
##
##  Welch Two Sample t-test
##
## data:  jellysource.db and jellya.db
## t = 4.5122, df = 52.424, p-value = 3.653e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.04444174 0.11560077
## sample estimates:
## mean of x mean of y
## 0.3449336 0.2649123
```

```
t.test(jellysource.db, jellyb.db)
```

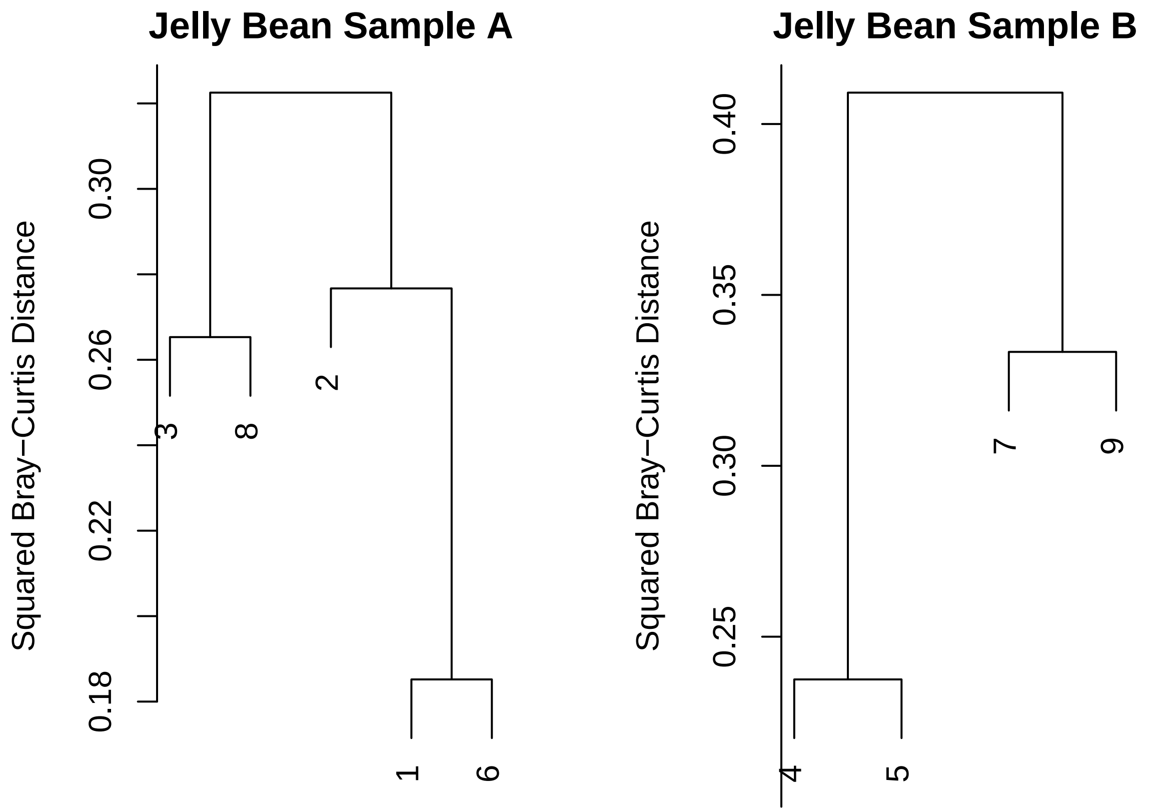
```
##
##  Welch Two Sample t-test
##
## data:  jellysource.db and jellyb.db
## t = 0.55387, df = 9.344, p-value = 0.5927
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.04480740  0.07407919
## sample estimates:
## mean of x mean of y
## 0.3449336 0.3302977
```

Using a t-test, the vicariance event may have affected the compositional similarity; p = 0.05041, which is very close to significance; I cannot confidently say that the vicariance event affected composition, but it's so close to significance that I'm not comfortable ruling it out, either.

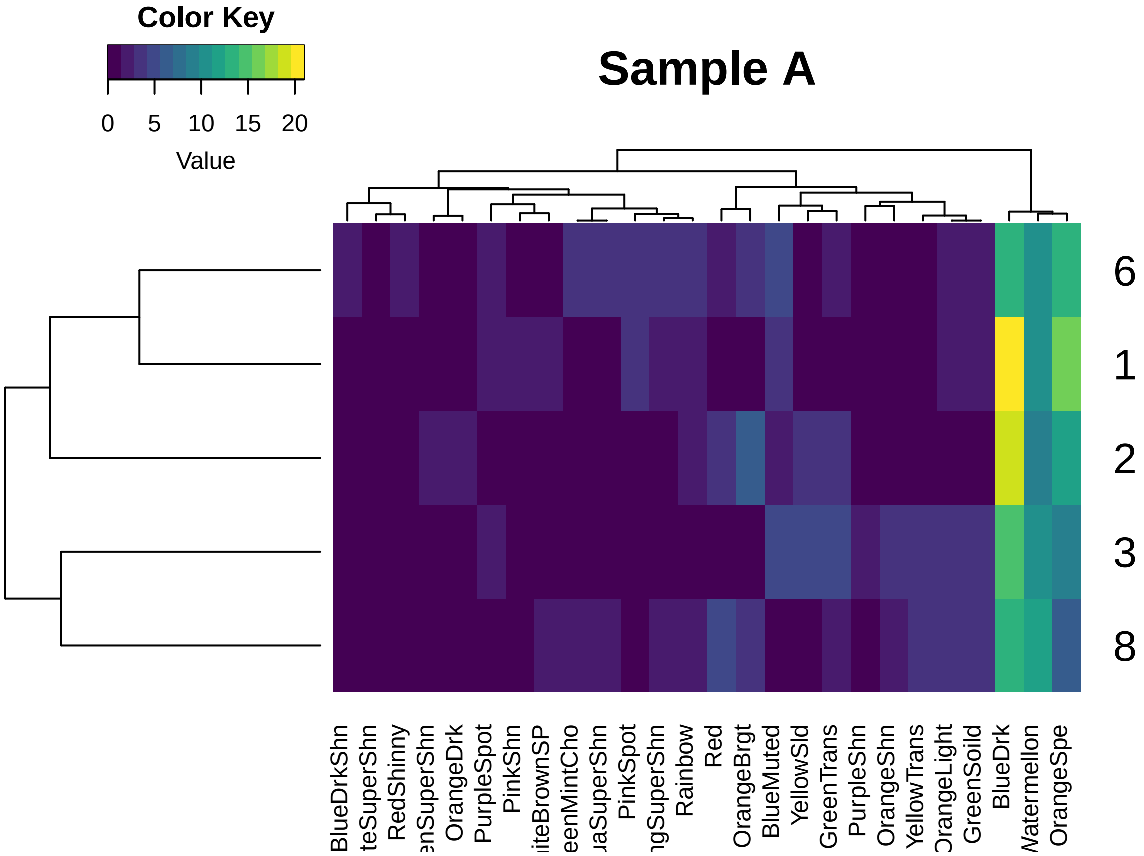
The composition of sample A varied significantly from the source community, with a very low p-value, but the composition of sample B had a very high p-value and did not differ significantly from the source community.

2) Create a cluster diagram or ordination using the jelly bean data. Are there any visual trends that would suggest a difference in composition between group A and group B?

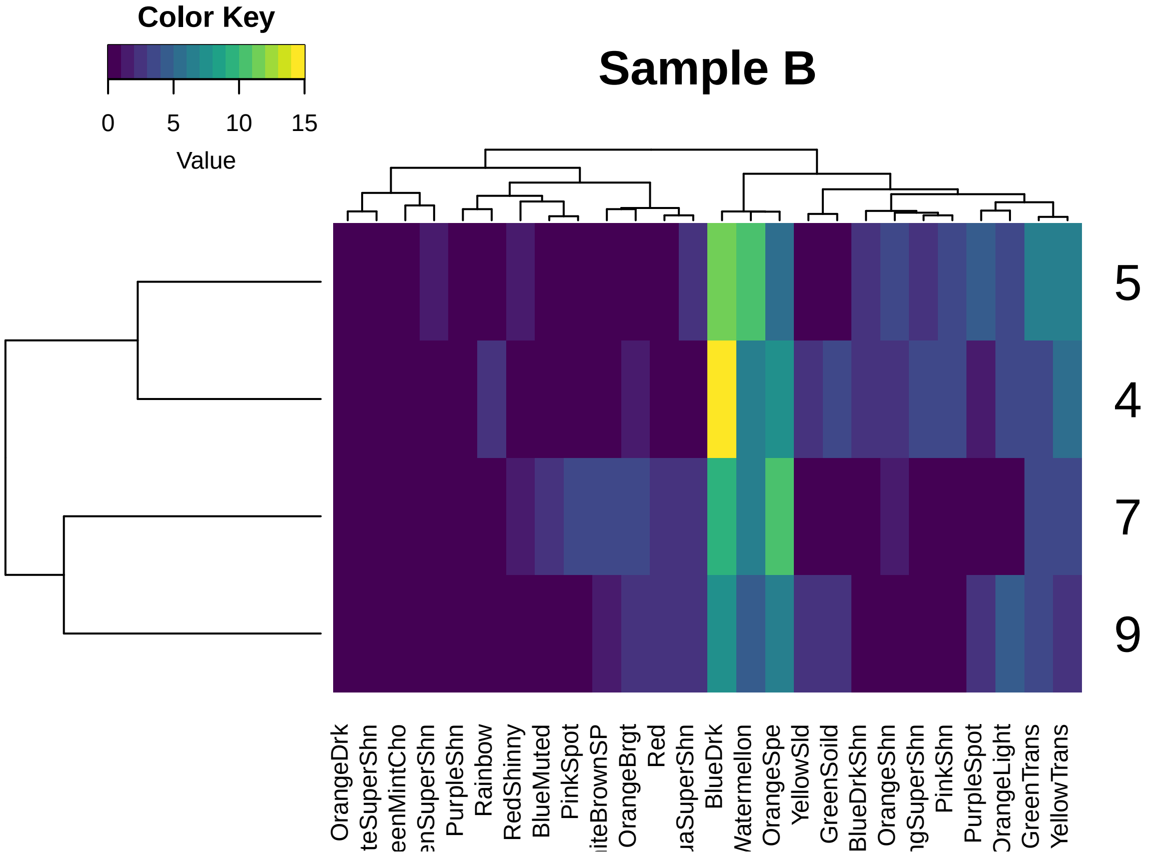
```
jellya.ward <- hclust(jellya.db, method = "ward.D2")
jellyb.ward <- hclust(jellyb.db, method = "ward.D2")

par(mar = c(1, 5, 2, 2) + 0.1, mfrow = c(1, 2))
plot(jellya.ward, main = "Jelly Bean Sample A", ylab = "Squared Bray-Curtis Distance")
plot(jellyb.ward, main = "Jelly Bean Sample B", ylab = "Squared Bray-Curtis Distance")
```



```
gplots::heatmap.2(as.matrix(jellya), distfun = function(x) vegdist(x, method = "bray"),
                  hclustfun = function(x) hclust(x, method = "ward.D2"),
                  col = viridis, trace = "none", density.info = "none", main = "Sample A")
```



```
gplots::heatmap.2(as.matrix(jellyb), distfun = function(x) vegdist(x, method = "bray"),
                  hclustfun = function(x) hclust(x, method = "ward.D2"),
                  col = viridis, trace = "none", density.info = "none", main = "Sample B")
```



The two heat maps are very similar in that the same three ‘species’ (BlueDrk, Watermellon, OrangeSld) are the most heavily represented in both samples. However, the two heat maps show the species arranged in quite different ways; in Sample A, all three very abundant species are an outgroup to the rest of the species, and this is not true in Sample B.