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

Diego Rios; Z620: Quantitative Biodiversity, Indiana University

06 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 6$^{\text{th}}$, 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).

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

```
## function ()
## .Internal(getwd())
## <bytecode: 0x7fbacbeb0040>
## <environment: namespace:base>
```

```
setwd("~/GitHub/QB2019_Rios/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
```

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

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

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

```
package.list
```

```
## [1] "vegan"         "ade4"          "viridis"       "gplots"
## [5] "BiodiversityR" "indicspecies"
```

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.

```
# note, please do not print the dataset when submitting
data("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:
```

```
?doubs
head(doubs$env)
```

```
##   dfs alt   slo flo pH har pho nit amm oxy bdo
## 1   3 934 6.176  84 79  45   1  20   0 122  27
## 2  22 932 3.434 100 80  40   2  20  10 103  19
## 3 102 914 3.638 180 83  52   5  22   5 105  35
## 4 185 854 3.497 253 80  72  10  21   0 110  13
## 5 215 849 3.178 264 81  84  38  52  20  80  62
## 6 324 846 3.497 286 79  60  20  15   0 102  53
```

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***: doubs is a list of four objects ***Answer 1b***: 27 species ***Answer 1c***: 30 sites

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 increases towards downstream ***Answer 2b***: Brown Trout abundance is higher Upstream than Downstream ***Answer 2c***: Richness weights each species as equally important, thus a lot information about the community's biology is being lost.

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.

```
beta.w <- function(site.by.species = ""){
  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 the region
  a.bar <- mean(specnumber(SbyS.pa))                   # average richness at each site
  b.w <- round(S/a.bar, 3)
  return(b.w)
}

beta.w <- function(site.by.species = "", sitenum1 = "", sitenum2 = "", pairwise = FALSE){
  if (pairwise == TRUE){
    if (sitenum1==""|sitenum2==""){
      print("Error: please specify sites to compare")
    return(NA)}
    site1 = site.by.species[sitenum1,]
    site2 = site.by.species[sitenum2,]
    site1 = subset(site1, select = site1 > 0)
    site2 = subset(site2, select = site2 > 0)
    gamma = union(colnames(site1), colnames(site2))
    s = length(gamma)
    a.bar = mean(c(specnumber(site1),specnumber(site2)))
    b.w = round(s/a.bar - 1,3)
    return(b.w)
  }
  else{
    SbyS.pa <- decostand(site.by.species, method = "pa")
    S <- ncol(SbyS.pa[,which(colSums(SbyS.pa)> 0)])
    a.bar <- mean(specnumber(SbyS.pa))
    b.w <- round(S/a.bar, 3)
    return(b.w)
  }
}

head(doubs$env)
```

```
##   dfs alt   slo flo pH har pho nit amm oxy bdo
## 1   3 934 6.176  84 79  45   1  20   0 122  27
## 2  22 932 3.434 100 80  40   2  20  10 103  19
## 3 102 914 3.638 180 83  52   5  22   5 105  35
## 4 185 854 3.497 253 80  72  10  21   0 110  13
## 5 215 849 3.178 264 81  84  38  52  20  80  62
## 6 324 846 3.497 286 79  60  20  15   0 102  53
```

```
doubs$env[1,]
```

```
##   dfs alt   slo flo pH har pho nit amm oxy bdo
## 1   3 934 6.176  84 79  45   1  20   0 122  27
```

```
beta.w(site.by.species = doubs$fish,sitenum1 = 1,sitenum2 = 2,pairwise = TRUE)
```

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

```
beta.w(site.by.species = doubs$fish,sitenum1 = 1,sitenum2 = 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***:
> ***Answer 3b***: fish assemblage at site 1 is more similar to the one at site 2 than site 10 ***Answer 3c***:

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 metrics give the same weight (1 or 0) to species, regardless of their abundance. Abundance metrics do take into account the relative importance of species per site

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 data

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

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

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

```
## [1] 0.5000000 0.6000000 0.7777778 0.8333333 0.8181818 0.6666667
```

```
fish.db <- vegdist(fish, method = "bray", upper = TRUE, diag = TRUE)
head(fish.db)
```

```
## [1] 0.6000000 0.6842105 0.7500000 0.8918919 0.7500000 0.6842105
```

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: dissimilarity. Sites downstream-upstream sites have higher values than downstream-downstream sites ***Answer 5b***: Sorensen's index seems to inflate dissimilarities, so it would confirm my interpretation. However, information is lost in the process.

4) VISUALIZING BETA-DIVERSITY

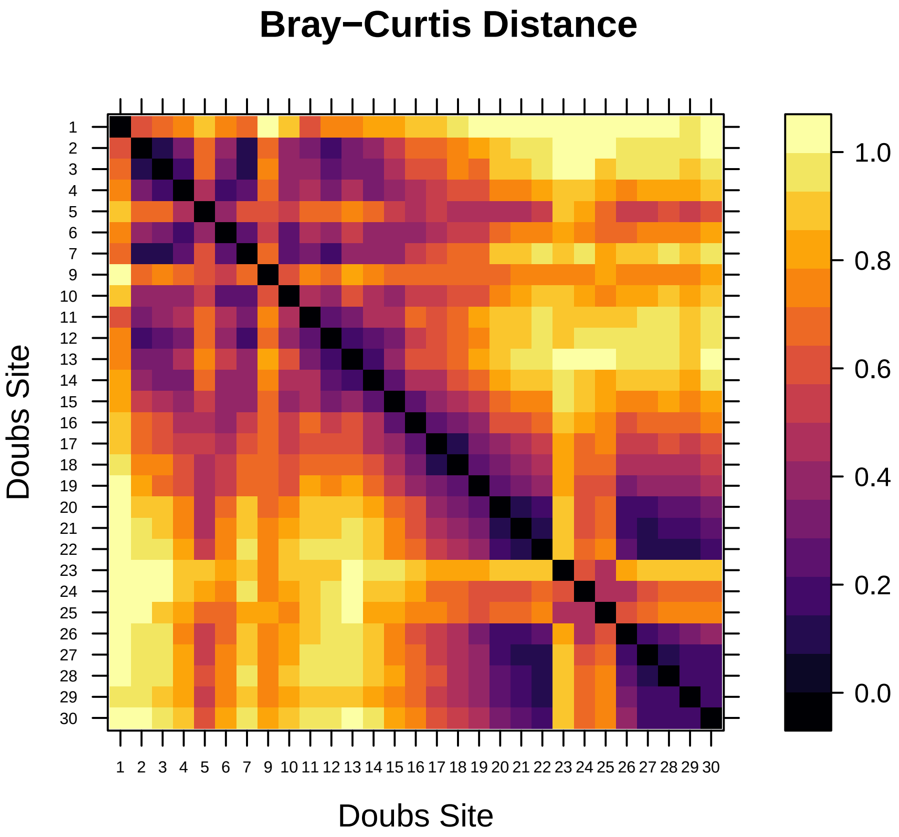
A. Heatmaps

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.

```
order <- rev(attr(fish.db, "Labels"))

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

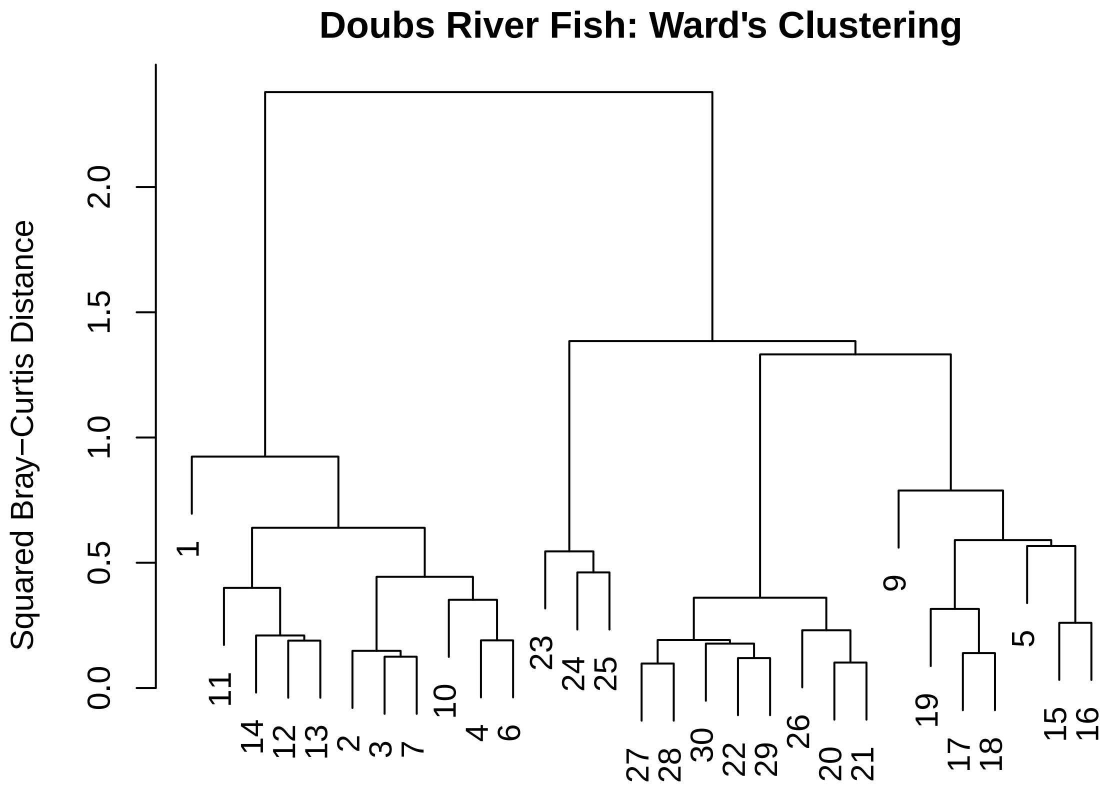


B. Cluster Analysis

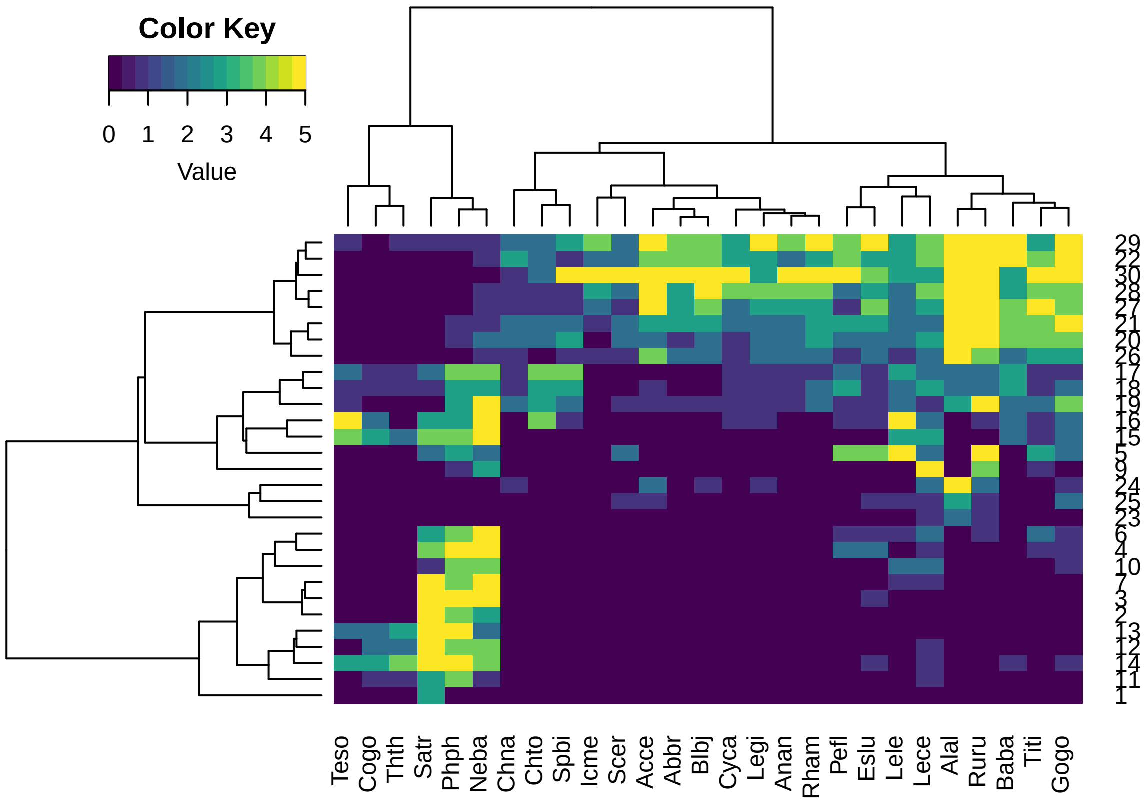
In the R code chunk below, do the following:

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

```
fish.ward <- hclust(fish.db, method = "ward.D2")
par(mar = c(1,5,2,2)+0.1)
plot(fish.ward, main = "Doubs River Fish: Ward's Clustering",
     ylab = "Squared Bray-Curtis Distance")
```



```
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***: direction towards the source of the river affects fish diversity

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.

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

```
## [1] 81.3
```

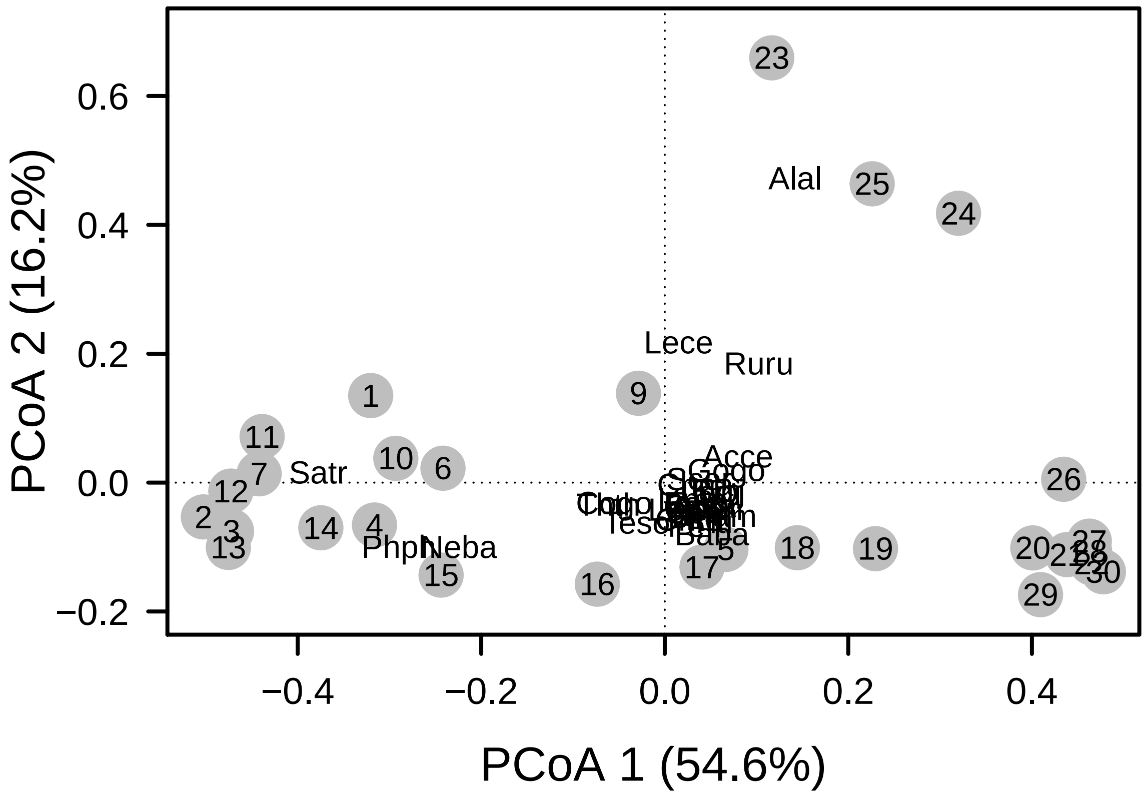
```
par(mar = c(5,5,1,2) + 0.1)
plot(fish.pcoa$points[,1], fish.pcoa$points[,2], ylim = c(-0.2, 0.7),
```

```r
     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(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))

fishREL <- fish
  for(i in 1:nrow(fish)){
    fishREL[i, ] = fish[i,] / sum(fish[i,])
  }

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.

```r
spe.corr <- add.spec.scores(fish.pcoa, fishREL, method = "cor.scores")$cproj

corrcut <- 0.7
imp.spp <- spe.corr[abs(spe.corr[, 1]) >= corrcut | abs(spe.corr[, 2]) >= corrcut, ]
imp.spp
```

```
##            Dim1       Dim2        Dim3
## Phph -0.8674640 -0.1699316 -0.12463098
## Neba -0.7674114 -0.1855678 -0.36963830
## Rham  0.8088751 -0.4192567  0.14136301
## Legi  0.8201759 -0.1701803  0.12423941
## Cyca  0.7595122 -0.4442926  0.17313658
## Abbr  0.7704744 -0.3452714  0.29277803
## Acce  0.7635195  0.2155765  0.10288179
## Blbj  0.8118483 -0.1324698  0.25581178
## Alal  0.4471283  0.8119843 -0.05167131
## Anan  0.7974122 -0.3918972  0.20944968
```

```r
fit <- envfit(fish.pcoa, fishREL, perm = 999)
fit
```

```
##
## ***VECTORS
##
##          Dim1     Dim2     r2 Pr(>r)
## Cogo -0.83884 -0.54438 0.2982  0.012 *
## Satr -0.99904  0.04371 0.4326  0.005 **
## Phph -0.94110 -0.33813 0.7814  0.001 ***
## Neba -0.91413 -0.40543 0.6234  0.001 ***
## Thth -0.87692 -0.48063 0.2634  0.023 *
## Teso -0.44704 -0.89452 0.1700  0.076 .
## Chna  0.99707 -0.07644 0.4612  0.001 ***
## Chto  0.42032 -0.90738 0.2579  0.029 *
## Lele  0.33041 -0.94384 0.0495  0.547
## Lece  0.06856  0.99765 0.3399  0.013 *
## Baba  0.54118 -0.84091 0.6752  0.001 ***
## Spbi  0.57341 -0.81927 0.4138  0.002 **
## Gogo  0.97507  0.22188 0.3753  0.003 **
## Eslu  0.72044 -0.69352 0.1673  0.084 .
## Pefl  0.43762 -0.89916 0.3048  0.008 **
## Rham  0.72476 -0.68901 0.8301  0.001 ***
## Legi  0.93461 -0.35568 0.7016  0.001 ***
## Scer  0.98569  0.16858 0.3533  0.010 **
## Cyca  0.68181 -0.73153 0.7743  0.001 ***
## Titi  0.64378 -0.76521 0.4586  0.002 **
## Abbr  0.77254 -0.63497 0.7128  0.001 ***
## Icme  0.75626 -0.65427 0.5270  0.001 ***
## Acce  0.88799  0.45986 0.6294  0.001 ***
## Ruru  0.48379  0.87518 0.5177  0.001 ***
## Blbj  0.95802 -0.28671 0.6766  0.001 ***
## Alal  0.28755  0.95777 0.8592  0.001 ***
## Anan  0.74277 -0.66954 0.7894  0.001 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Permutation: free
## Number of permutations: 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***: One cluster is mostly defined by the high abundance of species Satr, Phph, Neba. The next cluster is defined by a low abundance of the previously-mentioned species. Then, there are three groups: a subgroup with high abundance of Lece, Alal, and Ruru; another group with high abundance of Teso, Cogo and Thth, and Chto and Sppbi; the last group is comprised by sites with a high abundance of the remaining species. ***Answer 7b***: I would argue that species in the downstream sites might be more important for assesing river quality because that is where most nutrients accumulate. Thus, I would pick a rare species that can only be found downstream, such as Chna (Chondrostoma nasus)

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?

```
JB <- read.delim("JellyBeans.txt", header = T)
BirthdayCakeMix <- JB$WhiteSolid + JB$Rainbow
Lime <- JB$GreenTrans + JB$GreenTrans2
row.names(JB) <- JB$Site
JB <- JB[,-c(2,14,15,27,30)]
JB <- cbind(JB,Lime,BirthdayCakeMix)
JBa <- JB[-c(4,5,7,9),]
JBb <- JB[-c(1,2,3,6,8),]

JB.ds <- vegdist(JB[,-1], method = "bray", binary = TRUE)
JBa.ds <- vegdist(JBa[,-1], method = "bray", binary = TRUE)
JBb.ds <- vegdist(JBb[,-1], method = "bray", binary = TRUE)

t.test(JBa.ds,JBb.ds) # group A vs Group B
```

```
##
##  Welch Two Sample t-test
##
## data:  JBa.ds and JBb.ds
## t = -2.4743, df = 6.9736, p-value = 0.04269
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.140338491 -0.003127327
## sample estimates:
## mean of x mean of y
## 0.1510523 0.2227852
```

```
t.test(JB.ds,JBa.ds) # group A vs Whole community
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  JB.ds and JBa.ds
## t = 1.3424, df = 18.555, p-value = 0.1957
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.01069067  0.04875423
## sample estimates:
## mean of x mean of y
## 0.1700840 0.1510523
```

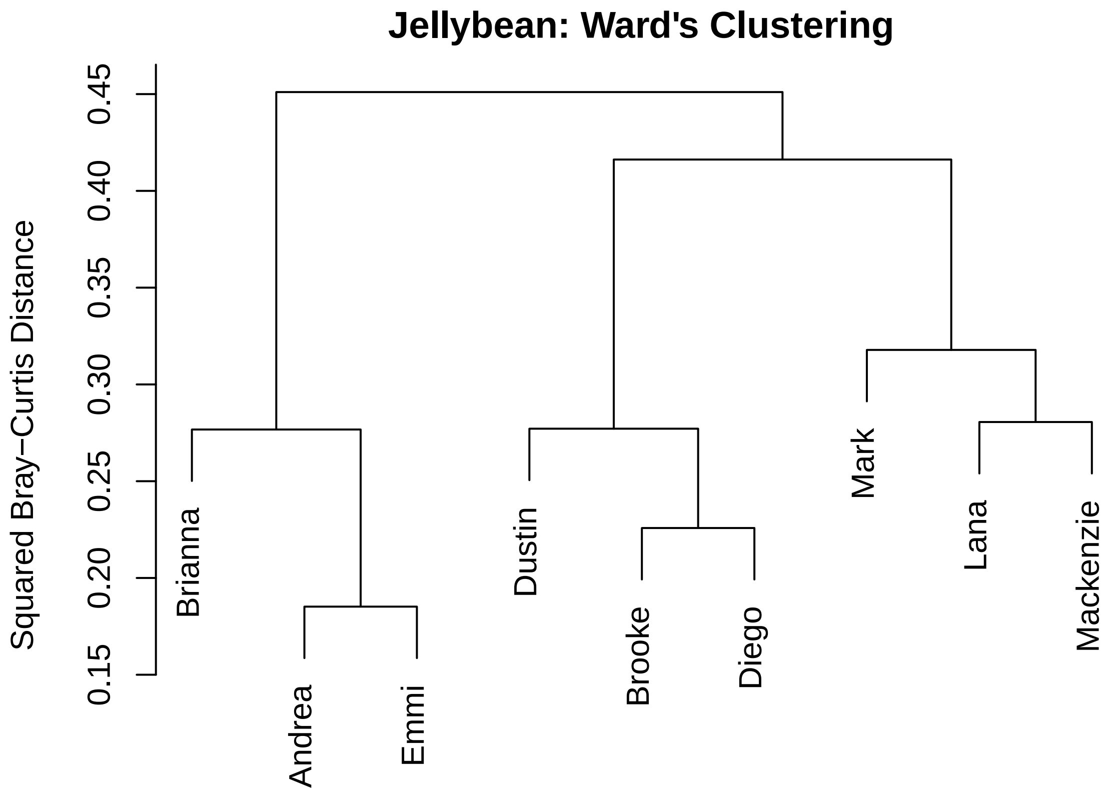
```
t.test(JB.ds,JBb.ds) # group B vs Whole community
```

```
##
##  Welch Two Sample t-test
##
## data:  JB.ds and JBb.ds
## t = -1.9, df = 5.9572, p-value = 0.1065
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.12069030  0.01528803
## sample estimates:
## mean of x mean of y
## 0.1700840 0.2227852
```

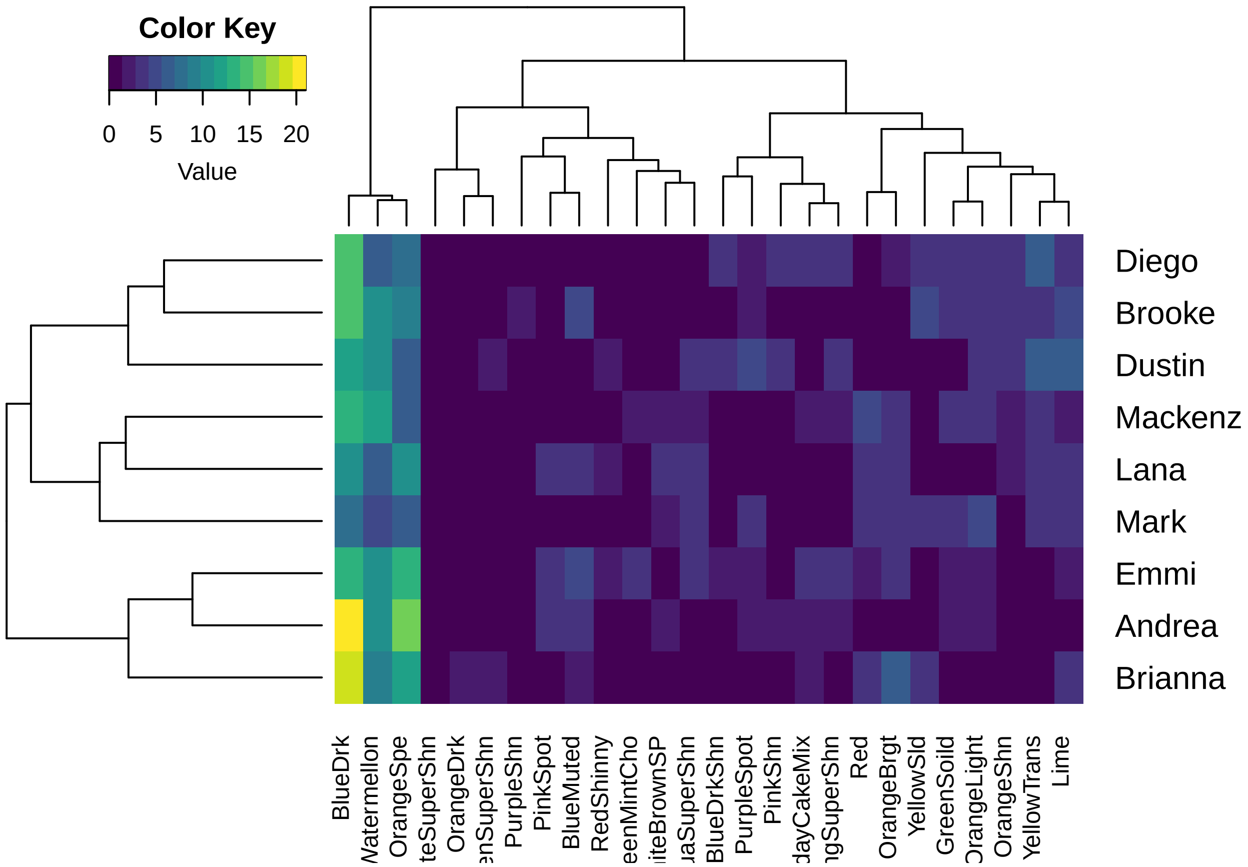
the species composition was significantly different between groups; but there wasn't a significant difference between groups and the source community.

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?

```
JB.db <- vegdist(JB[,-1], method = "bray", upper = TRUE, diag = TRUE)
JB.ward <- hclust(JB.db, method = "ward.D2")
par(mar = c(1,5,2,2)+0.1)
plot(JB.ward, main = "Jellybean: Ward's Clustering",
     ylab = "Squared Bray-Curtis Distance")
```



```
gplots::heatmap.2(as.matrix(JB[,-1]), distfun = function(x) vegdist(x, method= "bray"),
                  hclustfun = function(x) hclust(x, method = "ward.D2"),
                  col = viridis, trace = "none", density.info="none")
```



There seems to be no differences in Jellybean composition between groups A and B. Three sites of group A (Brianna, Andrea, and Emmi) were clustered together, while the other two sites were clustered within a cluster that is comprised of sites from Group B. The species that contribute the most to the clustering are BlueDrk, Watermellon, and OrangeSpe, which were more abundant in the main Group A cluster.