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

Mark Hibbins; 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).

```
remove(list=ls())
getwd()
```

```
## [1] "/Users/mark/Box Sync/Courses/Quantitative Biodiversity/QB2019_Hibbins/2.Worksheets/8.BetaDivers
```

```
setwd('/Users/mark/Box Sync/Courses/Quantitative Biodiversity/QB2019_Hibbins/2.Worksheets/8.BetaDiversi
library(vegan)
```

```
## Loading required package: permute
```

```
## Loading required package: lattice
```

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

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, pleae do not print the dataset when submitting
library(ade4)
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:
```

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: Four ***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 richess appears to be higher for downstream sites than upstream sites. ***Answer 2b***: Trout abundance appears to be higher at upstream sites than at downstream sites. ***Answer 2c***: Richness doesn't capture information about the identity of the species, so we have no way of knowing if the species composition of the downstream sites is the same even if they have similar richness. Also, the lack of identity information means we can't understand the processes that are driving the differences in diversity across sites.

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 = '', sitenum1 = '', sitenum2 = '', pairwise = FALSE){

  # if pairwise is true, calculate turnover
  if (pairwise == TRUE){

    if (sitenum1 == '' | sitenum2 == ''){
      print('No sites provided for comparison')
      return(NA)
    }

    #if it passes the check, we can estimate the statistic
    site1 = site.by.species[sitenum1,]
    site2 = site.by.species[sitenum2,]
    site1 = subset(site1, select = site1 > 0) #remove absences
    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)
  }

  #if pairwise is left at false, estimate diversity
  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)
  }
}

beta.w(site.by.species = doubs$fish, pairwise = FALSE)
```

```
## [1] 2.16
```

```
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: Regional diversity is 2.16 times higher than the average local diversity across sites, suggesting that beta diversity plays an important role in contributing. ***Answer 3b***: Site 1 is more similar to site 2, because there is less turnover between these two sites. ***Answer 3c***: The multiplicative definition of beta diversity can be thought of as how many times higher gamma diversity is than would be expected if it was shaped by diversity within sites alone. If it was defined additively instead, it would be the additional contribution of beta diversity to gamma diversity after alpha diversity is considered.

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:

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,]
fish.ds <- vegdist(fish, method = 'bray', binary = TRUE, upper = TRUE, diag = TRUE)
fish.db <- vegdist(fish, method = 'bray', upper = TRUE, diag = TRUE)
```

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: It represents dissimilarity, because all the diagonal values (which are comparing the site with itself) have values of 0. ***Answer 5b***: There do not appear to be glaring differences in dissimilarity between the two matrices; values are different, but reasonably within the same range. One might argue that the Bray-Curtis distance is more informative because it incorporates data on species abundances.

4) VISUALIZING BETA-DIVERSITY

A. Heatmaps

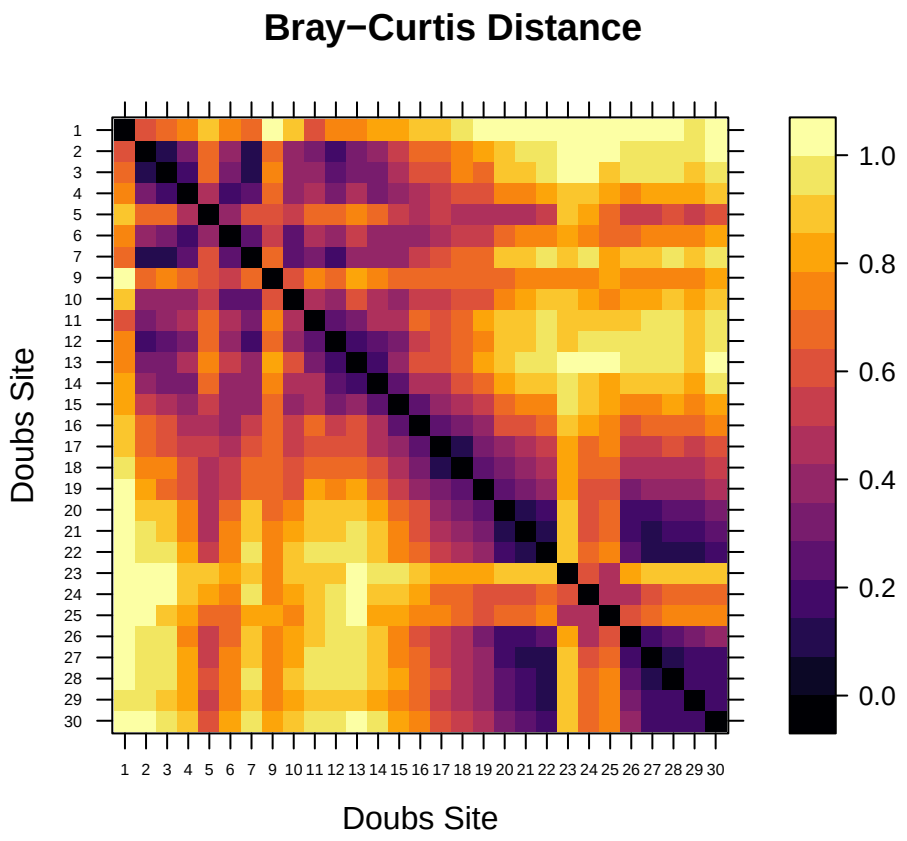
In the R code chunk below, do the following:

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.

```
library(lattice)
library(viridis)
```

```
## Loading required package: viridisLite
```

```r
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')
```

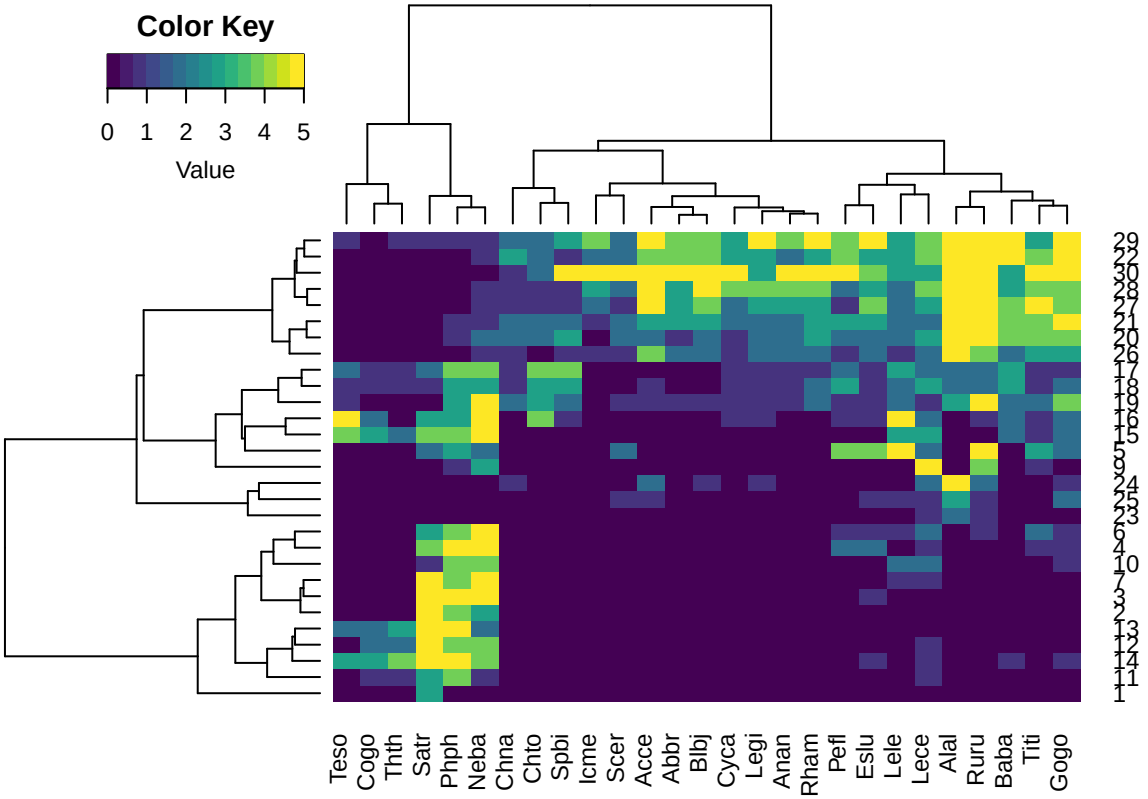


B. Cluster Analysis

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`).

```r
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***: Based on the Bray-Curtis heatmap and the cluster plot above, there appear to be two distinct ecological communities; one upstream (lower number sites), and one downstream (higher number sites). There is also somewhat of a transition zone between the upstream and downstream sites where fish from both communities can be found. Since downstream sites also appear to be more abundant, this pattern may be driven by differences in nutrient quality (with more abundant nutrients downstream).

C. Ordination

Principal Coordinates Analysis (PCoA)

In the R code chunk below, do the following:

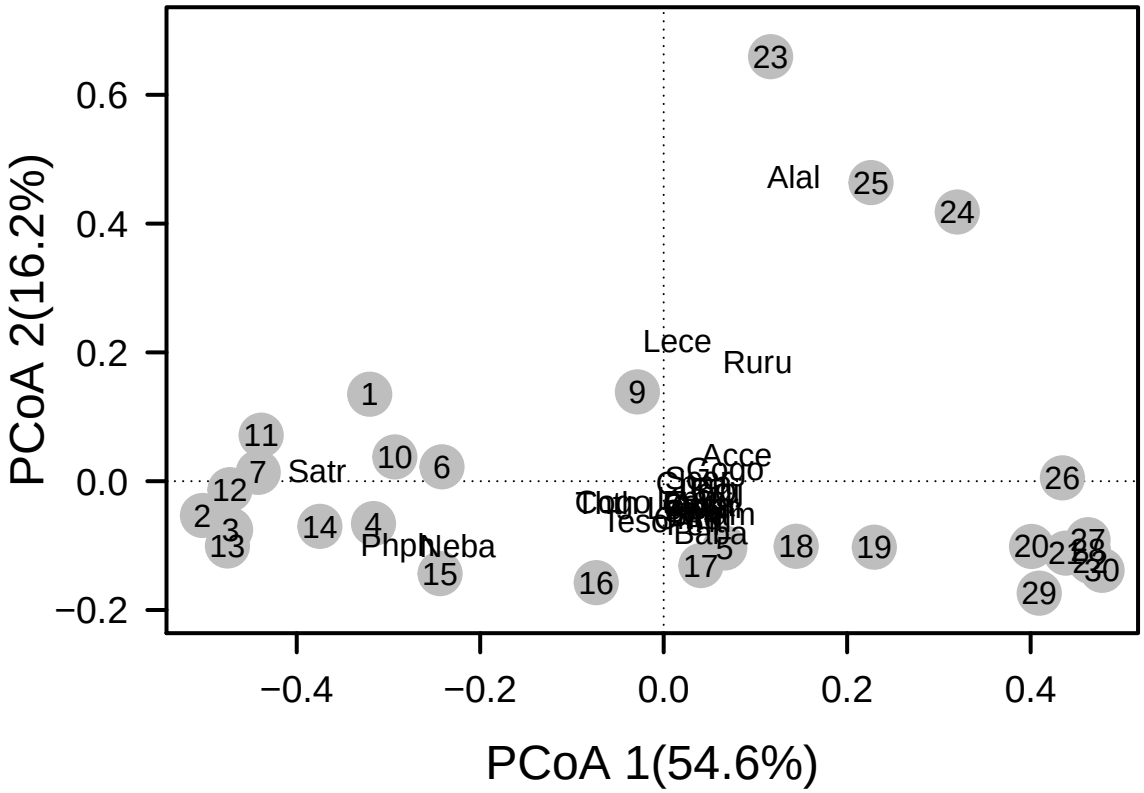
1. perform a Principal Coordinates Analysis to visualize beta-diversity
2. calculate the variation explained by the first three axes in your ordination
3. plot the PCoA ordination,
4. label the sites as points using the Doubs River site number, and
5. identify influential species and add species coordinates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

```
#Doing the PCA
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)
```

```r
#Plotting ordination

par(mar = c(5, 5, 1, 2) + 0.1)

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)

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')

#label points by site number
text(fish.pcoa$points[ ,1],
     fish.pcoa$points[ ,2],
     labels = row.names(fish.pcoa$points))

#adding coordinates for species contributions
library(tcltk)
library(BiodiversityR)
```

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

```r
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.

```
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, ]

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***: The PCoA plot confirms the trends in the cluster analysis; there appear to be roughly two clusters which are separated primarily along the first axis. This axis is probably related to the geographical position of the site (ie. upstream or downstream). There are also a few downstream sites that appear to be in their own grouping, differentiated on the second axis. ***Answer 7b***: Most of the fish species appear to cluster around the center of the ordination space. There are a few that seem highly influential; Salmo trutta fario, Phoxinus phoxinus and Nemacheilus barbatulus for axis 1; Alburnus alburnus, Leuciscus cephalus cephalus and Rutilus rutilus for axis 2. Axis 1 is probably related to nutrient quality associated with the stream location, so the species that drive differentiation along axis 1 may be good bioindicators.

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?

```
#Comparison between groups A and B

JellyBeans.Source <- read.delim("~/Box Sync/Courses/Quantitative Biodiversity/QB2019_Hibbins/2.Workshee
JellyBeans <- read.delim("~/Box Sync/Courses/Quantitative Biodiversity/QB2019_Hibbins/2.Worksheets/6.Di

#Lump jellybean categories
JellyBeans$Rainbow <- rowSums(JellyBeans[, c(27, 30)])
JellyBeans <- JellyBeans[1:29]

JellyBeans$GreenTrans <- rowSums(JellyBeans[, c(14, 15)])
JellyBeans <- JellyBeans[c(1:14, 16:29)]

#Subset by group
JellyBeans_A <- subset(JellyBeans, Group == 'A')[3:28]
JellyBeans_B <- subset(JellyBeans, Group == 'B')[3:28]

#Compare composition
jb_a_db <- vegdist(JellyBeans_A, method = 'bray')
jb_b_db <- vegdist(JellyBeans_B, method = 'bray')

mean(jb_a_db)
```

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

```
mean(jb_b_db)
```

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

```
t.test(jb_a_db, jb_b_db)
```

```
##
##  Welch Two Sample t-test
##
## data:  jb_a_db and jb_b_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
```

There is a slightly significant difference ($p = 0.033$) in the average pairwise similarity between sites in group A vs. group B.

```
#Comparison to source community
jb_a_alpha <- mean(specnumber(JellyBeans_A))
jb_b_alpha <- mean(specnumber(JellyBeans_B))

26 / jb_a_alpha
```

```
## [1] 1.181818
```

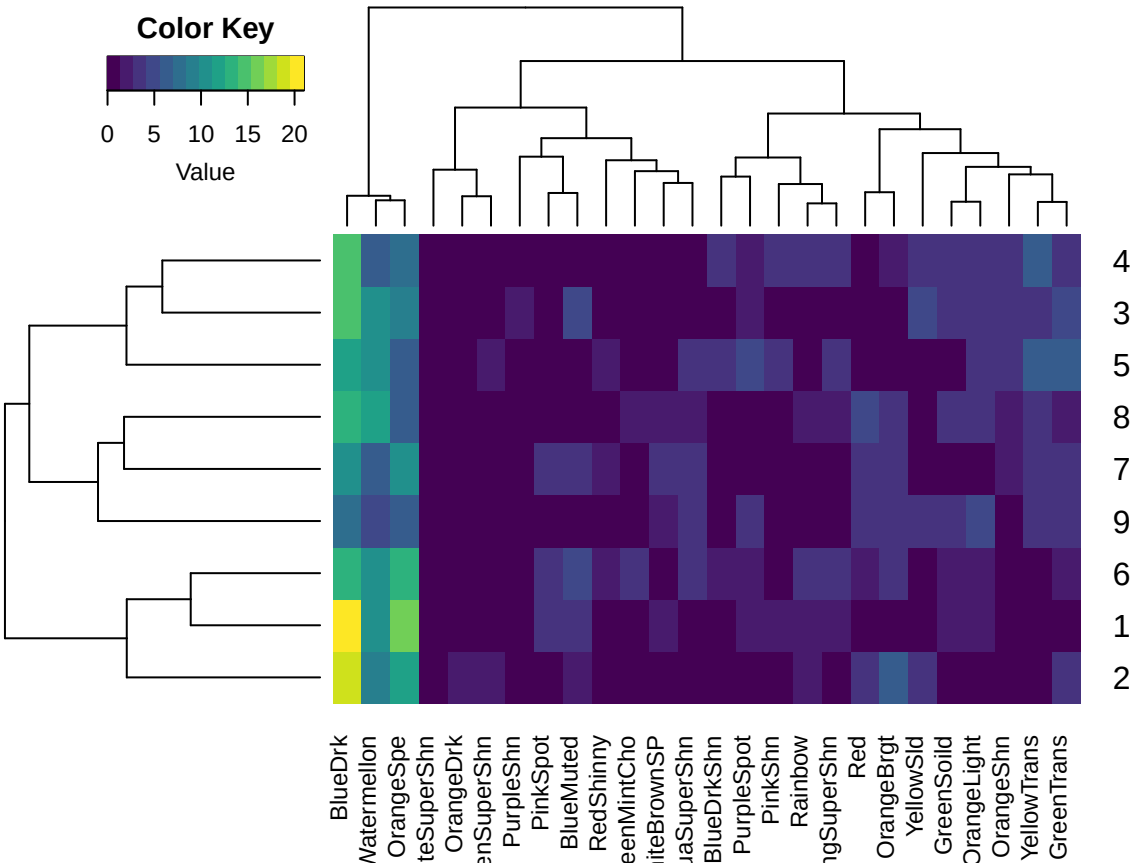
```
26 / jb_b_alpha
```

```
## [1] 1.253012
```

Both groups have approximately the same amount of beta diversity, with group B having slightly more variation between sites.

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?

```
gplots::heatmap.2(as.matrix(JellyBeans[3:28]),
                  distfun = function(x) vegdist(x, method = 'bray'),
                  hclustfun = function(x) hclust(x, method = 'ward.D2'),
                  col = viridis,
                  trace = 'none',
                  density.info = 'none')
```



This cluster diagram does not appear to indicate any distinctions in the composition of groups A and B. The most common species does appear to be more common in sites 1 and 2, which are both part of group A, but it is not really a strong pattern.