

Microbiome Tutorial

Tutorial aims:

1. Understand the relevancy of linking genomes to phenomes to understand physiology leveraging knowledge of collaborative research projects in the gut microbiome of herbivores
2. Demonstrate basic analyses of microbiota data
3. Determine if and how communities differ by variables of interest.
4. Perform various measurements characterizing microbial community diversity, composition, and structure

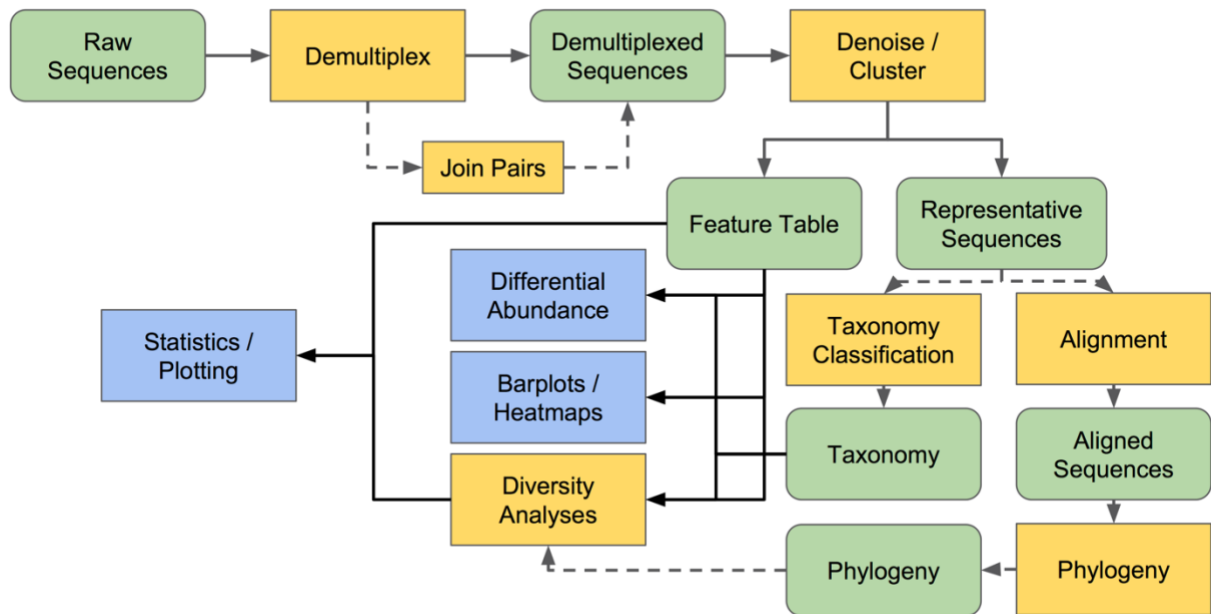
Background

In order to assess how plant secondary compounds influence diet selection, nutritional status, and detoxification in large mammalian herbivores during winter months, a population of moose (*Alces alces*) on Isle Royale National Park, Michigan was followed over a 6-year period. Microbial communities in the intestinal tract of herbivores may play vital roles in detoxification of plant secondary compounds. The coevolutionary arms race refers to the reciprocal coevolution of plants and herbivores where in which plants develop secondary compounds (PSCs) toxic to their herbivore predators. Digestion and detoxification of these volatile compounds is thought to be primarily associated with important microbial interactions. Experimental diet manipulation can be impractical for wild herbivore systems as it may not capture true interactions and variations naturally occurring.

We captured fecal and urine samples from moose on east and west sides of Isle Royale National Park in Lake Superior, Michigan in the Winter season over a three year period. The eastern and western regions of the island provide a naturally occurring variation in diet by location due to available vegetation. One such species is the balsam fir (*Abies balsamea*) which has not only been shown to produce PSCs toxic to moose, but is also the most important food source for *A.alces* in the winter months (when available vegetation is limited). Relative abundance of this plant species has been diminishing in the western portion of the island over time while the population in the east has remained stable. Microbial community diversity can be determined via 16S metagenomic analysis of fecal samples . We found that the nutritional and chemical properties of the diet varied between years and between locations on the island. Additionally, analysis revealed richness of microbial species within individual moose was altered by this diet change. The natural formation of these two distinct non-interactive moose populations divided by region provided a great study system to explore 1) shifts in microbial diversity in response to diet shifts as well as 2) provide a link to connect microbial diversity to population ecology. The data provides future opportunities for understanding the role of specific taxonomic groups with specific chemical compounds. This information can provide insight for conservation of this species as well as other herbivorous mammals that participate in coevolution with plant species.

Polymerase chain reactions (PCR) using primers targeting the V3-V4 hypervariable regions of the 16S rRNA gene (Klindworth et al., 2013) were performed to isolate microbial metagenomes. Sequencing was conducted by the University of California San Diego Microbiome Institute using Illumina EMP protocol 515fbc, 806r amplification of 16S rRNA V4 repeated in all rows.

The Microbiome



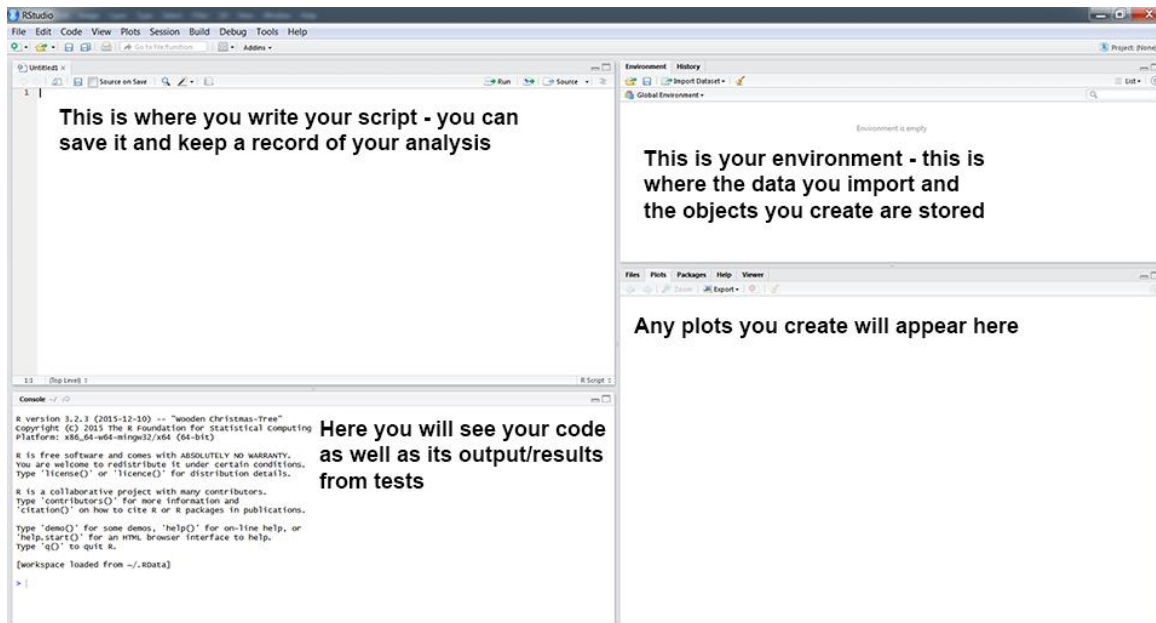
In microbiome experiments, investigators frequently wonder about things like:

- How many different species/OTUs/ASVs are present in a set of samples?
- How much phylogenetic diversity is present in each sample?
- How similar/different are individual samples and groups of samples?
- What factors (e.g., pH, elevation, blood pressure, body site, or host species just to name a few examples) are associated with differences in microbial composition and biodiversity?

At this point of analysis you have a feature table, taxonomy classification results, a phylogenetic tree, and a metadata file.

Open R Studio

Open RStudio. Click on “File/New File/R script”.



Step 1 and 2: Download and load packages

Start by recording who is writing, the date, and the main goal. Follow this by setting a random seed number at the start of the workflow (Keep this number). This action allows for the reproducibility of your code when randomized methods are performed. It also may not be saved when you exit R. Always make sure you consistently set.seed prior to running your code.

```
##Olivia Rodriguez  
#Zool 421  
#Microbiome Tutorial
```

```
set.seed(2863)
```

The next few lines of code involve installing **and loading** the packages.

Remember: To install a package, type `install.packages("package-name")`. You only need to install packages once, but after installing the package you need to load the package by calling upon it via the function `library(package-name)`. Notice how there are no quotation marks when you call upon the package, but there are when you install the package.

```
###Step 1: Install packages###
```

```
if (!requireNamespace("BiocManager", quietly = TRUE))
```

```
  install.packages("BiocManager")
```

```
BiocManager::install("phyloseq")
```

```

BiocManager::install("microbiome")

install.packages("ggplot2")
install.packages("dplyr")

###Step 2: Load Packages###

library(phyloseq)
library(ggplot2)
library(dplyr)
library(microbiome)

## Installing Q2 plugin##

if (!requireNamespace("devtools", quietly =
TRUE)){install.packages("devtools")}
devtools::install_github("jbisanz/qiime2R")

library(qiime2R)

```

Let's make sure we've defined our **working directory**.

Remember: To find out where your working directory is now, run the code `getwd()`. If you want to change it, you can use `setwd()`.

```

getwd()
[1] "/Users/mac/Desktop/Microbiome_tutorial"

#My working directory points to a folder on my desktop named
Microbiome_Tutorial

```

Step 4: Import and Check Data

The data used here was created using 2x250 bp amplicon sequencing of the bacterial V4 region of the 16S rRNA gene on the Illumina MiSeq platform. We will be correlating the fecal bacterial microbiota of 326 moose to variables like region of habitat and year.

First, you will need to download the data. Go to your blackboard course site → Lab modules → Mammal Microbiome → download all files in the folder titled "data"

```

###Step 4: Reading in data###

#Read in data table of sample reads#
SVs<-read_qza("feature_table.qza")

asv_table<-(SVs$data)

```

```

Asv_table<-data.frame(asv_table)

colnames(Asv_table) <- colnames(asv_table)

#Read in phylogenetic tree#
tree<- read_qza("rooted_tree.qza")

Tree<-tree$data

#Read in metadata (data table of information about samples)#

Metadata_final<-read.csv("moose-metadata.csv")

#Only do this step If you have a windows system. You need to rename the
SampleID column because R in windows renames the SampleID column when import#
names(Metadata_final)[1] <- " SampleID"

# get column names to assure renaming worked#
colnames(Metadata_final)[1]

# Others without Windows systems, skip to this step#

metadata_finalnames <- as.character(Metadata_final$SampleID)

rownames(Metadata_final) <- metadata_finalnames

#Read in Taxonomy Table#
taxonomy<-read_qza("classified_rep_seqs.qza")
head(taxonomy$data)

##               Feature.ID
## 1 ff09a02e713739b2fd8691060c448a8c
## 2 44fc97e2afa5cb0ad899cfff1514b024
##
Taxon
## 1
D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Lachnos
piraceae
## 2
D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Ruminoc
occaceae;D_5__Ruminococcaceae UCG-005
## Confidence
## 1 0.9999996
## 2 0.9982597

```

#when taxonomy is imported, a single string is returned along with a confidence score.

#For many analysis we will want to break up this string and for that purpose the parse_taxonomy() function is provided#

```
taxonomy<-parse_taxonomy(taxonomy$data)
head(taxonomy)
```

```
##                               Kingdom      Phylum      Class
Order
## ff09a02e713739b2fd8691060c448a8c Bacteria Firmicutes Clostridia
Clostridiales
## 44fc97e2afa5cb0ad899cfff1514b024 Bacteria Firmicutes Clostridia
Clostridiales
## 894d1b85fa0596dbecea9e122b7175f1 Bacteria Firmicutes Clostridia
```

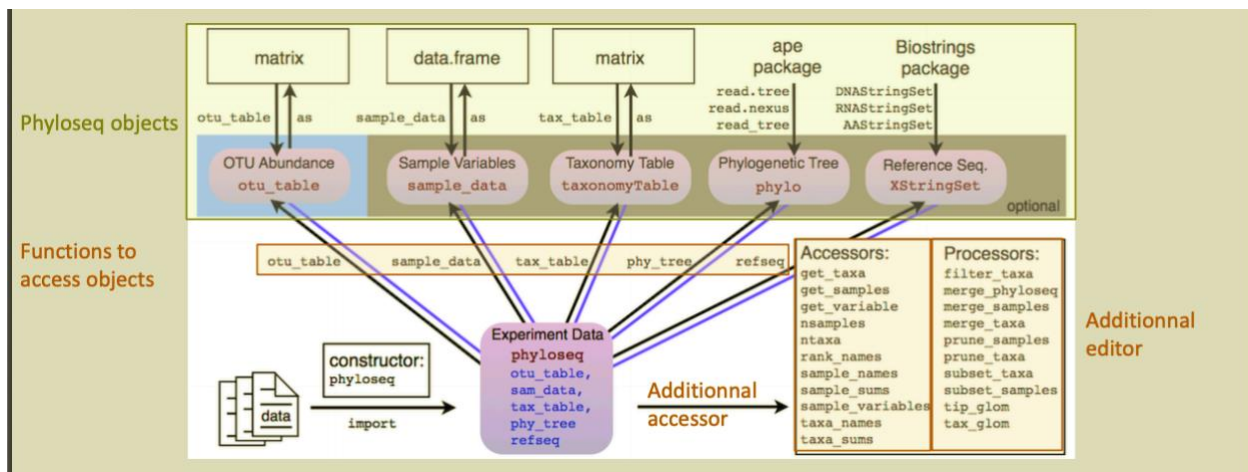
```
taxonomy$Kingdom <- as.character(taxonomy$Kingdom)
taxonomy$Phylum <- as.character(taxonomy$Phylum)
taxonomy$Class <- as.character(taxonomy$Class)
taxonomy$Order <- as.character(taxonomy$Order)
taxonomy$Family <- as.character(taxonomy$Family)
taxonomy$Genus <- as.character(taxonomy$Genus)
taxonomy$Species <- as.character(taxonomy$Species)
```

```
taxonomy <- as.matrix(taxonomy)
```

Step 5: Phyloseq

Create and combine sequencing, phylogenetic, and sample data into a condensed object for use in downstream analyses in Phyloseq. The Phyloseq object you create serves as the foundation for the remainder of this pipeline. Phyloseq is a nice data structure to store the count table, taxonomic information, contextual data and phylogenetic tree as different components of a single R object. A phyloseq object is made of up to 5 components:

- otu_table: an OTU abundance table;
- sample_data: a table of sample metadata, like sequencing technology, location of sampling, etc;
- tax_table: a table of taxonomic descriptors for each OTU, typically the taxonomic assignment at different levels (phylum, order, class, etc.);
- phy_tree: a phylogenetic tree of the OTUs;
- refseq: a set of reference sequences (one per OTU).



You can summarize the phyloseq object by entering the object name

###Step 5: Create phyloseq object###

```
phyloseq_16S_moose <- phyloseq(otu_table(Asv_table, taxa_are_rows = TRUE,
errorIfNULL = TRUE), sample_data(Metadata_final), tax_table(taxonomy),
phy_tree(Tree))
```

```
phyloseq_16S_moose
```

```
## phyloseq-class experiment-level object
## otu_table() OTU Table: [ 2382 taxa and 324 samples ]
## sample_data() Sample Data: [ 324 samples by 90 sample variables ]
## tax_table() Taxonomy Table: [ 2382 taxa by 7 taxonomic ranks ]
## phy_tree() Phylogenetic Tree: [ 2382 tips and 2319 internal nodes ]
```

###Step 6: On your own###

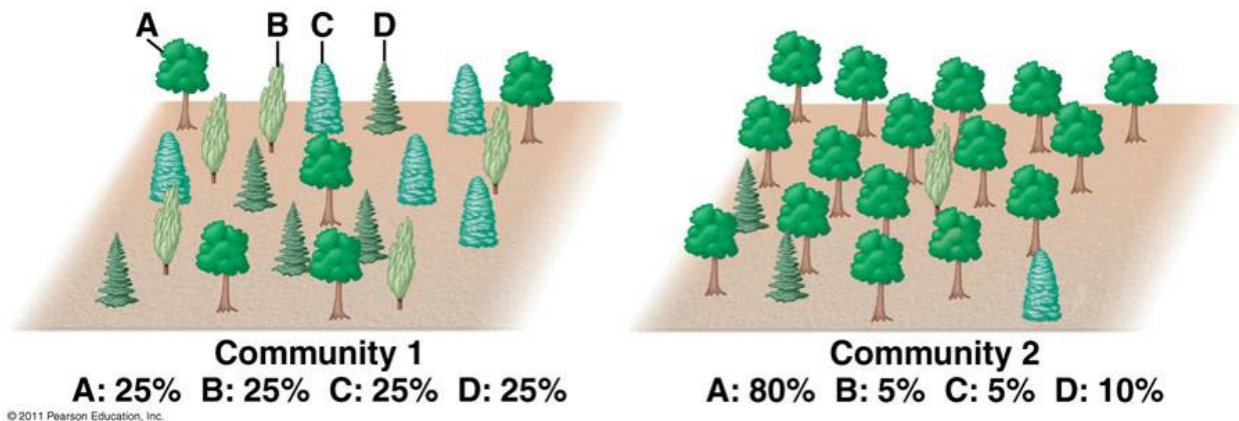
```
#Skip this when running tutorial first time and start here to begin the
deliverable assignment.#
#(You will Subset the object to only include moose samples from a certain
year. Example below for 2013)#
phyloseq_16S_2013 <- subset_samples(phyloseq_16S_moose, year == "2013")

# Create a sample metadata dataframe
meta <- data.frame(sample_data(phyloseq_16S_moose))
```

Step 7: Diversity Analysis

Alpha Diversity

- This is the diversity in a single sample. These values are used to compare between samples or groups of samples.
 - There's 2 categories of alpha metrics
 - Richness: These describes the number of bacterial species in an environment.
 - ACE, Chao1: Richness estimates
 - Observed richness: literal count of the number of taxa present
 - Diversity: These consider both the sample richness and the sample evenness.
 - Shannon (#of species)): higher value = greater diversity
 - Simpson: lower number= greater diversity
 - Phylogenetic diversity: sum of all branch lengths for phylogenetic tree of asvs
 - The image below illustrates richness v. diversity. Although both forests have the same richness (4 species of tree), Community 1 has a more even distribution of the 4 species while Community 2 is dominated by 1 type of tree species. Community 1 is more diverse.



Tests:

- Shannon's diversity index (a quantitative measure of community richness) (Shannon & Weaver, 1949)
- Observed features (a quantitative measure of community richness, called "observed OTUs" here for historical reasons);
- Evenness (or Pielou's Evenness; a measure of community evenness) (Pielou, 1966);
- Faith's Phylogenetic Diversity (a qualitative measure of community richness that incorporates phylogenetic relationships between the features) (Faith, 1992); this metric is sometimes referred to as PD_whole_tree, but we discourage the use of that name in favor of Faith's Phylogenetic Diversity or Faith's PD

Measurements of alpha diversity are often the first major calculations produced

in a microbiome pipeline. In this section, we will focus on the Shannon (overall community diversity) diversity index, however, other metrics are calculated as well and provide additional information. Additionally, we will construct plots to visualize these values across treatments.

Beta Diversity

To compare the structure of the microbiome communities. Beta diversity is between sample diversity. Essentially, how different each sample is from every other sample. So each sample will have more than one value. There are two types of beta diversity metrics:

- Abundance (diversity)
 - Bray-Curtis
 - Weighted UniFrac
- Presence/Absence (richness)
 - Jaccard
 - Unweighted UniFrac
 - The UniFrac take into account the phylogenetic information. So if two samples don't have any ASVs that match but are closely related in phylogeny, the distance between these two in the UniFrac metric would be lower.

Tests

- Jaccard distance (a qualitative measure of community dissimilarity) (P. Jaccard, 1908);
- Bray-Curtis distance (a quantitative measure of community dissimilarity) (Sørensen, 1948);
- unweighted UniFrac distance (a qualitative measure of community dissimilarity that incorporates phylogenetic relationships between the features) (C. Lozupone & Knight, 2005); Implementation based on Striped UniFrac (McDonald et al., 2018) method.
- weighted UniFrac distance (a quantitative measure of community dissimilarity that incorporates phylogenetic relationships between the features) (C. A. Lozupone, Hamady, Kelley, & Knight, 2007); Implementation based on Striped UniFrac (McDonald et al., 2018) method.

Our phyloseq object (phyloseq_16S_moose) is now ready for comparative analyses of microbial community composition and structure. In this section, we will calculate and compare the beta diversity metrics Bray-Curtis dissimilarity (compositional dissimilarity and the alpha diversity metrics Chao1 and Shannon.

###Step 7: Diversity Analysis###

```
##Step 7A: Alpha diversity measurements##
```

```
#Calculate alpha diversity metrics then create a table containing those values
```

```
alpha <- data.frame(estimate_richness(phyloseq_16S_moose, split = TRUE, measures = NULL))
```

```
row.names(alpha) <- row.names(meta)
```

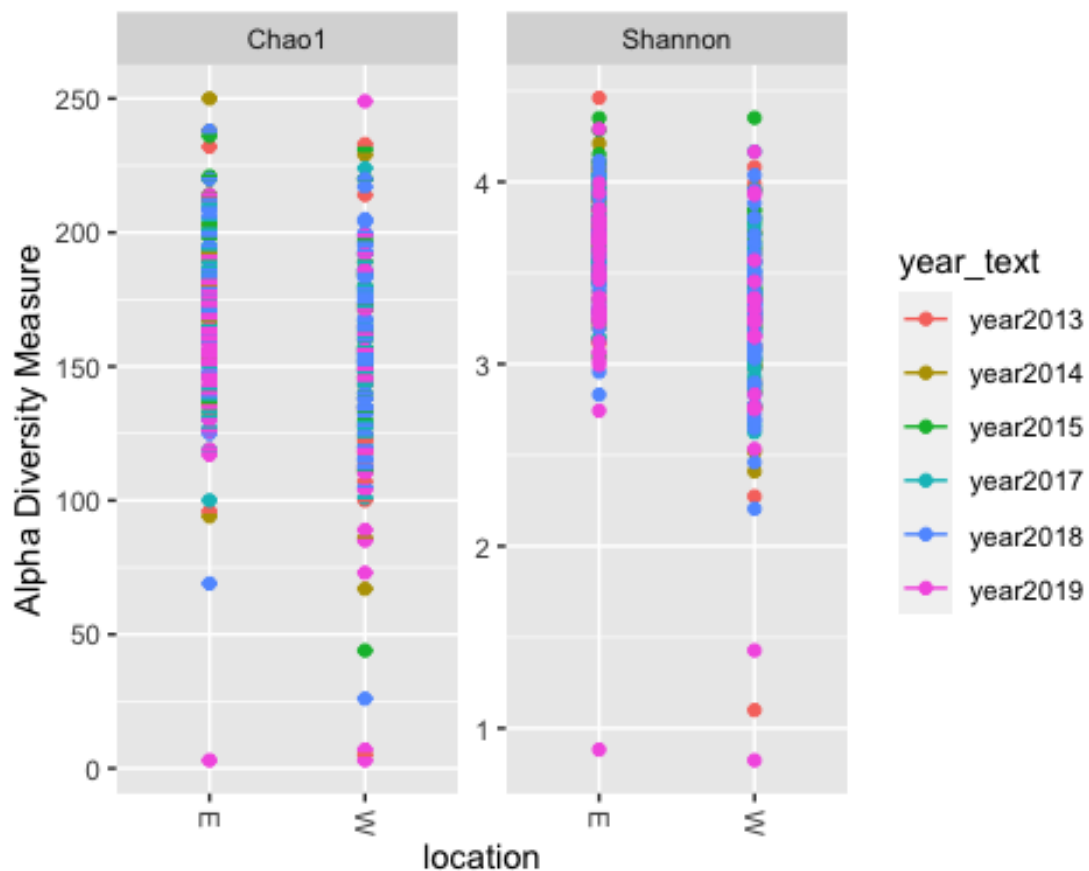
```
alpha$SampleID <- rownames(alpha)
```

```
alpha_meta <- left_join(meta, alpha)
```

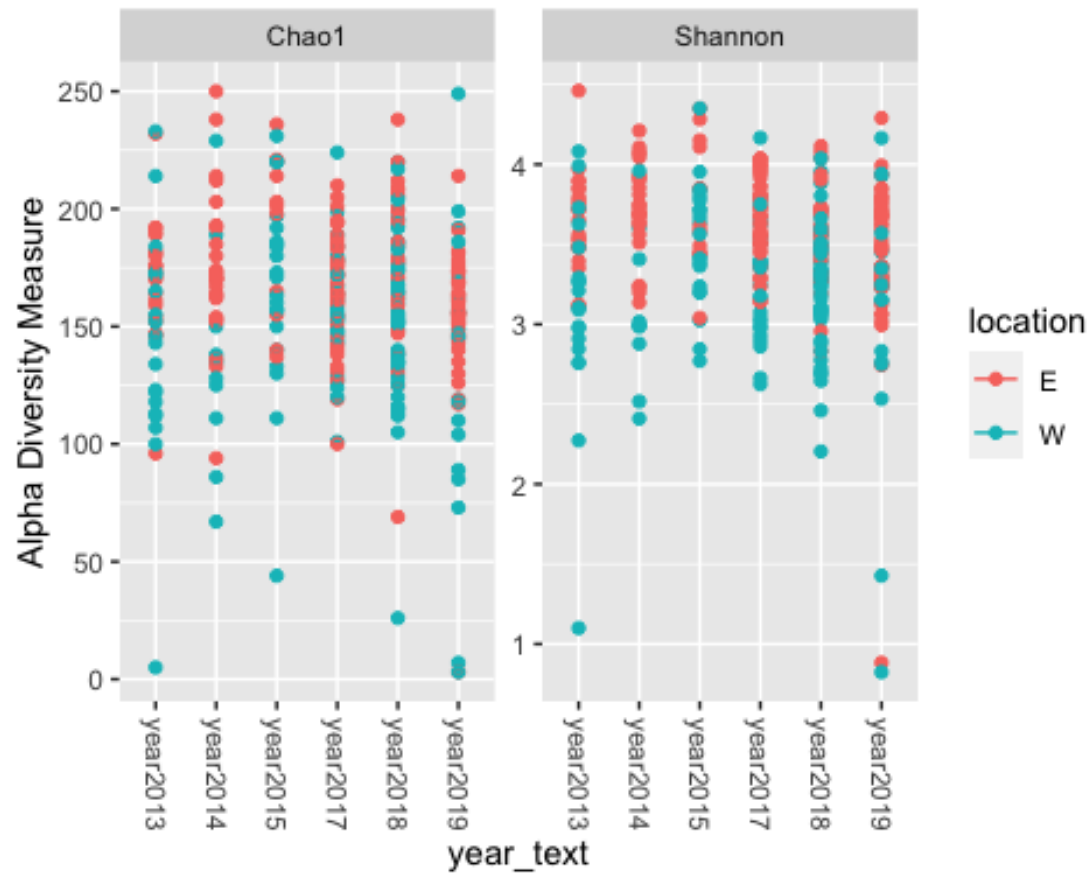
```
## Joining, by = "SampleID"
```

```
#Estimate Species Richness by year then by Location#
```

```
plot_richness(phyloseq_16S_moose, x="location", measures = c("Chao1", "Shannon"), color = "year_text")
```

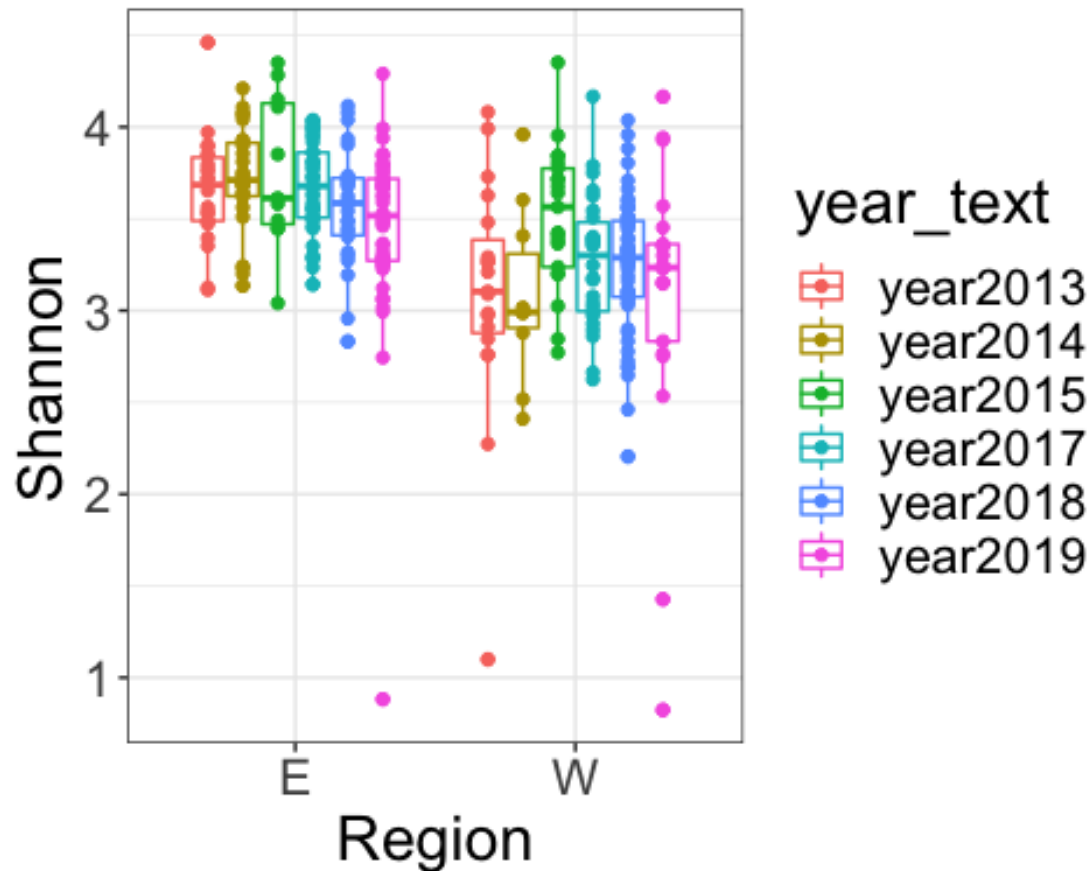


```
plot_richness(phyloseq_16S_moose, x="year_text", measures = c("Chao1", "Shannon"), color = "location")
```



#Plot a Shannon value boxplot for samples group by location and colored by year#

```
shannon_location <-  
  ggplot(alpha_meta, aes(location, Shannon, color = year_text)) +  
  geom_boxplot() + labs(x = "Region") +  
  geom_jitter(position = position_jitterdodge(jitter.width = 0)) +  
  theme_bw()  
shannon_location + theme(text=element_text(size=20))
```



```
##Step 7B: Beta diversity measures##
```

```
#Transform data: Needs to be compositional#
```

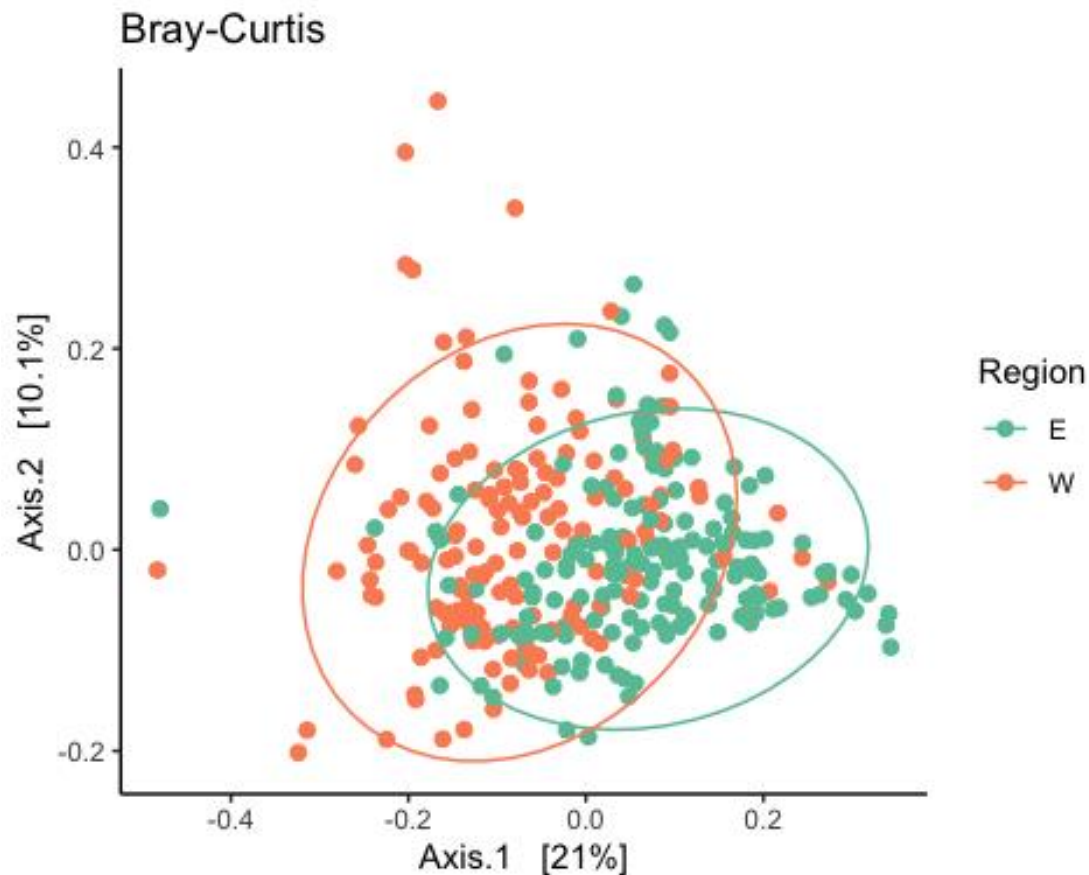
```
moose.comp <- microbiome::transform(phyloseq_16S_moose, "compositional")
```

```
#ordinate the data to PCoA design#
```

```
ord.pcoa.bray <- ordinate(moose.comp, "PCoA", "bray")
```

```
#plot the ordination with samples colored by location#
```

```
moose.bray <- plot_ordination(moose.comp,
                             ord.pcoa.bray, color="location")
moose.bray <- moose.bray + ggtitle("Bray-Curtis") + geom_point(size = 2)
moose.bray <- moose.bray + theme_classic() + scale_color_brewer("Region",
palette = "Set2")
print(moose.bray + stat_ellipse())
```



Step 8: Taxonomy

Now that we've done diversity analysis, we can take a look at the data and create compositional barplots to identify key bacteria.

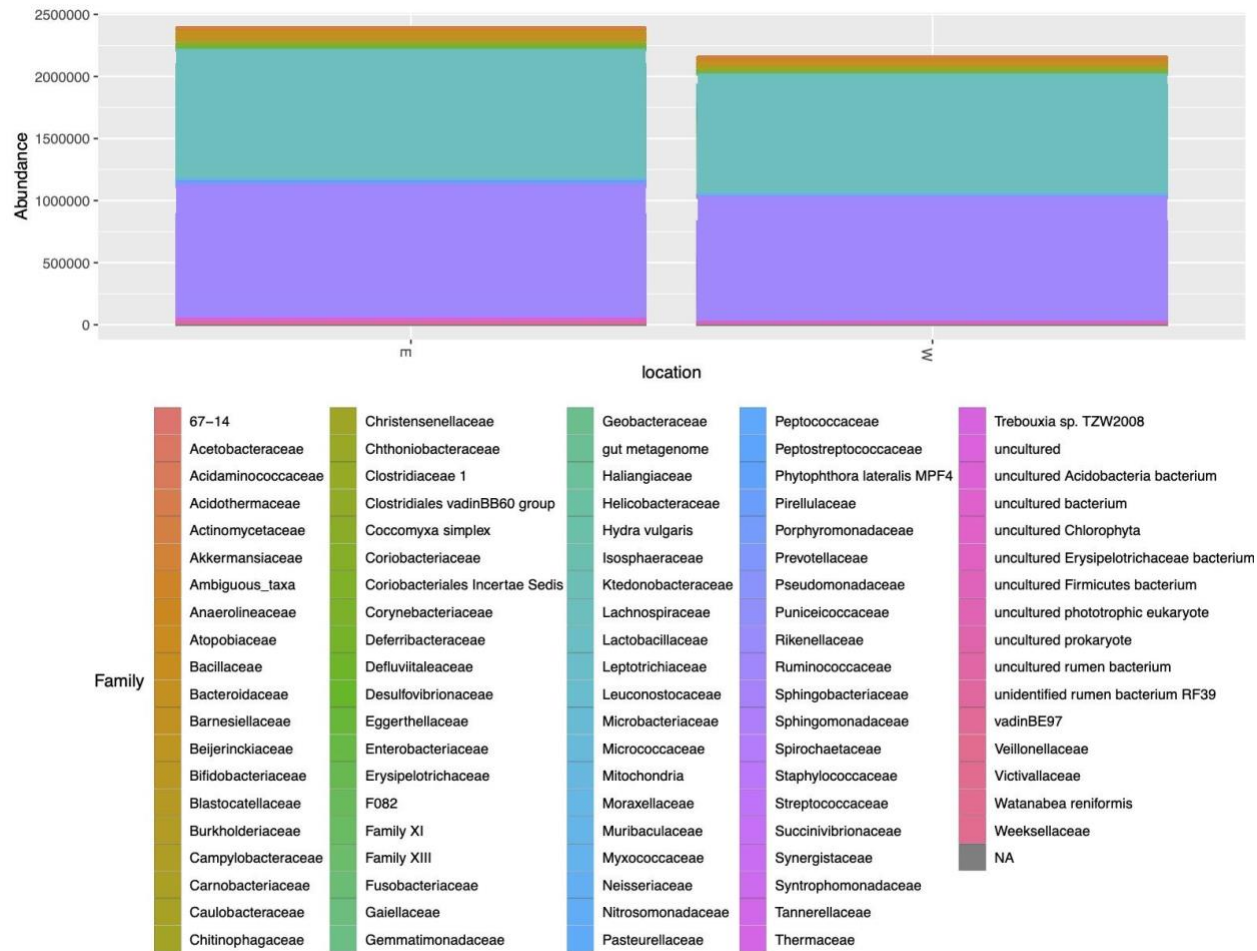
###Step 8: Taxonomy Exploration###

#Subset moose dataset to just the observed Bacteria and remove any other microorganisms#

`moose_bacteria <- subset_taxa(phyloseq_16S_moose, Kingdom=="Bacteria")`

#Step 8A: Create an Initial basic plot of taxonomy grouped by location and identifying all bacterial families#

`plot_bar(moose_bacteria, x="location", fill="Family") +
geom_bar(aes(color=Family, fill=Family), stat="identity", position="stack") +
theme(legend.position="bottom")`



#Sort the Families by abundance and pick the top 10 most abundant families#

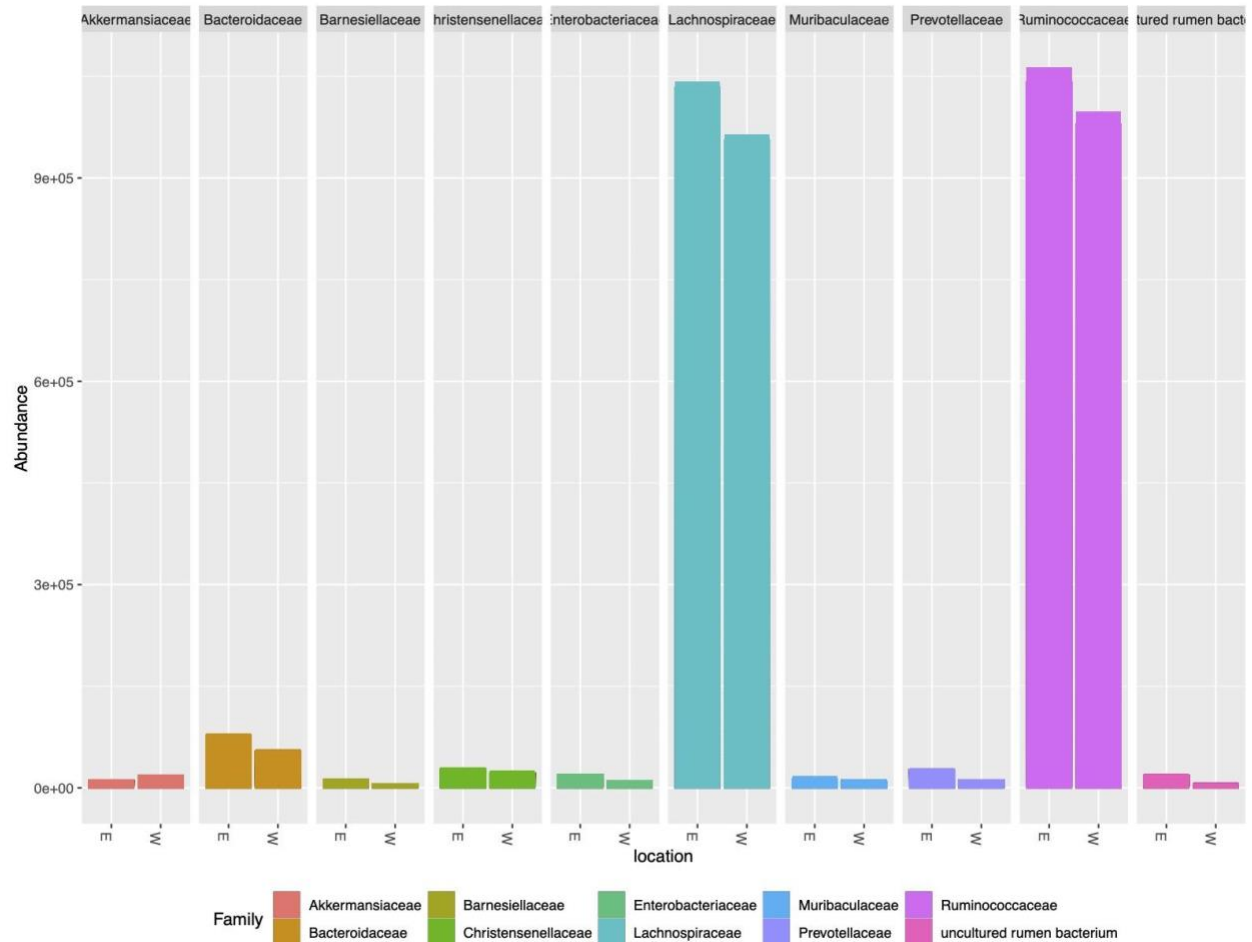
```
top10F.names = sort(tapply(taxa_sums(moose_bacteria),
tax_table(moose_bacteria)[, "Family"], sum), TRUE)[1:10]
```

#Cut down the physeq.tree data to only include the top 10 Families#

```
top10F = subset_taxa(moose_bacteria, Family %in% names(top10F.names))
```

#Step 8B: Plot top10 families

```
plot_bar(top10F, x="location", fill="Family", facet_grid = ~Family) +
geom_bar(aes(color=Family, fill=Family), stat="identity", position="stack")+
theme(legend.position="bottom")
```



###Step 9: Optional##

#In order to create reproducible data and save all your work, you can export your objects#

#Export taxonomy#

```
tax<-as(tax_table(phyloseq_16S_moose),"matrix")
tax_cols <- colnames(tax)
tax<-as.data.frame(tax)
tax$taxonomy<-do.call(paste, c(tax[tax_cols], sep=";"))
for(co in tax_cols) tax[co]<-NULL
write.table(tax, "tax.txt", quote=FALSE, col.names=FALSE, sep="\t")
```

#export OTU table as biom file#

```
library(biomformat);packageVersion("biomformat")
```

```
## Warning: package 'biomformat' was built under R version 4.0.3
```

```
## [1] '1.18.0'
```

```
otu<-as(otu_table(phyloseq_16S_moose), "matrix")
otu_biom<-make_biom(data=otu)
write_biom(otu_biom, "otu_biom.biom")

#Export metadata file

write.table(sample_data(phyloseq_16S_moose), "sample-metadata.txt", sep="\t",
row.names=FALSE, col.names=TRUE, quote=FALSE)

#export tree file#
tree.seq = phy_tree(phyloseq_16S_moose)

write.tree(tree.seq, "phy_tree.tree")

#Save your entire phyloseq object#

saveRDS(phyloseq_16S_moose, "phyloseq_16S_moose.rds")
```