The Human Microbiome Biodiversity in Health and Disease
Data Analysis Tutorial

Angela Peña-Gonzalez, PhD and Emily Weigel, PhD
School of Biological Sciences, Georgia Institute of Technology

Summary

High-throughput sequencing of PCR-amplified taxonomic markers like the 16S rRNA gene has enabled a new level of analysis of the human body's complex microbial communities, which are collectively known as the human microbiome. Many tools exist to quantify and compare abundance levels of organisms via operational taxonomic units (OTUs) and to visualize the structure and composition of microbial communities in different conditions (e.g. healthy and diseased, or sick, states). In this introductory tutorial, we provide a simple workflow in R to perform basic analysis of gut and vaginal microbiome data in healthy and diseased states. This workflow is based on software packages from the open-source Bioconductor project (https://www.bioconductor.org/) and some CRAN packages.

Introduction

The human body is home to trillions of microorganisms that colonize the body inside and outside. Collectively, they are regarded as **'The Human Microbiome'**. The human microbiome can be determined from the genetic material from all the microbes that live on and inside the human body (bacteria, fungi, protozoa and viruses). In a healthy person, these "bugs" coexist peacefully throughout the body, with the largest numbers found in the small and large intestines (1, 2). The microbiome is even considered by some as a supporting organ because it plays many key roles in promoting the smooth daily operations of the human body, including digestion and immune support.

The human gut microbiome corresponds to the total collection of microbes (and their genomic content) that reside inside the digestive tract of humans. It is an extremely variable, dynamic, and complex habitat, containing up to 10^{12}-10^{15} microbial cells per gram of fecal material (3). The composition of this complex microbial population is highly variable over time and susceptible to both environmental and host-specific modifications (4). Several studies have highlighted the critical role that the gut microorganisms play in human health. Gut microbes are involved in a variety of essential activities for the host including energy harvest and storage, fermentation and metabolism of otherwise indigestible carbohydrates, maturation of host immune-defense, and even cognitive processes (5, 6).

The vaginal tract is also home to more than 250 microbial species that have been associated with the healthy state; they are regarded as **the vaginal microbiome** (VM). It is believed that the VM plays an important role in regulating environmental disturbances and protecting against infections of the urogenital tract, including sexually transmitted diseases such as the human papillomavirus (HPV) (7).

Key distinguishing features of the vaginal microenvironment in healthy women at reproductive age includes low pH (ranging from 3.8 to 4.5), high abundance of *Lactobacillus spp.*, and low microbial diversity compared to other body sites such as the gut (8, 9). Shifts in structure and composition of the vaginal microbiota have been correlated with host-associated and environmental factors, including changes in estrogen levels, menstrual cycle, age, pregnancy, sexual behavior, ancestry, use of personal hygiene products, and even probiotic intake or diet.

The human microbiome in health and disease.

The human microbiome consists of microbes that are both helpful and potentially harmful (opportunistic). Most are symbiotic (where both the human body and microbiota benefit) and some, in smaller numbers, are pathogenic (promoting disease). In a healthy body, pathogenic and symbiotic microbiota coexist without problems. However, if there is a disturbance in that balance, perhaps brought on by infectious illnesses, chronic diseases, certain diets, or the prolonged use of antibiotics, dysbiosis occurs, which stops these normal interactions (10). As a result, the body may become more susceptible to disease.

Next Generation Sequencing (NGS) technologies are powerful tools to study human microbiomes in health and disease. Bacterial communities can now be identified through the use of next generation sequencing applied at several levels. Shotgun sequencing of all bacteria in a sample delivers knowledge of all the microbes and genes present in a given sample.

In this tutorial, we will focus in the detection and quantification of individual taxa (or phylogroups) that were identified through the sequencing of a fingerprint gene called 16SrRNA which is present in all bacteria. We will describe ‘*how many*’ different phylogroups can be detected in microbial ecosystems as the gut and the vaginal cavity and ‘*how differently*’ distributed are they in a healthy *vs*. disease states.

Objectives

1. To perform microbial ecology analysis of healthy and disease individuals based on the sequencing of the 16S rRNA gene.
2. To perform data analysis and visualization using open source, statistical tools in R
3. To test statistical differences in community composition and structure within and between samples using computational tools.

Glossary

Microbiome = the collection of all microorganisms and their genes found in a particular environment (including the body or a part of the body).

16SrRNA = the 16S ribosomal RNA. The component of the 30S small subunit of a prokaryotic ribosome that binds to the Shine-Dalgarno sequence. It is widely used for phylogenetic reconstructions due to the slow rates of evolution in certain regions.

OTUs = Operational Taxonomic Unit. It is an operational definition used to classify groups of closely related objects based on genetic material. In microbial ecology, OTUs are pragmatic proxies for microbial "species" at different taxonomic levels.

Richness = A metric describing the number of different species (or OTUs) in a community.

Evenness = A metric describing how even in numbers each species (or OTU) in a community are.

Diversity = A metric describing how variable the specie (or OTUs) in a community are.

Dysbiosis = A term describing an “imbalance” in a microbial community that is associated with disease.

Data Analysis Workflow

Part A: Taxonomic profiling of microbial communities in gut and vaginal cavity

Part B: Alpha-diversity analysis (within sample analysis)

Part C: Beta-diversity analysis (between samples and/or groups)

Part A: Taxonomic profiling of microbial communities in gut and vaginal body sites

We will initiate our microbiome data analysis by graphing taxa distributions at two different taxonomic levels: Phylum (L2) and Family (L5) using the library `enveomics.R`. The enveomics collection is a toolbox for specialized analysis of microbial genomes and metagenomes. The enveomics collection is freely available at *https://github.com/lmrodriguezr/enveomics*

To install the latest version of `enveomics.R` uploaded to CRAN, execute in R:

```
> install.packages ('enveomics.R')
```

To load the library `enveomics.R` into the working space, simply execute:

```
> library ('enveomics.R')
```

We will initiate our microbiome analysis by studying the microbial community composition in the human gut in two groups of individuals: healthy and disease (individuals with acute diarrhea).

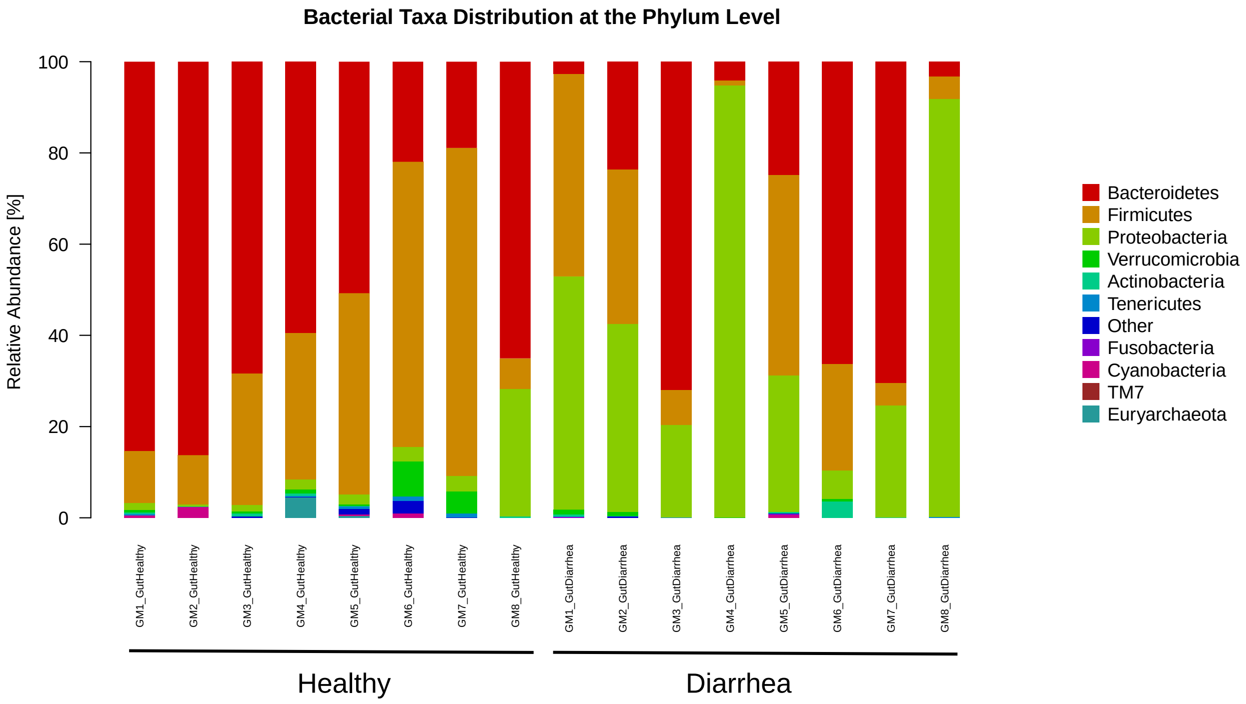
1. Bacterial taxa distribution in the gut microbiome at the phylum level (L2)

To load the initial dataset for the gut microbiome (Phylum level-L2), execute:

```
> Gut_L2 <-read.table('Gut_L2.txt', header=TRUE, sep ='\t', row=1)
```

The function ‘`enve.barplot`’ produces nice barplots from tab-delimitated tables. To graph barplots showing the estimated relative abundance percent of bacterial phylogroups at the phylum level, execute:

```
> pdf(file="Gut_L2_Taxonomy.pdf", width=12, height=6)

> enve.barplot(Gut_L2, bars.width = 3, add.trend = TRUE, organic.trend =
TRUE, cex.names=0.6, las=2, main="Bacterial Taxa Distribution at the Phylum
Level",ylab="Relative Abundance [%]")


> dev.off()
```



Now, let’s estimate the average relative abundance percent of the three most abundant bacterial phylogroups in the healthy set and compare them with the diarrhea set of samples.

To estimate the average relative abundance in the set of healthy (C1:C8) and diarrhea (C9:C16) individuals, execute this:

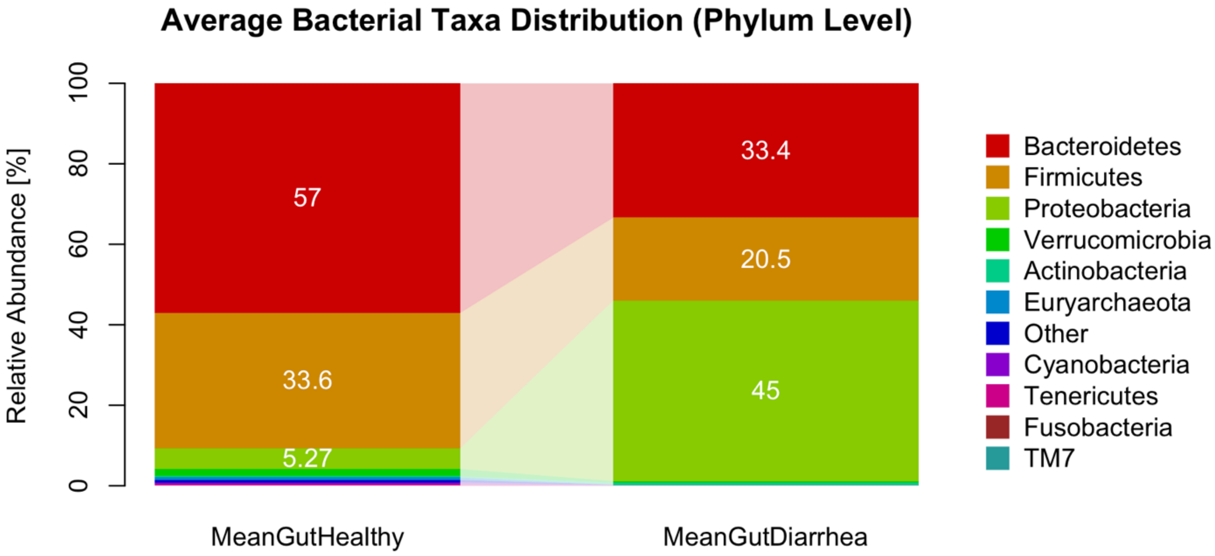
```
> Merged_dataframe <-
cbind(data.frame(MeanGutHealthy=rowMeans(Gut_L2[,1:8])),
data.frame(MeanGutDiarrhea=rowMeans(Gut_L2[,9:16])))
```

Visualize the new data frame that was created:

```
> Merged_dataframe
```

Graph the new data frame containing the estimated average relative abundance percent for healthy and sick individuals with the function ‘`enve.barplot`’:

```
> enve.barplot(Merged_dataframe, bars.width = 3, add.trend = TRUE,
min.report = 5, main="Average Bacterial Taxa Distribution (Phylum
Level)",ylab="Relative Abundance [%]")
```



Analyze the two graphs that you just produced and answer the following questions:

1. How many individuals compose the healthy and sick groups?

   The healthy gut microbiome group is composed by eight (8) individuals; the diarrhea gut microbiome group is also composed by eight (8) individuals.

2. How many phyla were identified in total with the 16S rRNA sequencing in the gut microbiome?

   In total, 10 bacterial phyla were identified in the gut microbiome of all individuals.

3. Which are the three most abundant phylogroups observed in all individuals?

The three most abundant phylogroups detected in all individuals (both sick and healthy) are Bacteroidetes, Firmicutes and Proteobacteria.

4. Which is the first most abundant (dominant) phylum observed in the set of healthy individuals?

   The most dominant phylum observed in the group of healthy individuals is Bacteroidetes.

5. Which is the first most abundant (dominant) phylum identified in the set of sick (diarrhea) individuals?

   The most dominant phylum observed in the group of healthy individuals is Proteobacteria.

6. How the average taxonomic composition of *Bacteroidetes*, *Firmicutes* and *Proteobacteria* varies between healthy and sick groups?

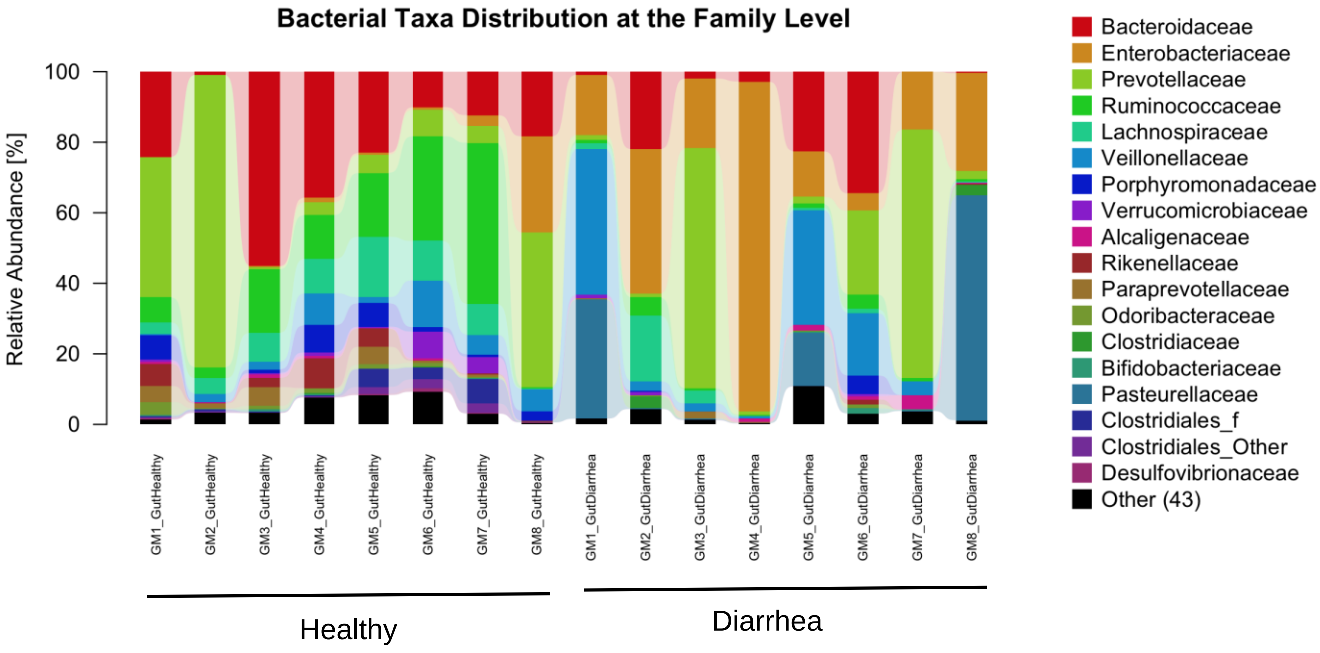
   The average relative abundance of the phyla Bacteroidetes (57%) and Firmicutes (33.6%) was significantly reduced in the group of sick individuals relative to healthy individuals, where they represented 33.4% and 20.5%, respectively. In the other hand, the phylum Proteobacteria, which is observed in a relatively low abundance in the set of healthy individuals (5.27%) was increased to an average relative abundance of 45% when individuals were experiencing episodes of acute diarrhea. The majority of samples in the diarrhea group exhibited a large proportion of sequences belonging to Proteobacteria, a bacterial group comprising several enteric pathogens, which indicates a possible underlying gastrointestinal infection.

7. Do you observe ‘high’, ‘medium’ or ‘low’ variation in the taxonomic composition and abundance within healthy and sick groups?

   Although overall, all individuals in both healthy and sick groups are dominated by three phylogroups (Bacteroidetes, Firmicutes and Proteobacteria) some individuals within each groups also display signatures of sample-specific phylogroups. Therefore, I observe a ‘medium’ variation in the taxonomic composition within heathy and sick groups.

Independent work:
Repeat the same community composition analysis for the dataset Gut_L5.txt (family level)

```
> Gut_L5 <-read.table('Gut_L5.txt', header=TRUE, sep ='\t', row=1)

> enve.barplot(Gut_L5, bars.width = 3, add.trend = TRUE, , top=18,
organic.trend = TRUE, cex.names=0.6, las=2, main="Bacterial Taxa
Distribution at the Family Level",ylab="Relative Abundance [%]")
```



2. Bacterial taxa distribution in the vaginal cavity at the family level (L5)

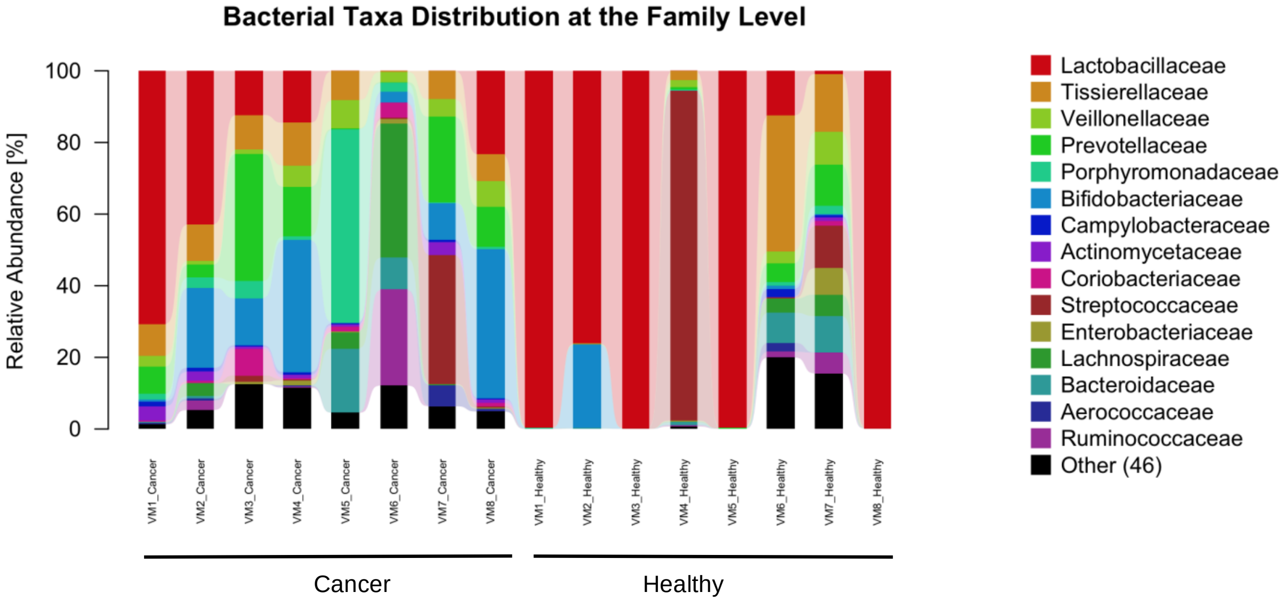
Now, let’s visualize the vaginal microbiome taxonomic profile at the family level (L5) in healthy and sick women with cervical cancer

To load the initial dataset for the Vaginal Microbiome (Family level-L5), execute:

```
> Vaginal_L5 <-read.table('Vaginal_L5.txt', header=TRUE, sep ='\t', row=1)
```

To graph barplots showing the estimated relative abundance percent of bacterial phylogroups at the family level, execute:

```
> pdf(file="Vaginal_L5_Taxonomy.pdf", paper="a4")

> enve.barplot(Vaginal_L5 , bars.width = 2, top=15, add.trend = TRUE,
organic.trend = TRUE, cex.names=0.5, las=2, main="Bacterial Taxa
Distribution at the Family Level",ylab="Relative Abundance [%]")

> dev.off()
```



Analyze the graph that you just produced and answer the following questions:

1. What is the most dominant OTU (phylogroup) detected in the healthy vaginal microbiome group?

The most dominant bacterial phylogroup detected in the vaginal microbiome of healthy women is the family Lactobacillaceae. Some individuals (e.g. VM1, VM3, VM5 and VM8) in the healthy group are completely dominated by bacteria belonging to this family group.

2. Do you observe more phylogroups detected in the healthy or disease group?

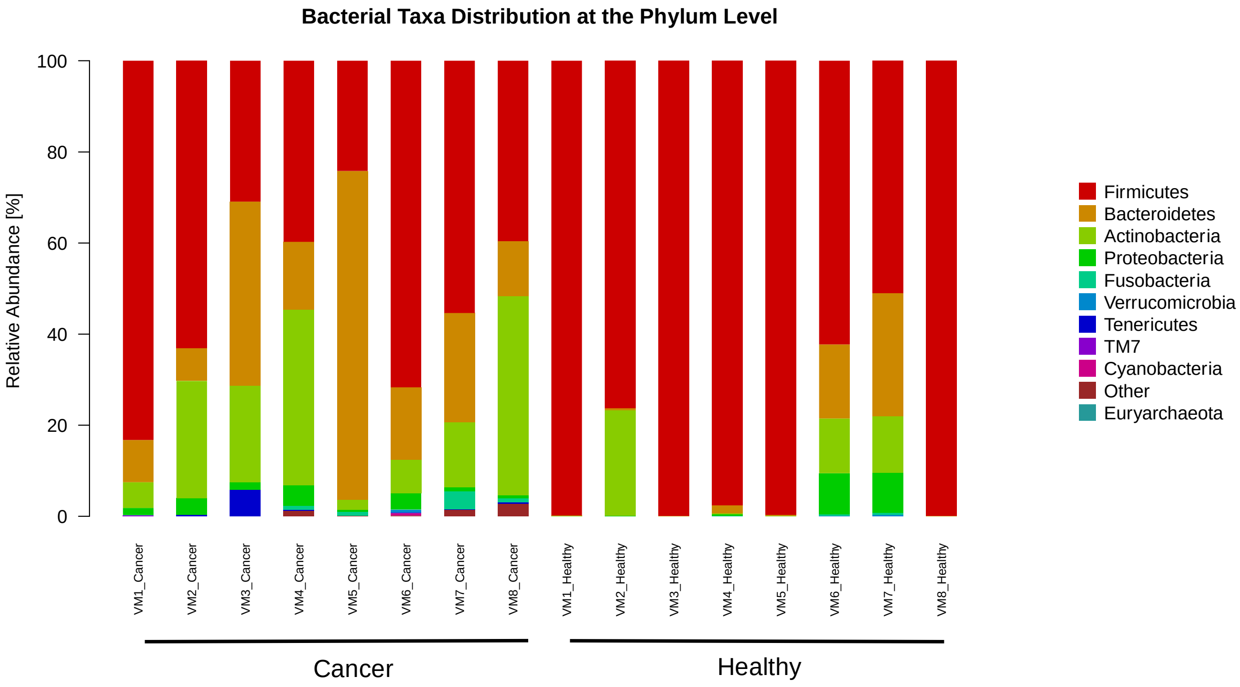
I can observe more OTUs detected in the cancer group. It seems to be more biodiversity in composition compared with the healthy group.

3. Which individuals showed a relative high abundance of the family *Bifidobacteriaceae*? are they mostly observed in healthy or disease groups?

Individuals VM2, VM3, VM4, VM6, VM7 and VM8 in the cancer group show a relatively large abundance of the family Bifidobacteriaceae. Individual VM2 in the healthy group also showed a large signature of the same bacterial group. However, the Bifidobacteriaceae family seems to be more abundant in the sick group.

Independent work:
Repeat the same community composition analysis for the dataset Vaginal_L2.txt (phylum level)

```
> Vaginal_L2 <-read.table('Vaginal_L2.txt', header=TRUE, sep ='\t', row=1)
> pdf(file="Vaginal_L2_Taxonomy.pdf", width=12, height=6)
> enve.barplot(Vaginal_L2 , bars.width = 3, add.trend = TRUE, organic.trend
= TRUE, cex.names=0.7, las=2, main="Bacterial Taxa Distribution at the
Phylum Level",ylab="Relative Abundance [%]")
> dev.off()
```



Part B: Alpha-diversity Analysis (within sample analysis)

Species diversity is a valuable tool for describing the ecological complexity within a sample (alpha diversity) or between samples (beta diversity). However, diversity is not a physical quantity that can be measured directly, and many different metrics have been proposed to quantify diversity. Species diversity consists of three components: species richness, species evenness. and taxonomic (or phylogenetic) diversity.

We will continue our microbiome data analysis by comparing alpha diversity metrics (richness, evenness and Shannon diversity) in microbiome data obtained from gut and vaginal body sites in healthy and disease individuals using **boxplots**.

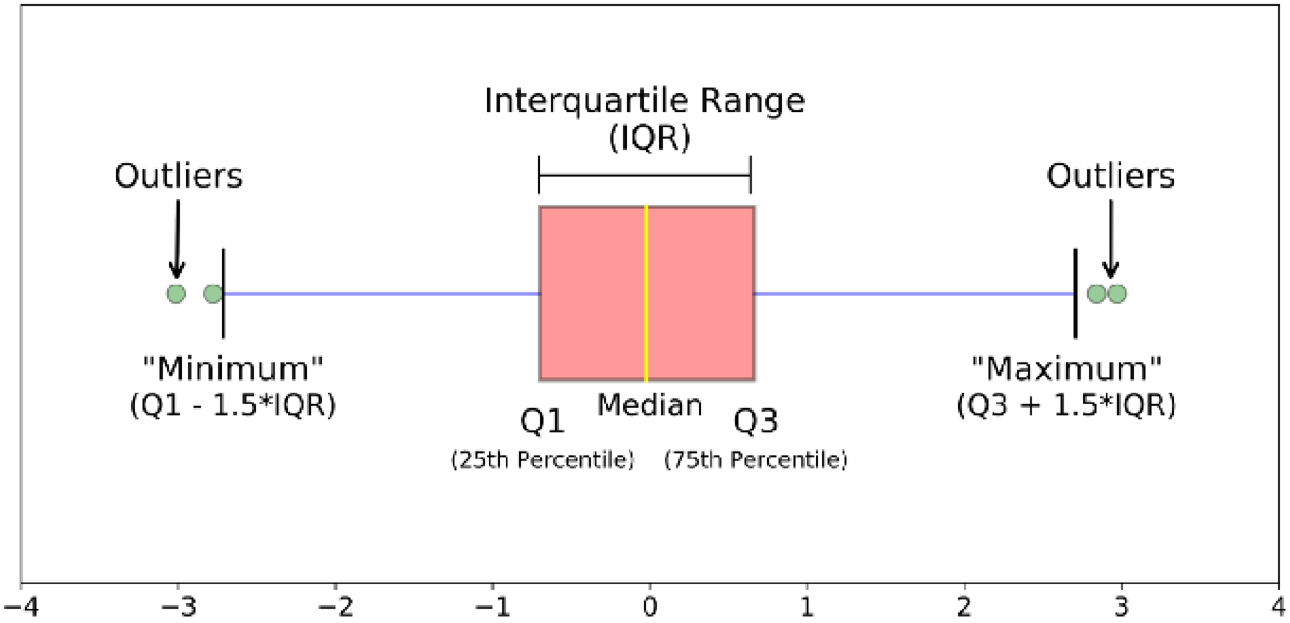


Figure taken from: *https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51*

A boxplot (figure above) is a standardized way of displaying the distribution of a dataset based on five numbers summary:

- Median (Q2/50th percentile) = The middle value of the dataset
- First quartile (Q1/25th Percentile): the middle number between the smallest number (not the “minimum”) and the median of the dataset.
- Third quartile (Q3/75th Percentile): the middle value between the median and the highest value (not the “maximum”) of the dataset.
- Interquartile range (IQR): 25th to the 75th percentile.
- Whiskers (shown in blue)
- Outliers (shown as green circles)
- Maximum: Q3 + 1.5*IQR
- Minimum: Q1 -1.5*IQR

In each boxplot, the horizontal thick black band represents the median value, and the boxplot margins indicate first and third quartiles.

We will use the package `'ggplot2'` to display the distribution of the different diversity metrics with boxplots

To install the package `'ggplot2',` execute in R:

```
> install.packages ('ggplot2')
```

Next, load the library into your working space

```
> library(ggplot2)
```

Next, load the initial alpha diversity metric dataset (tab-delimitated table in txt format) for gut and vaginal microbiomes. Execute in the console:

```
> alpha<-read.table('alpha_diversity_metrics.txt', check.names=F,
header=TRUE, sep = '\t', row=1)

> attach(alpha)
```

The `‘attach’` function attach a set of objects to a search path. This means that the data frame is searched by R when evaluating a variable, so objects in the data frame can be accessed by simply giving their column names.

1. Richness

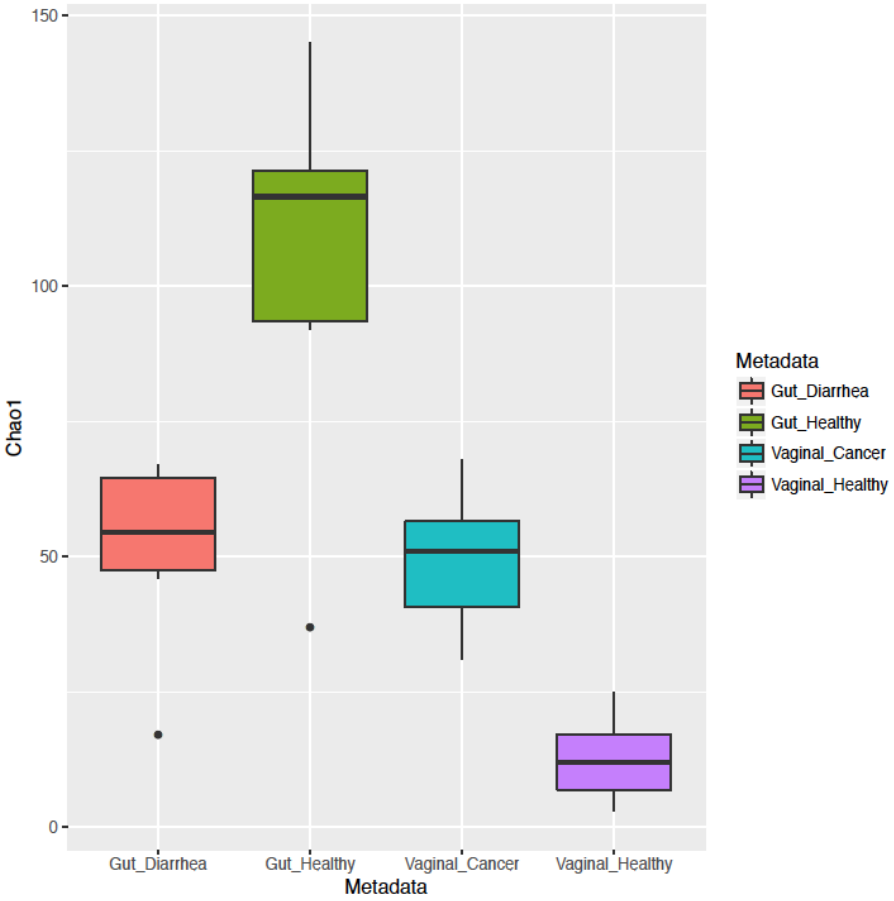
Richness, as measured by Chao1 (Chao 1984) is a measure of abundance of species. In short, it uses the number of rare species found just once, and weights those against more common species, to measure how diverse a community is. The higher the Chao1 (richness), the more diverse a community.

Now let’s create a PDF file with boxplots containing the richness values comparing the categories described in the column named ‘Metadata’. For this purpose, we will use the package `‘ggplot2’`.

Run the following lines in the R console:

```
> pdf(file="Richness_Chao1.pdf", paper="a4")

> ggplot(data=alpha,
aes(x=Metadata,y=Chao1))+geom_boxplot(aes(fill=Metadata))

> dev.off()
```

RICHNESS (Chao1)



2. Evenness

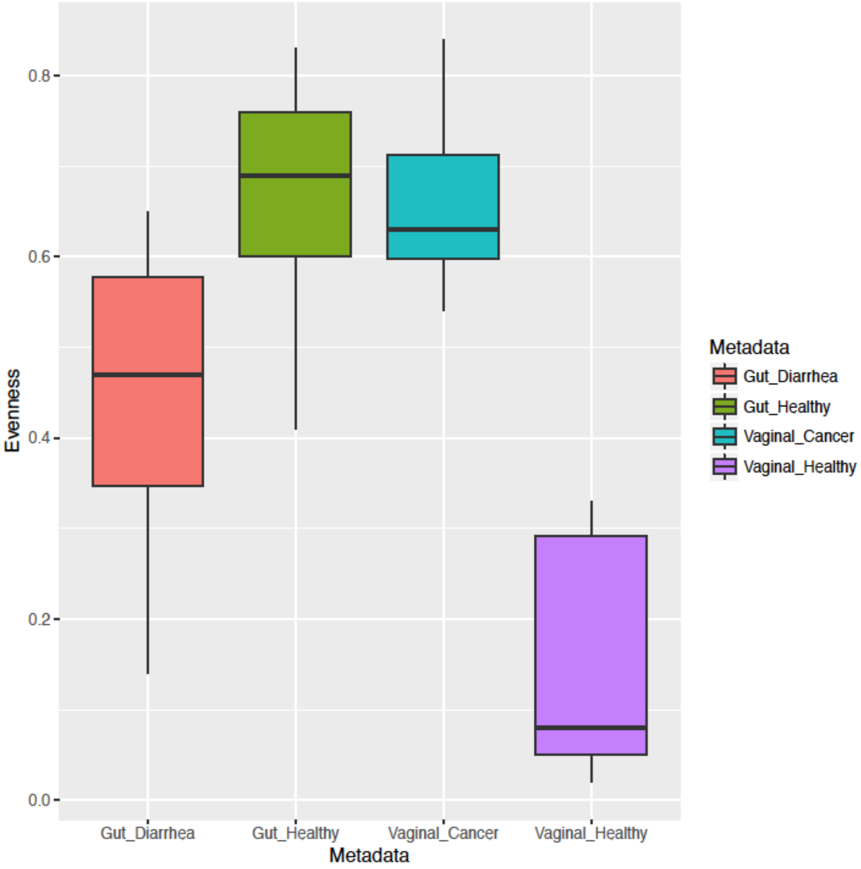
Evenness describes how abundant species are in a community relative to one another. The closer the evenness to 1, the more equally the species are represented among all individuals in a community (or, said another way, the more equal the probability of pulling an individual of any species in that community at random).

Now let’s create a PDF file with boxplots containing the values of evenness comparing the categories described in the column named ‘Metadata’.

Run the following lines in the R console:

```
> pdf(file="Evenness.pdf", paper="a4")
> ggplot(data=alpha,
aes(x=Metadata,y=Evenness))+geom_boxplot(aes(fill=Metadata))
> dev.off()
```

EVENNESS



3. Diversity (Shannon)

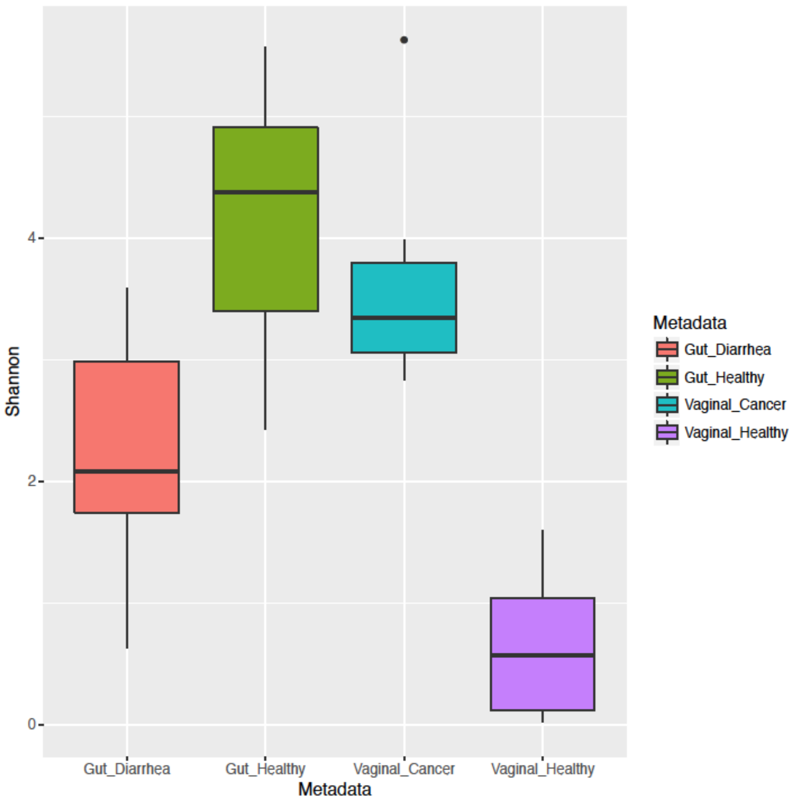
Lastly, we'll take a look at diversity, as measured by the Shannon Diversity Index (H). This measurement accounts for both species abundance and evenness. The higher the H, the higher the diversity.

Now let's create a PDF file with boxplots containing the values of diversity comparing the categories described in the column named 'Metadata'.

Run the following lines in the R console:

```
> pdf(file="Shannon.pdf", paper="a4")
> ggplot(data=alpha,
aes(x=Metadata,y=Shannon))+geom_boxplot(aes(fill=Metadata))
> dev.off()
```

DIVERSITY (SHANNON)



Statistics

Now let's test whether the observed differences in richness within and between groups are significant. For this purpose, we will use the function '`subset()`' to cut our current data frame into subgroups.

Execute the following lines in the R console:

```
> Gut_Healthy <- subset(alpha, Metadata == 'Gut_Healthy')
> Gut_Diarrhea <- subset(alpha, Metadata == 'Gut_Diarrhea')
```

To determine whether the observed differences in alpha diversity metrics between groups are significant, we will use a non-parametric test. A statistical method is called **non-parametric** if it makes no assumption on the population distribution or sample size. R provides functions for carrying out non-parametric tests, such as the Wilcoxon Rank Sum Test (WRST), which we will use here. Because it does not rely on having a normal distribution, you can think about the WRST as a more general version of a parametric t-test.

To perform statistical analysis within each body site, execute:

```
> wilcox.test(Gut_Healthy$Chao1, Gut_Diarrhea$Chao1, paired=FALSE)

Wilcoxon rank sum test
data:  Gut_Healthy$Chao1 and Gut_Diarrhea$Chao1
W = 57, p-value = 0.006993
alternative hypothesis: true location shift is not equal to 0
```

Repeat the same analysis for the remaining alpha diversity metrics:

```
> wilcox.test(Gut_Healthy$Evenness, Gut_Diarrhea$Evenness, paired=FALSE)

Wilcoxon rank sum test with continuity correction
data:  Gut_Healthy$Evenness and Gut_Diarrhea$Evenness
W = 54.5, p-value = 0.02077
alternative hypothesis: true location shift is not equal to 0
```

```
> wilcox.test(Gut_Healthy$Shannon, Gut_Diarrhea$Shannon, paired=FALSE)

Wilcoxon rank sum test
data:  Gut_Healthy$Shannon and Gut_Diarrhea$Shannon
W = 58, p-value = 0.004662
alternative hypothesis: true location shift is not equal to 0
```

Analyze the different alpha diversity plots that you just produced with ALL statistical tests performed for gut microbiomes and answer the following questions:

1. Compare the **richness** (chao1) boxplots for **gut microbiome** in healthy and disease groups. Which of these communities has the highest number of OTUs? Are the differences statistically significant?

   According with the boxplots generated, the largest estimated species richness is observed in the group of healthy gut microbiomes. The richness differences observed between healthy and diarrhea groups are significantly different (W = 57, p-value = 0.006993).

2. Compare the **evenness** boxplots for **gut microbiome** in healthy and disease groups. Which of these communities has the highest equitability of OTUs? Are the differences statistically significant?

Bacterial phylogroups detected in the gut microbiome are more evenly represented in the group of healthy individuals compared to disease individuals. The observed differences are statistically significant (W=54, p-value=0.02077).

3. Compare the **diversity** boxplots for **gut microbiome** in healthy and disease groups. Which of these communities has the highest diversity? Are the observed differences statistically significant?

   Bacterial community detected in the gut micro of healthy individuals is more diverse compared to disease individuals. The observed differences are statistically significant (W=58, p-value=0.004662).

Now let's perform the same non-parametric statistical analysis for the vaginal microbiome dataset:

First, subset the original data frame executing:

```
> Vaginal_Healthy <-subset(alpha, Metadata == 'Vaginal_Healthy')
> Vaginal_Cancer <-subset(alpha, Metadata == 'Vaginal_Cancer')
```

And perform the different non-parametric statistical analyses:

```
> wilcox.test(Vaginal_Healthy$Chao1, Vaginal_Cancer$Chao1, paired=FALSE)

Wilcoxon rank sum test
data:  Vaginal_Healthy$Chao1 and Vaginal_Cancer$Chao1
W = 0, p-value = 0.0001554
alternative hypothesis: true location shift is not equal to 0


>wilcox.test(Vaginal_Healthy$Evenness, Vaginal_Cancer$Evenness,
paired=FALSE)

Wilcoxon rank sum test with continuity correction
data:  Vaginal_Healthy$Evenness and Vaginal_Cancer$Evenness
W = 0, p-value = 0.0009148
alternative hypothesis: true location shift is not equal to 0


>wilcox.test(Vaginal_Healthy$Shannon, Vaginal_Cancer$Shannon, paired=FALSE)

Wilcoxon rank sum test
data:  Vaginal_Healthy$Shannon and Vaginal_Cancer$Shannon
W = 0, p-value = 0.0001554
alternative hypothesis: true location shift is not equal to 0
```

Analyze the different alpha diversity plots that you just produced with ALL statistical tests performed for vaginal microbiomes and answer the following questions:

4. Compare the **richness** (chao1) boxplots for **vaginal microbiome** in healthy and disease groups. Which of these communities has the highest number of OTUs? Are the differences statistically significant?

   The estimated species richness is higher in the vaginal microbiome of women with cervical cancer compared to healthy group. The median estimated species richness in the healthy group is lower than 25 OTUs, which means that very few species occur naturally in the vaginal cavity of healthy women. The observed differences are statistically significant (W=0, p-value=0.0001).

5. Compare the **evenness** boxplots for **vaginal microbiome** in healthy and disease groups. Which of these communities has the highest equitability of OTUs? Are the differences statistically significant?

   Bacterial phylogroups are less evenly represented in the vaginal microbiome of healthy women compared with disease women. This means that the healthy vaginal microbiome is dominated by very few species and this pattern is altered when there is chronic disease as cervical cancer. The observed differences are statistically significant (W=0, p-value=0.0009).

6. Compare the **diversity** boxplots for **vaginal microbiome** in healthy and disease groups. Which of these communities has the highest diversity? Are the observed differences statistically significant?

   Bacterial community detected in the vaginal microbiome of healthy women is significantly less diverse compared to disease women. The observed differences are statistically significant (W=0, p-value=0.0001).

Critical thinking:

7. How does the healthy vaginal microbiome compare to the healthy gut microbiome? Explain

   When comparing healthy gut microbiome to healthy vaginal microbiome, the healthy gut microbiome is characterized by having a higher species richness, evenness, and diversity than the healthy vaginal microbiome.

8. How does the disease vaginal microbiome compare to the disease gut microbiome? Do trend follow the same pattern as above? Explain.

When comparing the structure and diversity of the disease gut microbiome to disease vaginal microbiome, diseased guts have higher richness, lower evenness, and lower Shannon diversity. Generally, guts have higher species richness, but a healthy microbiome in vaginas means low evenness and diversity, but the opposite is true for guts.

Part C: Beta-diversity Analysis (between samples and/or groups)

1. Dissimilarity analysis based on Bray-Curtis distances (multidimensional analysis with NMDS plots)
2. Hierarchical clustering

1. Dissimilarity analysis ~~based on Jaccard distances~~

Now, we will finalize our microbiome data analysis by comparing *all versus all* individuals using multidimensional scaling plots (NMDS). A multidimensional scaling analysis is a **visual representation of distances or dissimilarities between sets of objects** (11). “Objects” can be colors, faces, map coordinates, etc. In this case, ‘objects’ represent a particular taxonomic configuration (statistical arrangement) of the microbial community per individual. Microbiomes that are more similar are closer together (or separated by shorter distances) on the graph than less similar communities.

To install the package `'vegan'` and `'ggplot2',` execute in R:

```
> install.packages ('vegan')
> install.packages ('ggplot2')
```

Next, load the library into your working space

```
> library(vegan)
> library(ggplot2)
```

Next, load both the Abundance-weighted Jaccard distance matrix and metadata files

```
>   distances<-read.table('abund_jaccard_dm.txt',   head=TRUE,   row.name=1,
sep="\t")
> metadata <-read.table('mappingFile2.txt', head=TRUE, sep="\t", check.name
= F, row.names=1)
```

Now we are ready to perform an NMDS plot with our microbiome data. The following lines will calculate an NMDS and use the results to compose a data frame with NMDS coordinates and associated metadata for each sample

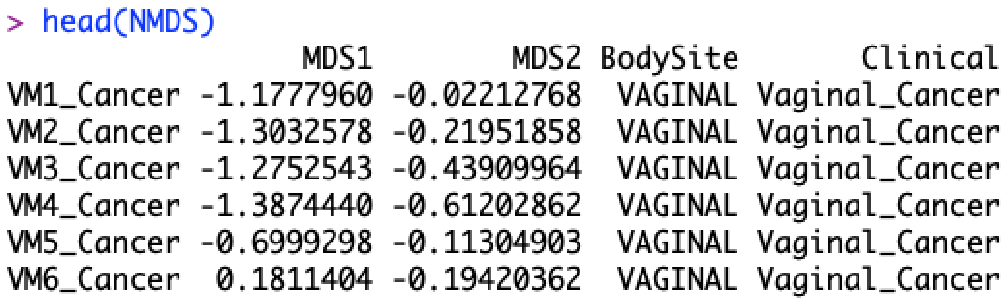
```
> dist_mat<- dist(distances, method = 'euclidean')
> NMDS <- metaMDS(dist_mat, distance = "bray")
> MDS1 = NMDS$points[,1]
> MDS2 = NMDS$points[,2]
> NMDS = data.frame(MDS1 = MDS1, MDS2 = MDS2, BodySite = metadata$BodySite,
Clinical = metadata$Clinical)
```

We always want to check if the figure we made shows 'good' data. We do this by looking at the ‘stress’, or a measure of the degree to which distances between samples in our reduced dimensional space capture actual differences with more dimensions. Lower stress values mean we’ve captured this information well, whereas higher values indicate that the arrangement of the data points doesn’t truly reflect their similarities very well.

By a general rule, stress < 0.05 provides an excellent representation in reduced dimensions, < 0.1 is great, <0.2 is good/ok, and stress < 0.3 provides a poor representation. *Look at the stress values generated when creating the NDMS figure above. What are they generally?*

Here is a peak at what our data frame should look like:

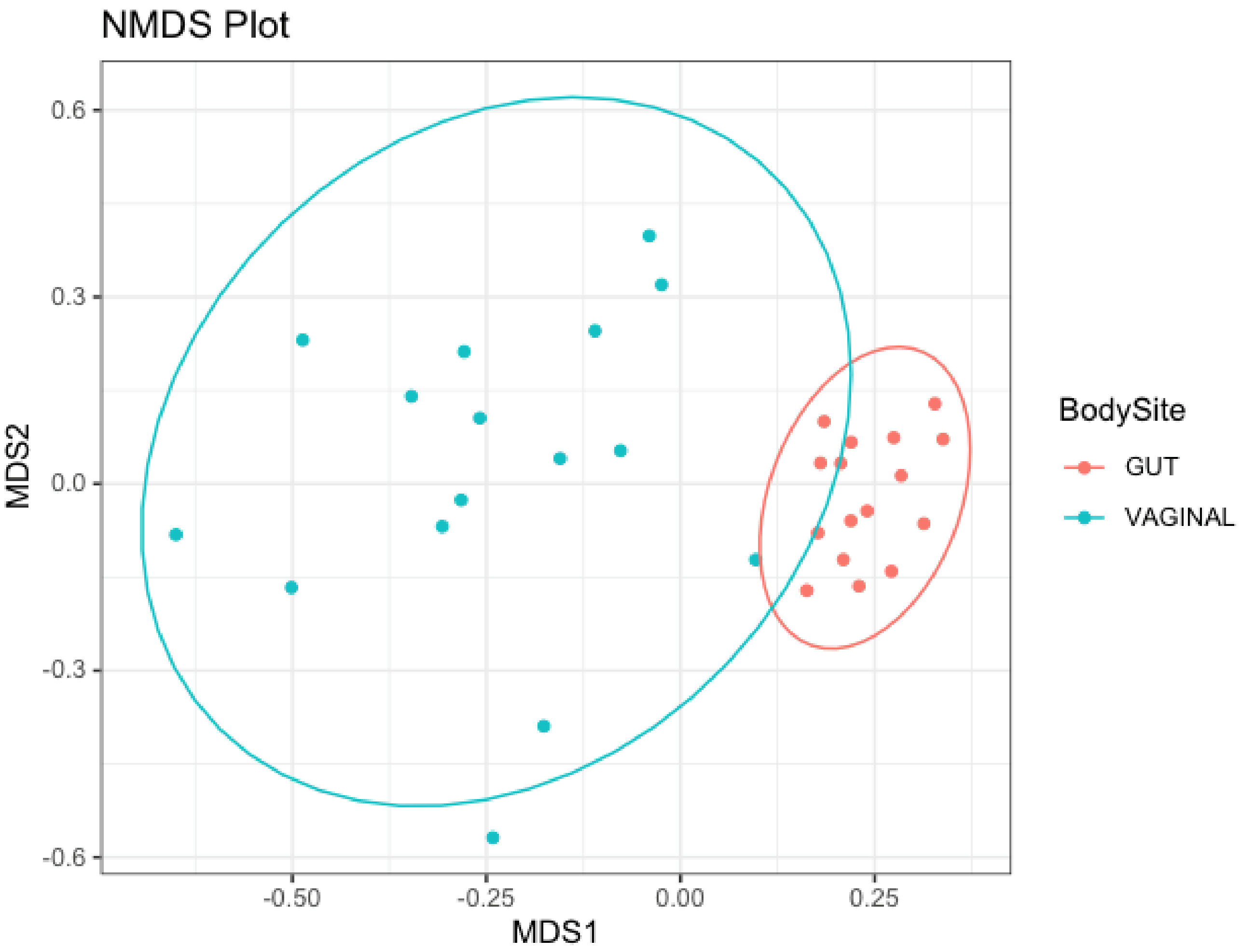
```
> head(NMDS)
```



```
> head(NMDS)
                  MDS1        MDS2 BodySite       Clinical
VM1_Cancer -1.1777960 -0.02212768  VAGINAL Vaginal_Cancer
VM2_Cancer -1.3032578 -0.21951858  VAGINAL Vaginal_Cancer
VM3_Cancer -1.2752543 -0.43909964  VAGINAL Vaginal_Cancer
VM4_Cancer -1.3874440 -0.61202862  VAGINAL Vaginal_Cancer
VM5_Cancer -0.6999298 -0.11304903  VAGINAL Vaginal_Cancer
VM6_Cancer  0.1811404 -0.19420362  VAGINAL Vaginal_Cancer
```

Now, let’s plot the NMDS and take a look at where our samples fall in the ‘ordination space’ based on the ‘body site’. We will use the package ‘`ggplot2`’ for this purpose.

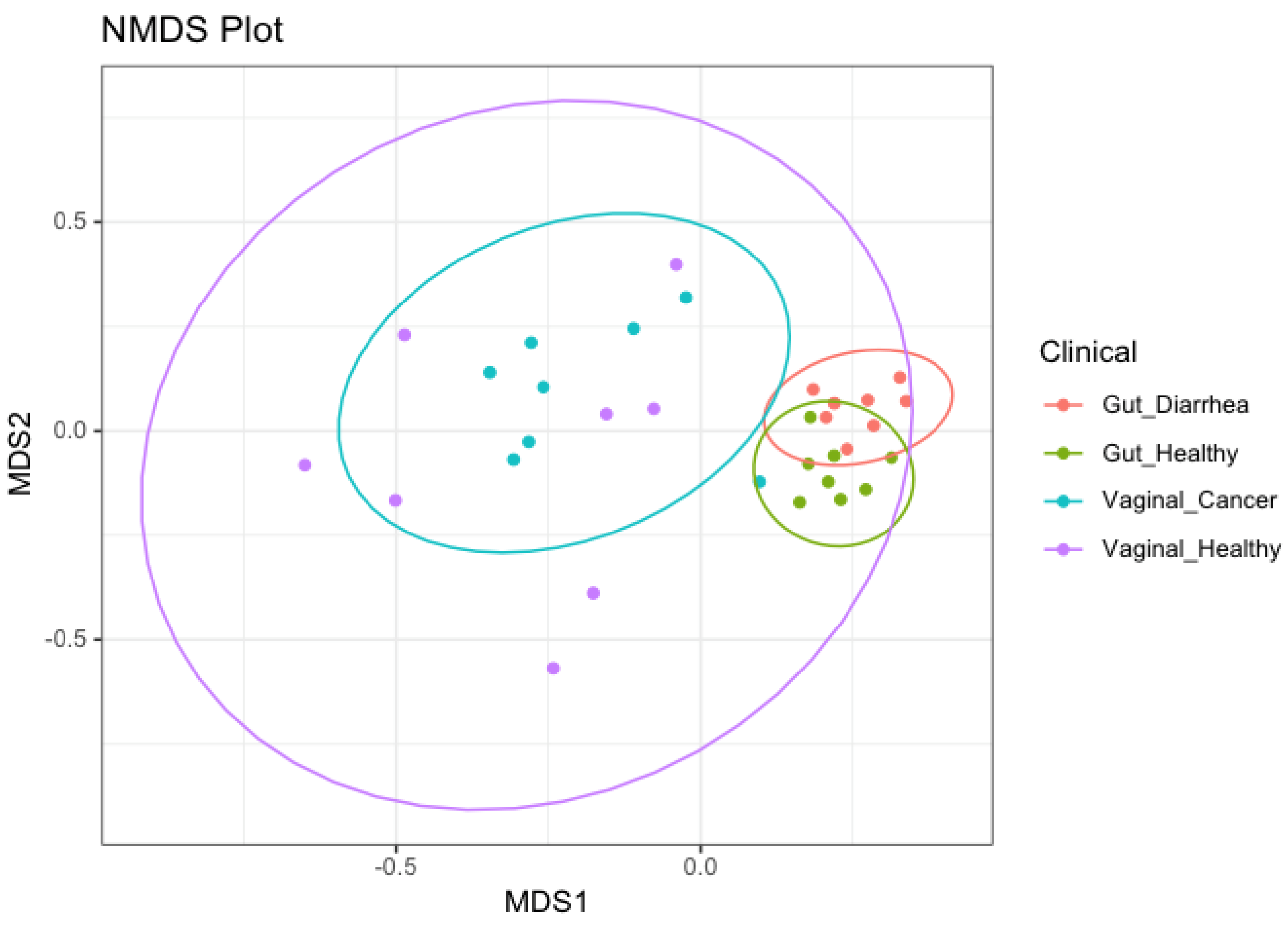
```
> ggplot(NMDS, aes(x=MDS1, y=MDS2, col=metadata$BodySite)) + geom_point() +
stat_ellipse() + theme_bw() + labs(title = "NMDS Plot")
```



Note: Ellipses represent 95% CI around the centroid. Overall, we can observe that there is very little overlap between gut and vaginal microbial communities. This means the particular species present in each community are not that similar between communities.

Now, let’s repeat the analysis based on ‘clinical status’.

```
> ggplot(NMDS, aes(x=MDS1, y=MDS2, col=metadata$Clinical)) + geom_point() +
stat_ellipse() + theme_bw() + labs(title = "NMDS Plot")
```



Note: Ellipses represent 95% CI around the centroid.

2. Hierarchical Clustering Analysis

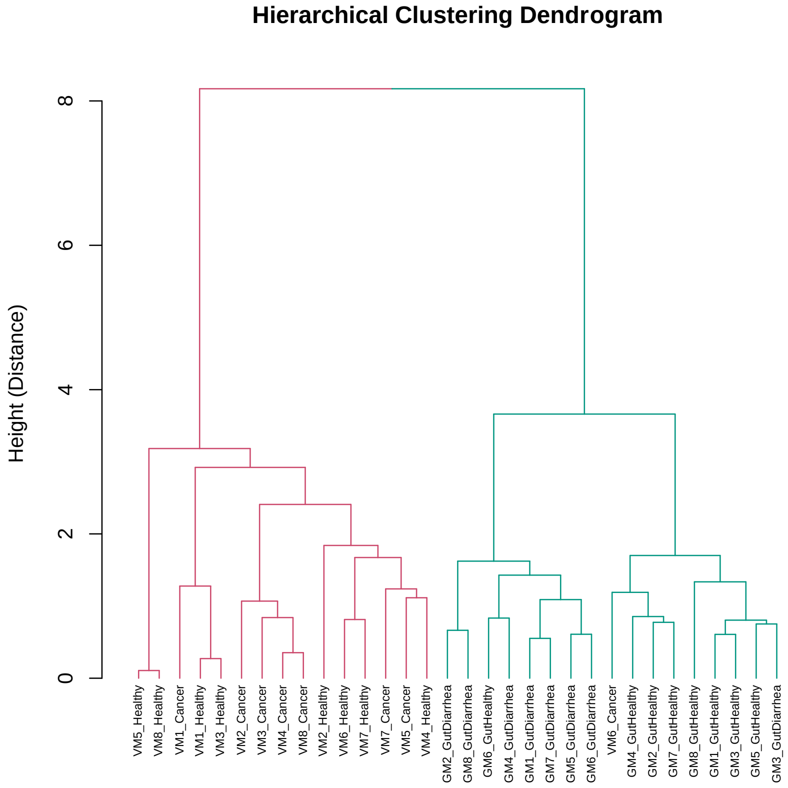
Hierarchical clustering is an analysis in data science that groups similar objects into groups called '*clusters*'. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly more similar to each other than with other clusters (12).

Hierarchical clustering can be performed with either a *distance matrix* or on *raw data.* When raw data are provided, the software will automatically compute a distance matrix in the background. Clustering analysis can be visualized through dendrograms. A *dendrogram* is a diagram that shows the hierarchical relationship between objects. It is most commonly created as an output from *hierarchical clustering*. The main use of a dendrogram is to work out the best way to allocate objects to clusters.

Now let's perform a hierarchical clustering analysis based on microbiome distance matrices

```
> distances<-read.table('abund_jaccard_dm.txt', head=TRUE, row.name=1, sep="\t")
```

```
> dist_mat<- dist(distances, method = 'eucledian')
> hclust_avg <- hclust(dist_mat, method = "ward.D2")

> library(dendextend)

> avg_dend_obj <- as.dendrogram(hclust_avg)

> avg_col_dend <- color_branches(avg_dend_obj, k = 2)%>% set("labels_cex", 0.6)

> pdf(file="Clustering.pdf", paper="a4")

> plot(avg_col_dend, ylab = "Height (Distance)", main = "Hierarchical Clustering
Dendrogram")

> dev.off()
```



Finally, we will run a statistical test to determine whether two or more groups of samples are significantly different, while taking into account how they are grouped. For this purpose, we will run an ADONIS model.

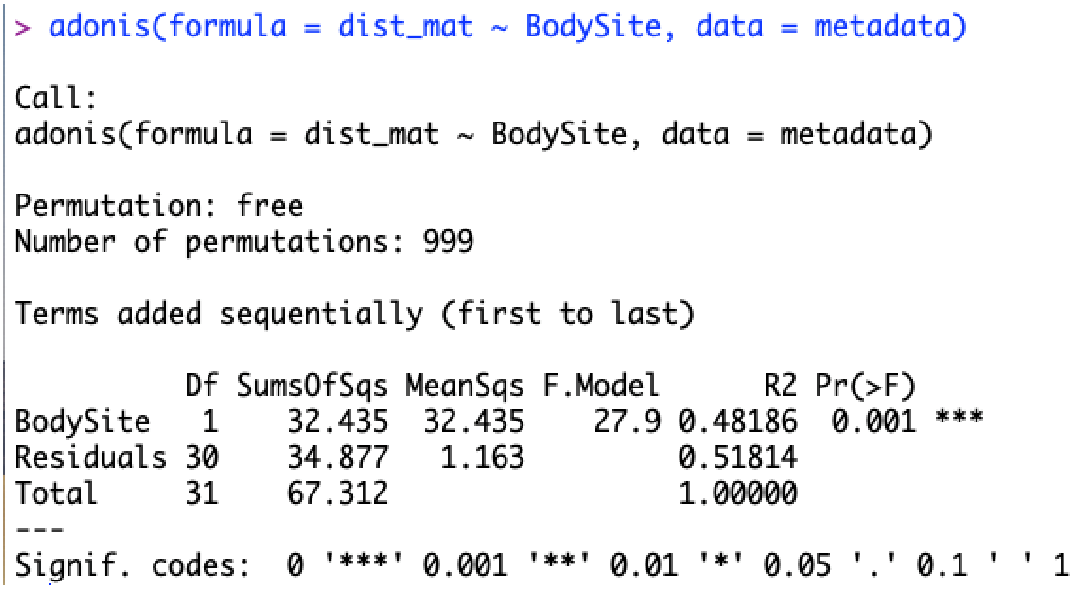
Adonis is a nonparametric statistical test that takes a distance matrix and a category to determine sample grouping. It computes an R^2 value denominated ‘effect size’ which shows the percent of variation explained by the supplied category, as well as a *p-value* to determine the statistical significance.

Adonis creates a set by first identifying the relevant centroids of the data and then calculating the square deviations from these points. After that, significant tests are performed using F-tests based on sequential sums of squares from permutations of the raw data. In short, you can think of this as comparing clusters to give F values you interpret like an ANOVA.

To run an ADONIS model with the category ‘BodySite’, execute in the terminal

```
> adonis(formula = dist_mat ~ BodySite, data = metadata)
```

You will get these results. The percent of microbiome variation explained by the body site is 48.18% and it is significantly different based on the site (*p-val = 0.001*).

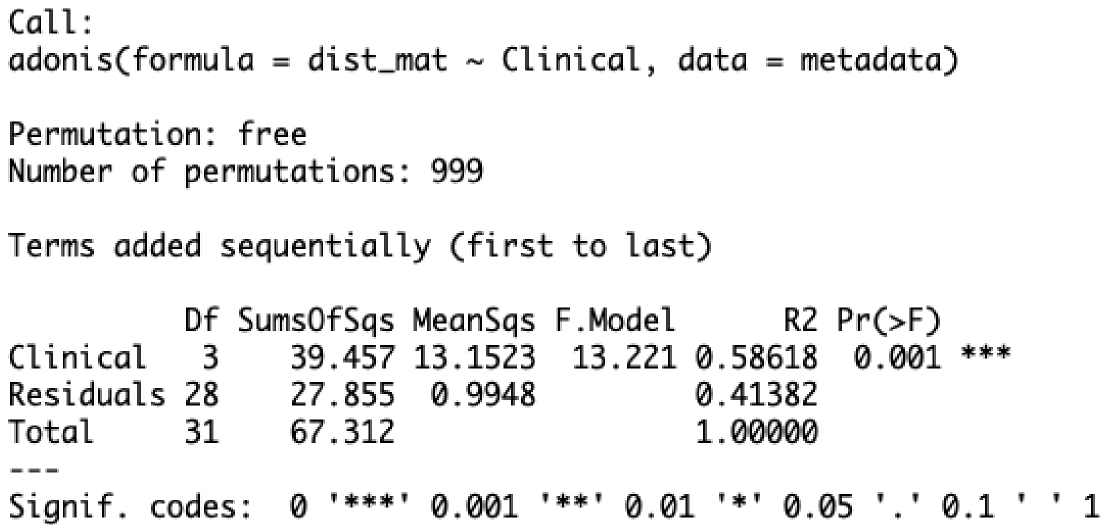


```
> adonis(formula = dist_mat ~ BodySite, data = metadata)

Call:
adonis(formula = dist_mat ~ BodySite, data = metadata)

Permutation: free
Number of permutations: 999

Terms added sequentially (first to last)

          Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)    
BodySite   1    32.435  32.435    27.9 0.48186  0.001 ***
Residuals 30    34.877   1.163         0.51814           
Total     31    67.312                 1.00000           
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Almost half of the microbial community variation in these samples is explained by the body site where they were obtained (gut or vaginal cavity).

Independent work:
Repeat the same analysis to run an ADONIS model using this time the category ‘Clinical’

```
> adonis(formula = dist_mat ~ Clinical, data = metadata)
```



```
Call:
adonis(formula = dist_mat ~ Clinical, data = metadata)

Permutation: free
Number of permutations: 999

Terms added sequentially (first to last)

          Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)    
Clinical   3    39.457 13.1523  13.221 0.58618  0.001 ***
Residuals 28    27.855  0.9948         0.41382           
Total     31    67.312                 1.00000           
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Take-home points:

1. The human microbiome is a complex ecosystem that varies considerably across the body and between individuals.

2. The microbiota of the gut is more diverse compared to other body sites, such as vaginal cavity

3. There is still considerable variation in the constituents of the gut and vaginal microbiota among apparently healthy individuals (High inter-sample variation)

4. Different factors (indigenous and exogenous) shape the structure and composition of the human microbiome in different body sites.

5. Both infectious and chronic diseases can disrupt the ecology of the healthy human microbiome producing a state of polymicrobial dysbiosis that can be studied and described using metagenomic tools.

References

1. Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., & Gordon, J. I. (2007). The human microbiome project. *Nature*, *449*(7164), 804
2. Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., ... & Giglio, M. G. (2012). Structure, function and diversity of the healthy human microbiome. *nature*, *486*(7402), 207.
3. Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., ... & Heath, A. C. (2012). Human gut microbiome viewed across age and geography. *nature*, *486*(7402), 222.
4. Tremaroli, V., & Bäckhed, F. (2012). Functional interactions between the gut microbiota and host metabolism. *Nature*, *489*(7415), 242.
5. Nicholson, J. K., Holmes, E., Kinross, J., Burcelin, R., Gibson, G., Jia, W., & Pettersson, S. (2012). Host-gut microbiota metabolic interactions. *Science*, *336*(6086), 1262-1267.
6. Heijtz, R. D., Wang, S., Anuar, F., Qian, Y., Björkholm, B., Samuelsson, A., ... & Pettersson, S. (2011). Normal gut microbiota modulates brain development and behavior. *Proceedings of the National Academy of Sciences*, *108*(7), 3047-3052.
7. Green, K. A., Zarek, S. M., & Catherino, W. H. (2015). Gynecologic health and disease in relation to the microbiome of the female reproductive tract. *Fertility and sterility*, *104*(6), 1351-1357.
8. Ma, B., Forney, L. J., & Ravel, J. (2012). Vaginal microbiome: rethinking health and disease. *Annual review of microbiology*, *66*, 371-389.
9. Huang, B., Fettweis, J. M., Brooks, J. P., Jefferson, K. K., & Buck, G. A. (2014). The changing landscape of the vaginal microbiome. *Clinics in laboratory medicine*, *34*(4), 747-761.
10. Belizário, J. E., & Napolitano, M. (2015). Human microbiomes and their roles in dysbiosis, common diseases, and novel therapeutic approaches. *Frontiers in microbiology*, *6*, 1050.
11. Agarwal, S., Wills, J., Cayton, L., Lanckriet, G., Kriegman, D., & Belongie, S. (2007, March). Generalized non-metric multidimensional scaling. In *Artificial Intelligence and Statistics*(pp. 11-18).
12. Galili, T. (2015). dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics*, *31*(22), 3718-3720.