Course Name

July 04, 2025

Table of Contents

# About

What is the essence/purpose of this course or module (2-3 sentences).



Figure 0.1: C-MOOR logo

### Audience and Prerequisites

What is the target audience?

**Prerequisites**:

* Prereq 1
* Prereq 2

Why are these needed? Provide a brief explanation that will help instructors decide whether it’s a good fit for their students.

### Format

* **Class Type:**
* **Lesson Length:**

### Learning Goals

1. Goal 1
2. Goal 2
3. Goal 3

### Core Competencies

This activity addresses the following core concepts and competencies:

**Vision and Change**

**Genetics**

**Bioinformatics**

Core concepts and competencies are taken from the following sources:

* [Vision and Change in Undergraduate Biology Education](https://visionandchange.org/) AAAS report
* [Genetics Core Competencies](https://genetics-gsa.org/education/genetics-learning-framework/) by [GSA](https://genetics-gsa.org/)
* [Bioinformatics core competencies for undergraduate life sciences education](https://doi.org/10.1371/journal.pone.0196878) by [NIBLSE](https://qubeshub.org/community/groups/niblse)

### C-MOOR Content Collection

This content is part of a collection of teaching resources developed by C-MOOR. C-MOOR works to break down barriers to scientific participation and build pathways for the next generation of data scientists through authentic research experiences. Learn more about C-MOOR by [viewing our projects](https://github.com/c-moor), or read about how C-MOOR is [integrating research experience into undergraduate biology courses](https://www.cloviscollege.edu/alumni-and-community/c-moor/c-moor.html) at Clovis Community College.

# 1 Introduction

## 1.1 Motivation

## 1.2 Target Audience

The course is intended for …

## 1.3 Curriculum

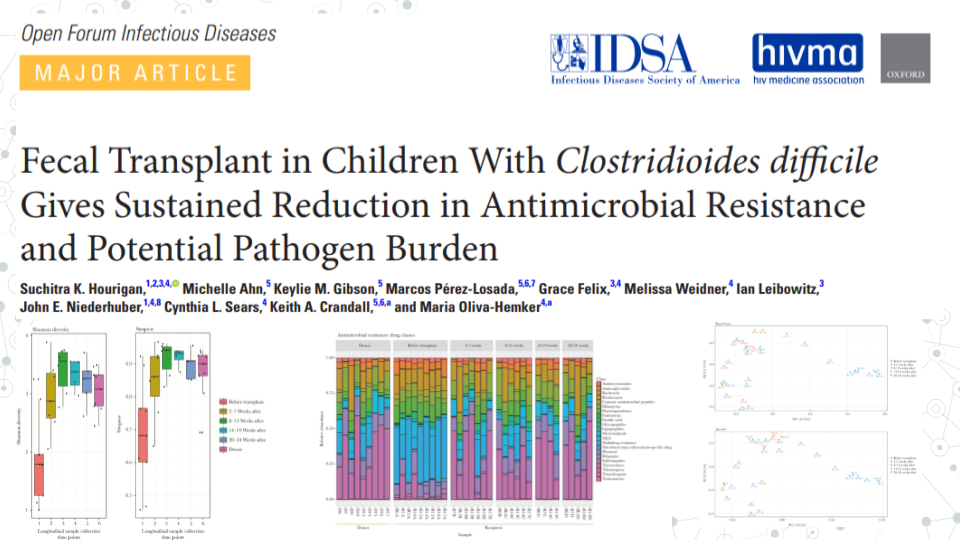
The course covers…

devtools::session\_info()

## ─ Session info ───────────────────────────────────────────────────────────────  
## setting value  
## version R version 4.3.2 (2023-10-31)  
## os Ubuntu 22.04.4 LTS  
## system x86\_64, linux-gnu  
## ui X11  
## language (EN)  
## collate en\_US.UTF-8  
## ctype en\_US.UTF-8  
## tz Etc/UTC  
## date 2025-07-04  
## pandoc 3.1.1 @ /usr/local/bin/ (via rmarkdown)  
##   
## ─ Packages ───────────────────────────────────────────────────────────────────  
## package \* version date (UTC) lib source  
## askpass 1.2.0 2023-09-03 [1] RSPM (R 4.3.0)  
## bookdown 0.41 2024-10-16 [1] CRAN (R 4.3.2)  
## cachem 1.0.8 2023-05-01 [1] RSPM (R 4.3.0)  
## chromote 0.3.1 2024-08-30 [1] CRAN (R 4.3.2)  
## cli 3.6.2 2023-12-11 [1] RSPM (R 4.3.0)  
## devtools 2.4.5 2022-10-11 [1] RSPM (R 4.3.0)  
## digest 0.6.34 2024-01-11 [1] RSPM (R 4.3.0)  
## dplyr 1.1.4 2023-11-17 [1] RSPM (R 4.3.0)  
## ellipsis 0.3.2 2021-04-29 [1] RSPM (R 4.3.0)  
## evaluate 0.23 2023-11-01 [1] RSPM (R 4.3.0)  
## fansi 1.0.6 2023-12-08 [1] RSPM (R 4.3.0)  
## fastmap 1.1.1 2023-02-24 [1] RSPM (R 4.3.0)  
## fs 1.6.3 2023-07-20 [1] RSPM (R 4.3.0)  
## generics 0.1.3 2022-07-05 [1] RSPM (R 4.3.0)  
## glue 1.7.0 2024-01-09 [1] RSPM (R 4.3.0)  
## hms 1.1.3 2023-03-21 [1] RSPM (R 4.3.0)  
## htmltools 0.5.7 2023-11-03 [1] RSPM (R 4.3.0)  
## htmlwidgets 1.6.4 2023-12-06 [1] RSPM (R 4.3.0)  
## httpuv 1.6.14 2024-01-26 [1] RSPM (R 4.3.0)  
## httr 1.4.7 2023-08-15 [1] RSPM (R 4.3.0)  
## janitor 2.2.0 2023-02-02 [1] RSPM (R 4.3.0)  
## jsonlite 1.8.8 2023-12-04 [1] RSPM (R 4.3.0)  
## knitr 1.48 2024-07-07 [1] CRAN (R 4.3.2)  
## later 1.3.2 2023-12-06 [1] RSPM (R 4.3.0)  
## lifecycle 1.0.4 2023-11-07 [1] RSPM (R 4.3.0)  
## lubridate 1.9.3 2023-09-27 [1] RSPM (R 4.3.0)  
## magrittr 2.0.3 2022-03-30 [1] RSPM (R 4.3.0)  
## memoise 2.0.1 2021-11-26 [1] RSPM (R 4.3.0)  
## mime 0.12 2021-09-28 [1] RSPM (R 4.3.0)  
## miniUI 0.1.1.1 2018-05-18 [1] RSPM (R 4.3.0)  
## openssl 2.1.1 2023-09-25 [1] RSPM (R 4.3.0)  
## ottrpal 1.3.0 2024-10-23 [1] Github (jhudsl/ottrpal@2e19782)  
## pillar 1.9.0 2023-03-22 [1] RSPM (R 4.3.0)  
## pkgbuild 1.4.3 2023-12-10 [1] RSPM (R 4.3.0)  
## pkgconfig 2.0.3 2019-09-22 [1] RSPM (R 4.3.0)  
## pkgload 1.3.4 2024-01-16 [1] RSPM (R 4.3.0)  
## processx 3.8.3 2023-12-10 [1] RSPM (R 4.3.0)  
## profvis 0.3.8 2023-05-02 [1] RSPM (R 4.3.0)  
## promises 1.2.1 2023-08-10 [1] RSPM (R 4.3.0)  
## ps 1.7.6 2024-01-18 [1] RSPM (R 4.3.0)  
## purrr 1.0.2 2023-08-10 [1] RSPM (R 4.3.0)  
## R6 2.5.1 2021-08-19 [1] RSPM (R 4.3.0)  
## Rcpp 1.0.12 2024-01-09 [1] RSPM (R 4.3.0)  
## readr 2.1.5 2024-01-10 [1] RSPM (R 4.3.0)  
## remotes 2.4.2.1 2023-07-18 [1] RSPM (R 4.3.0)  
## rlang 1.1.4 2024-06-04 [1] CRAN (R 4.3.2)  
## rmarkdown 2.25 2023-09-18 [1] RSPM (R 4.3.0)  
## rprojroot 2.0.4 2023-11-05 [1] CRAN (R 4.3.2)  
## sessioninfo 1.2.2 2021-12-06 [1] RSPM (R 4.3.0)  
## shiny 1.8.0 2023-11-17 [1] RSPM (R 4.3.0)  
## snakecase 0.11.1 2023-08-27 [1] RSPM (R 4.3.0)  
## stringi 1.8.3 2023-12-11 [1] RSPM (R 4.3.0)  
## stringr 1.5.1 2023-11-14 [1] RSPM (R 4.3.0)  
## tibble 3.2.1 2023-03-20 [1] CRAN (R 4.3.2)  
## tidyselect 1.2.0 2022-10-10 [1] RSPM (R 4.3.0)  
## timechange 0.3.0 2024-01-18 [1] RSPM (R 4.3.0)  
## tzdb 0.4.0 2023-05-12 [1] RSPM (R 4.3.0)  
## urlchecker 1.0.1 2021-11-30 [1] RSPM (R 4.3.0)  
## usethis 2.2.3 2024-02-19 [1] RSPM (R 4.3.0)  
## utf8 1.2.4 2023-10-22 [1] RSPM (R 4.3.0)  
## vctrs 0.6.5 2023-12-01 [1] RSPM (R 4.3.0)  
## webshot2 0.1.1 2023-08-11 [1] CRAN (R 4.3.2)  
## websocket 1.4.2 2024-07-22 [1] CRAN (R 4.3.2)  
## xfun 0.48 2024-10-03 [1] CRAN (R 4.3.2)  
## xml2 1.3.6 2023-12-04 [1] RSPM (R 4.3.0)  
## xtable 1.8-4 2019-04-21 [1] RSPM (R 4.3.0)  
## yaml 2.3.8 2023-12-11 [1] RSPM (R 4.3.0)  
##   
## [1] /usr/local/lib/R/site-library  
## [2] /usr/local/lib/R/library  
##   
## ──────────────────────────────────────────────────────────────────────────────

# 2 Scientific Literature

## 2.1 Pre-lab - Scientific literature



### 2.1.1 Purpose

We’ve all heard the saying. The best way to learn is by doing. So, let’s jump right in and read a science paper! As you read through the paper, keep in mind that **you are not expected to understand everything the first time through**. We will go through the paper in detail in the lab.

Reading this paper will give us a first taste of scientific literature and start familiarizing ourselves with the 16S rRNA analyses you will complete as a part of your research project.

### 2.1.2 Grading criteria

Download the assignment to your local computer as a .docx, complete it, and upload the assignment to your LMS (Blackboard, Canvas, Google Classroom).

### 2.1.3 Activity

*Estimated time: 30 minutes*

#### 2.1.3.1 Instructions

Read the paper [“Fecal transplant in children with Clostridioides difficile gives sustained reduction in antimicrobial resistance and potential pathogen burden” by Hourigan et al. (2019)](https://pmc.ncbi.nlm.nih.gov/articles/PMC6790402/) and answer the following questions.

#### 2.1.3.2 Questions

| 1. What is one thing you find interesting in the paper? |
| --- |
|  |

| 2. Define a term that is new to you. |
| --- |
|  |

| 3. Ask a question about the paper. |
| --- |
|  |

### 2.1.4 Footnotes

#### 2.1.4.1 Resources

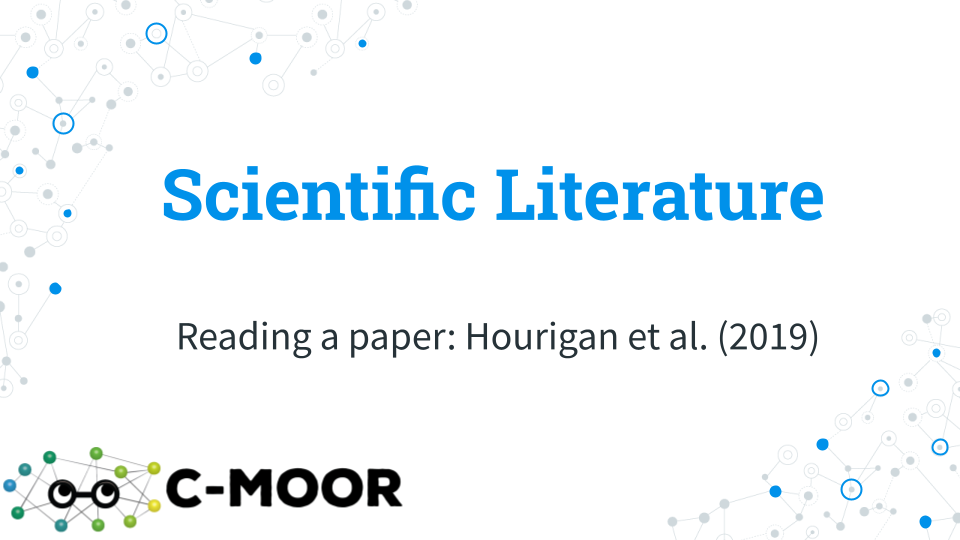
* [Google Doc](https://docs.google.com/document/d/1Uzw-dPwx4kxrcDKLTM7hjxN09oFkTQWg/edit?usp=sharing&ouid=117014088395158370475&rtpof=true&sd=true)
* [Paper](https://pmc.ncbi.nlm.nih.gov/articles/PMC6790402/)

#### 2.1.4.2 Contributions and affiliations

* Valeriya Gaysinskaya, Johns Hopkins University
* Frederick Tan, Johns Hopkins University
* Katherine Cox, Johns Hopkins University
* Sayumi York, Notre Dame of Maryland University
* Stephanie Coffman, Clovis Community College
* Rosa Alcazar, Clovis Community College

## 2.2 Lecture - Reading a paper

*Estimated time:*



[Lecture](https://docs.google.com/presentation/d/1mibD1XUpgRZcDPbYj5XOMUASuScuJoUBVV7RkT5nnSc/edit?usp=sharing)

## 2.3 Activity - Scientific Literature

### 2.3.1 Purpose

Examine research where a 16S rRNA-based approach is used to compare metagenomic diversity in children with Clostridioides difficile infection (CDI), before and after fecal microbiota transplantation (FMT) by reading: [“Fecal transplant in children with Clostridioides difficile gives sustained reduction in antimicrobial resistance and potential pathogen burden” by Hourigan et al. (2019)](https://pmc.ncbi.nlm.nih.gov/articles/PMC6790402/)

### 2.3.2 Learning Objectives

1. Understand the purpose and experimental setup of the paper
2. Understand the presented evidence (Figures and Tables) of the paper

### 2.3.3 Introduction

In this study, the effect of fecal transplantation on microbial composition and diversity in children with CDI is examined using short read shotgun metagenomics. Data visualization and taxonomic comparison between the samples was performed using R-based phyloseq, a tool you will use for 16S rDNA amplicon-seq analysis later in the course. Using phyloseq authors measure alpha diversity between samples, examine beta diversity with Principal Coordinate Analysis(PCoA) plots and use taxonomy barplots for comparing relative abundance of microbes.

### 2.3.4 Instructions

Based on the study, answer the following questions. The main text of the paper can be found below, in the ‘Resources’ section of this assignment.

### 2.3.5 Activity 1 - Overview of the paper

*Estimated time: 30 minutes* Determine the main objectives and purpose of the paper.

| 1. What is the purpose of this study? |
| --- |
|  |

| 2. What is the hypothesis in this study? |
| --- |
|  |

| 3. Describe the knowledge gap. In essence, what did the scientific community not know that this study was trying to answer? |
| --- |
|  |

| 4. What were the main conclusions the authors made in this study? |
| --- |
|  |

### 2.3.6 Activity 2 - Methods and Figure Analysis (in class and homework)

*Estimated time: 45 minutes*

#### 2.3.6.1 Activity 2.1 – Examine the Methods and Data presented in this study

| 1. Summarize one aspect of sample-associated metadata used for this study? e.g. What is the range of subjects’ age; what is the spectrum of subjects’ medical problems? |
| --- |
|  |

| 2. Discuss one method or analysis used in this paper. |
| --- |
|  |

| 3. Summarize one figure presented in this paper. |
| --- |
|  |

#### 2.3.6.2 Activity 2.2 - Figures and a Table

**Describe the method/approach used to generate this data**

|  | Method/Approach |
| --- | --- |
| Table 1 |  |
| Fig. 2 |  |
| Fig. 3A |  |
| Fig. 3B |  |

**Describe the main result/observation from each figure**

|  | Main result/Observation |
| --- | --- |
| Table 1 |  |
| Fig. 2 |  |
| Fig. 3A |  |
| Fig. 3B |  |

#### 2.3.6.3 Activity 3 Conclusions

*Estimated time: 10 min*

**1. Compare your conclusions to author conclusions.**

* How similar are the conclusions you came up with to the authors?
* Were there any differences in interpreting the data?
* Do you think they are over-stating their findings? Do you think their conclusions are accurate and appropriate?

| Summarize one figure presented in this paper. |
| --- |
|  |

#### 2.3.6.4 Activity 4 Future Directions

*Estimated time: 10 min*

**1. Scientific work builds on previous studies. What do you believe could be the next step to further the work these researchers did?** - What follow-up question(s) do you have for the authors?

|  |
| --- |
|  |

**2. What is the impact of this research area in general (or this study in particular?)** - Do you believe further research in this area may benefit society? Can we build on what this study found?

|  |
| --- |
|  |

### 2.3.7 Grading criteria

Download the assignment to your local computer as a .docx, complete it, and upload the assignment to your LMS (Blackboard, Canvas, Google Classroom).

### 2.3.8 Footnotes

#### 2.3.8.1 Resources

* [Google Doc](https://docs.google.com/document/d/1Mg9DGVk0EsSfYwl8VsRHb-_QdtZqiBmhWvbhUWkQERc/edit?usp=sharing)
* [Paper](https://pmc.ncbi.nlm.nih.gov/articles/PMC6790402/)

#### 2.3.8.2 Contributions and affiliations

* Valeriya Gaysinskaya, Johns Hopkins University
* Frederick Tan, Johns Hopkins University
* Sayumi York, Notre Dame of Maryland University

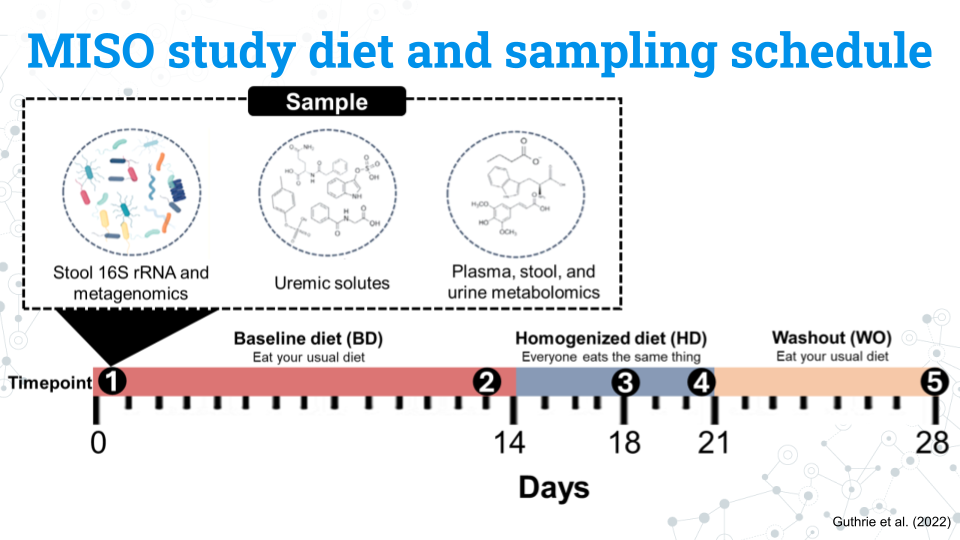
Let’s get prepared to do some research using 16S rRNA data! In this section, we’ll be exploring data from *Impact of a 7-day homogeneous diet on interpersonal variation in human gut microbiomes and metabolomes* by Guthrie et al. (2022). This study is one of many that explores the relationship between diet and the human gut microbiome and will help familiarize us with the format that 16S rRNA data takes. We’ll start slow and look at just a few lines of data in Google Sheets and then use phyloseq and DESeq2 in R to take our analysis to the next level.

# 3 Meet the MISO study

Human diet has been implicated heavily in the establishment and maintenance of the gut microbiome [REFs]. For example, human babies undergo a drastic change in the gut microbiome following the transition to solid food [REF], and microbes that assist in the breakdown of seaweed are found in Japanese gut microbiomes.

It’s difficult to capture the mechanisms and effects of diet on the gut microbiome given the sheer number of variables involved. The microbiome individuality and stability over time (MISO) study aimed to explore the connection between diet, microbiome, and metabolites by looking at the effect of feeding a standardized diet to people over 7 days.

### 3.0.1 MISO study diet and sampling schedule



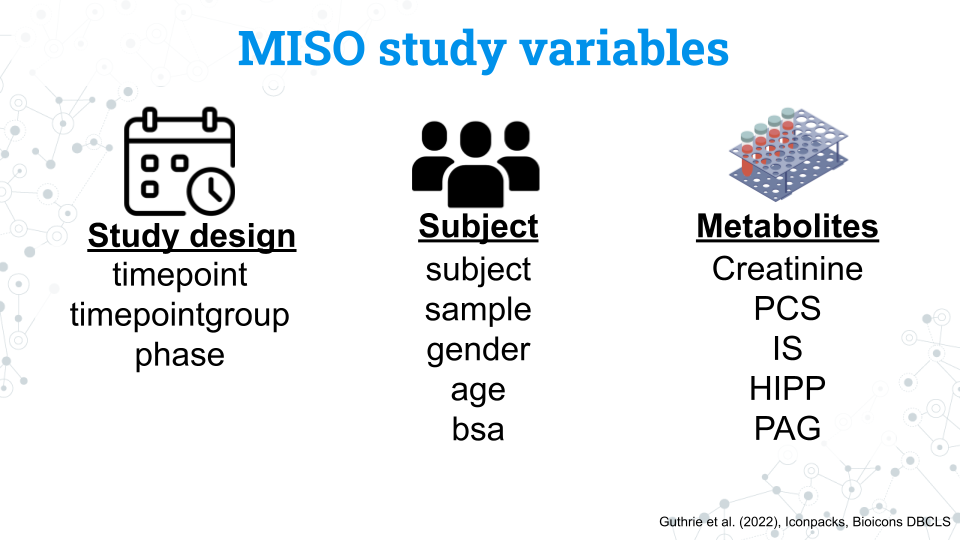
The figure above shows the study design for the MISO study. You do not need to memorize all these details; feel free to refer to this page throughout the project:

* Participants eat their usual, **baseline diet (BD) for 14 days**
* Participants all eat the same diet, the **homogenized diet (HD), for 7 days**
* Participants return to their usual diet during the **washout (WO) period for 7 days**

The study lasts a total of 28 days. Samples from the blood, stool, and urine, our metabolite and 16S rRNA data are taken at 5 different timepoints:

* **Timepoint 1:** Day 0 (the start of the study)
* **Timepoint 2:** Day 13
* **Timepoint 3:** Day 17
* **Timepoint 4:** Day 21
* **Timepoint 5:** Day 28

### 3.0.2 MISO study variables and factors



We have a number of variables we can use in our analysis. Notice that the variables for the study and subjects are in lowercase. This is also how you will access these variables in R.

#### 3.0.2.1 Study design variables

| Variable | What is it? | Factors |
| --- | --- | --- |
| timepoint | The 5 samplings that occur on days 0, 13, 17, 21, and 28 coded as timepoints 1 through 5 | 1, 2, 3, 4, 5 |
| timepoint group | The diet the subject was on during the sampling | BD, HD, WO |
| phase | Whether the subject was on their usual diet or the homogenized diet during sampling | non-miso, miso |

#### 3.0.2.2 Subject variables

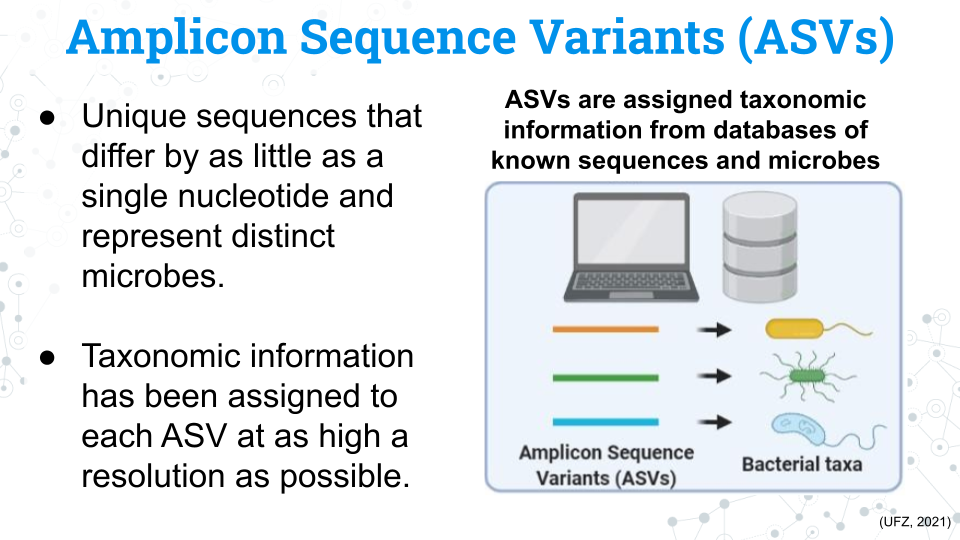
| Variable | What is it? | Factors |
| --- | --- | --- |
| subject | A unique ID given to each subject (participant) | S## (ex. S02 is subject 2. Note that while there are a total of 21 subjects in the study, but their subject numbers are not 1 through 20 |
| sample | A unique ID given to every sample taken during the study that includes the subject and timepoint of the sample | MISO-Subject##-Sample# (ex. MISO1-S02-1 is the sample from subject 2 at timepoint 1) |
| gender | The gender of the subject | M, F. All subjects were cisgender |
| age | The age in years of the subject | A continuous variable from 23 to 75 years old |
| bsa | Body surface area; a measure of body size | A continuous variable from 1.6 to 2.8 |

#### 3.0.2.3 Metabolite variables

| Variable | What is it? | Factors |
| --- | --- | --- |
| Creatinine | Creatinine | A continuous variable from 1072 to 3971 |
| PCS | *p*-cresol sulfate | A continuous variable from 2 to 95 |
| IS | Indoxyl sulfate | A continuous variable from 3 to 58 |
| HIPP | Hippuric acid | A continuous variable from 16 to 1119 |
| PAG | Phenylacetylglutamine | A continuous variable from 16 to 318 |

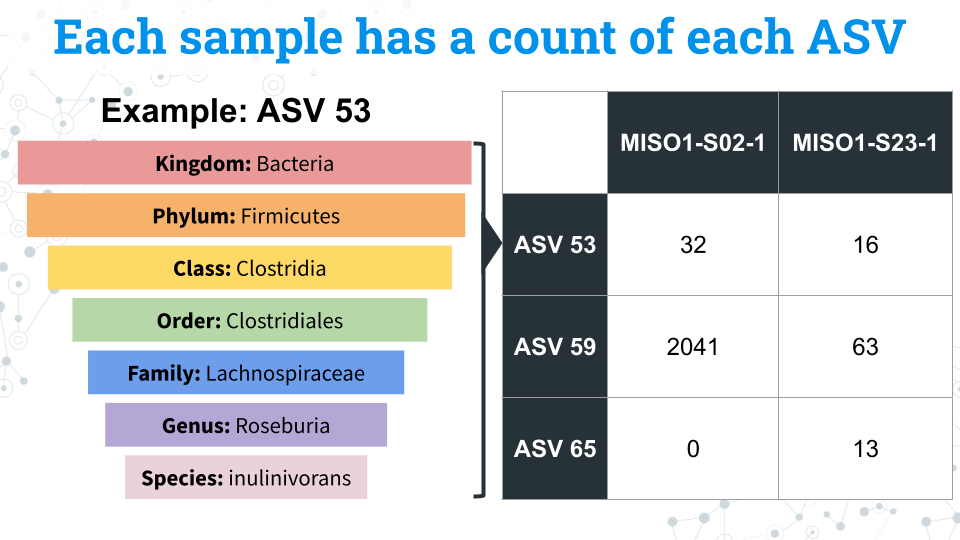
### 3.0.3 Amplicon sequence variants (ASVs) data

Finally we have our microbe count data, or our *Amplicon Sequence Variants (ASVs)*. Each ASV is a unique sequences that differs by as little as a single nucleotide from other ASVs and represents a specific microbe. We have assigned taxonomy to these microbes, but not all microbes have taxonomic data through the species level. These missing fields will appear as NA in the data.



You may see the word Operational Taxonomic Unit (OTU) in phyloseq and in other published studies. **For our purposes, we will generally consider OTUs and ASVs as interchangable**, although they have some key differences.

Each sample is associated with a count of each ASV. We can compare the counts of these ASVs between samples to determine differences in the composition of the microbiome between samples.



### 3.0.4 Footnotes

#### 3.0.4.1 Resources

#### 3.0.4.2 Contributions and affiliations

* Sayumi York, Notre Dame of Maryland University

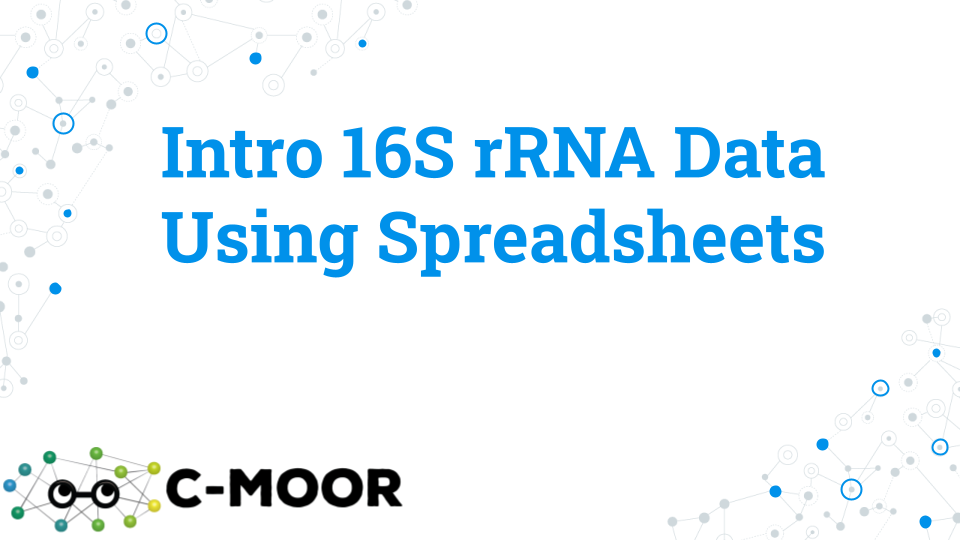
# 4 16S rRNA Data in spreadsheets

## 4.1 Exploring 16S rRNA data in spreadsheets

In this section we will look at 16S data in a spreadsheet via Google Sheets. We will familiarize ourselves with the format of 16S data and manipulate just a few ASVs to better understand what kinds of research questions we can ask with the data. Through this exercise we will also come to understand the necessity of more powerful analysis tools in R; sorting through trends in a spreadsheet could take years!

## 4.2 Lecture - Introducing 16S rRNA Data Using Spreadsheets

*Estimated time:*



[Lecture](https://docs.google.com/presentation/d/1nGKvSYGHJu0JZLMWZiFu_iGLIC16IKlnWsVyNyjdyek/edit?usp=sharing)

## 4.3 Activity - Introducing 16S rRNA Data Using Spreadsheet

### 4.3.1 Learning objectives

1. Explore Amplicon Sequence Variants (ASVs) through the taxonomy profile of select ASVs, and sample metadata for a subset of ASVs.
2. Generate initial research questions based on the above exploratory data analysis

### 4.3.2 Introduction

The most popular sequencing technique for the analysis of bacterial diversity is targeted sequencing, or sequencing of a specific gene (or region of a gene, e.g. a hypervariable region of the bacterial 16S ribosomal rRNA gene) using Polymerase Chain Reaction (PCR) to create sequences called amplicons. Sequence variation in the resulting amplicons creates Amplicon Sequence Variants (ASVs). ASVs varying from as little as one single nucleotide are defined as separate ASVs and as little as 1% difference in ASV sequence can be associated with different species. In this activity you will be exploring ASVs generated with sequencing 250 nucleotides (nt) Illumina sequencing, where Amplicon sequence variants (ASVs) were identified.

### 4.3.3 Activity 1 – Quick data overview based on 1 sample

*Estimated time: 10 min*

Although the full dataset contains 105 samples from 21 subjects and 5 timepoints (representing 3 dietary conditions), in this activity we will familiarize ourselves with the data by exploring data from 1 sample only.

#### 4.3.3.1 Activity 1-1. Explore data contents using one sample only, namely MISO1-S02-1 (subject 2, timepoint 1).

Access [‘MISO\_PS\_Leah\_SY\_Sample1’](https://docs.google.com/spreadsheets/d/11_OgIKbxnIiL8TsegxTu2GNtF9wr7MNGT_SLWq4Q8lU/edit?gid=1418834244#gid=1418834244) file and open it with Google Sheets.

1. Rows = ASVs
2. Columns = Samples and taxonomy:

* Col A = ASV id
* Cols B-H = Taxonomic information
* Col I = ASV counts (for sample MISO1-SO2-1)

| 1-1.1 How many ASVs are there? Scroll down to see how many rows are there |
| --- |
|  |

**1-1.2 How many microbes are there in an individual? In other words, what is the total number of ASVs per sample?**

IMPORTANT: An ASV is associated with an individual if it is NOT a zero. Zero means that particular ASV was not found in the individual, so you need to exclude zeroes from your calculations.

Create an empty new first row for temporary calculations by clicking to insert 1 row above. In the new row, calculate the number of ASVs per sample 1 (excluding the zeroes) using the following formula: =COUNTIF(range, “<>0”). For sample 1 (MISO1-S02-1) enter formula =COUNTIF(I2:I2185, “<>0”).

|  |
| --- |
|  |

| 1-1.3 What taxonomy is associated with ASV1? Include all taxonomic ranks. |
| --- |
|  |

| 1-1.4 What taxonomy is associated with ASV2? Include all taxonomic ranks. |
| --- |
|  |

| 1-1.5 For sample 1 (MISO1-S02-1) what is the ASV1 count? |
| --- |
|  |

| 1-1.6 For sample 1 (MISO1-S02-1) what is the ASV2 count? |
| --- |
|  |

**1-1.7 Plot ASV counts distribution (range of ASV abundance) for sample 1 ( (MISO1-S02-1).**

1. Make a bar graph of ASV distribution. To make a bar graph, hold Command key (or Ctrl on Windows, allows you to select data that are not next to each other) and then highlight columns A, with ASV ids (column “ASV”) and I, with ASV counts (column “MISO1-S02-1”) and go to Insert Chart.

* Make sure your graph is a bar plot: click on the 3 dots in the top right corner of the plot → edit chart and in the Chart editor, under Chart type, selecting Column Chart.
* Copy the bar graph into a box below: click on the 3 dots in the top right corner of the plot, and select copy chart, then, paste it below. |:–| | |

| B) What is the highest abundance ASV in sample 1? A simple way to get an answer is to hover over the highest bar in your ASV distribution plot from activity at its highest point. |
| --- |
|  |

| C) Describe abundance pattern between ASVs. E.g., is abundance even or spotty and what does that say about ASV/taxa variation? |
| --- |
|  |

### 4.3.4 Activity 2 – Explore variation between individuals

*Estimated time: 10 min*

One aspect of Guthrie et al. study is the extent to which diet contributes to interpersonal microbiome variation (variation in microbiome between individuals). To this end, authors ‘normalize’ the diet for all individuals with the same Homogenized Diet (HD) to evaluate how much of an effect the diet has. **The study period is 28 days.** The study has 3 dietary conditions: 1) **BD** (Baseline Diet) which lasted 14 days (2 time points collected), followed by 2) **HD** which lasted 7 days (2 time points collected), followed by 3) **WO** (washout diet), which lasted for 7 days and simply marks the return to baseline, regular diet (1 time point).

**To explore variation between individuals, here we simplify the dietary variation to a single condition, WO.**

* For this activity, the data file MISO\_PS\_Leah\_SY\_WO-select contains samples pre-filtered for WO condition only (and excludes samples corresponding to BD and HD).
* Since there is only 1 WO datapoint per individual, each WO sample corresponds to a different individual, so you should see 21 samples.

#### 4.3.4.1 Activity 2-1 - What are the most common ASVs between individuals?

IMPORTANT: An ASV is associated with an individual if it is NOT a zero. Zero means that particular ASV was not found in the individual, so you need to exclude zeroes from your calculations.

Calculate the number of samples associated with each ASVs. Record first 5 values below.

* Create an empty new first column for temporary calculations by inserting new column A (via clicking on column A and selecting to Insert 1 column to the left). Label new column “Occurrence”.
* Starting with the first ASV in row 3, in an empty cell A3 use the following formula: =COUNTIF(range, “<>0”). For example, for ASV1 enter formula = COUNTIF(C3:W3, “<>0”). The answer should be 17 (so 17 of 21 WO samples have ASV1).
* Extend the function to all ASVs

|  |
| --- |
|  |

**B) Calculate % of samples associated with each ASV and record the first top 5.** - Create another new column A and label it “% Occurrence”. | | |:–| | |

**C) How many ASVs are present in all 21 individuals and what are their ids? Simply find which ASVs are present in 100% of samples.** - Copy columns A (% samples) and C (ASV id) into a new sheet and Paste Special → Values only - Delete empty row 1 - Right-click on Row 1 to Create a filter, then Sort Z to A - Count the ASVs that have 100 in % samples  
| | |:–| | |

**Activity 2-2. Explore most common ASV taxonomy**

* Go the original MISO\_PS\_Leah\_SY\_Sample1 google docs and record taxonomic information for most common ASV from the activity above.
* In MISO\_PS\_Leah\_SY\_Sample1, Columns B-H correspond to taxonomic information

| 2-2.1 What is the taxonomy assigned to the most common ASV you found in Activity above? |
| --- |
|  |

| 2-2.2 According to the taxonomy for most common ASV above, what might be the most abundant phyla in the human gut? |
| --- |
|  |

**Activity 2-3. Plot individual variation in abundance of the most common ASV.**

1. Make a bar graph of most common ASV distribution across 21 individuals
2. In the original MISO\_PS\_Leah\_SY\_WO-select find the row corresponding to the most common ASV, from Activity 2-1 above.
3. Highlight the header row (corresponding to sample IDs) and the row corresponding to most common ASV and go to Insert Chart. Make sure you don’t include the columns with the # of samples with that ASV and % samples columns.

* Make sure your bar graph is a bar plot: click on the 3 dots in the top right corner of the plot → edit chart and in the Chart editor, under Chart type, selecting Column Chart.
* Copy the bar graph into a box below: click on the 3 dots in the top right corner of the plot, and select copy chart, then, paste it below.

|  |
| --- |
|  |

| B) Comment on the chart above - how similar is the ASV abundance between individuals? |
| --- |
|  |

### 4.3.5 Activity 3 – Explore variation within a single individual associated with diet

*Estimated time: 10 min*

Although the full dataset contains 105 samples from 21 subjects and 5 timepoints (representing 3 dietary conditions), in this activity we will zoom into ASV counts data for 1 individual (comprised of 5 samples representing the 5 timepoints collected from that individual).

* Timepoints 1 and 2: BD (Baseline diet) prior to normalization with homogenized diet
* Timepoints 3 and 4: HD (Homogenized diet): normalized to be same for all
* Timepoint 5: WO (Washout diet), which simply involves individuals going back to their baseline, regular diet

**Activity 3-1 Access ‘MISO\_PS\_Leah\_SY\_Subject1’ file and open it with Google Sheets, then explore differences between dietary changes in individual 1 by exploring differences between the 5 sample timepoints.**

**3-1.1 Calculate variation in ASV abundance within individual 1 (MISO1-S02) and record the first 5 values below.** This involves plotting counts for ASV 1-10 for MISO1-SO2 timepoints 1 through 5. To calculate variation within a sample, we will use the VAR.S function in google sheets to calculate variance of a sample.

1. Label an empty Column G “VAR”. Use this column to calculate variance.
2. Starting with the first ASV use the following formula: =VAR.S(range) to calculate ASV abundance variation in the 5 samples/timepoints collected for individual MISO1-SO2. For example, for ASV1 enter formula =VAR.S(B2:F2). The answer should be 14947547.8.
3. Extend the function to all ASVs and record the first 5 variance values below. | | |:–| | |

**3-1.2 Plot ASVs with highest variance.** Sort Variance column and record the top 5 values below. To sort the variance column: 1. Click on the 1st (header) row 1. Right-click to select ‘Create a filter’ 1. Sort ‘VAR’ column Z to A (click on the new arrow in the VAR column). | | |:–| | |

After sorting ASVs based on high to low variance make a plot of ASV counts corresponding to top 15 ASVs with the highest variance. Insert the plot in the grey box below. Here, variance within a sample might suggest variation (difference) between conditions. For example, high variance may come from the first 2 samples (BD diet) being different from samples 3 and 4 (HD diet).

* To make a bar graph highlight top 16 rows corresponding to sample names and the 15 ASVs with highest variance, for samples MISO1-S02-1 through 5. Exclude highlighting the variance column itself.
* Go to Insert Chart. For this chart you can highlight the rows starting with title row so as to include sample names in your chart.
* Make sure your graph is a bar plot: click on the 3 dots in the top right corner of the plot → edit chart and in the Chart editor, under Chart type, selecting Column Chart.
* Copy the bar graph into a box below: click on the 3 dots in the top right corner of the plot, and select copy chart, then, paste it below. | | |:–| | |

**Activity 3-2. Examine the top 15 most variable ASVs and in the box below comment/speculate on source of variation for one of the ASVs. Do you think the source of variation is change of diet, inconsistent replicates or something else?** - E.g., Variance associated with different diets would produce a plot where ASV counts for samples 1 and 2 (BD diet) might be different from samples 3 and 4 (HD diet) or sample 5 (WO diet).  
- E.g., Variance observed between samples 1 and 2 (both BD diets) might suggest difference between biological replicate 1 and 2 due to unknown factors. | | |:–| | |

### 4.3.6 Activity 4 – Try it Out!

*Estimated time: 60 min*

Now it’s time to explore the full dataset with all 105 samples! Working with your group, explore one of the questions below or come up with another question. When you are done, copy and paste your results into the class poster. Here are the files that you will need for this Activity:

* [ASV counts (2185 rows by 114 columns)](drive.google.com/file/d/1plBUMPNPpxRN_UCyBc23Fdb_eZjaUCsA)
* [Sample metadata (106 rows by 28 columns)](drive.google.com/file/d/1FUolpv4GnS0pTGX_wLw2IksbL5L4ppsP)
* [Class Poster](https://docs.google.com/presentation/d/1eE5AltbkW6km0r3uQO8KKnUVRLCLx70SDNuBHp6aASU/edit#slide=id.g33ef02f472a_0_0)

Remember, to relate sample IDs from MISO\_PS\_Leah\_SY\_FULL, to the metadata in the MISO\_metadata\_FULL file.

For example, the MISO study timepoints are: - Timepoints 1 and 2: **BD** (Baseline diet) prior to normalization with homogenized diet - Timepoints 3 and 4: **HD** (Homogenized diet): normalized to be same for all - Timepoint 5: **WO** (Washout diet), which simply involves individuals going back to their baseline, regular diet

**Question 1 - Profile and compare ASV abundance and taxonomy for 2 diets, BD and HD. Your final analysis should result in a plot and/or a table.** - Question 1A: Compare 1 BD timepoint and 1 HD timepoint (males): n = 20 - Question 1B: Compare 1 BD timepoint and 1 HD timepoint (females) : n = 22

Sample approach: Compare the 5 most abundant ASVs between BD1 and HD1. You’ll have to think about:

* How will you identify the 5 most abundant ASVs? Will you identify the top 5 in the BD group, and top 5 in the HD group? Should you sort on 1, 2, or all samples within a group? Or just average all samples within a group?)
* How will you compare the multiple BD1 samples against the multiple HD1 samples?

1. ASV counts - plot average or boxplot of abundance for top ASVs?
2. For taxonomy - is there an overlap between 5 BD and 5 HD taxa? Should you consider a pie chart, a barplot or no plot?)

**Question 2 - Profile and compare ASV abundance and taxonomy of 21 individuals using a single homogenized dietary timepoint. Your final analysis should result in a plot and/or a table.**

* Question 2A: Examine HD1 timepoint (males and females): n = 21
* Question 2B: Examine HD2 timepoint (males and females): n = 21

Sample approach: You can compare the 5-10 most abundant ASVs between 21 individuals for HD1. You’ll have to think about - How will you identify the 5 most abundant ASVs? Should you show the top 5 per each sample? Sort on 1, 2, or all samples? Or just average all samples?) - How will you compare the 21 samples? 1. ASV counts - plot average or boxplot of abundance for top ASVs? 1. For taxonomy - Barplot top taxa for all individuals? Which rank - Phyla, Genus, Species?)

### 4.3.7 Grading criteria

* Download this assignment as Microsoft Word (.docx) and upload on Canvas
* Download your Google Sheet as Microsoft Excel (.xlsx) and upload on Canvas

### 4.3.8 Footnotes

#### 4.3.8.1 Resources

[Google Doc](https://docs.google.com/document/d/1fTy4D3KShevTMuQ0ct8HZ60fWqUSUPcy7e-eHqY3N3w/edit?usp=sharing) [Google Sheet](https://docs.google.com/spreadsheets/d/11_OgIKbxnIiL8TsegxTu2GNtF9wr7MNGT_SLWq4Q8lU/edit?gid=1418834244#gid=1418834244)

#### 4.3.8.2 Contributions and affiliations

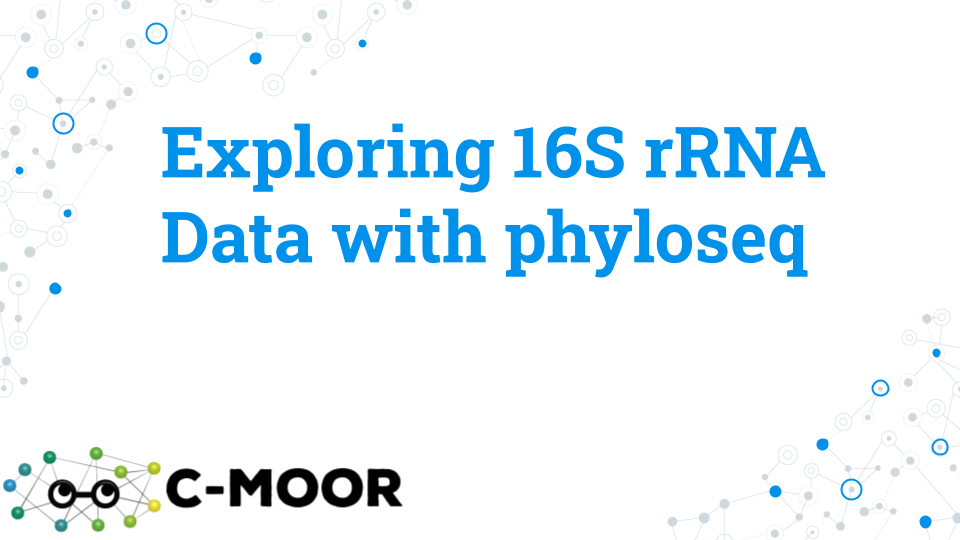
* Valeriya Gaysinskaya, Johns Hopkins University
* Gauri Paul, Clovis Community College
* Frederick Tan, Johns Hopkins University
* Sayumi York, Notre Dame of Maryland University

# 5 Exploring 16S rRNA Data with phyloseq

## 5.1 Exploring 16S rRNA Data with phyloseq

## 5.2 Lecture - Exploring 16S rRNA Data with phyloseq

*Estimated time:*



[Lecture](https://docs.google.com/presentation/d/1gJ_4YoqDJD5-40MArbpgTr_5652JpHI7cgnHHo4Vsws/edit?usp=sharing)

## 5.3 Activity - Exploring 16S rRNA Data with phyloseq

To explore bacterial diversity based on 16S rRNA gene sequencing using data from the “Impact of a 7-day homogeneous diet on interpersonal variation in human gut microbiomes and metabolomes” by Guthrie et al., 2023. This study is also referred to as the MISO study for “Microbiome Individuality and Stability Over Time” because the study aims to understand variation (or stability) of microbiomes across individuals.

This activity aims to explore and profile the metagenomic diversity using the R/Bioconductor package phyloseq which leverages various R-based tools to produce publication-quality taxonomy profiling and analysis graphics.

### 5.3.1 Learning objectives

1. Use phyloseq to explore data associated with Amplicon Sequence Variants (ASVs).
2. In phyloseq, subset data based on metadata and taxonomy.
3. In phyloseq, profile and plot taxonomy.

### 5.3.2 Introduction

The most popular sequencing technique for the analysis of bacterial diversity is targeted sequencing, or sequencing of a specific gene (or region of a gene, e.g. a hypervariable region of the bacterial 16S ribosomal rRNA gene) using polymerase chain reaction (PCR) to create sequences called amplicons. Sequence variation in the resulting amplicons creates amplicon sequence variants (ASVs). ASVs varying from as little as one single nucleotide can be defined as separate ASVs, which can be further clustered into OTUs (Operational Taxonomic Units) based on sequence similarity; e.g. as ASVs within 1% sequence difference can clustered into the same species/OTU.

### 5.3.3 Activity 1 – Explore 16S rRNA Data with phyloseq tutorial

*Estimated time: 25 min*

Explore a phyloseq object through the “Explore 16S rRNA Data with phyloseq” tutorial on SciServer. Log into SciServer, click on compute, and create a new C-MOOR LearnR container. When creating a container, remember to: 1. Use the “C-MOOR LearnR (Bioconductor 3.17)” image (not the Bioconductor 3.16) 2. Check the box next to Data volume “C-MOOR Data” 3. Start the “Explore 16S rRNA Data with phyloseq” tutorial. Visit SciServer Guides and FAQs. If you need assistance accessing the tutorial. 4. To move through the activities click “Continue” at the bottom of the screen. When you are done with a topic, click “Next Topic” to move on. 5. This tutorial has small boxes in which you can enter and run short lines of code to analyze the data. As you work through the tutorial, take snapshots of your work and paste your answers in the grey boxes below:

| 1-1. OTU\_table Exercise – Take a snapshot and paste your code and the output for the following question in the tutorial: “What is the normalized count for ASV115 in sample 10?” |
| --- |
|  |

| 1-2. Visualizing Taxonomy Exercise – Take a snapshot and paste your code and the output for the following question in the tutorial: “What timepoint has the lowest proportion of phylum Verrucomicrobia?” |
| --- |
|  |

| 1-3. Using subset\_taxa() Exercise – Take a snapshot and paste your code and the output for the following question in the tutorial: “What are the 2 most abundant orders in the class Gammaproteobacteria?” |
| --- |
|  |

### 5.3.4 Activity 2 – Try it out!

*Estimated time: 60 min*

With your group, perform some exploratory data analysis selecting from one of the four questions below or coming up with your own question. When you are done, copy and paste your results into the class poster.  
[Class Poster](https://docs.google.com/presentation/d/1eE5AltbkW6km0r3uQO8KKnUVRLCLx70SDNuBHp6aASU/edit?slide=id.g343cb6dd56c_2_0#slide=id.g343cb6dd56c_2_0)

#### 5.3.4.1 Question 1 - What has more impact on human microbial variation, diet or individuality?

Approach: Plot high level taxa (Phyla) for all timepoints and subjects and compare the plots to determine which factor has a stronger effect on microbiome composition.

**Step 1A: Plot high-level taxonomy (Phylum level) for each timepoint using the plot\_bar() function.** Specifically, your plot should show Phyla diversity across all 5 timepoints. - Refer to the “Explore 16S rRNA Data with phyloseq” tutorial for help using the plot\_bar() function. - Ensure your plot has a title - Remember, the BD diet includes timepoints 1 and 2, the HD diet includes timepoints 3 and 4, and the WO diet includes timepoint 5. - Use the following code as a template:

plot\_bar(miso, "fill in the blank", fill = "fill in the blank", title = "choose a name for your graph") +  
 geom\_bar(aes(color = fill in the blank, fill = fill in the blank), stat = "identity", position = "stack")

| 1A-1. Paste figure 1A below. |
| --- |
|  |

**1A-2. State your conclusion for Figure 1A below.** - Include a list of the identified Phyla and their names - Include if you think there is a difference between conditions and your reasoning why or why not || |:–| | |

Step 1B: Plot high-level taxonomy (Phylum level) for each individual, using only normalized homogenized diet “HD” timepoints to minimize contribution of diet to the analysis using the plot\_bar() function. Specifically, your plot should show Phyla diversity across all 21 individuals at HD timepoints.

* Use the command below to create a phyloseq object with only “HD” timepoints

misoHD = subset\_samples(miso, timepointgroup == "HD")

* Refer to the “Explore 16S rRNA Data with phyloseq” tutorial for help using the plot\_bar() function.
* Ensure your plot has a title

| 1B-1. Paste figure 1B below. |
| --- |
|  |

**1B-2. State your conclusion for Figure 1B below.** - Include a list of the identified Phyla and their names - Include if you think there is a difference between conditions and your reasoning why or why not || |:–| | |

A suggestion for your poster: 1. Combine plot 1A-1 and 1B-1 into a single figure. Include their titles and short figure description (figure legend). 1. Paste the code you used to generate your final figure(s) below: || |:–| | |

#### 5.3.4.2 Question 2 - What are some of the most abundant microbes when viewed at different taxonomic resolutions?

Approach: Plot progressively lower level taxa for all individuals, subsetting most abundant from each rank.

**Step 2A. Plot Phyla across all subjects and identify the most abundant Phyla.** Specifically, your plot should show Phyla diversity across all 21 individuals. - Refer to the “Explore 16S rRNA Data with phyloseq” tutorial for help using the plot\_bar() function. - Use the following code as a template:

plot\_bar(miso, "fill in the blank", fill = "fill in the blank", title = "choose a name for your graph") + geom\_bar(aes(color = fill in the blank, fill = fill in the blank), stat = "identity", position = "stack")

| 2A-1. Insert the resulting plot below: |
| --- |
|  |

| 2A-2. Enter the name of the most abundant Phyla below: |
| --- |
|  |

**Step 2B. Subset the most abundant Phyla, and plot Order (Skip Class!)** - Refer to the “Subset taxonomy” section of the “Explore 16S rRNA Data with phyloseq” tutorial for help using the subset\_taxa and plot\_bar() functions. - Use the following code as a template:

subset = subset\_taxa(miso, Phylum == "fill in the blank")  
plot\_bar(subset, "subject", fill = "Order", title = "choose a name for your graph") +   
 geom\_bar(aes(color = "fill in the blank", fill = "fill in the blank"), stat = "identity", position = "stack")

| 2B-1. Insert the resulting plot below: |
| --- |
|  |

| 2A-2. Enter the name of the most abundant Order below: |
| --- |
|  |

**Step 2C. Subset the most abundant Order, and plot associated Family** - Refer to the “Subset taxonomy” section of the “Explore 16S rRNA Data with phyloseq” tutorial for help using the subset\_taxa and plot\_bar() functions. - Use the following code as a template:

subset = subset\_taxa(miso, Order == "fill in the blank")  
plot\_bar(subset, "fill in the blank", fill = "Family", title = "choose a name for your graph") +   
 geom\_bar(aes(color = fill in the blank, fill = fill in the blank), stat = "identity", position = "stack")

| 2C-1. Insert the resulting plot below: |
| --- |
|  |

| 2C-2. Enter the name of the most abundant Family below: |
| --- |
|  |

**Step 2D. Subset most abundant Family, and plot associated Genus** - Refer to the “Subset taxonomy” section of the “Explore 16S rRNA Data with phyloseq” tutorial for help using the subset\_taxa and plot\_bar() functions. - Use the following code as a template:

subset = subset\_taxa(miso, Family == "fill in the blank")  
plot\_bar(subset, "fill in the blank", fill = "Genus", title = "choose a name for your graph") +   
 geom\_bar(aes(color = fill in the blank, fill = fill in the blank), stat = "identity", position = "stack")

| 2D-1. Insert the resulting plot below: |
| --- |
|  |

| 2D-2. Enter the name of the most abundant Genus of your chosen Family below: |
| --- |
|  |

**Step 2E. Subset the most abundant Genus, and plot associated Species** - Refer to the “Subset taxonomy” section of the “Explore 16S rRNA Data with phyloseq” tutorial for help using the subset\_taxa and plot\_bar() functions. - Use the following code as a template:

subset = subset\_taxa(miso, Genus == "fill in the blank")  
plot\_bar(subset, "fill in the blank", fill = "fill in the blank", title = "choose a name for your graph") +   
 geom\_bar(aes(color = fill in the blank, fill = fill in the blank), stat = "identity", position = "stack")

* Note, given the many different Species associated with an abundant Genus, we may need to reduce the legend text and key size in order to visualize the plot better.
* You can do so by adding the following lines of code:

theme(legend.text=element\_text(size=6)) +   
theme(legend.key.size = unit(6, "pt"))

| 2E-1. Insert the resulting plot below: |
| --- |
|  |

| 2E-2. Enter the name of the some of the most abundant species below: |
| --- |
|  |

| 2E-3. After looking through all your plots, summarize your conclusion regarding individual microbiome composition diversity. Explain what reasoning led you to your conclusion. |
| --- |
|  |

A suggestion for your poster: - Combine 4 plots into a single Figure showing your subsetting of the most abundant Order, Family, Genus, and Species. Skip showing Phyla because the group working on Activity 1 will have Phyla data and show a plot for that. Include title and short figure description (figure legend). - Paste the code you used to generate your final figure(s) below: | | |:–| | |

#### 5.3.4.3 Question 3 - Does gender contribute to individual microbiome variation?

Approach: First survey the differences in microbiome across gender at the level of phyla and species, and identify potential candidates that may differ between males and females. Then, plot your candidates of interest and additional candidate microbes previously associated with gender-based variation.

**Step 3A. Perform an initial survey to see if there are any differences in phyla and species composition between genders.** Specifically, your plot should show phyla diversity within male and female groups. - Refer to the “Explore 16S rRNA Data with phyloseq” tutorial for help using the plot\_bar() function. - Use the following code as a template:

plot\_bar(miso, "fill in the blank", fill = "fill in the blank", title = "fill in the blank") +   
geom\_bar(aes(color = fill in the blank, fill = fill in the blank), stat = "identity", position = "stack")

| 3A-1. Insert the resulting plot below: |
| --- |
|  |

Consider any initial differences you see, then choose one high-abundant and one low-abundant candidate Phylum that shows hints of being differentially abundant between genders.

| 3A-2. Record the phylum you chose: |
| --- |
|  |

**Step 3B. Test gender-based difference in Species for the high abundant phylum you selected.** Specifically, your plot should show species diversity within male and female groups for only your phylum of choice.

* Refer to the “Subset taxonomy” section of the “Explore 16S rRNA Data with phyloseq” tutorial for help using the subset\_taxa and plot\_bar() functions.
* Use the following code as a starter:

subset = subset\_taxa(miso, Phylum == "fill in the blank")  
  
plot\_bar(subset, "fill in the blank", fill = "Species", title = "choose a name for your graph") +   
 geom\_bar(aes(color = fill in the blank, fill = fill in the blank), stat = "identity", position = "stack")

* Note, given the many different Species associated with an abundant Genus, we may need to reduce the legend text and key size in order to visualize the plot better.
* You can do so by adding the following lines of code:

theme(legend.text=element\_text(size=6)) +   
theme(legend.key.size = unit(6, "pt"))

| 3B-1. Insert the resulting plot below: |
| --- |
|  |

| 3B-2. Do you observe any differences in species based on gender? |
| --- |
|  |

**Step 3C. Test gender-based difference in Species for the low abundant phylum you selected.** Specifically, your plot should show species diversity within male and female groups for only your phylum of choice. - Refer to the “Subset taxonomy” section of the “Explore 16S rRNA Data with phyloseq” tutorial for help using the subset\_taxa and plot\_bar() functions. - Use the code you used in Step 3B as a template for your code

| 3C-1. Insert the resulting plot below: |
| --- |
|  |

**Step 3D: Plot the individual abundance of 2 candidate microbes for gender differences based on your analysis above, and/or the curated menu of microbes below (previously associated with gender variation), with color denoting gender.** Specifically, your plot should show the chosen microbial group across 21 subjects with the bars colored by gender.

* Refer to the “Subset taxonomy” section of the “Explore 16S rRNA Data with phyloseq” tutorial for help using the subset\_taxa and plot\_bar() functions.
* Use the following code as a template:

subset = subset\_taxa(miso, Phylum == "fill in the blank")  
plot\_bar(subset, "subject", fill = "gender", title = "fill in the blank") +   
 geom\_bar(aes(color = gender, fill = gender), stat = "identity", position = "stack")

Here is the candidate list of different microbes previously implicated in gender-based variation. - Phylum: Proteobacteria - Order: Clostridiales - Family: Ruminococcaceae - Genus: Prevotella - Genus: Fusobacterium - Species: muciniphila

| 3D-1. Insert the plot from your first chosen microbial group below: |
| --- |
|  |

| 3D-2. Insert the plot from your second chosen microbial group below: |
| --- |
|  |

| 3D-3. Based on the all graphs you made, did you identify any candidate Phylum or Species that may differ between genders? |
| --- |
|  |

| 3D-4. Based on the graphs you made, what conclusion can you make about microbiome differences between genders? |
| --- |
|  |

| 3D-5. Which seems to have more effect on microbial compositions? Gender or individual subject? |
| --- |
|  |

**A suggestion for your poster:** - Combine the 2 plots that cover the entire dataset (3A-1 and 3B-1) and 2 plots from exploring microbial groups (3D) to make one figure summarizing your findings for gender based differences in microbial composition. - Include titles and a short figure description (figure legend). - Paste the code you used to generate your final figure(s) below: | | |:–| | |

#### 5.3.4.4 Question 4 - Does age contribute to individual microbiome variation?

Approach: Survey the differences in microbiome across age, refining the plot to accommodate a continuous variable (age) and unequal sample sizes (some individuals in the study are the same age). Then, test if there is a shift in microbiome abundance of phyla Firmicutes and Bacteroidetes, two phyla suggested previously to change with age.

**Step 4A. Perform an initial survey to see if there are any differences in phyla related to age.** Specifically, your plot should show the relationship phyla diversity and age of an individual. - Refer to the “Explore 16S rRNA Data with phyloseq” tutorial for help using the plot\_bar() function. - Use the following code as a template plot\_bar(miso, “fill in the blank”, fill = “fill in the blank”, title = “fill in the blank”) + geom\_bar(aes(color = fill in the blank, fill = fill in the blank), stat = “identity”, position = “stack”)

plot\_bar(miso, "fill in the blank", fill = "fill in the blank", title = "fill in the blank") +   
geom\_bar(aes(color = fill in the blank, fill = fill in the blank), stat = "identity", position = "stack")

| 4A-1. Insert the resulting plot below: |
| --- |
|  |

* You probably found it difficult to discern any differences.
* Abundance is additive, so having more samples of the same age will result in adding the abundances together, so if there are say 2 samples with age 27 (as is the case), the abundance of the bar will be much higher than other ages that have only one individual.

**Step 4B. Merge samples by age, then re-normalize and finally re-plot Phyla by age.**

1. Merge samples of the same age, saving the resulting phyloseq object as ‘merge’. You will probably get a warning message letting you know “NAs were introduced by coercion.” That is OK.

merge <- merge\_samples(miso, "age")

1. Check your merged phyloseq object. Your new ‘merge’ object should have 16 samples corresponding to 16 unique ages in the MISO dataset.

merge  
  
phyloseq-class experiment-level object  
otu\_table() OTU Table: [ 1702 taxa and 16 samples ]  
sample\_data() Sample Data: [ 16 samples by 32 sample variables ]  
tax\_table() Taxonomy Table: [ 1702 taxa by 9 taxonomic ranks ]

1. Normalize the sample counts using the function transform\_sample\_counts. Our function, (100\* x/sum(x)) will transform all the samples into proportions.

merge <- transform\_sample\_counts(merge, function(x) 100 \* x/sum(x))

1. Re-plot Phyla by age.

plot\_bar(merge, x= "age", fill = "fill in the blank", title = "fill in the blank") +  
 geom\_bar(aes(color = fill in the blank, fill = fill in the blank), stat = "identity", position = "stack")

1. If you need to decrease the size of your legend key and text to fit, add code below to your plot after adding a + to the end of the geom\_bar line:

theme(legend.text=element\_text(size=6)) +   
theme(legend.key.size = unit(6, "pt"))

| 4B-1. Insert the resulting plot below: |
| --- |
|  |

| 4B-2. Do you observe any candidate Phyla that changes with age? |
| --- |
|  |

**Step 4C. Candidate Phyla implicated in age variation include Firmicutes and Bacteroidetes. Plot these specific phyla on their own, separately, to observe any shift that occurs with age.** Use your new merged and normalized ‘merge’ phyloseq object as input. - Refer to the “Subset taxonomy” section of the “Explore 16S rRNA Data with phyloseq” tutorial for help using the subset\_taxa and plot\_bar() functions. - Use the following code as a template:

subset = subset\_taxa(merge, Phylum == "fill in the blank")  
plot\_bar(subset, "age", fill = "fill in the blank", title = "fill in the blank") +   
 geom\_bar(aes(color = fill in the blank, fill = fill in the blank), stat = "identity", position = "stack")

| 4C-1. Insert the plot that graphs the relationship of age to the Phylum Firmicutes below: |
| --- |
|  |

| 4C-2. Do you observe any changes in Phyla based on age? |
| --- |
|  |

| 4C-3. Insert the plot that graphs the relationship of age to the Phylum Bacteroidetes below: |
| --- |
|  |

| 4C-4 Do you observe any changes in Phyla based on age? |
| --- |
|  |

| Step 4C-5: Based on the graphs you made, did you identify any candidate Phylum that may differ between genders? |
| --- |
|  |

**A suggestion for your poster:** - Combine 2-4 of your plots to make one figure summarizing Phylum and Species age-based differences you observed in the MISO study. Include titles and short figure description (figure legend). - Paste the code you used to generate your final figure(s) below: | | |:–| | |

### 5.3.5 Grading criteria

* Download this assignment as Microsoft Word (.docx) and upload on Canvas

### 5.3.6 Footnotes

#### 5.3.6.1 Resources

[Google Doc](https://docs.google.com/document/d/1dTRjAHHXeuzXqBOl-h80MRvNxxpMKfT-F_YLIKaxzmo/edit?usp=sharing)

#### 5.3.6.2 Contributions and affiliations

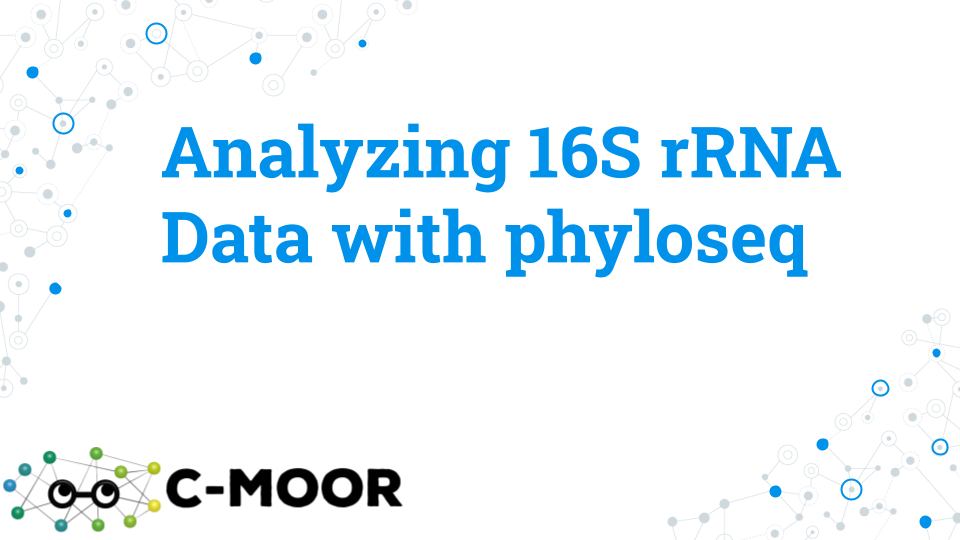
* Valeriya Gaysinskaya, Johns Hopkins University
* Gauri Paul, Clovis Community College
* Frederick Tan, Johns Hopkins University
* Sayumi York, Notre Dame of Maryland University

# 6 Analyzing 16S rRNA Data with phyloseq

## 6.1 Analyzing 16S rRNA Data with phyloseq

## 6.2 Lecture - Analyzing 16S rRNA Data with phyloseq

*Estimated time:*



[Lecture](https://docs.google.com/presentation/d/1befeRj4IAML556qaN9OQEb1gvqZo0lj6aOPlpgQkzvE/edit?usp=sharing)

## 6.3 Activity - Exploring 16S rRNA Data with phyloseq

To analyze bacterial diversity based on 16S rRNA gene sequencing using data from the “Impact of a 7-day homogeneous diet on interpersonal variation in human gut microbiomes and metabolomes” by Guthrie et al., 2023. This study is also referred to as the MISO study for “Microbiome Individuality and Stability Over Time” because the study aims to understand variation (or stability) of microbiomes across individuals.

This activity aims to analyze metagenomic diversity using the following R packages:

* **phyloseq**: A popular package for taxonomy profiling
* **DESeq2**: A package originally designed for differential expression which we will use for differential abundance
* **ggplot2**: A tidyverse package that produces high-quality figures

### 6.3.1 Learning objectives

1. Use phyloseq to cluster samples based on the similarity of their microbial compositions using multidimensional scaling methods (NMDS or PCoA)
2. Use phyloseq to profile alpha diversity (Shannon and Simpson’s indices)
3. Use DESeq2 to perform differential abundance analysis

### 6.3.2 Introduction

16S data can be manipulated and visualized in a variety of ways. In Google Sheets, we explored the data manually to gain an understanding at the level of 1-10 ASVs. Through phyloseq we collapsed ASVs with the same taxonomic annotation to survey microbial diversity and perform the same exploration across the entire dataset. In this activity, we will use the functions of phyloseq and DESeq2 that are more advanced. These analyses would be extremely difficult and tedious to do manually and will allow us deeper insights into our dataset. We will cluster samples based on similarity of ASV counts, survey alpha diversity, and search for differentially abundant ASVs between different groups.

### 6.3.3 Activity 1 – Analyze 16S rRNA Data with phyloseq tutorial

*Estimated time: 35 min*

**Activity 1. Explore a phyloseq object through the “Analyze 16S rRNA Data with phyloseq” tutorial on SciServer.**

1. Log into SciServer, click on compute, and create a new C-MOOR LearnR container. When creating a container, remember to:

* Use the “C-MOOR LearnR (Bioconductor 3.17)” image (not the Bioconductor 3.16)
* Check the box next to Data volume “C-MOOR Data”

1. Start the “Analyze 16S rRNA Data with phyloseq” tutorial. Visit SciServer Guides and FAQs. If you need assistance accessing the tutorial.
2. To move through the activities click “Continue” at the bottom of the screen. When you are done with a topic, click “Next Topic” to move on.
3. This tutorial has small boxes in which you can enter and run short lines of code to analyze the data.
4. As you work through the tutorial, take snapshots of your work and paste your answers in the grey boxes below:

**1-1. Take a snapshot and paste the code for an alpha diversity plot (Simpson) from the quiz question:** What male subject has the data point for the LOWEST alpha diversity? HINT: Use subset\_samples() to subset males, and specify individuals (subject) on x-axis of the alpha diversity plot || |:–| | |

**1-2. Take a snapshot and paste the code for a PCoA plot from the quiz question:** In a PCoA with only data from ASVs with the class Bacteroidia, what is the percent of variance in the dataset explained by principal coordinate 1? HINT: You will need to change the code subsetting phylum and Firmicutes - Subsets the Bacteroidetes phylum - Color is by subject - Shape is by timepointgroup || |:–| | |

**1-3. Take a snapshot and paste the code for the differential abundance plot for the quiz question:** Which of the following Phylum have ASVs that are differentially abundant between the subject S02 and subject S03? || |:–| | |

### 6.3.4 Activity 2 – Try it out!

*Estimated time: 60 min*

With your group, perform some exploratory data analysis selecting from one of the four questions below or coming up with your own question. When you are done, copy and paste your results into the [class poster](https://docs.google.com/presentation/d/1eE5AltbkW6km0r3uQO8KKnUVRLCLx70SDNuBHp6aASU/edit?slide=id.g343cb6dd56c_2_18#slide=id.g343cb6dd56c_2_18)

#### 6.3.4.1 Question 1 - How sensitive is microbial diversity to variables like diet, age and gender?

Approach: Plot Simpson alpha diversity using plot\_richness() command in phyloseq and assess the impact of different study variables on changes in microbial diversity. Identify the variables that impact alpha diversity. Visible shifts in alpha diversity measure suggests a shift in microbial diversity, and higher alpha diversity value indicates an increase in alpha diversity.

**Note, when plotting this data you may get a warning below. That is ok!**

Warning: The data you have provided does not have any singletons. This is highly suspicious. Results of richness estimates (for example) are probably unreliable, or wrong, if you have already trimmed low-abundance taxa from the data.We recommended that you find the un-trimmed data and retry.

Refer to the “alpha diversity” section of the “Analyze 16S rRNA Data with phyloseq” tutorial for help using the plot\_richness() function.

**Step 1A. Plot alpha diversity of the full MISO dataset by subject (individuals).**

* subject is on the x-axis
* Color is by subject
* Use the following code as a template:

plot\_richness(miso\_counts, x="fill in the blank",   
 color="subject",   
 measures= "Simpson")

| 1-A. Paste the resulting plot below: |
| --- |
|  |

**1A-2. Based on the alpha plot above, is individuality a strong determinant of microbial alpha diversity (richness and evenness) within an individual? Explain.** How close are the 5 points for an individual? Are individual data points closer to each other or to the 5 points of another individual? Give examples of individuals in your explanation (e.g. compare S26 and S30, or S02 and S09). | | |:–| | |

Some individual alpha diversity profiles appear to have “outliers”, such that e.g. one of the sample points appears separate from the rest. E.g., individuals S03, S10 and S31 seem to have 1 outlier sample. **Speculate in general on what such singleton “outliers” could mean including recalling what the 5 points represent, evoking concepts of replicates, thinking about sampling and real biology.** | | |:–| | |

**Step 1B. Plot alpha diversity of the full MISO dataset, by timepointgroup.** - timepoint group is on the x-axis - Color is by subject - Use the following code as a template:

plot\_richness(miso\_counts, x="fill in the blank",   
 color="fill in the blank",   
 measures= "Simpson")

| 1B-1. Paste your plot below: |
| --- |
|  |

| 1B-2. Based on the alpha plot above, does diet have an impact on microbial diversity (richness and evenness) of individuals? Explain - do you observe a shift up or down in data distribution? |
| --- |
|  |

**Step 1C. Plot alpha diversity of the full MISO dataset by age.** - age is on the x-axis - Color is by subject - Use the following code as a template:

plot\_richness(miso\_counts, x="fill in the blank",   
 color="fill in the blank",   
 measures= "Simpson")

| 1C-1. Paste your plot below: |
| --- |
|  |

| 1C-2. Based on the alpha plot above, does age have an impact on microbial diversity of individuals? Explain - do you observe a shift up or down in data during ageing? Do you observe a tighter data distribution in younger or older individuals? Which group, younger (< 50 yo) or older (>50 yo) has higher and tiger overall distribution? |
| --- |
|  |

**Step 1D. Plot alpha diversity of the full MISO dataset by gender.**

* gender is on the x-axis
* Color is by subject
* Use the following code as a template:

plot\_richness(miso\_counts, x="fill in the blank",   
 color="fill in the blank",   
 measures= "Simpson")

| 1D-1. Paste your plot below: |
| --- |
|  |

| 1D-2. Based on the alpha plot above, does gender have an impact on microbial diversity? Explain. Which group has a tighter distribution? Which group overall has a more diverse microbiome, males or females? Which microbiomes are more similar to each other, male or female microbiomes? |
| --- |
|  |

**Step 2. Plot alpha diversity of the full MISO dataset, grouping based on levels for 5 different metabolites: Creatinine, PCS, IS, HIPP, PAG.** Plot alpha diversity for the 5 metabolites independently then choose your favorite metabolite (e.g. one with most difference) and show plot below.

* metabolite on the x-axis (one at a time: Creatinine, PCS, IS, HIPP, PAG)
* Color is by subject
* Use the following code as a template:

plot\_richness(miso\_counts, x="fill in the blank metabolite name",   
 color="fill in the blank",   
 measures= "Simpson")

| 2A-1. Paste your metabolite plot(s) below: |
| --- |
|  |

| 2A-2. Which metabolite did you choose to show and why? Any difference between microbiomes for individuals with low versus high metabolite levels? |
| --- |
|  |

**Step 3. Select data for poster and include the code used to generate the figure.** 1. Prepare your final figure and associated code for the poster. 1. Figure suggestion: A 2-part figure with Figure A showing alpha diversity based on your favorite variable (subject, timepointgroup, age, or gender), and Figure B, showing alpha diversity with your favorite metabolite (Creatinine, PCS, IS, HIPP, or PAG). - Ensure you have a figure legend that explains your final figure. - Ensure you submit the code you used to generate your final figure.

| 3A-1. Paste final poster figure and code below: |
| --- |
|  |

#### 6.3.4.2 - Question 2 - Do diet, age, gender and levels of metabolites correlate with microbe variation between individuals?

Approach: Perform multidimensional scaling (also known as principal component analysis) to establish a relationship between the samples given multivariate data (metadata variables). Using a PcOA plot (via commands ordinate() and plot\_ordination() in phyloseq), you will condense the original high-dimensional data into a low-dimensional one by converting data to distance map (matrix) with 2 dimensions, x and y, that best explain variability in your data. From your PCoA plot you will assess the contribution of different study variables to sample diversity and identify variables that help explain sample diversity. In a PcOA plot, samples with similar microbial profiles will be plotted close and may appear as “clusters”.

**Step 1. Make a PcOA plot, ordinating on the entire miso dataset, and coloring by individual. Investigate resulting plot shape.** - Ordination is a term used to summarize a multidimensional dataset when projected onto a low-dimensional space (like X & Y axes) and then observing any pattern the data may possess with a visual inspection. - Subsequent coloring of the pattern with metadata variables can reveal underlying relationships between data and experiment variables.

* color - by subject
* Include a title for your PCoA plot
* Use the following code as a template:

#Ordinate   
miso.pcoa <- ordinate(miso, method="PCoA", distance="bray")   
  
#Plot PCoA  
plot\_ordination(miso, miso.pcoa,   
 color = "fill in the blank",   
 title="fill in the blank")

| 1A-1. Paste your plot below: |
| --- |
|  |

| 1A-2. Describe the shape of your PCoA plot. Does the PCoA organize the into groups? Is there any pattern? |
| --- |
|  |

| 1A-3. Do you observe sample clustering by individual? Provide examples and an explanation. |
| --- |
|  |

| 1A-4. Offer one question you want to ask about this PCoA plot? |
| --- |
|  |

**Step 2. Using the PCoA plot from Step 1 above, look for any correlations between the shape of your PCoA plot and diet, gender and age.**

Step 2A. Using the PCoA plot from Step 1, look for any correlations between the shape of your PCoA plot and diet? How well is your data explained by the variation in the diet - look for signs of correlation between PCoA shape and variable “timepointgroup” representing BD, HD and WO diets. Note that since we are using the same plot, we do not need to re-ordinate (no ordinate() function).

* Color - by timepointgroup
* Include a title for your PcOA plot
* Use the following code as a template:

plot\_ordination(miso, miso.pcoa,   
 color = "fill in the blank",   
 title="fill in the blank")

| 2A-1. Paste your plot below: |
| --- |
|  |

| 2A-2 Did coloring the ‘miso’ PCoA plot by timepointgroup explain any variation in your data? E.g. did each of the 3 data groups/clusters correspond to each of the 3 timepointgroups or not? E.g. does the coloring appear randomly distributed or not? |
| --- |
|  |

**Step 2B. Using the PCoA plot from Step 1, look for any correlations between the shape of your PCoA plot and gender?**  How well is your data explained by the variation in the gender - look for signs of correlation between PcOA shape and variable “gender”.

* Color - by gender
* Include a title for your PcOA plot
* Use the following code as a template.

plot\_ordination(miso, miso.pcoa,   
 color = "fill in the blank",   
 title="fill in the blank")

| 2B-1. Paste your plot below: |
| --- |
|  |

| 2B-2 Did coloring of the ‘miso’ PCoA plot by gender explain any variation in your data? E.g. does the coloring appear randomly distributed or not? |
| --- |
|  |

Step 2C. Using the PCoA plot from Step 1, look for any correlations between the shape of your PCoA plot and age? How well is your data explained by the variation in the age - look for signs of correlation between PcOA shape and variable “age”.

* Color - age
* Include title name for your PcOA plot
* Use the following code as a template. To enhance the color range of the age variable, additionally included below (in blue) is the command to specify the color gradient with yellow–to-blue gradient.
* Try running the code with and without the scale\_colour\_gradient() command to appreciate the difference.

plot\_ordination(miso, miso.pcoa,   
 color = "fill in the blank",   
 title="fill in the blank") +   
scale\_colour\_gradient(low = "red", high = "green")

| 2C-1. Paste your plot below: |
| --- |
|  |

| 2C-2 Did coloring of the ‘miso’ PCoA plot by age explain any variation in your data? E.g. Does the coloring appear randomly distributed or not? Is there any hint of a correlation between age and groups of data on the left or the right? |
| --- |
|  |

Step 3. Using the PCoA plot from Step 1, look for any correlations between the shape of your PCoA plot and the level of metabolites. How well is your data correlated with the variation in the 5 metabolites - Creatinine, PCS, IS, HIPP, or PAG? Test each metabolite by coloring each metabolite at a time, then, choose your favorite metabolite (e.g. one with most difference) and show the plot below.

* Color - metabolite one at a time: Creatinine, PCS, IS, HIPP, PAG
* Include a title for your PcOA plot
* Use the following code as a template:

plot\_ordination(miso, miso.pcoa fill in the blank metabolite name,   
 color = "fill in the blank",   
 title="fill in the blank") +   
scale\_colour\_gradient(low = "yellow", high = "blue")

| 3A-1. Paste your metabolite plot(s) below: |
| --- |
|  |

| 3A-2. Which metabolite did you choose to show and why? Any difference between microbiomes for individuals with low versus high metabolite levels? Any association between metabolite level and microbiome diversity? |
| --- |
|  |

Step 4. One way of checking if the metabolite levels indeed correlate with your data, is to see if subsetting smaller chunks or specific chunks of the data will still maintain the metabolite-data relationship or break it. Using subset() command, subset out e.g. HD diet specifically, and then BD diet. Then, generate new PcOA plots and see if the correlation with your metabolite of interest still holds.

Step 4A. Subset “HD”, ordinate on “HD”, and make a new PCoA plot.

* Subset data - by the timepointgroup “HD”
* Ordinate - misoHD
* Color - by your selected metabolite from part 3
* Include a title for your PcOA plot
* Use the following code as a template:

1. Subset only “HD” timepoint from your miso data, creating a new phyloseq object called “misoHD”

misoHD = subset\_samples(miso, timepointgroup == "fill in the blank")

1. Ordinate using only “HD” subset, creating a new ordination matrix called “pcoa.misoHD”

pcoa.misoHD <- ordinate(misoHD, method="PCoA", distance="bray")

1. Make a PCoA plot using your “HD” subset.

plot\_ordination(misoHD, pcoa.misoHD,   
 color = "fill in the blank",   
 title="fill in the blank")

| 4A-1. Paste your plot below for HD subset: |
| --- |
|  |

| 4A-2 Did subsetting by the timpointgroup “HD” maintain your data’s relationship to your chosen metabolite? Did the shape of the plot change or stay the same with “HD” subsetting? Can you conclude if your metabolite of choice correlates with HD diet? |
| --- |
|  |

Step 4B. Repeat the analysis on only the BD timepoint group samples. Subset “BD”, ordinate on “BD”, and make a new PCoA plot.

* Subset data - by the timepointgroup “BD”
* Ordinate - misoBD
* Color - by your selected metabolite from part 3
* Include a title for your PcOA plot
* Use the code above as a template.

| 4A-1. Paste your plot below for BD subset: |
| --- |
|  |

| Did subsetting “BD” maintain your data’s relationship to your chosen metabolitelevels? Did the shape of the plot change or stay the same with “BD” subsetting? Explain your results. Can you conclude if your metabolite of choice correlates with sample groups when only BD timepointgroup data is plotted? |
| --- |
|  |

Step 5. Select data for poster and include the code used to generate the figure. - Prepare your final figure and associated code for the poster. - Figure suggestion: A 2-part figure with Figure A showing PCoA plot based on subject and Figure B, showing PCoA plot with your favorite metabolite (Creatinine, PCS, IS, HIPP, or PAG). - Ensure you have a figure legend that explains your final figure. - Ensure you submit the code you used to generate your final figure.

| 5A-1. Paste final poster figure and code below: |
| --- |
|  |

#### 6.3.4.3 Question 3 - What microbes (ASVs) differ between males and females, and does age have an impact?

Approach: Perform DESeq analysis between females and males and determine how many differentially abundant microbes are there between the sexes. Then examine if age has a further impact on the abundance of the microbes between the sexes.

**Step 3A. Perform differential abundance analysis between genders.**

* Design - is based on gender (no double quotes)
* Groups to compare - females, F and males, M (baseline)
* Plot Phylum on the X-axis and color by Class
* Make sure your graph has a title that includes the comparison you are making (ex. WO vs HD)
* Refer to the “Differential abundance” section of the “Analyze 16S rRNA Data with phyloseq” tutorial for help using the plot\_ordination() function.
* Use the code below as a template:

STEP 1: Convert the phyloseq object to a DESeq2 object and specify experimental design

DESeq2 <- phyloseq\_to\_deseq2(miso\_counts, design = ~ fill in the blank)

STEP 2: Select the groups to compare, where the latter group is your baseline

my\_comparision <-c("gender", "fill in the blank", "fill in the blank (baseline)")

STEP 3: Run the differential abundance analysis

Significant\_DEseq2\_ASVs<-Differential\_Abundance(DESeq2, my\_comparision, 0.05)

STEP 4: Retrieve the list of ASVs with a significant difference in abundance between the chosen groups

Significant\_DEseq2\_ASVs

STEP 5: Plot the results with your chosen x axis and legend

ggplot(significant\_ASVs, aes(x = fill in the blank, y=log2FoldChange, color= fill in the blank)) + geom\_point(size=4, position="jitter") +  
 theme(axis.text.x = element\_text(angle = -90, hjust = 0, vjust=0.5))+  
 ggtitle("fill in the blank")

| 3A-1. Insert the resulting plot below: |
| --- |
|  |

| 3A-2. How many differentially abundant ASVs were identified? |
| --- |
|  |

| 3A-3. Give an example of a differentially abundant ASVs that was significantly higher in females than males. Include ASV ID and taxa. HINT: If male was your baseline, then an example of an ASV that is higher in females will have a positive log2FC. |
| --- |
|  |

| 3A-4. Give an example of a differentially abundant ASVs that was significantly higher in males than females. Include ASV ID and taxa. HINT: If male was your baseline, then an example of an ASV that is higher in males will have a negative log2FC. |
| --- |
|  |

**Step 3B. Determine the impact of younger age on differential abundance in males and females, by subsetting age < 50 data from the phyloseq object, and performing DESeq2 analysis on this subset.**

* Subset phyloseq object to only include younger individuals (age < 50)
* For DESeq2 analysis, keep the design the same as in part A - based on gender
* Groups to compare: females (F) and males (M, baseline)
* Plot Phylum on the X-axis and color by Class
* Use template code below:

STEP 1: subset by age < 50

ageA = subset\_samples(miso\_counts, age <50)

STEP 2: Convert the phyloseq object to a DESeq2 object and specify experimental design

DESeq2 <- phyloseq\_to\_deseq2(ageA, design = ~ fill in the blank)

STEP 3: Select the groups to compare, where log2FoldChange reported will correspond to y/x.

my\_comparision <-c("gender", "fill in the blank", "fill in the blank (baseline)")

STEP 4: Run the differential abundance analysis

significant\_ASVs<-Differential\_Abundance(DESeq2, my\_comparision, 0.05)

STEP 5: Retrieve the list of ASVs with a significant difference in abundance between the chosen groups

significant\_DEseq2\_ASVs

STEP 6: Plot the results with your chosen x axis and legend

ggplot(significant\_ASVs, aes(x = fill in the blank, y=log2FoldChange, color= fill in the blank)) + geom\_point(size=4, position = "jitter") +  
 theme(axis.text.x = element\_text(angle = -90, hjust = 0, vjust=0.5))+  
 ggtitle("fill in the blank")

| 3B-1. Insert the resulting plot below: |
| --- |
|  |

| 3B-2. How many differentially abundant ASVs were identified? |
| --- |
|  |

| 3B-3. Give an example of a differentially abundant ASVs that was significantly higher in younger females than younger males. Include ASV ID and taxa. HINT: If male was your baseline, then an example of an ASV that is higher in females will have a positive log2FC. |
| --- |
|  |

| 3B-4. Give an example of a differentially abundant ASVs that was significantly higher in younger males than younger females. Include ASV ID and taxa. HINT: If male was your baseline, then an example of an ASV that is higher in males will have a negative log2FC. |
| --- |
|  |

| 3B-5. How does younger age impact microbe abundance between females and males in general? |
| --- |
|  |

**Step 3C. Determine the impact of older age on differential abundance in males and females, by subsetting age >= 50 data from the phyloseq object, and performing DESeq2 analysis on this subset.**

| 3C-1. Insert the resulting plot below: |
| --- |
|  |

| 3C-2. How many differentially abundant ASVs were identified? |
| --- |
|  |

| 3C-3. How does older age impact microbe abundance between females and males in general? |
| --- |
|  |

| 3C-4. Comparing fold change plots between younger and older females and males, what is one difference between Phyla distribution? |
| --- |
|  |

**Step 4. Select data for poster and include the code used to generate the figure.**

Prepare your final figure and associated code for the poster. - Figure suggestion: A 3-part figure with Figure A showing fold change differences between males and females in general, Figure B showing gender differences in younger individuals, and Figure C showing gender differences in older individuals.

* Ensure you have a figure legend that explains your final figure.
* Ensure you submit the code you used to generate your final figure.

| 4A-1. Paste final poster figure and code below: |
| --- |
|  |

#### 6.3.4.4 Question 4 - Is there diet and age interaction and what microbes (ASVs) correlate with changes in diet-age interaction?

Approach: Use alpha diversity measure and DESeq2 tools to answer this question. Using alpha diversity, determine if there is an interaction between diet and age. Then use DESeq2 to see if any ASVs are associated with changes in die-age interaction. Using Simpson alpha diversity measure (or specifically, Gini-Simpson) evaluate how microbial diversity changes with age and diet in general, or age and BD, HD and WO diet specifically. Remember, shifts in alpha diversity measure suggest a shift in microbial diversity, with high alpha diversity suggesting high microbial diversity. Then, use DESeq2 to evaluate if ageing changes ASVs associated with diet.

Note, when plotting this data you may get a warning below. That is ok!

Warning: The data you have provided does not have any singletons. This is highly suspicious. Results of richness estimates (for example) are probably unreliable, or wrong, if you have already trimmed low-abundance taxa from the data.We recommended that you find the un-trimmed data and retry.

Refer to the “alpha diversity” section of the “Analyze 16S rRNA Data with phyloseq” tutorial for help using the plot\_richness() function.

**Step 1. Plot alpha diversity based on age for all dietary groups, and then for BD, HD and WO.**

1A-1. Plot alpha diversity based on age for all dietary groups.

* Subset data - None
* age is on the x-axis
* Color is by subject
* Use the following code as a template:

plot\_richness(miso\_counts, x="age",  
 color="subject",  
 title = "miso",  
 measures= c("Simpson"))

| Paste your plot below: |
| --- |
|  |

| 1A-2. Based on the alpha plot above, is age a good determinant of microbial alpha diversity? Explain. Is there an age-based trend in alpha diversity? |
| --- |
|  |

1B-1. Plot alpha diversity based on age for BD diet only.

* Subset data - timepointgroup BD
* age is on the x-axis
* Color is by subject
* Use the following code as a template:

# Subset only BD samples  
miso\_counts\_BD = subset\_samples(miso\_counts, timepointgroup == "BD")  
  
plot\_richness(miso\_counts\_BD, x="fill in the blank",  
 color="fill in the blank",  
 title = "fill in the blank",  
 measures= c("Simpson"))

| Paste your plot below: |
| --- |
|  |

| 1B-2. Based on the alpha plot above, how did subsetting for BD alter the age-based alpha diversity plot? Did the trend in age-based alpha diversity change? Explain if this makes sense. |
| --- |
|  |

1C-1. Plot alpha diversity based on age for HD diet only.

* Subset data - timepointgroup HD
* age is on the x-axis
* Color is by subject
* Use the following code as a template:

# Subset only HD samples  
miso\_counts\_BD = subset\_samples(miso\_counts, timepointgroup == "HD")  
  
plot\_richness(miso\_counts\_BD, x="fill in the blank",  
 color="fill in the blank",  
 title = "fill in the blank",  
 measures= c("Simpson"))

| Paste your plot below: |
| --- |
|  |

| 1C-2. Based on the alpha plot above, how did subsetting for HD alter the age-based alpha diversity plot? Did the trend in age-based alpha diversity change? Explain if this makes sense. |
| --- |
|  |

1D-1. Plot alpha diversity based on age for WO diet only.

* Subset data - timepointgroup WO
* age is on the x-axis
* Color is by subject
* Use the following code as a template:

# Subset only WO samples  
miso\_counts\_BD = subset\_samples(miso\_counts, timepointgroup == "WO")  
  
plot\_richness(miso\_counts\_BD, x="fill in the blank",  
 color="fill in the blank",  
 title = "fill in the blank",  
 measures= c("Simpson"))

| Paste your plot below: |
| --- |
|  |

| 1D-2. Based on the alpha plot above, how did subsetting for WO alter the age-based alpha diversity plot? Did the trend in age-based alpha diversity change? Explain if this makes sense. Remember, WO has twice as few datapoints as BD and HD since there is only 1 WO timepoint (versus 2 for BD and HD). |
| --- |
|  |

**Step 2. Perform differential abundance analysis comparing HD to BD for the population in general, and then for older individuals specifically (because older individuals produced the biggest diet–age interaction based on alpha diversity).**

Step 2A. Perform differential abundance analysis between HD and BD. - Design - is based on timepointgroup (no double quotes) - Groups to compare - HD and BD (baseline) - Plot Phylum on the X-axis and color by Class

* Make sure your graph has a title that includes the comparison you are making (ex. WO vs HD)
* Refer to the “Differential abundance” section of the “Analyze 16S rRNA Data with phyloseq” tutorial for help using the plot\_ordination() function.
* Use the code below as a template:

STEP 1: Convert the phyloseq object to a DESeq2 object and specify experimental design

DESeq2 <- phyloseq\_to\_deseq2(miso\_counts, design = ~ timepointgroup)

STEP 2: Select the groups to compare, where the latter group is your baseline

comparision <-c("timepointgroup", "HD", "fill in the blank (baseline)")

STEP 3: Run the differential abundance analysis

Significant\_DEseq2\_ASVs<-Differential\_Abundance(DESeq2, comparision, 0.05)

STEP 4: Retrieve the list of ASVs with a significant difference in abundance between the chosen groups

Significant\_DEseq2\_ASVs

STEP 5: Plot the results with your chosen x axis and legend

ggplot(significant\_ASVs, aes(x = Phylum, y=log2FoldChange, color= fill in the blank)) + geom\_point(size=4) +  
 theme(axis.text.x = element\_text(angle = -90, hjust = 0, vjust=0.5))+  
 ggtitle("fill in the blank")

| 2A-1. Insert the resulting plot below: |
| --- |
|  |

| 2A-2. How many differentially abundant ASVs were identified? |
| --- |
|  |

| 2A-3. Discuss the meaning for your differential abundance plot above. Did you observe few or many differences between BD and HD for the general population (all individuals)? What does that indicate in terms of diet-induced microbiome changes? |
| --- |
|  |

**Step 2B. Perform differential abundance analysis between HD and BD for older individual (> 50 yo.)**

* Subset data - older age (> 50 yo)
* Design - is based on timepointgroup (no double quotes)
* Groups to compare - HD and BD (baseline)
* Plot Phylum on the X-axis and color by Class
* Make sure your graph has a title that includes the comparison you are making (ex. WO vs HD)
* Refer to the “Differential abundance” section of the “Analyze 16S rRNA Data with phyloseq” tutorial for help using the plot\_ordination() function.
* Use the code below as a template:

STEP 1: Subset miso\_counts object to only include older individuals over age 50

subset = subset\_samples(miso\_counts, age >50)

STEP 2: Convert the phyloseq object to a DESeq2 object and specify experimental design

DESeq2 <- phyloseq\_to\_deseq2(subset, design = ~ timepointgroup)

STEP 3: Select the groups to compare, where the latter group is your baseline

comparision\_subset <-c("timepointgroup", "HD", "fill in the blank (baseline)")

STEP 4: Run the differential abundance analysis

Significant\_ASVs\_subset <- Differential\_Abundance(DESeq2, comparision\_subset, 0.05)

STEP 5: Retrieve the list of ASVs with a significant difference in abundance between the chosen groups

Significant\_ASVs\_subset

STEP 6: Plot the results with your chosen x axis and legend

ggplot(significant\_ASVs\_subset, aes(x = Phylum, y=log2FoldChange, color= fill in the blank)) + geom\_point(size=4) +  
 theme(axis.text.x = element\_text(angle = -90, hjust = 0, vjust=0.5))+  
 ggtitle("fill in the blank")

| 2B-1. Insert the resulting plot below: |
| --- |
|  |

| 2B-2. How many differentially abundant ASVs were identified? |
| --- |
|  |

| 2B-3. Discuss the meaning for your differential abundance plot above. Did you observe any differences between BD and HD for the older individuals? How does that compare to general population? Note, by comparison, when subsetting the younger population (<= 50yo), the number of differential ASVs produced is 3. What does that indicate in terms of diet-age interaction? |
| --- |
|  |

**Step 3. Select data for poster and include the code used to generate the figure.**

* Prepare your final figure and associated code for the poster.
* Figure suggestion: A 4-part figure with top panel: Figure A showing alpha diversity in general (includes all 3 dietary interventions), and Figure B showing alpha diversity for HD diet only. The bottom panel: Figure C showing differentially abundant ASVs between HD and BD for all ages, and Figure D showing differentially abundant ASVs for older individuals.
* Ensure you have a figure legend that explains your final figure.
* Ensure you submit the code you used to generate your final figure.

| Paste final poster figure and code below: |
| --- |
|  |

### 6.3.5 Grading criteria

* Download this assignment as Microsoft Word (.docx) and upload on Canvas

### 6.3.6 Footnotes

#### 6.3.6.1 Resources

[Google Doc](https://docs.google.com/document/d/1fiMHAB-BtHwYBKQ_pnjZrehIK7cooaVS_UT9bDNomrw/edit?usp=sharing)

#### 6.3.6.2 Contributions and affiliations

* Valeriya Gaysinskaya, Johns Hopkins University
* Gauri Paul, Clovis Community College
* Frederick Tan, Johns Hopkins University
* Sayumi York, Notre Dame of Maryland University

# About the Authors

These credits are based on our [course contributors table guidelines](https://www.ottrproject.org/more_features.html#giving-credits-to-contributors).

| Credits | Names |
| --- | --- |
| **Pedagogy** |  |
| Lead Content Instructor(s) | [FirstName LastName](link%20to%20personal%20website) |
| Lecturer(s) (include chapter name/link in parentheses if only for specific chapters) - make new line if more than one chapter involved | Delivered the course in some way - video or audio |
| Content Author(s) (include chapter name/link in parentheses if only for specific chapters) - make new line if more than one chapter involved | If any other authors besides lead instructor |
| Content Contributor(s) (include section name/link in parentheses) - make new line if more than one section involved | Wrote less than a chapter |
| Content Editor(s)/Reviewer(s) | Checked your content |
| Content Director(s) | Helped guide the content direction |
| Content Consultants (include chapter name/link in parentheses or word “General”) - make new line if more than one chapter involved | Gave high level advice on content |
| Acknowledgments | Gave small assistance to content but not to the level of consulting |
| **Production** |  |
| Content Publisher(s) | Helped with publishing platform |
| Content Publishing Reviewer(s) | Reviewed overall content and aesthetics on publishing platform |
| **Technical** |  |
| Course Publishing Engineer(s) | Helped with the code for the technical aspects related to the specific course generation |
| Template Publishing Engineers | [Candace Savonen](https://www.cansavvy.com/), [Carrie Wright](https://carriewright11.github.io/), [Ava Hoffman](https://www.avahoffman.com/) |
| Publishing Maintenance Engineer | [Candace Savonen](https://www.cansavvy.com/) |
| Technical Publishing Stylists | [Carrie Wright](https://carriewright11.github.io/), [Ava Hoffman](https://www.avahoffman.com/), [Candace Savonen](https://www.cansavvy.com/) |
| Package Developers ([ottrpal](https://github.com/jhudsl/ottrpal)) [Candace Savonen](https://www.cansavvy.com/), [John Muschelli](https://johnmuschelli.com/), [Carrie Wright](https://carriewright11.github.io/) |  |
| **Art and Design** |  |
| Illustrator(s) | Created graphics for the course |
| Figure Artist(s) | Created figures/plots for course |
| Videographer(s) | Filmed videos |
| Videography Editor(s) | Edited film |
| Audiographer(s) | Recorded audio |
| Audiography Editor(s) | Edited audio recordings |
| **Funding** |  |
| Funder(s) | Institution/individual who funded course including grant number |
| Funding Staff | Staff members who help with funding |

## ─ Session info ───────────────────────────────────────────────────────────────  
## setting value  
## version R version 4.3.2 (2023-10-31)  
## os Ubuntu 22.04.4 LTS  
## system x86\_64, linux-gnu  
## ui X11  
## language (EN)  
## collate en\_US.UTF-8  
## ctype en\_US.UTF-8  
## tz Etc/UTC  
## date 2025-07-04  
## pandoc 3.1.1 @ /usr/local/bin/ (via rmarkdown)  
##   
## ─ Packages ───────────────────────────────────────────────────────────────────  
## package \* version date (UTC) lib source  
## bookdown 0.41 2024-10-16 [1] CRAN (R 4.3.2)  
## cachem 1.0.8 2023-05-01 [1] RSPM (R 4.3.0)  
## cli 3.6.2 2023-12-11 [1] RSPM (R 4.3.0)  
## devtools 2.4.5 2022-10-11 [1] RSPM (R 4.3.0)  
## digest 0.6.34 2024-01-11 [1] RSPM (R 4.3.0)  
## ellipsis 0.3.2 2021-04-29 [1] RSPM (R 4.3.0)  
## evaluate 0.23 2023-11-01 [1] RSPM (R 4.3.0)  
## fastmap 1.1.1 2023-02-24 [1] RSPM (R 4.3.0)  
## fs 1.6.3 2023-07-20 [1] RSPM (R 4.3.0)  
## glue 1.7.0 2024-01-09 [1] RSPM (R 4.3.0)  
## htmltools 0.5.7 2023-11-03 [1] RSPM (R 4.3.0)  
## htmlwidgets 1.6.4 2023-12-06 [1] RSPM (R 4.3.0)  
## httpuv 1.6.14 2024-01-26 [1] RSPM (R 4.3.0)  
## knitr 1.48 2024-07-07 [1] CRAN (R 4.3.2)  
## later 1.3.2 2023-12-06 [1] RSPM (R 4.3.0)  
## lifecycle 1.0.4 2023-11-07 [1] RSPM (R 4.3.0)  
## magrittr 2.0.3 2022-03-30 [1] RSPM (R 4.3.0)  
## memoise 2.0.1 2021-11-26 [1] RSPM (R 4.3.0)  
## mime 0.12 2021-09-28 [1] RSPM (R 4.3.0)  
## miniUI 0.1.1.1 2018-05-18 [1] RSPM (R 4.3.0)  
## pkgbuild 1.4.3 2023-12-10 [1] RSPM (R 4.3.0)  
## pkgload 1.3.4 2024-01-16 [1] RSPM (R 4.3.0)  
## profvis 0.3.8 2023-05-02 [1] RSPM (R 4.3.0)  
## promises 1.2.1 2023-08-10 [1] RSPM (R 4.3.0)  
## purrr 1.0.2 2023-08-10 [1] RSPM (R 4.3.0)  
## R6 2.5.1 2021-08-19 [1] RSPM (R 4.3.0)  
## Rcpp 1.0.12 2024-01-09 [1] RSPM (R 4.3.0)  
## remotes 2.4.2.1 2023-07-18 [1] RSPM (R 4.3.0)  
## rlang 1.1.4 2024-06-04 [1] CRAN (R 4.3.2)  
## rmarkdown 2.25 2023-09-18 [1] RSPM (R 4.3.0)  
## sessioninfo 1.2.2 2021-12-06 [1] RSPM (R 4.3.0)  
## shiny 1.8.0 2023-11-17 [1] RSPM (R 4.3.0)  
## stringi 1.8.3 2023-12-11 [1] RSPM (R 4.3.0)  
## stringr 1.5.1 2023-11-14 [1] RSPM (R 4.3.0)  
## urlchecker 1.0.1 2021-11-30 [1] RSPM (R 4.3.0)  
## usethis 2.2.3 2024-02-19 [1] RSPM (R 4.3.0)  
## vctrs 0.6.5 2023-12-01 [1] RSPM (R 4.3.0)  
## xfun 0.48 2024-10-03 [1] CRAN (R 4.3.2)  
## xtable 1.8-4 2019-04-21 [1] RSPM (R 4.3.0)  
## yaml 2.3.8 2023-12-11 [1] RSPM (R 4.3.0)  
##   
## [1] /usr/local/lib/R/site-library  
## [2] /usr/local/lib/R/library  
##   
## ──────────────────────────────────────────────────────────────────────────────

# References