miniCURE: Human Gut 16S rRNA

September 23, 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 C-MOOR Overview

### 1.0.1 About C-MOOR

C-MOOR is a project to invite students to join the data science revolution and be part of the next generation of data scientists. This project provides online materials to help students and instructors incorporate authentic research experiences in lower division courses.

Over the years, C-MOOR has expanded its curricula to include RNA-seq, scRNA-seq, 16S amplicon sequencing, and WGS modules. Here’s some statistics about C-MOOR students:

* **Students taught:** 700+.
* **90% of our students are first years:** We break the myth that only advanced-level students can perform and excel in real research.
* **Student research posters:** 100+. We choose a poster as our final deliverable to facilitate participation in research symposiums. Posters also allow us to evaluate the efficacy of our curricula over the years and make it easy for instructors to quickly and objectively grade student work. You can see a handful of posters [here](https://help.c-moor.org/c/look-at-this/8)!
* **50% + Students choose to present posters at optional symposiums:** C-MOOR posters have served as the impetus for establishing a new research symposium at Clovis Community College. Our students have attended and won awards at their institution’s research days and at conferences such as GRADS-4C and the ASM Maryland Branch Meeting.

### 1.0.2 Why data science?

Data Science is an evolving career path for individuals (data scientists) in fields that need to manipulate large amounts of data. The job of a Data Scientist was called by the Harvard Business Review “[the sexiest job of the 21st century](https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century)” Data Scientists are individuals with a curiosity to look through data, identify patterns and develop testable hypotheses. In the Biological Sciences, data scientists dig through data to answer questions about health, disease, evolution, ecology, drug development, and much more. A currently relevant example is compiling and looking at the similarities and differences between the different SARS-CoV-2 strains and identify differences between strains that may lead to differences in rates of infection.

### 1.0.3 Learning Goals

1. Engage with real and current scientific data
2. Explore available research resources online
3. Recognize the interdisciplinary nature of biological sciences
4. Synthesize findings from scientific literature
5. Summarize findings and discuss results with your peers
6. Collaborate with peers on a data exploration activity

# 2 For new instructors

Welcome to new and seasoned instructors alike! We are excited to have you use a C-MOOR module in your class. This section of our OTTR book is specifically for instructors and will help you prepare for teaching a C-MOOR offering as a part of your class.

### 2.0.1 Course-based undergraduate research experiences

A **course-based undergraduate research experience (CURE)** is a high impact practice (HIP) where students learn through authentic research. In addition to the standard concepts of the material and the technical skills required to complete lab exercises students gain critical thinking skills troubleshooting their research projects. They don’t just learn about science; they practice it!

### 2.0.2 How C-MOOR supports you

Our curriculum is designed to be accessible for anyone to teach and aims to remove many of the barriers that prevent instructors from trying computational biology methods.

* **No previous experience in bioinformatics required. We’ll help with the code:** We provide all code templates required to complete the analyses. You just bring your expertise as a researcher! For support with code and other technical issues reach out to us with a post on the [C-MOOR Academy Discussion Forum](https://help.c-moor.org/). Someone will assist you as soon as you are available.
* **Our curriculum comes with lectures, assignments, and rubrics for you to use**: We do as much as possible to take the workload of creating these materials off your back. But you’re not beholden to use them; if you have materials you’ve used over the years and prefer to use or want to use only parts of ours you’re free to do so.
* **We designed our miniCUREs to be modular to fit into your course**: Our miniCURE curricula are designed to be compact, with just 6 main sessions followed by project work. We’ve taught our RNA-seq miniCURE in pre-established classes, such as introductory biology. You don’t have to design your entire course around them and the modules fit easily into your lessons about gene expression and microbiology.

### 2.0.3 New instructor to-do list

If you haven’t taught a C-MOOR curriculum in the past, here’s a list of tasks you should do, ideally before the semester begins:

1. **Make an introductory post on the** [**C-MOOR Academy Discussion Forum**](https://help.c-moor.org/) in the general category. We recommend including:

* Your name
* Your institution
* The class you’re teaching that uses C-MOOR modules
* Your area of research expertise
* Any previous experience with bioinformatics
* The cloud platform you will be using to teach (see next item)

1. **Set up on a cloud platform**: C-MOOR uses either [SciServer](https://www.sciserver.org/) or [AnVIL](https://anvil.terra.bio/) for our cloud computing needs. If you’re not sure which is right for you, please mention this in your forum post so we can get you set up on whichever is more appropriate. Once you’ve chosen your cloud platform, follow the instructions in the respective section to access modules and learn how to grant your students the same access. While you can technically run the modules locally, running them on a cloud platform will prevent the need for students to download all the neccessary components and provides you easier access to updates.
2. **Go through the modules yourself**: Familiarize yourself the curriculum and how to use the cloud platform and our fill-in-the-blank code chunks. Look through our resources and see what support is currently available; for example, our [C-MOOR RNA-seq guide](https://docs.google.com/presentation/d/1ic09dhbt6WLEAgBTOqfT_tveknFF936BPVC2JhjutmY/edit?usp=sharing) and our student made [How To Get Started With Your C-MOOR Project Guide](https://docs.google.com/document/d/1WxlN2XG5_VEDN0u-JpAO7dJQrG8P0YG6Svm-eEvagxY/edit?usp=sharing).
3. **Look over our generic** [**C-MOOR Poster Rubric**](https://docs.google.com/document/d/1Aq4wVqO6S-PbKuV-rBiiRySVor_FLaRHp_CPuWBJPm4/edit?usp=sharing) and alter it as neccessary. The rubric is designed so that you can choose what sections you want included in your students’ posters and what weight (point values) you want to give to each element. Students will use the rubric as a checklist as they create their posters. Some instructors share this rubric at the start of their class. Others may leave it until later. If you have a specific symposium or conference in mind, tailor the rubric to suit that venue (ex. if your school’s research day requires posters with abstracts and no introductions, change the rubric to reflect this requirement).
4. **Consider your syllabus and where each module fits into your class schedule.** We provide estimated times for each module and the activities within as well as sample schedules from instructors who have taught the curriculum before. Some instructors teach a wet lab alongside the C-MOOR curriculum, using downtime (ex. waiting for the results of a gel electrophoresis lab) to do computational activities.
5. **Have students create an account on your chosen cloud platform and provide the neccessary information to you:** Do this well in advance of the first day you’re using the C-MOOR curriculum to avoid technical issues on the day of (ex. students without Google associated email addresses on AnVIL, students who have unknown errors logging into SciServer). You will need to collect the students’ email addresses (AnVIL) or their usernames (SciServer). We recommend using tools in your LMS or a Google form to expedite this process.

# 3 Human gut 16S rRNA

### 3.0.1 Content overview

### 3.0.2 Sample Schedules

### 3.0.3 What to know

### 3.0.4 Where to get help

### 3.0.5 Grades

### 3.0.6 Footnotes

**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

Last Revised: July 2025

# 4 AnVIL

## 4.1 About AnVIL

AnVIL (The Genomic Data Science **An**alysis, **V**isualization, and **I**nformatics **L**ab-space) is a platform created by the National Human Genome Research Institute (NHGRI) in collaboration with cloud computing platform providers like Google and Microsoft. Using AnVIL we you can access computing resources on the cloud through your browser without need for any fancy physical equipment. Through AnVIL you will also have access to all the software and data necessary to complete your research project.

In this section, we will set up our accounts on AnVIL and go through the entire lifecycle of an RStudio environment from creation to deletion. You will repeat this process throughout the semester; feel free to refer back to this section if you need a refresher on how to use AnVIL.

## 4.2 Sign up for an AnVIL account

#### 4.2.0.1 Purpose

You will need an account on AnVIL in order to use the platform. In this section we’ll go over the specifics of account creation.

#### 4.2.0.2 Learning Objectives

1. Create an account on AnVIL
2. Login to AnVIL
3. Share the email you used to sign up for AnVIL with your instructor (if applicable)

### 4.2.1 Create an AnVIL account

Follow the written steps below or refer to the [slides](https://docs.google.com/presentation/d/1uwlG7uaTOnItdpd4Ll6nNQiBJKBivsvR-erupicAwJM/edit?usp=sharing) or video guide.

1. Open [anvil.terra.bio](https://anvil.terra.bio/) in **Google Chrome** . Google Chrome is the only officially supported web browser for AnIVL. Because of this, while you can run AnVIL in other browsers you strongly suggest using Chrome.
   1. It is a good idea to bookmark this page so that you can easily access it throughout the course.
2. Click the hamburger icon (3 lines) in the top left corner of the screen
3. Click “Sign in”
4. Click “Sign in with Google”.
5. Sign in with a **Google associated email address** such as an institutional email that uses Gmail or a personal Gmail account. You must use a Google associated email address to gain access to Google Cloud computing resources.
6. If you are a student, share the email you used to sign up for AnVIL with your instructor following their instructions. If you are an instructor on a C-MOOR billing project, share the email you used to sign up with someone from C-MOOR.

**Until your account is associated with a billing project you will be unable to use computational resources on AnVIL.**

## 4.3 Set up billing in AnVIL

**This section is only for instructors. Students do NOT have to set up billing**

AnVIL will charge you for computing costs; as of the writing of this guide, the cost for running RStudio with the default settings is $0.06 per hour. RStudio will also cost $0.01 per hour while paused and not in use. Additionally, the persistent storage the environment comes with (50GB) costs $2.00 per month if not deleted.

To minimize costs, we ask students to delete their RStudio Environment and persistent disk at the end of every session. While you can also reduce the amount of CPUs and memory allotted per session this will also slow down your computation.

We will not go over setting up a Google billing account which you will need to setting up a billing project on AnVIL. For assistance with setting up your billing account we suggest refering to the Terra (the platform on which AnVIL runs) [guide to billing](https://support.terra.bio/hc/en-us/articles/360048632271-Terra-costs-and-billing-GCP-details) and speaking with your institution’s information technology and finance departments.

#### 4.3.0.1 Purpose

We will learn about the billing structure of AnVIL and how to attach users to billing projects. We will then cover how to create groups on AnVIL which you may find helpful in organizing billing.

#### 4.3.0.2 Learning Objectives

1. Distinguish between a billing account and a billing project
2. Understand how billing projects are connected to workspaces
3. Add users to a billing project
4. Learn how to create groups that can be used to control users’ access

### 4.3.1 What is a billing project?

A billing project is used to connect a workspace - where students will be executing code - to a billing account, which is where your actual payment information is stored. As you can see in the above diagram, a billing account can have multiple billing projects, and each billing project can be used by multiple workspaces.

We suggest each student uses their own workspace and attaching all those workspaces to a billing project for the class. Currently, AnVIL only allow you to monitor costs from billing projects, not workspaces. But having different billing projects shown above for each class can help you learn how much computational resources reach class is using.

### 4.3.2 Adding students to a billing account

Follow the written steps below or refer to the [slides](https://docs.google.com/presentation/d/1yyH3DZb8Et19galJhNPiUnCevrFRC1acvMG2PB52FCo/edit?usp=sharing) or video guide.

1. Access billing by clicking on the hamburger icon in the top left corner of the window, click on your name, and select billing. That brings you to your billing projects page.
2. On the billing project you’d like to use, click on the Members tab.
3. Click Add users. A new window will open. You are then able to add students to the billing project with the same email they used to sign up for their AnVIL account.

### 4.3.3 Using groups to manage classes

If you have a lot of students and classes, you may find it helpful to organize your students with groups. Everyone in the group is controlled by a single email address.

1. Access billing by clicking on the hamburger icon in the top left corner of the window, click on your name, and select groups. This will take you to this groups page.
2. Click Create New group. You will be prompted to give your group a unique name.
3. Click on the name of the newly created group to enter it.
4. Click add users. Add users to the group using the same email they used to sign up for AnVIL.
5. You can use the group email created for the group seen on the group management page to manage everyone in that group. For example, you can add everyone in the group to a billing project by adding this group email to the billing project.

### 4.3.4 Preventing runaway costs

Our team at C-MOOR is still piloting our curriculum on AnVIL. As such, we don’t have much information on the approximate cost per student. We hope to provide this data in the future.

The best way to conserve costs is to make sure that students close out their session on AnVIL properly, including the deletion of the persistent disk after they are done working. AnVIL will continue to bill the billing project even if the environment is not in active use.

For more information on how to control computing charges, please see the Terra guide: [How to cut off GCP charges](https://support.terra.bio/hc/en-us/articles/360042023952-Runaway-costs-How-to-cut-off-GCP-charges).

## 4.4 Running a module on AnVIL

#### 4.4.0.1 Purpose

In this section we will go over how to run C-MOOR modules on AnVIL. First we will need to clone the workspace that correlates to the research project we want to do (RNA-seq or 16S). We will only need to do this step once. Then we will go over how to create an RStudio environment in that workspace to run the module and properly end a session on AnVIL to prevent runaway costs.

#### 4.4.0.2 Learning Objectives

1. Clone a public workspace for the research project you want to do (RNA-seq or 16S)
2. Launch a module through the cloned workspace
3. Close out a session on AnVIL properly to prevent runaway costs

### 4.4.1 What is a workspace?

The workspace is the heart of AnVIL. Here are some key points about workspaces:

* Every workspace comes with its own Google Bucket (our cloud storage). Your bucket will be empty.
* Every workspace has its own billing project. Students who are not yet associated with a billing project will not be able to compute on their workspace.
* We can control access levels of users and set them either as owners, writers, or readers. Students will be writers with compute access.

### 4.4.2 Clone a workspace on AnVIL

**We only have to do this once.**

Follow the written steps below or refer to the [slides](https://docs.google.com/presentation/d/11wb3b7i9SwrDX_WO3mWNAycd2mbY4Moy8SuT0X3XvXo/edit?usp=sharing) or video guide.

1. While logged into AnVIL, using the hamburger icon in the top left corner of the screen, navigate to the workspaces page
2. Select the public tab
3. Search for the desired workspace. Your instructor will tell you which workspace to look for (miniCURE-RNA-seq or miniCURE-16S-microbiome).
4. Click on the more options icon on the right side of the desired workspace and click clone
5. Give the cloned workspace a unique name, such as by adding your initials or last name. All workspaces must have unique names; if someone has already taken the workspace name you initially wanted, please try a different name.
6. Confirm the billing project is the one your instructor has chosen.
7. The rest of the options should be left as is. Clone the workspace. It may take a few minutes to clone.

**You only have to clone the workspace once. From now on, use your cloned workspace.** After you clone the workspace you will automatically be directed to it. For all other times, your workspaces can be found using the hamburger icon in the top left.

### 4.4.3 Running modules on AnVIL

#### 4.4.3.1 Starting a module on AnVIL

When you open the workspace, you will be on the dashboard tab by default. The dashboard contains the instructions on how to use the workspace, links to C-MOOR websites, and the startup script. Let’s try running a module.

1. Take note of the container image for the custom environment. We recommend copying this to a word document or notepad. Make sure there are no spaces before or after what you copy. You will need to input this URL soon.
2. Take note of the startup script. Make sure there are no spaces before or after what you copy. This script is held in the original workspace everyone cloned. It does not have to be in your own workspace for it to work. You will need to input this URL soon.
3. Click on the Environment Configuration button , the cloud with a thunderbolt.
4. In the RStudio section, click Settings.
5. Make sure you have the following settings matching these instructions. Under Application configuration, choose “Custom environment”. In the container image field that appears, paste the container image URL that we copied earlier from the workspace. The URL should end with Bioconductor 3.19.1. In the startup script field, paste the URL for the startup script. This URL contains the words C-MOOR Startup Script.
6. Select 4 CPUs and 15 gigabytes of memory.
7. Confirm that the cloud compute cost is 20 cents per hour. If it is not 20 cents per hour, reselect CPUs and memory allocation in part 6. This is a known bug in AnVIL at the writing of this guide.
8. Scroll to the bottom of the window and click “Create”.

It will take some time for the RStudio Environment to be created. You can keep track of the status of the environment based on the colored dot next to the RStudio icon. The dot will turn green when the environment is ready. While it is loading (blue), you cannot interact with it.

1. When the environment is ready, use the Open RStudio button that will pop up. You can also access RStudio through the Analyses tab. If you hold down Ctrl as you click, you can open RStudio in a new window.
2. Use the file explorer in RStudio to navigate to your module of choice. From the folder called cure-rnaseq, go to tutorials, and then the folder of the module you want.
3. In the module’s directory, open the .Rmd file by double clicking its name.
4. Click Run Document in the open .Rmd file

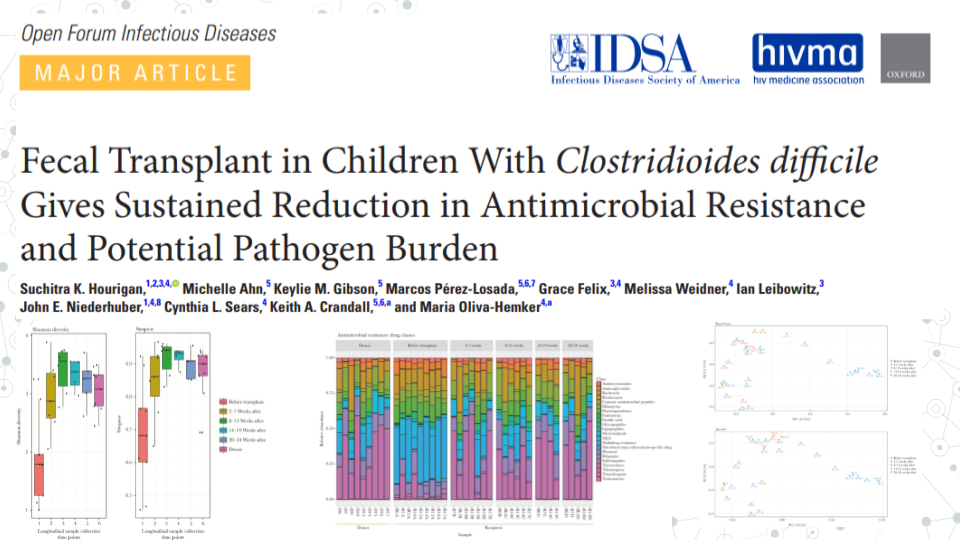
**When you are finished, make sure you close out your session properly to prevent runaway costs!**.

### 4.4.4 Closing out a session on AnVIL

1. On the right side of the screen, click the Cloud Environment button. This is the Cloud with the lighting symbol.
2. Under the RStudio section, click settings.
3. Scroll to the bottom of the new window and click delete environment.
4. Check **Delete everything, including the persistent disk or your instructor’s billing account will incur costs for storage**.

# 5 Scientific Literature

## 5.1 Pre-lab - Scientific literature



### 5.1.1 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.

### 5.1.2 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).

### 5.1.3 Activity

*Estimated time: 30 minutes*

#### 5.1.3.1 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.

#### 5.1.3.2 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. |
| --- |
|  |

### 5.1.4 Footnotes

#### 5.1.4.1 Resources

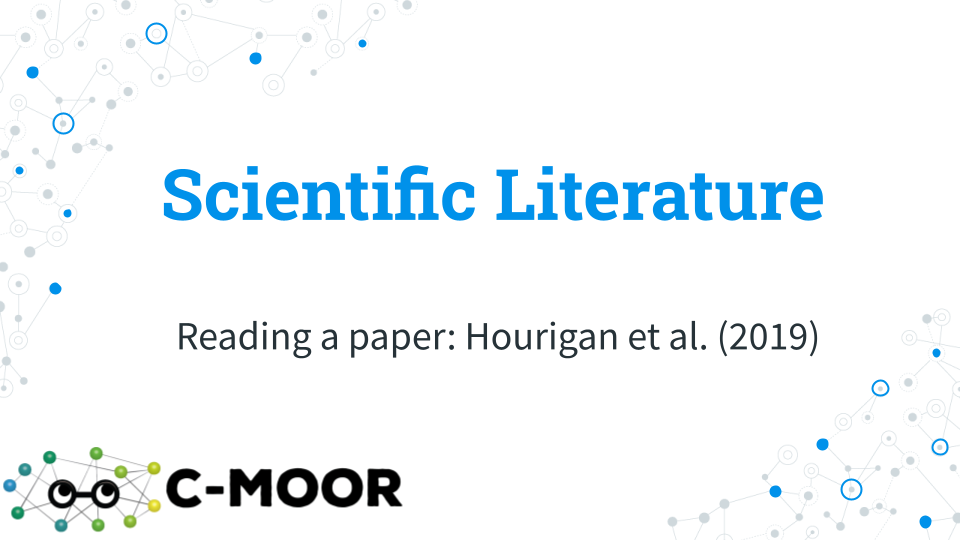
* [Google Doc](https://docs.google.com/document/d/1Uzw-dPwx4kxrcDKLTM7hjxN09oFkTQWg/edit?usp=sharing&ouid=117014088395158370475&rtpof=true&sd=true)
* [Paper](https://pmc.ncbi.nlm.nih.gov/articles/PMC6790402/)

#### 5.1.4.2 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

## 5.2 Lecture - Reading a paper

*Estimated time:*



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

## 5.3 Activity - Scientific Literature

### 5.3.1 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/)

### 5.3.2 Learning Objectives

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

### 5.3.3 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.

### 5.3.4 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.

### 5.3.5 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? |
| --- |
|  |

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

*Estimated time: 45 minutes*

#### 5.3.6.1 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. |
| --- |
|  |

#### 5.3.6.2 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 |  |

#### 5.3.6.3 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. |
| --- |
|  |

#### 5.3.6.4 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?

|  |
| --- |
|  |

### 5.3.7 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).

### 5.3.8 Footnotes

#### 5.3.8.1 Resources

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

#### 5.3.8.2 Contributions and affiliations

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

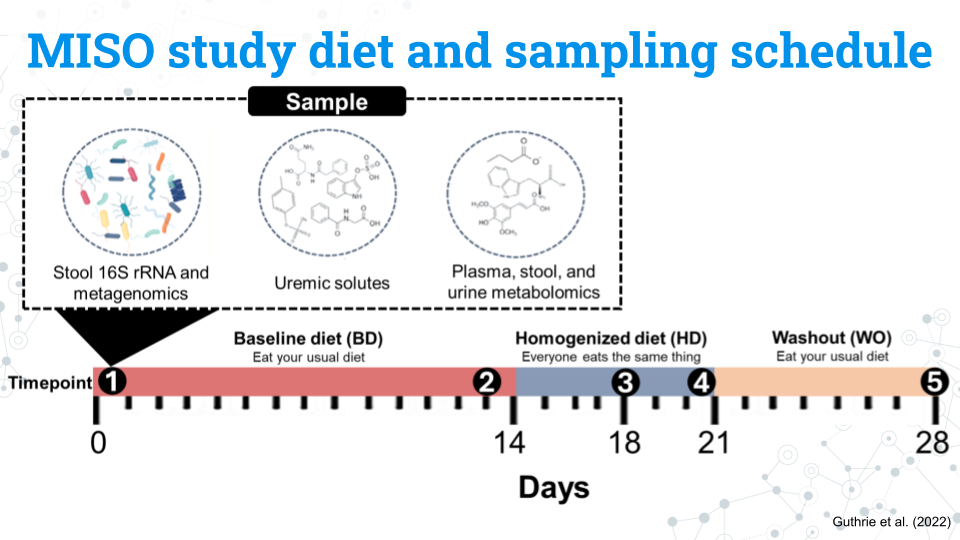
# 6 Reference - Meet the MISO study

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.

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.

### 6.0.1 MISO study diet and sampling schedule



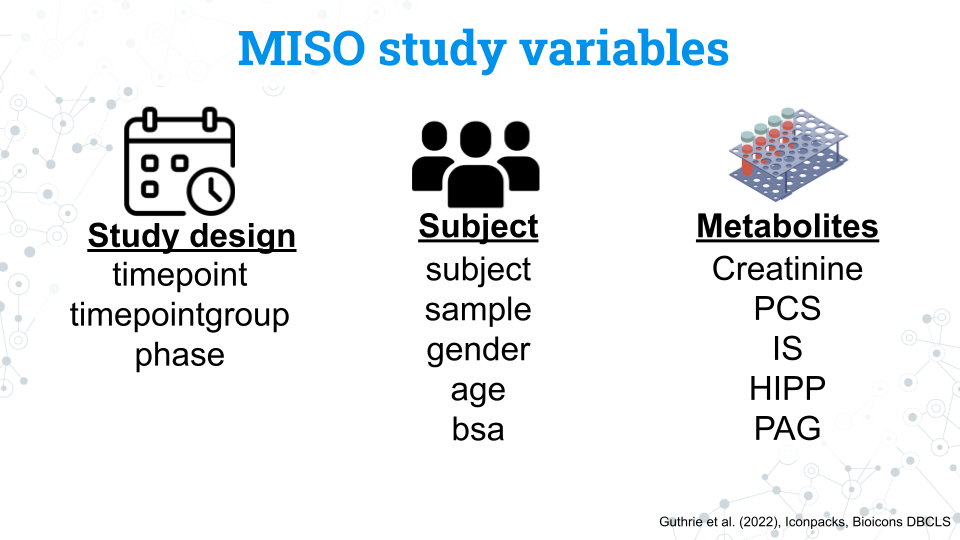
The figure above shows the study design for the MISO study. You do not need to memorize all these details; feel free to refer to this page throughout the project:

* 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

### 6.0.2 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.

#### 6.0.2.1 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 |

#### 6.0.2.2 Subject variables

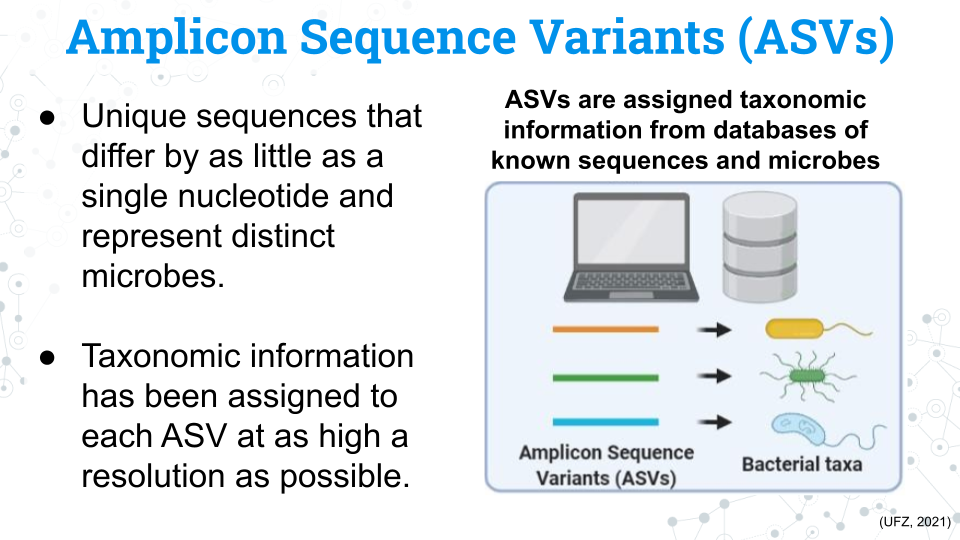
| Variable | What is it? | Factors |
| --- | --- | --- |
| subject | A unique ID given to each subject (participant) | S## (ex. S02 is subject 2. Note that while there are a total of 21 subjects in the study, but their subject numbers are not 1 through 20 |
| sample | A unique ID given to every sample 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 |

#### 6.0.2.3 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 |

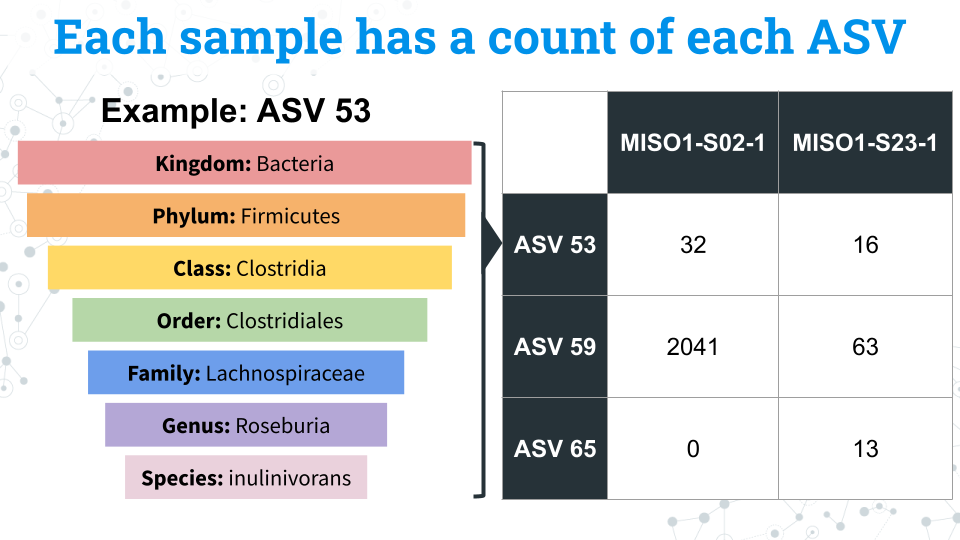
### 6.0.3 Amplicon sequence variants (ASVs) data

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. **ASVs and OTUs have some differences between them but they both represent distinct units of microbial taxa**, 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.



### 6.0.4 Footnotes

#### 6.0.4.1 Resources

#### 6.0.4.2 Contributions and affiliations

* Sayumi York, Notre Dame of Maryland University

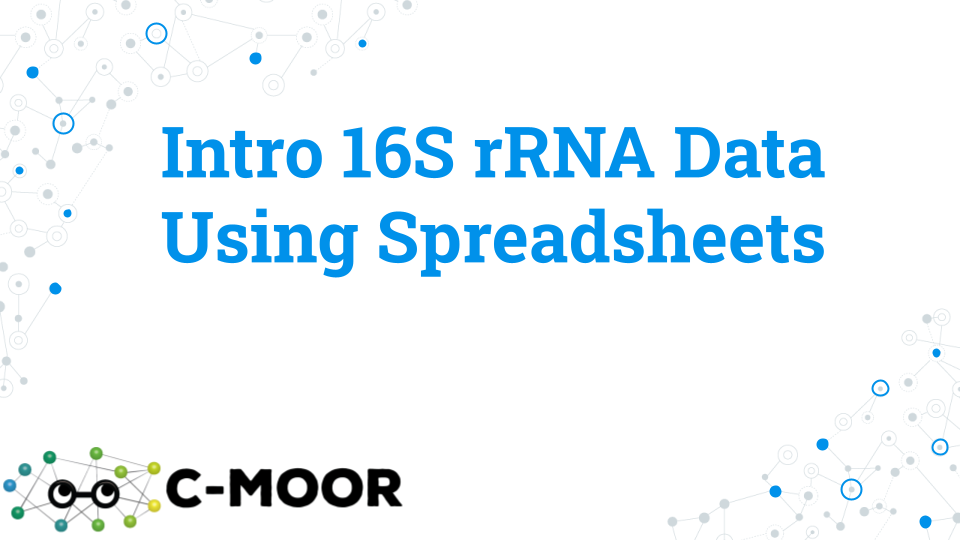
# 7 16S rRNA Data in spreadsheets

## 7.1 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!

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

*Estimated time:*



[Lecture](https://docs.google.com/presentation/d/14E26SbzaLEDJxAaZ1qjaZcJehAI8q2GHuTokNPKKDTc/edit?usp=sharing)

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

### 7.3.1 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

### 7.3.2 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.

### 7.3.3 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.

#### 7.3.3.1 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/1hpSyjO0H8sXq6C1xM_oGMyv9vCY2qTvfYHVdqO9ZVhk/edit?usp=sharing) file and open it with Google Sheets.

* Rows = ASVs
* 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? HINT: 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).**

**A) 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? |
| --- |
|  |

### 7.3.4 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](https://docs.google.com/spreadsheets/d/1vD0wbxgYuDlfKA0DF2Km-N4GHL2gRoUaEN_oCrSb2jU/edit?usp=sharing) 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.

#### 7.3.4.1 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.

**A) 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

|  |
| --- |
|  |

#### 7.3.4.2 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? |
| --- |
|  |

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

**A) Make a bar graph of most common ASV distribution across 21 individuals** 1. In the original MISO\_PS\_Leah\_SY\_WO-select find the row corresponding to the most common ASV, from Activity 2-1 above. 2. 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? |
| --- |
|  |

### 7.3.6 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

#### 7.3.6.1 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.**

**A) Sort Variance column and record the top 5 values below.** To sort the variance column:

1. Click on the 1st (header) row
2. Right-click to select ‘Create a filter’
3. Sort ‘VAR’ column Z to A (click on the new arrow in the VAR column).

|  |
| --- |
|  |

**B) Plot ASVs with highest variance.** 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.

|  |
| --- |
|  |

#### 7.3.6.2 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.

|  |
| --- |
|  |

### 7.3.7 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)](https://docs.google.com/spreadsheets/d/1mDk_enQzQVQVplbLzPMYJBF5whsvZ62gOPDwkqIpJvM/edit?usp=sharing)
* [Sample metadata (106 rows by 28 columns)](https://docs.google.com/spreadsheets/d/1TFXtaantSgDnMwnUvcAllQErWcOXOUwNCwRb_vyk2f0/edit?usp=sharing)
* [Class Poster](---)

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

#### 7.3.7.1 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:

1. 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?)
2. How will you compare the multiple BD1 samples against the multiple HD1 samples?

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

#### 7.3.7.2 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 1. 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?) 1. How will you compare the 21 samples? - ASV counts - plot average or boxplot of abundance for top ASVs? - For taxonomy - Barplot top taxa for all individuals? Which rank - Phyla, Genus, Species?)

### 7.3.8 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).
* Download the Google Sheets to your local computers as a .xlsx, complete all the steps for the assignment, and upload the .xlsx to your LMS (Blackboard, Canvas, Google Classroom).

### 7.3.9 Footnotes

#### 7.3.9.1 Resources

* [Google Doc](https://docs.google.com/document/d/12pKSIwcVXisaf6dniyoG_AGdCQh1dG1dDwZ5s1cz92c/edit?usp=sharing)
* [‘MISO\_PS\_Leah\_SY\_Sample1’](https://docs.google.com/spreadsheets/d/1hpSyjO0H8sXq6C1xM_oGMyv9vCY2qTvfYHVdqO9ZVhk/edit?usp=sharing)
* [MISO\_PS\_Leah\_SY\_WO-select](https://docs.google.com/spreadsheets/d/1vD0wbxgYuDlfKA0DF2Km-N4GHL2gRoUaEN_oCrSb2jU/edit?usp=sharing)
* [ASV counts (2185 rows by 114 columns)](https://docs.google.com/spreadsheets/d/1mDk_enQzQVQVplbLzPMYJBF5whsvZ62gOPDwkqIpJvM/edit?usp=sharing)
* [Sample metadata (106 rows by 28 columns)](https://docs.google.com/spreadsheets/d/1TFXtaantSgDnMwnUvcAllQErWcOXOUwNCwRb_vyk2f0/edit?usp=sharing)
* [Class Poster](---)

#### 7.3.9.2 Contributions and affiliations

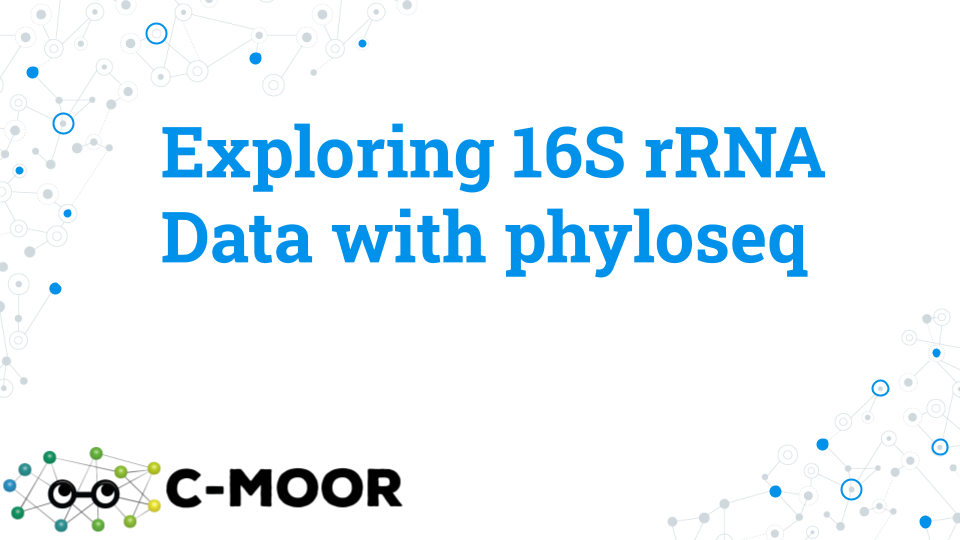
* Valeriya Gaysinskaya, Johns Hopkins University
* Gauri Paul, Clovis Community College
* Frederick Tan, Johns Hopkins University
* Sayumi York, Notre Dame of Maryland University

# 8 Exploring 16S rRNA Data with phyloseq

## 8.1 Exploring 16S rRNA Data with phyloseq

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

*Estimated time:*



[Lecture](https://docs.google.com/presentation/d/15tv_IS177gF_px3JYKcQZmOsJg2d-GkjVXi7a-Skgs8/edit?usp=sharing)

## 8.3 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.

### 8.3.1 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.

### 8.3.2 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.

### 8.3.3 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?” |
| --- |
|  |

### 8.3.4 Activity 2 – Try it out questions

*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](---)

There are 4 different questions for groups to explore. Take a look at the four questions below and go to the section of your chosen question for more instructions!

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

Diet is suggested to play an important role in shaping the human gut microbiome. However, other factors that are specific to an individual such as their physical fitness, metabolism, and genotype are also suspected to play a strong role. Here we will explore the impact of diet and individuality on human gut microbiome.

Approach: Examine high-level (Phylum) taxonomy for 5 timepoints and 21 individuals and evaluate if the change of diet or inter-individual differences have more impact on microbe variation.

1. Evaluate the effect of diet – Plot the three dietary groups BD, HD, and WO corresponding to 5 experimental timepoints (where timepoints 1 and 2 are BD, timepoints 3 and 4 are HD and timepoint 5 is WO).
2. Evaluate the effect of individuality – Plot the microbiome for all 21 individuals but focus on just the HD timepointgroup (where diet is the same).

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

Microbes within the core microbiome - the most common and abundant species across samples in a given group - are likely to be involved in key functions of the holobiont. However, the thousands of species present in the human gut is difficult to visualize and interpret when done all at once. One approach is the focus on the most abundant taxa at a given taxonomic rank.

Approach: Examine microbial composition for most abundant taxonomic ranks in the human gut of the 21 individuals. To do so, you will use progressive sub-setting of the most abundant taxonomic ranks ranging from Phylum down to Species.

1. Plot all the Phyla and identify the most abundant Phylum
2. Subset the most abundant Phylum, and plot all the Orders
3. Subset the most abundant Order, and plot all the Families
4. Subset the most abundant Family, and plot all the Genera
5. Subset the most abundant Genus, and plot all the Species

#### 8.3.4.3 Question 3. Does gender have an impact on the human gut microbiome?

The biological sex of the host has been suggested to help shape or influence its gut microbiome. Some candidate microbes have been implicated in sex differences. Here we will explore the potential gender-based differences in the human gut microbiome.

Approach: Test if Phyla and/or species vary based on gender. First you will survey the differences between male and female subjects in the composition of their Phylum and species. Then you will select some candidate microbes of interest to examine in further detail.

1. Survey differences in phylum and species composition by gender.
2. Plot differences in species for a high abundance phylum by gender.
3. Plot differences in species for a low abundance phylum by gender.
4. Plot two candidate microbes reported in the literature to vary with gender.

#### 8.3.4.4 Question 4. Does age have an impact on human gut microbiomes?

The human gut microbiome has been associated with age-related disease states, immune-system changes, and metabolic function. Here we will explore the potential microbiome changes associated with aging.

Approach: Test if Phylum and/or species composition differs by age across the dataset. First you will plot the relationship between age and phylum; we will see that age is a more challenging variable to work with than our categorical variables. You will collapse and transform the data to fix this issue to create a final plot for your interpretation.

1. Plot all Phyla based on age.
2. Normalize age that includes multiple individuals to avoid overestimation (e.g., 2 individuals are of age 27, 46 and 58, and 3 individuals are 54).
3. Profile candidate Phyla Firmicutes based on age.
4. Profile candidate Phyla Bacteroidetes based on age.

### 8.3.5 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).

### 8.3.6 Footnotes

#### 8.3.6.1 Resources

* [Class Poster](---)
* [Google Doc](https://docs.google.com/document/d/1d1co_BVzZ1P5RuUYnPWrkB-Lvo7f4lYvahXhoCOjvIg/edit?usp=sharing)
* [16S rRNA R Quick Reference](https://docs.google.com/document/d/1pYTq_U-e3Fath5yB6N9b5ic-cMynUtclkEgdojkAXwE/edit?usp=sharing)

#### 8.3.6.2 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

## 8.4 Try it 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.

### 8.4.1 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

|  |
| --- |
|  |

### 8.4.2 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).
2. Paste the code you used to generate your final figure(s) below:

|  |
| --- |
|  |

## 8.5 Try it 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.

### 8.5.1 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: |
| --- |
|  |

### 8.5.2 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: |
| --- |
|  |

#### 8.5.2.1 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: |
| --- |
|  |

#### 8.5.2.2 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: |
| --- |
|  |

#### 8.5.2.3 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:

|  |
| --- |
|  |

## 8.6 Try it 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.

#### 8.6.0.1 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: |
| --- |
|  |

#### 8.6.0.2 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? |
| --- |
|  |

### 8.6.1 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: |
| --- |
|  |

### 8.6.2 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:

|  |
| --- |
|  |

## 8.7 Try it 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.

### 8.7.1 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.

### 8.7.2 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? |
| --- |
|  |

### 8.7.3 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:

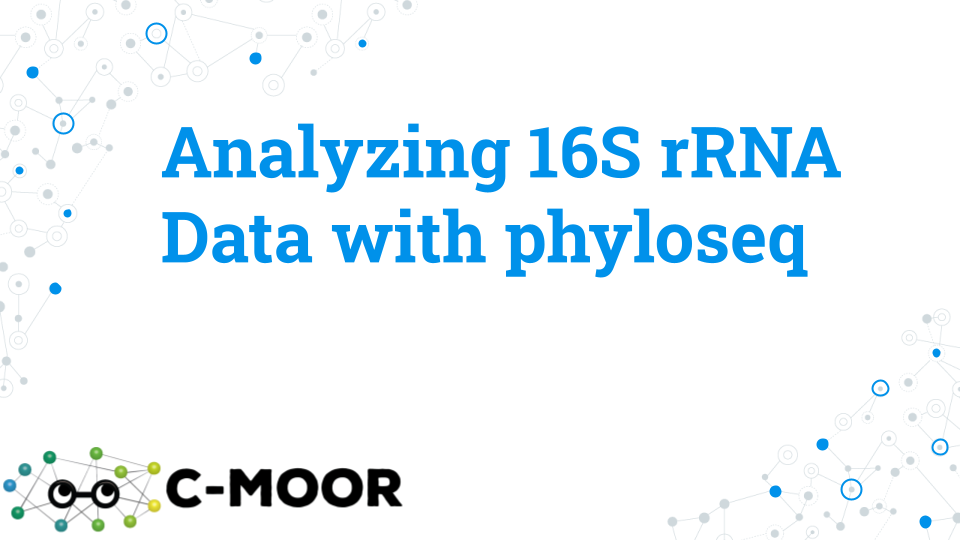
|  |
| --- |
|  |

# 9 Analyzing 16S rRNA Data with phyloseq

## 9.1 Analyzing 16S rRNA Data with phyloseq

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

*Estimated time:*



[Lecture](https://docs.google.com/presentation/d/1QZbSBPOGkBeizh1L45C6EaktOxl6pqrTDGNIncHQ5lY/edit?usp=sharing)

## 9.3 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

### 9.3.1 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

### 9.3.2 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.

### 9.3.3 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?

|  |
| --- |
|  |

### 9.3.4 Activity 2 – Try it out questions

*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](---).

There are 4 different questions for groups to explore. Take a look at the four questions below and go to the section of your chosen question for more instructions!

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

Alpha diversity is a measure that estimates how the distribution of microbes changes due to a variable (or metadata category). Alpha diversity measures changes in the richness (the number of different organisms or ASVs) and evenness (how evenly are these organisms distributed in terms of their abundance). Using the “MISO” study dataset we will use Simpson (or specifically, Gini-Simpson) alpha diversity to evaluate changes in microbial diversity in individuals due to different metadata variables.

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 variables that impact alpha diversity. Visible shifts in alpha diversity suggest a shift in microbial diversity, and a higher alpha diversity value indicates an increase in alpha diversity (either richness or abundance).

1. Evaluate the impact of diet on alpha diversity by plotting ASVs based on “timepointgroup” variable.
2. Evaluate the impact of individuality on alpha diversity by plotting ASVs based on “subject” variable.
3. Evaluate the impact of gender on alpha diversity by plotting ASVs based on “age” variable.
4. Evaluate the impact of gender on alpha diversity by plotting ASVs based on “gender” variable.
5. Evaluate the impact of gender on alpha diversity by plotting ASVs based on the 5 different levels of metabolites in the study (Creatinine, PCS, IS, HIPP, PAG).

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

PCoA plot is a principal coordinate analysis used to represent similarity between samples (sample microbiomes in our case). Using the “MISO” study dataset, we will use the PCoA plot to summarize individuals based on ASVs and plot the resulting relationships between individuals. Based on how well color-coding the different variables matches the sample distribution on the PcOA plot, we will aim to help explain potential sources of sample similarity.

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 correlate with sample diversity. In a PcOA plot, samples with similar microbial profiles will be plotted close and may appear as groups.

1. Make a PCoA plot, ordinate on the entire dataset (all ASVs) and color by individual. Investigate the shape of the resulting PCoA plot.
2. Correlate PCoA plot shape with metadata variables by coloring the PCoA plot with different variables including diet, subject, age and gender. Do any of the variables correlate with the shape of PCoA plot and data groupings?
3. Correlate PCoA plot shape with the levels of metabolites in the study (Creatinine, PCS, IS, HIPP, PAG). From the color-coding pattern, identify 1. which variables help potentially explain the data groups formed in the PCoA plot. Subset “HD”, ordinate on “HD”, and make a new PCoA plot.

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

Approach: Perform DESeq2 differential abundance 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 differential abundance of the microbes between the sexes.

1. Perform DESeq2 analysis between females and males and identify differences.
2. Test if younger age contributes to differential microbe abundance between females and males by subsetting younger (< 50 years old) individuals.
3. Test if older age contributes to differential microbe abundance between females and males by subsetting older (> 50 years old) individuals.

#### 9.3.4.4 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. Then, use DESeq2 to evaluate if younger or older age changes ASVs associated with diet.

1. Plot alpha diversity based age for the population in general, and then for individuals subsetted for BD, HD and WO diets. Look for shifts in alpha diversity with change in diet.
2. Perform DESeq2 analysis based on the diet for the population in general, establishing baseline differential abundance between dietary groups HD and BD.
3. Perform DESeq2 analysis based on the diet for the younger (<= 50 yo) and older (>=50 yo) individuals, evaluating age-associated changes in differential abundance between dietary groups HD and BD.

### 9.3.5 Grading criteria

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

### 9.3.6 Footnotes

#### 9.3.6.1 Resources

[Google Doc](https://docs.google.com/document/d/1OWgUwaT2MlSd-qq7wSlB3BnM-cq_GTxEyy38HV5ajoU/edit?usp=sharing)

#### 9.3.6.2 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

## 9.4 Try it 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.

### 9.4.1 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.**

|  |
| --- |
|  |

### 9.4.2 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? |
| --- |
|  |

### 9.4.3 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?

|  |
| --- |
|  |

### 9.4.4 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? |
| --- |
|  |

### 9.4.5 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? |
| --- |
|  |

### 9.4.6 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.
2. 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: |
| --- |
|  |

## 9.5 Try it 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”.

### 9.5.1 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? |
| --- |
|  |

### 9.5.2 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? |
| --- |
|  |

### 9.5.3 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? |
| --- |
|  |

### 9.5.4 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 5A. 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: |
| --- |
|  |

## 9.6 Try it 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.

### 9.6.1 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. |
| --- |
|  |

### 9.6.2 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? |
| --- |
|  |

### 9.6.3 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? |
| --- |
|  |

### 9.6.4 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: |
| --- |
|  |

## 9.7 Try it Question 4 - Is there an interaction between diet and age and the microbiome

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.

### 9.7.1 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). |
| --- |
|  |

### 9.7.2 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? |
| --- |
|  |

### 9.7.3 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? |
| --- |
|  |

### 9.7.4 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: |
| --- |
|  |

# 10 Community analysis and feedback

## Warning: Chunk option fig.align is not supported for docx output  
## Warning: Chunk option fig.align is not supported for docx output

## Warning: Chunk option fig.alt is not supported for docx output

## Warning: Chunk option fig.align is not supported for docx output

## Warning: Chunk option fig.alt is not supported for docx output

## Warning: Chunk option fig.align is not supported for docx output

## Warning: Chunk option fig.alt is not supported for docx output

## Warning: Chunk option fig.align is not supported for docx output

## Warning: Chunk option fig.alt is not supported for docx output

## Warning: Chunk option fig.align is not supported for docx output

## Warning: Chunk option fig.alt is not supported for docx output

## Warning: Chunk option fig.align is not supported for docx output

## Warning: Chunk option fig.alt is not supported for docx output  
## Warning: Chunk option fig.alt is not supported for docx output  
## Warning: Chunk option fig.alt is not supported for docx output  
## Warning: Chunk option fig.alt is not supported for docx output  
## Warning: Chunk option fig.alt is not supported for docx output

## 10.1 Lecture - Scientific Posters

*Estimated time: —*



[Scientific Posters](https://docs.google.com/presentation/d/1-orSi8DpN22hMt9-6p_rHZnte1YXXLe-a132HDSyd0U/edit?usp=sharing)

## 10.2 Activity - Scientific Posters

### 10.2.1 Activity

*Estimated time: —*

#### 10.2.1.1 Instructions

1. Review [**Presentations Guidelines for Posters**](https://ur.umbc.edu/urcad/resources/posters-guidelines)
2. Skim three posters from among the following (must be Biology if from UMBC Posters)

* [**Look at This! Category on the Academy Discussion Forum**](https://help.c-moor.org/c/look-at-this/8)
* [**UMBC Biology Posters**](https://ur.umbc.edu/poster-presentation-examples)
* [**The example PacBio Poster**](https://www.pacb.com/wp-content/uploads/PAG-Portik.pdf)

1. Pick one poster and address the following points.

* **Notice** – What about this poster most interests you?
* **Wonder** – Two or three questions you would ask the authors.
* **Support** – Two or three suggestions on how the poster could be improved.

1. Post your answers by replying to the “Project Work: Scientific Posters” topic in the Discussion Forum

### 10.2.2 Grading Criteria

* Submit URL to your reply on Canvas.

### 10.2.3 Footnotes

**Contributions and Affiliations**

* Valeriya Gaysinskaya, Johns Hopkins University
* Frederick Tan, Johns Hopkins University

Last Revised: February 2025

## 10.3 Activity - Share Your Poster

### 10.3.1 Introduction

It’s almost time! Prepare for your final presentation by thinking about how to describe the great work that you’ve done this semester in both written and oral form. Assume that your audience has a basic scientific background but does not know a lot about your particular field. Provide enough context to give your listener a reason to care about your project and each of the results you will describe. You have five minutes to get your story across so have a plan and practice.

#### 10.3.1.1 Activity 1 – Share Your Poster

*Estimated time: 20 min*

#### 10.3.1.2 Instructions

1. Download your poster as a .png file

* <https://drive.google.com/drive/folders/1y_GCJl7VIYTS_5y7057u2s58ZCdm_PxM>

1. (One person on behalf of the group) Create a New Topic on the Discussion Forum with the title of your poster as the topic title and your .png file as the contents in either the
2. Look at This! category if every member of your group is ok sharing your work publicly <https://help.c-moor.org/c/look-at-this/8>
3. JHU 2025 Spring category if you need to keep your work private <https://help.c-moor.org/c/jhu-2025-spring/46>
4. (Each person in the group) Reply to your topic using your own words with the following two bolded sections using the following template

|  |
| --- |
| **What I Did** |
|  |
| **How You Can Help** |
|  |

#### 10.3.1.3 Questions

| 1. Provide URL of your Discussion Forum post |
| --- |
|  |

### 10.3.2 Activity 2 – Present Your Poster

*Estimated time: 40 min*

#### 10.3.2.1 Instructions

1. Create a plan for a 5 min presentation taking into consideration
2. Who will speak when
3. What each person will cover
4. Practice your presentation
5. In your group
6. To the class

#### 10.3.2.2 Questions

| Create a bullet point outline of who will speak what when: |
| --- |
| Overview (question, background, hypothesis) |
| Approach (dataset, general methods) |
| Results (detailed methods, notable results) |
| Conclusions (summary, next steps) |

### 10.3.3 Grading Criteria

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

### 10.3.4 Footnotes

**Resources**

* [Google Doc]

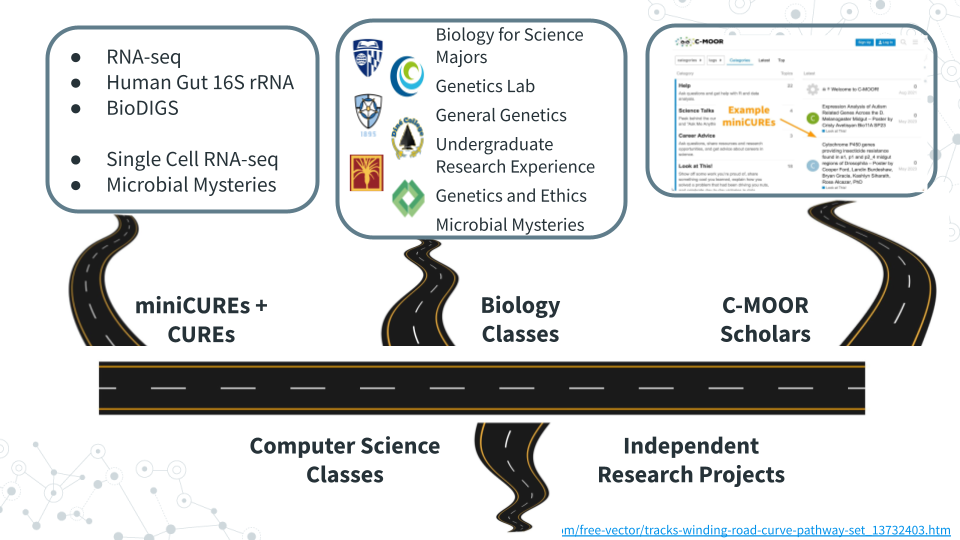
**Contributions and Affiliations**

* Valeriya Gaysinskaya, Johns Hopkins University
* Frederick Tan, Johns Hopkins University

Last Revised: April 2025

# 11 Professional Development

Now that you’ve finished your C-MOOR project, how do you close out your research experience or take it further? In this section we’ll explore how to publicize and use your experience to help build your professional career.



## 11.1 Lecture - Next Steps

*Estimated time: —*



[Slides: Next Steps](https://docs.google.com/presentation/d/1PMvZ19kSTK2ghdgLNFRS_t5p8R_GKFNbExPRxMsE1Ns/edit?usp=sharing)

## 11.2 Activity - Next Steps

### 11.2.1 Introduction

Congratulations on making it to the end of Microbial Mysteries! Take a moment and reflect on all that you’ve accomplished during this semester, especially those of you who reported in the Welcome Poll “No experience” with Microbiology, Galaxy, and Scientific Research. While the “Grad Student Motivation Level” PHD Comics spans years, it likely reflects the ups and downs as you developed and explored your hypothesis, including the thrill of presenting your work to people in our mini-symposium. The feedback has been great with faculty impressed at what you’ve accomplished, the depth that you understand your projects, and even one scientist who does metagenomics research reporting that they learned a new trick or two!

We’ll wrap up today with several next steps including completing feedback to improve the next offering of this course , researching opportunities to do more science, exploring communities where scientists openly share insights, and optionally sharing more work.

### 11.2.2 Activity 1 – Complete Feedback

*Estimated time: 15 min*

#### 11.2.2.1 Instructions

Help us improve this course as we work to grow enrollment to accommodate 32 to 64 students, incorporate graduate TAs and undergraduate course assistants and tutors, and promote more computational training and research opportunities throughout departmental courses <https://forms.gle/XkqiMERHc2PETmy4A>

### 11.2.3 Activity 2 – Research Opportunities

| 1. Find one LAB at YOUR INSTITUTION that you would work in and explain why. |
| --- |
| Your text here: |

| 2. REU – Which NSF Research Experiences for Undergraduates program would you attend and why? |
| --- |
| a. Overview – <https://www.nsf.gov/funding/initiatives/reu/students> |
| b. Search – <https://www.nsf.gov/funding/initiatives/reu/search> |
| - Consider filtering by Research Area (e.g. Biology) |
| Your text here: |

### 11.2.4 Activity 3 – Explore Communities

*Estimated time: 15 min*

#### 11.2.4.1 Instructions

Spend a little bit of time browsing communities that discuss genomics, data analysis, and more. Paste the URLs for a couple of posts that you find interesting with a brief comment.

#### 11.2.4.2 Questions

1. Specialized Communities
2. SEQanswers: the next-generation sequencing community

* <https://www.seqanswers.com>

1. Biostars: bioinformatics explained

* <https://www.biostars.org>

| Post URL of interest for Specialized Communities and your comments here: |
| --- |
|  |

1. General Communities
2. Reddit: The heart of the internet

* <https://www.reddit.com/r/metagenomics>

1. Twitter/X: It’s what’s happening

* <https://x.com/search?q=%23metagenomics>

1. Bluesky: Social media as it should be

* <https://bsky.app/search?q=metagenomics>

| Post URL of interest for General Communities and your comments here: |
| --- |
|  |

### 11.2.5 (Optional) Activity 4 – Share Work

*Estimated time: 15 min*

#### 11.2.5.1 Instructions

1. Upload photos – Especially photo of your group standing in front of your poster

* <https://drive.google.com/drive/folders/1y_GCJl7VIYTS_5y7057u2s58ZCdm_PxM>

1. Publicly share poster – If every member of your group is ok sharing your work publicly

* <https://help.c-moor.org/c/look-at-this/8>

### 11.2.6 Grading Criteria

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

### 11.2.7 Footnotes

**Resources**

* Google Doc

**Contributions and Affiliations**

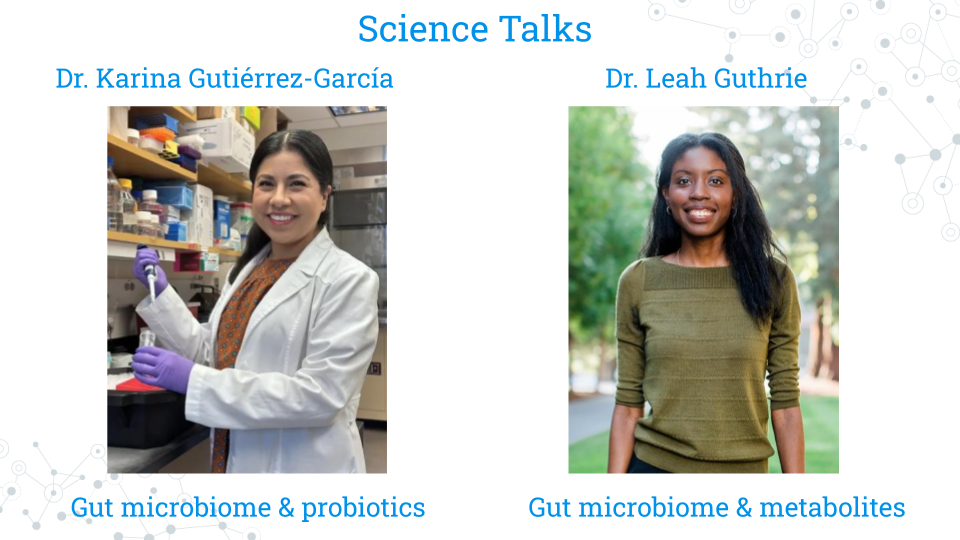
* Valeriya Gaysinskaya, Johns Hopkins University
* Frederick Tan, Johns Hopkins University

## 11.3 Science Talks

Science Talks feature a guest lecturer to showcase their research work, share their scientific journety and engage in Q and A with the students. It is a one of a kind opportunity for students to engage with a scientific scholar and/or expert from across departments, fields, and career stages and explore, broaden or narrow students’ scientific interests.

The two invited speakers for the Microbial Mysteries CURE Spring 2025 at the JHU were:

1. Dr. Karina Gutiérrez-García
2. Dr. Leah Guthrie



For each speaker, students were asked to prepare by completing the Science Talks Activities.

### 11.3.1 Science Talks - Activity 1

*Estimated time: 50 min*

1. Read/Browse the following information about **Dr. Karina Gutiérrez-García**
2. Postdoc Spotlight <carnegiescience.edu/news/postdoc-spotlight-karina-gutierrez-garcia>
3. “Home sweet home” perspective on study of fruit fly gut microbiome <pubmed.gov/39637006>
4. Abstract for “A conserved bacterial genetic basis for commensal-host specificity” <pubmed.gov/39636981>
5. Abstract for “Gut microbiomes of cycad-feeding insects tolerant to β-methylamino-L-alanine (BMAA) are rich in siderophore biosynthesis <pubmed.gov/37993724>
6. Post three questions to the speaker in the Discussion Forum at <https://help.c-moor.org/t/469>.
7. Science – What scientific question do you have about Speaker’s research, ranging from basic background questions to possible next steps?
8. Method – What methods would you like the Speaker insight on, whether computational, genomics, or other exciting technologies?
9. Career – What advice would you ask the Speaker for regarding how to get started, finding a mentor, etc.?

### 11.3.2 Science Talks - Activity 2

*Estimated time: 50 min*

1. Read/Browse the following information about **Dr. Leah Guthrie**
2. Faculty focus <qb3.berkeley.edu/news/faculty-focus-leah-guthrie>
3. Abstract for “Impact of a 7-day homogeneous diet on interpersonal variation in human gut microbiomes and metabolomes” <pubmed.gov/35643079>
4. Abstract for “Human microbiome signatures of differential colorectal cancer drug metabolism” <pubmed.gov/29104759>
5. Post three questions to the speaker in the Discussion Forum at [https://help.c-moor.org/t/469](https://help.c-moor.org/t/498).
6. Science – What scientific question do you have about Speaker’s research, ranging from basic background questions to possible next steps?
7. Method – What methods would you like the Speaker insight on, whether computational, genomics, or other exciting technologies?
8. Career – What advice would you ask the Speaker for regarding how to get started, finding a mentor, etc.?

## 11.4 Create your CV

*Estimated time: 10 min*

Everyone who completes a C-MOOR project can add their experience to their resume or CV! We advise you and your students to do this as soon as possible after the end of class while your research project is still on your mind.

#### 11.4.0.1 Resume or CV?

You may hear the terms resume and CV being used interchangeably. While both are drawn from the same information, they vary in brevity and focus.

A **resume** is a one-page (sometimes two-page) document of your experiences generally used for jobs outside of academia and research labs. It’s not uncommon for people to have more than one resume, as it’s best to tailor every resume to each specific role you’re applying to.

A **Curriculum Vitae (CV)** is an extensive document that can take as many pages as needed to list the entirety of your research and academic experience. It often includes sections that aren’t found in non-academic fields such as publications, conferences, and published papers.

Most positions will detail which they require from applicants and most colleges and universities will be able to help you build either at your library or career center. We suggest that current students and recent graduates interested add their C-MOOR research experience to either their classwork or projects section. Remember to acknowledge the contributions of your team members if applicable; this does not count against you in your applications! Here are some examples of how you might word your experiences; notice how they’re similar to an abstract in just one or two sentences:

* **An RNA-seq project:** Along with a partner, examined differential expression data of the *period (per)* gene along the *Drosophila melanogaster* midgut from a published dataset (Marianes & Spradling 2013) using DESeq2 and ClusterProfiler. We discovered an elevated expression of per in the anterior region of the midgut where the primary digestion and absorption of carbohydrates occur. Our results were presented as a research poster at the 2025 Kuntz Research Day symposium and have implications for using Drosophila as a model for the connection between the circadian rhythm and metabolic disorders.
* **A 16S project:** My group and I profiled the abundance of Archean ASVs in the gut microbiome from a published 16S study comparing the gut microbiota of people before, during, and after a standardized diet (Guthrie et al. 2022). As a part of this study, we created a PCoA and performed an alpha diversity analysis in phyloseq. Archean community members are often overlooked in microbiome studies; we found differential abundance in *Methanobacteria* between male and female participants regardless of diet and hypothesize this may be due to consumption of fiber reported by male subjects pre- and post-diet.
* **A WGS project**: Public data (SRA#######) from Robinson et al. (2021) was analyzed in Galaxy using Trimmomatic, Megahit, and Krona to determine the viral abundance in samples from tap water from various North American Lakes and make comparisons to lake depth. Viral abundance was correlated with geographic distance between samples; future studies could examine if this pattern persists in samples sourced from Europe and Oceana.

Your blurb is a strong blurb when:

* ☐   You credit the source of your data (Citation, SRA #, Project #)
* ☐   You describe why your research project matters
* ☐   You list some of the computational platforms or software you used

**Q: Will people think I am an expert in genomics and bioinformatics if I put my project on my resume/CV? I don’t want to look like a phony.**

A: As a current student or new graduate, you are not expected to have mastery over any scientific domain. What you are trying to show is you have taken an idea and brought it to completion, your ability to learn new concepts and skills, and your previous exposure to research. You can explain more about your role in the project in an interview.

**Q: Can I list things like R or Galaxy in my skills section?**

A: You can, however we suggest being explicit in your experience and add the level at which you feel comfortable with each respective platform. Are you R (Beginner) or R (Proficient)? Can you teach someone else how to do your project? Can you perform your project with novel data? How good are you at troubleshooting? This answer may seem contrary to the previous one, but remember that other people who are listing these platforms in their skills section includes computer/data scientists (majors) and bioinformaticians (majors), so people may assume mastery. Think about it the same way you would a foreign language. Spanish (Beginner) describes a different level of skill than ‘just’ Spanish.

# 12 C-MOOR Scholars

## 12.1 C-MOOR Scholars



C-MOOR Scholars are research clubs that function as an extension of the C-MOOR curriculum. You can explore techniques and concepts from modules you’ve yet to take to learn about different approaches we can use to explore different questions through genomics, connect with other scholars and mentors, and experience the climate of a research lab.

Fill out our [C-MOOR Scholars Interest Form](https://docs.google.com/forms/d/1GJDpfG28k2utCVVcRXQraE-1coLS_GCX83irUyio9KE/viewform?edit_requested=true) if interested, and meet some [C-MOOR Scholars at Clovis Community College](https://www.cloviscollege.edu/alumni-and-community/c-moor/c-moor-scholars.html)!

**What C-MOOR Scholars is**:

* A place to grow your skills
* A place to network
* A place to connect science with the greater community

**What C-MOOR Scholars is not**:

* A commitment to pursuing research as a career
* A graded class where we score work
* A drop-in drop-out popular science club

Less formal than a class, more formal than just a gathering of like-minded minds, C-MOOR Scholars provides structure for your growth as a scientist. We are actively looking to recruit new scholars for chapters at our established sites (CCC, JHU, & NDMU); speak with your instructor for more information or reach out to us at the [C-MOOR Academy Discussion Forum](help.c-moor.org) to hear about any remote opportunities we have.

### 12.1.1 C-MOOR Scholars links

Share your interests via C-MOOR Scholars Interest Form and learn how C-MOOR community can support you.

* <https://docs.google.com/forms/d/1GJDpfG28k2utCVVcRXQraE-1coLS_GCX83irUyio9KE/edit>

Meet the C-MOOR Scholars and learn how you can support them

* <https://www.cloviscollege.edu/alumni-and-community/c-moor/c-moor-scholars.html>

## 12.2 Pursue further research

Perhaps you’re interested in pursuing research beyond what C-MOOR currently provides. Maybe you’re more interested in wet lab work, or a specific model organism. We encourage you to try all sorts of different opportunities even if you feel totally comfortable working with C-MOOR so you can better understand what kind of roles suit you best. This section serves as a guide for some common jumping off points to find outside research opportunities.



### 12.2.1 Research opportunities at your institution

Many colleges and universities have research programs for their students or resources available to connect you to some. You can ask your instructor, adviser, or library for more information. Some popular programs include:

* **Undergraduate Research Opportunity Programs (UROP):** Which often provide a database or pipeline of research opportunities for students at the univeristy or from surrounding areas.
* **Research experiences for undergraduates (REU):** Similar to UROP, but generally open for a wider audience of students from outside the community. See the [NSF REU website](https://www.nsf.gov/funding/initiatives/reu) for more information.

Opportunities come and go and often follow the academic-year in terms of availability. If you can’t find something immediately, keep trying!

### 12.2.2 Find a research lab to join

You can choose to pursue additional research opportunities in what we consider the more traditional way: by reaching out to a professor or research scientist whose work you are interested in and asking if they might be willing to have you in their lab. As these opportunities are at the discretion of the lab leader they will vary widely in their scope, duration, and ability to provide financial support.

An email to a potential research lab should contain:

* ☐  An introduction about yourself (name, class year, institution, major)
* ☐  1-2 sentences about your previous experience in research
* ☐  1-2 sentences about why you’re interested in their lab specifically
* ☐  A polite, open ended request to work with them

Dear Professor Penguin,

My name is Student, and I am a junior at My University majoring in biology (or other). I read your chapter on the genetics yellow-eyed penguins in Penguins: Natural history and conservation, and then checked out your profile on the University website. My interests are broadly defined in environmental toxicology and I am interested in the work you are doing using eDNA and studies on the effects of plastic on developing penguin transcriptomes.

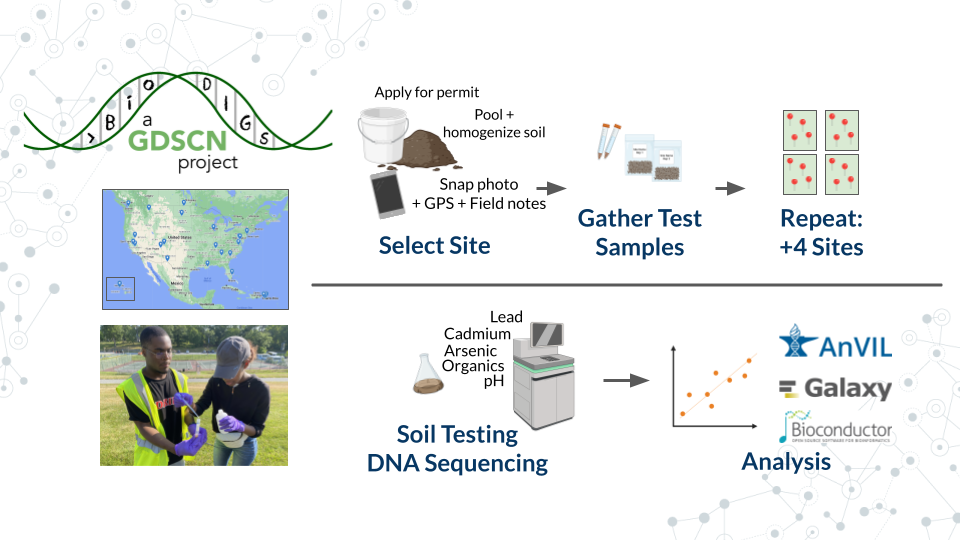
I previously completed a project as a part of an RNA-seq miniCURE and have some exposure to working with DESeq2, though I still have much to learn. My group and I profiled the gene expression of the genes XYZ in the Drosophila midugt and I presented our poster at our university’s research day poster session.

Thank you for your time and consideration in reading my email. Please let me know if you are currently available to mentor an undergraduate in your lab and if you would be willing to have me as a potential student. I hope to hear from you soon!

Don’t be surprised if it takes the researcher a while to get back to you or if they respond curtly. They likely get many requests and are very busy. Continue to reach out to labs and apply to programs; it is difficult to know when opportunities are available as funding fluctuates.

## 12.3 BioDIGS

This section is primarily geared toward instructors as students will be able to get access to the BioDIGS data through their local C-MOOR Scholars chapter.



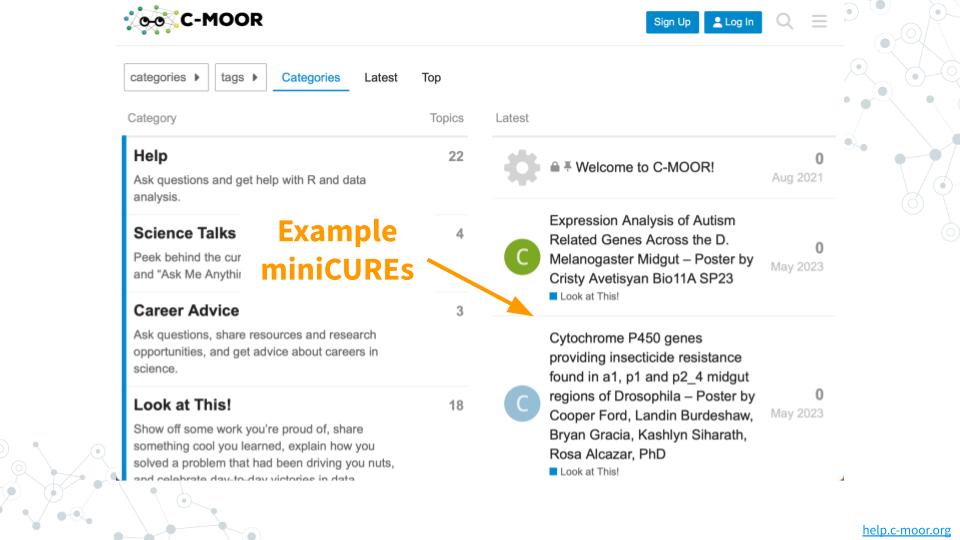
**What is BioDIGS?**: [BioDIGS](biodigs.org) is a project from the Genomic Data Science Community Network (GDSCN), which aims to characterize the microbiome of soil throughout sites in the US while connecting scientists to research.

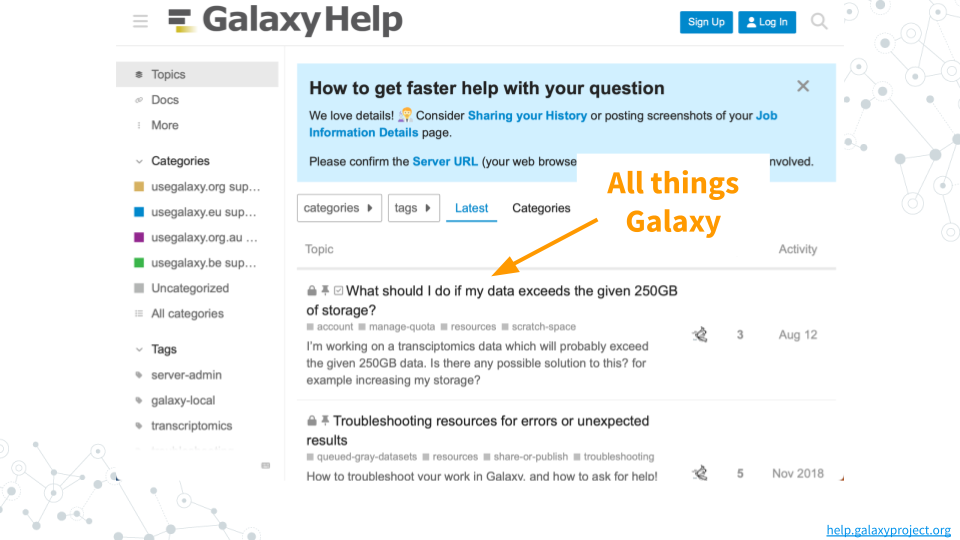
**What makes BioDIGS different?:** Soil is hypothesized to be the most diverse system on our planet, and BioDIGS is gathering sequence data and matching environmental data to elucidate the connection of soil microbes to abiotic variables and human health. Students at participating sites can take the entire project from soil sampling all the way through computational analysis. Curricula and more information on how to get a soil sampling kit are available on the BioDIGS website.

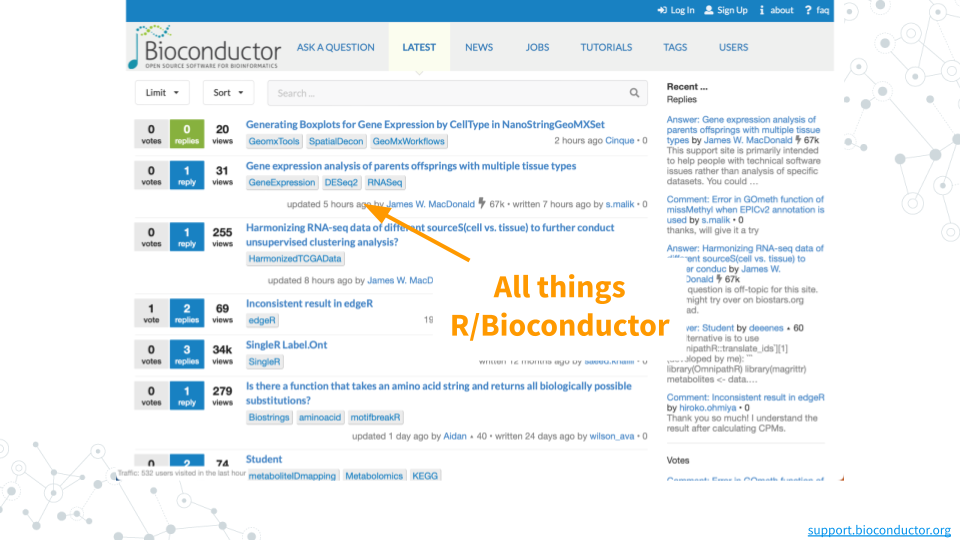
# 13 Online Community

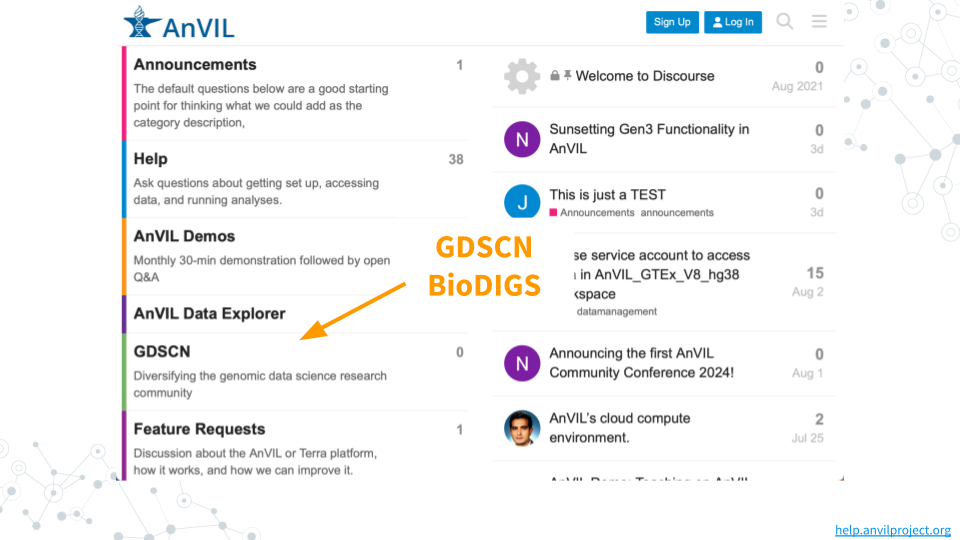
The genomic data science community is a remarkably open community that continually strives to increase opportunitites and broaden access to the scientific enterprise (e.g. [GDSCN 2022](https://pubmed.gov/35858750)). There are many opportunities to interact online with scientists that share a common research question, career stage, or even favorite tool! Below is a sampling of some discussion forums that you may find useful:

* [help.c-moor.org](https://help.c-moor.org) for help with and example miniCUREs
* [help.galaxyproject.org](https://help.galaxyproject.org) for all things Galaxy
* [support.bioconductor.org](https://support.bioconductor.org) for all things R/Bioconductor
* [help.anvilproject.org](https://help.anvilproject.org) for interactions with the GDSCN BioDIGS consortium <https://biodigs.org/#home>

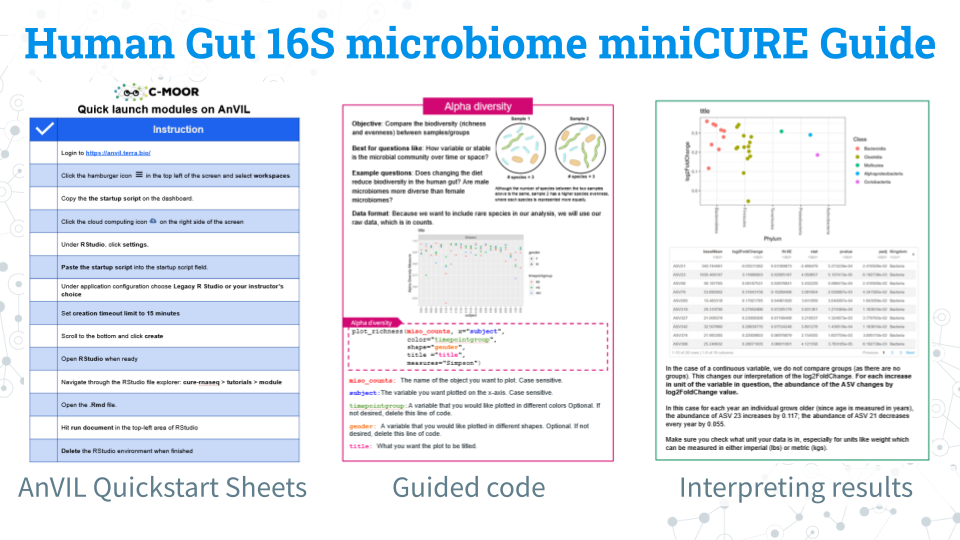








# 14 16S miniCURE Guide



We have a [comprehensive guide for this miniCURE](https://docs.google.com/presentation/d/1uipJ2uS2g8svKtWX-mN3MGvWYLWL9C0Pmv2nK6FjbCo/edit?usp=sharing)!

This guide is a complete packet that contains an overview of the learning objectives for ecah module, instructions for accessing modules on AnVIL, highlights of content, guided code blocks, and an overview of each dataset we have available for research projects!

Download the PDF for a digital reference and print out selected sheets for use in the classroom!

### 14.0.1 Footnotes

**Contributions and Affiliations**

* Sayumi York, Notre Dame of Maryland University

Last Revised: August 2025

# 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 2025-09-23  
## pandoc 3.1.1 @ /usr/local/bin/ (via rmarkdown)  
##   
## ─ Packages ───────────────────────────────────────────────────────────────────  
## package \* version date (UTC) lib source  
## bookdown 0.43 2025-04-15 [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 1.0.4 2025-06-18 [1] CRAN (R 4.3.2)  
## 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 1.50 2025-03-16 [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 1.1.6 2025-04-11 [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 0.52 2025-04-02 [1] CRAN (R 4.3.2)  
## xtable 1.8-4 2019-04-21 [1] RSPM (R 4.3.0)  
## yaml 2.3.10 2024-07-26 [1] CRAN (R 4.3.2)  
##   
## [1] /usr/local/lib/R/site-library  
## [2] /usr/local/lib/R/library  
##   
## ──────────────────────────────────────────────────────────────────────────────

# References