CURE: Microbial Mysteries

July 03, 2025

Table of Contents

# About

## 0.1 Course Description

Microorganisms surround us, from the microbes in our soil to the bacteria in our digestive systems. As part of the C-MOOR project <https://www.c-moor.org>, this Course-based Undergraduate Research Experience (CURE) introduces computational tools using Galaxy <https://usegalaxy.org> and SciServer <https://sciserver.org> to analyze metagenomic datasets that you will use to conduct independent projects. Dig into soil metagenomics datasets generated by the GDSCN BioDIGS consortium <http://biodigs.gdscn.org> and study the soil biodiversity and ecological factors that impact human health and ecological concerns or explore publicly accessible human gut metagenomics datasets <https://www.ncbi.nlm.nih.gov/bioproject> and study how our microbiomes are shaped by our diet, medications, and genetics.

This course progresses through a series of in-person and online learning experiences that introduce important concepts and tools in data analysis and genomics, necessary to subsequently complete a Capstone Project. Students work in groups and present frequently during class periods similar to research lab meetings in order to receive and provide community analysis and feedback.  
This course is open to all students, regardless of major. No prior computational experience necessary.

Microbial Mysteries CURE was offered at the Johns Hopkins University in the Spring of 2025 as AS.020.126.01.SP25 Microbial Mysteries: Genomics from Ground to Gut.

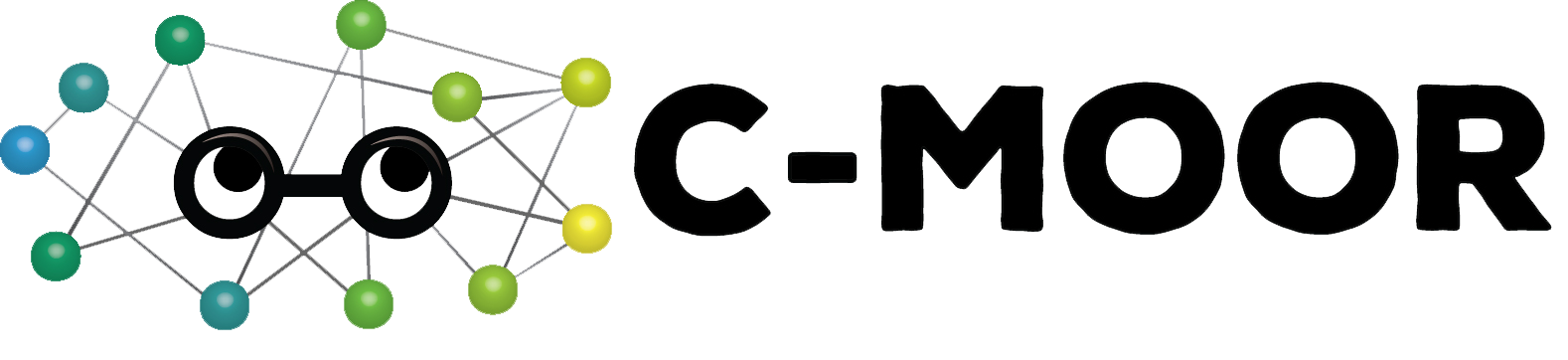


Figure 0.1: C-MOOR logo

### Audience and Prerequisites

This course is offered to any undergraduate student, with freshmen especially encouraged to take. The course is designed to introduce genomics and data science as apart of flexible reserach-based experience or CURE (Course-Based Undergraduate Research Experience). Students will learn to analyze metagenomic data from DNA sequencing using a variety of computational tools and resources. Students will then engage in their own project work using the concepts and tools they learned and utilizing publically available metagenomics datasets.

**Prerequisites**:

* Prereq 1: No biology presequisites are required, but basic concepts in biology and genetics are encouraged.
* Prereq 2: No programming, coding or data science experience are required - beginner level.
* Prereq 3: No research experience is required.

### 0.1.1 Format

This course is a combination of: - Lectures - Pre-labs and other homework assignments - Activities (homework or in-class) - On-line tutorials - Discussions - Presentations - Projects - Seminars (guest lectures) - Other

## 0.2 Course Goals

* **Navigate biological databases** - access the wealth of publically available scientific knowledge and data (e.g. NCBI).
* **Analyze Big Data** - e.g. plot and interpret results from next generation sequencing (NGS) experiment.
* **Gain skills in computational and bioinformatics data analysis** - e.g. web-based platforms or command-line bioingofmatics tools (e.g. Galaxy, R).
* **Practice the scientific process and critical thinking** - learing to read and evaluate scientific literature, ask questions, identify avenues for research, design experiments, analyze and interpret data, and integrating results into the broader scientific discourse.
* **Engage in scientific research** - Apply knowledge and skills learned to complete a project using existing datasets to ask new questions, generate novel conclusions and create and present a Final Poster.

### 0.2.1 Vision and Change

This course supports the change in vision of undergraduate science education by teaching through research experience, or Course-based Undergraduate Research Experiences (CUREs). This CURE is a genomic data science research experience with the curriculum that supports [Vision and Change](https://nabt.org/files/galleries/Vision-and-Change-Final-Report.pdf) national report that aims to improve and modernize undergraduate science education and promote competencies in the practices including:

- 1: Ability to apply the process of science  
 - 2: Ability to use quantitative reasoning  
 - 3: Ability to tap into the interdisciplinary nature of science  
 - 4: Ability to communicate and collaborate with other disciplines  
 - 5: Ability to understand the relationship between science and society

## 0.3 Available course formats

This course is available in multiple formats which allows you to take it in the way that best suites your needs. You can take it for certificate which can be for free or fee.

* The material for this course can be viewed without login requirement on this [Bookdown website](LINK%20HERE). This format might be most appropriate for you if you rely on screen-reader technology.
* This course can be taken for [free certification through Leanpub](LINK%20HERE).
* This course can be taken on [Coursera for certification here](LINK%20HERE) (but it is not available for free on Coursera).
* Our courses are open source, you can find the [source material for this course on GitHub](LINK%20HERE).

# Instructor Guide

C-MOOR Microbial Mysteries CURE content overview.

## 0.4 Course Content

The Microbial Mysteries course content was organized and presented by way of Unit modules below. Each unit has a main theme, reflected by the unit title. Practically, each unit is a week covering both, themed assignments and those from previous and next units, including any other additional material an instructor wishes to cover.

* Microbial Mysteries CURE was offered at the Johns Hopkins University over the Spring Semester of 2025 lasting a total of 13 weeks
* Each class/session lasted 1 hour and 15 minutes

Course content was organized broadly into 3 categories:

1. Structured (Core) Units - on weekly basis, Units 0 through 4  
2. Research Module (Project Work) - over the course of 7-8 weeks  
3. Professional Development Module - over the course of 4 weeks

* Core units 1, 2, 3 and 4, are fundamental knowledge and skills units that were taught over the coure of 4 weeks and 8 classes (1 unit per week and 2 classes/week)
* Each core unit was typically associated with 1.5-2.5 hours of homework per class, or 3-5 hours of homework per week. The amount of homework is based on:

1. For each hour of class students were generally expected to spend approximately 2 hours of independent study time (homework).
2. To ensure meaningful coverage of important concepts and scientific content and computational skills necessary to proceed to next step and subsequently execute independent group project work.

### 0.4.1 Structured Units

* Table below - shows an example of course organization for Core Units 0 through 4
* Spring 2025 - refers to **due date** for assignments based the Spring 2025 semester at the Johns Hopkins University.
* **Unit 0** - Introduction: Session 1 (Jan 21), Unit length (Jan 21 - Jan 22)

| Unit 0-1 | Type | Length | Spring 2025 |
| --- | --- | --- | --- |
| Lecture: Welcome to Your Genomics Adventure! | in-class | 20 min | Jan 21 |
| Create Accounts | Homework due | 30 min | Jan 21 |
| Lecture: The Scientific Process | in-class | 20 min | Jan 21 |
| Post Introductions | Homework next | 30 min | Jan 22 |

* **Unit 1** - Scientific Literature: Session 1 (Jan 23), Session 2 (Jan 28), Unit length (Jan 23 - Jan 30)

| Unit 1-1 | Type | Length | Spring 2025 |
| --- | --- | --- | --- |
| Lecture: What’s in your XYZ? | in-class | 20 min | Jan 23 |
| Activity: Taxonomy Profiling Spreadsheet | Homework due | 45 min | Jan 23 |
| Galaxy Tutorial: A short introduction to Galaxy | Homework due | 40 min | Jan 23 |
| Pre-lab: Scientific Literature | Homework next | 45 min | Jan 27 |
|  |  |  |  |
| **Unit 1-2** | **Type** | **Length** | **Spring 2025** |
| Discussion: Scientific Literature | in-class | 30 min | Jan 27 |
| Lecture: Scientific Literature | in-class | 20 min | Jan 28 |
| Activity: Scientific Literature | Homework next | 80 min | Jan 29 |

* **Unit 2** - Microbial Genomes: Session 1 (Jan 30), Session 2 (Feb 4), Unit length (Jan 30 - Feb 6)

| Unit 2-1 | Type | Length | Spring 2025 |
| --- | --- | --- | --- |
| **Presentation Unit 1**: Scientific Literature | in-class | 30 min | Jan 30 |
| Lecture: Microbial Genomes | in-class | 20 min | Jan 30 |
| Pre-lab: Microbial Genomes | Homework next | 90 min | Feb 3 |
| Galaxy Tutorial: QC & Galaxy Workflows | Homework next | 30 min | Feb 3 |
|  |  |  |  |
| **Unit 2-2** | **Type** | **Length** | **Spring 2025** |
| Discussion: Microbial Genomes | in-class | 30 min | Feb 4 |
| SciServer Tutorial: test-drive R | Homework due | 25 min | Feb 4 |
| Project: Microbial Genomes | Homework next | 90 min | Feb 5 |

* **Unit 3** - Taxonomy Profiling: Session 1 (Feb 6), Session 2 (Feb 11), Unit length (Feb 6 - Feb 13)

| Unit 3-1 | Type | Length | Spring 2025 |
| --- | --- | --- | --- |
| **Presentation - Unit 2**: Microbial Genomes Project | in-class | 30 min | Feb 6 |
| Lecture: Taxonomy Profiling | in-class | 20 min | Feb 6 |
| Pre-lab: Taxonomy Profiling | Homework next | 100 min | Feb 10 |
| Activity: Possible Datasets | Homework next | 50 min | Feb 10 |
| Galaxy Tutorial: Using dataset collections | Homework next | 45 min | Feb 10 |
|  |  |  |  |
| **Unit 3-2** | **Type** | **Length** | **Spring 2025** |
| Discussion: Taxonomy Profiling | in-class | 30 min | Feb 11 |
| Lecture: Possible Datasets | in-class | 20 min | Feb 11 |
| SciServer Tutorial: Meet R! | Homework due | 25 min | Feb 11 |
| Project: Taxonomy Profiling | Homework next | 100 min | Feb 12 |

* **Unit 4** - Finding AMRs: Session 1 (Feb 13), Session 2 (Feb 18), Unit length (Feb 13 - Feb 20)

| Unit 4-1 | Type | Length | Spring 2025 |
| --- | --- | --- | --- |
| **Presentation - Unit 3**: Taxonomy Profiling Project | in-class | 30 min | Feb 13 |
| Lecture: Finding AMRs | in-class | 20 min | Feb 13 |
| Pre-lab: Finding AMRs | Homework next | 90 min | Feb 17 |
| Lecture: Scientific Posters | in-class | 20 min | Feb 18 |
| Activity: Scientific Posters | Homework next | 50 min | Feb 17 |
| SciServer Tutorial: R for Data Science 1 | Homework | 50 min | Feb 17 |
|  |  |  |  |
| **Unit 4-2** | **Type** | **Length** | **Spring 2025** |
| Discussion: Finding AMRs | in-class | 30 min | Feb 18 |
| Project: Finding AMRs | Homework | 110 min | Feb 19 |
| SciServer Tutorial: R for Data Science 2 | Homework | 25 min | Feb 18 |

* **Unit 5** - Kickstart Project Work: Session 1 (Feb 20), Sessison 2 (Feb 25), Unit length (Feb 20 - Feb 27)

| Unit 5-1 | Type | Length | Spring 2025 |
| --- | --- | --- | --- |
| **Presentation - Unit 4**: Finding AMRs Project | in-class | 30 min | Feb 20 |
| Lecture: Kickstart Project Work | in-class | 20 min | Feb 20 |
| Activity: Kickstart Project Work | Homework next | 90 min | Feb 26 |

### 0.4.2 Research Module

Research Module is mainly dedicated to group project work, with intervening Professional Development Sessions such as Scientific Talks. During their project work that spans ~8 weeks, students do research work towards their project, prepare and update relevant documents (e.g. Written Doc, poster slide deck) and separately meet with instructor to get help with project direction and execution and ask questions.

* Students are expected to work on their project independently and with groups a minimum of 4 hours per week.
* When there is no assignment due, students are still expected to work on their projects and update their posters.
* Project Work activities consisted of **In-Class Round Table activities**, and **Written Updates + Advisory Meetings**.

**Table: List of Project Work activities**

| Project-related Activity | Format | Function |
| --- | --- | --- |
| Round Table Data | in-class slidedeck | Discuss data and ask questions |
| Round Table Troubleshooting | in-class slidedeck | Discuss methods and ask questions |
| Written Check-ins (Doc) | Doc text updates | Share project work updates on project work |
| Written Check-ins (Poster) | Poster updates | Update poster |
| Advisory Meetings | In-person with instructor | Discuss progress and issues |
| Peer-review (Doc) | Doc text updates | Provide peer-feedback on project and poster |

**Project Work In-Class Round Table**

Over the course of the Project work, students present:

* 4-5 times Round Table Data (once per week):
* 4-5 times Round Table Troubleshooting (once per week)
* Project work was intermixed with professional development (PD) activities including Science Talks, Scientific Communication activity, and Next Steps activity.

**Project Work Written Updates + Advisory Meetings**

Over the course of the Project work, students submit:

* 4 Written progress Check-ins (once per week)
* 2 Poster update Check-ins
* 2 Peer-review write-ups

**Project Work Sessions and Assignment overview**

February 25 | Session 1

* Round Table Data (slidedeck)

February 27 | Session 2

* Round Table Troubleshooting (slidedeck)

March 4 | Professional Development

* Science Talks: Guest Lecture and Q&A
* Homework: Work on Written Check-in #1 (Doc) (Due day before next class, March 5)

March 6 | Session 3 (slidedeck)

* Round Table Troubleshooting
* Homework: Work on Written Check-in #2 (Doc) (Due day before next class, March 10)

March 11 | Session 4

* Round Table Data (slidedeck)
* Homework: Work on Written Check-in #3 (Poster) (Due day before next class, March 12)

March 13 | Session 5

* Round Table Troubleshooting (slidedeck)
* Advisory Meeting #1 (Due next day, March 14)`

March 18

* Spring Break

March 20

* Spring Break

March 25 | Session 6

* Round Table Data (slidedeck)

March 27 | Sessison 7

* Round Table Troubleshooting (slidedeck)
* Homework: Work on Written Check-in #4 (Doc) (Due day before next class, March 31)

April 1 | Session 8

* Round Table Data (slidedeck)
* Homework: Work on Peer-review #1 (Doc) (Due day before next class, April 2)

April 3 | Session 9

* Round Table Troubleshooting (slidedeck)
* Advisory Meeting #1 (Due next day, April 4)`
* Homework: Work on Written Check-in #5 (Doc) (Due day before next class, April 7)

April 8 | Professional Development

* Science Talks: Guest Lecture and Q&A
* Homework: Work on Peer-review #2 (Doc) (Due day before next class, April 9)

April 10 | Session 10

* Round Table Troubleshooting (slidedeck)

April 15 | Session 11

* Round Table Data (slidedeck)

April 17 | Professional Development

* **Poster should be finalized and ready for printing**
* In class activity: Scientific Communication

April 22 | Professional Development

* **Project Work Final Poster Presentation**

April 24 | Professional Development

* Last Session
* Lecture: Next Steps
* In class activity: Next Steps

#### 0.4.2.1 Project Work Grading

| Project work activity | Format | points |  |
| --- | --- | --- | --- |
| Written Check-ins | Doc | each 50 pts |  |
| Advisory Meetings | In-person | each 50 pts |  |
| Peer-reviews | Doc | each 50 pts |  |
| Final Poster Presentation | 100 pts |  |  |

#### 0.4.2.2 Project Work Organizer

Use this organizer to find instructions and track documents.

* For setting up meeting times, we recommend using [when2meet](https://www.when2meet.com)

| Group | Check-in | Poster | Meeting #1 | Meeting # 2 |
| --- | --- | --- | --- | --- |
| Group A | Doc | Slides | Date 1 | Date 2 |
| Group B | Doc | Slides | Date 1 | Date 2 |
| Group C | Doc | Slides | Date 1 | Date 2 |
| Group D | Doc | Slides | Date 1 | Date 2 |

### 0.4.3 Professional Development Module

* **Weeks 9-13**

| Module: Project Work | Type |
| --- | --- |
| Activity: Science Talks | Homework |
| Science Talks | Guest Lecture |
| Activity: Science Talks | Homework |
| Science Talks | Guest Lecture |
| Activity: Scientific Communication | Homework |
| Activity: Next Steps | Homework |

## 0.5 Core Competencies

The activities in this CURE address several core concepts and competencies as identified by professional societies and working groups. The table below lists the relevant core concepts and competencies adopted from the following sources:

* [Scientific Thinking and Critical Thinking in Science Education](https://doi.org/10.1007/s11191-023-00460-5)
* [Genetics Learning Framework](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)
* [Establishing a baseline of science communication skills in an undergraduate environmental science course](https://stemeducationjournal.springeropen.com/articles/10.1186/s40594-021-00304-0)

| **Critical Thinking** |
| --- |
| - Enhancing higher-order cognitive skills that go beyond memorization |
| - Solve problems using quantitative and qualitative information and analysis |
| - Promote scientific evidence-based thinking and active learning |
|  |
| **Genetics** |
| - Locate, read, and comprehend primary literature research papers on genetics topics |
| - Implement observational strategies to formulate a question |
| - Generate testable hypotheses |
| - Generate and interpret graphs displaying experimental results |
| - Critique data and datasets and use bioinformatics to assess genetics data |
|  |
| **Bioinformatics** |
| - Explain the role of computation in addressing hypothesis-driven and hypothesis-generating questions within the life sciences |
| - Develop bioinformatics core competencies for undergraduate life sciences education |
| - Use bioinformatics tools to examine complex biological problems in genetics, evolution, information flow, and other important areas of biology |
| - Find, retrieve, and organize various types of biological data |
| use bioinformatics tools to analyze genomic and/or metagenomic data |
| - Explore and/or model biological interactions and networks a using bioinformatics |
|  |
| **Scientific Communication** |
| - Communicate experimental results effectively, including written summaries, verbal presentations and poster presentation |
| - Understand Essential Elements for Effective Science Communication (EEES) such as Who? (is the audience), Why? (e.g. Purpose), What? (content) and How? (e.g. style). |

## 0.6 What to know

As an instructor,

* What do I have to know to teach this course?
* What level do I need to be at?
* I am not a genomics expert, can I do it?
* I don’t know coding, can I do it?

## 0.7 Where to get help

As an instructor, where can I get help in teaching this course?

* Ask C-MOOR community for help [<https://help.c-moor.org>]

## 0.8 Grades

| % Effort | Content | Points |
| --- | --- | --- |
| **Core Labs** |  |  |
| 20% | Assignments | 200 pts |
| 20% | Class Discussions | 200 pts |
| **Project Work** |  |  |
| 20% | Written Check-ins | 200 pts |
| 10% | Advisory Meetings | 100 pts |
| 10% | Peer Reviews | 100 pts |
| 10% | Final Presentation | 100 pts |
| **Professional Development** |  |  |
| 10% | Portfolio Dev + Science Talks | 100 pts |

# 1 Introduction

## 1.1 Welcome to Your Genomics Adventure

[Slides: Welcome to Your Genomics Adventure!](https://docs.google.com/presentation/d/18hYo8xrYkyq3rG7RUy3n3-jWFK_JPs5NNW3HmM9HUeQ/edit?usp=sharing)

## 1.2 Activity - Create Accounts

### 1.2.1 Purpose

Over the course of this semester, we will use the following online platforms: - [C-MOOR Academy Discussion Forum](https://help.c-moor.org) – Join the community to get help and share your findings - [Google Docs](https://workspace.google.com/products/docs) – Collaborate on assignments and scientific posters - [Galaxy](https://usegalaxy.org) – Analyze data with >10,000 tools using a graphical user interface - [SciServer](https://sciserver.org) – Access virtual machines preinstalled with RStudio, Bioconductor, and more

### 1.2.2 Activity

*Estimated time: 30 min*

#### 1.2.2.1 Instructions

Create accounts on the following online platforms:

1. C-MOOR Academy Discussion Forum – <https://help.c-moor.org>

* Submit your username using [this form](https://docs.google.com/forms/d/e/1FAIpQLSctd0jPax7Ww9b9XGbzY0PTwmPgm6VQICmsOhVTl6OCDx18Hw/viewform)

1. Google Docs – <https://docs.google.com>

* Test by opening [tax-data-gut.tsv](https://drive.google.com/file/d/1vL6adVIrqxpONbae8rUsneK3tbdCpmR-) with Google Sheets

1. Galaxy – <https://usegalaxy.org>
2. SciServer – <https://sciserver.org>

#### 1.2.2.2 Questions

Fill out your username and insert a screenshot of that username in the boxes below.

| 1. C-MOOR Academy Discussion Forum. |
| --- |
| <Username: insert screenshot> |

| 2. Google Docs. |
| --- |
| <Username: insert screenshot> |

| 3. Galaxy. |
| --- |
| <Username: insert screenshot> |

| 4. SciServer. |
| --- |
| <Username: insert screenshot> |

#### 1.2.2.3 Grading Criteria

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

#### 1.2.2.4 Footnotes

**Resources**

* Google doc

**Contributions and Affiliations**

* Frederick Tan, Johns Hopkins University

Last Revised: January 2025

## 1.3 The Scientic Process

[Slides: The Scientic Process](https://docs.google.com/presentation/d/1VQE-rXASXIdf8rWLP5UTcrAhM_DcznARVigPTHTfw8M/edit?usp=sharing)

## 1.4 Homework - Post Introductions

### 1.4.1 Purpose

The purpose of this assignment is to learn how to post to the C-MOOR Academy Discussion Forum. This forum will be the primary place for students, their instructors, and experts to communicate about the student’s research project, professional development opportunities, and more.

### 1.4.2 Learning Objectives

1. Learn how to post to a discourse community.
2. Examine the differences between private and public discourse categories.

### 1.4.3 Introduction

Before beginning this assignment, you should have already made an account on the C-MOOR Academy Discussion Forum and been added to the course Category by your instructor. Within the C-MOOR community, you will find Categories, which can be either private or public. In this course we will use both. In our private category, students can talk to each other and their instructor and their conversations will not be publicly available. In the public category, students’ posts will be visible on the web. The value of public discourse communities is that you might get responses by experts that you do not know. In this assignment, students will post to both a public and private Category within the Community.

### 1.4.4 Activity 1 - Create a Topic in a Private Category

*Estimated time: 20 min*

#### 1.4.4.1 Instructions

1. Visit the C-MOOR Academy Discussion Forum (help.c-moor.org) and log in.
2. It is a good idea to bookmark this page so that you can easily access it throughout the course.
3. Read through the categories in the C-MOOR Academy Discussion Forum. Notice that some categories (at least one) have a lock next to them. This is a Private category, only visible to you, your classmates and your instructor. If you cannot see your course’s private category when you log in, email your instructor so that they can add you.
4. Click on the private category that belongs to your class: Spring 2025. In this private channel, we will add a “New Topic” in which you can introduce yourself to the class.
5. This first post will be an introduction to your class, including a bit about you and a photo.

**a. Tell your group about yourself. Answer the following questions.** 1. What is your name? 2. Why did you decide to take this class? What are you excited to learn about this semester? 3. Tell us one thing you like to do outside of school and work.

**b. Include a photo:** 1. Selfie: if you would like, post a photo of yourself. 2. Unselfie: or post a photo of something else. Maybe something that you feel represents your life right now, or a picture of something you love. A pet maybe. Just make sure it tells us something about you.

1. Play around with the platform to personalize your post. As you can see in the discourse topic, you can add emojis, upload images, embed content, etc. These features appear in plain text as you’re typing, but you can see a preview of your post on the right. For your first few posts you may see a Welcome Box on the right instead, you can close this to see the preview. Once you have posted a few times, the Welcome Box will stop appearing.
2. When you are done, click “Create Topic” and it will post for the class to see.
3. If other students have posted their introductions, read them and leave a couple replies. This is a discussion platform after all!

#### 1.4.4.2 Resources

* [C-MOOR Academy Discussion Forum](https://help.c-moor.org)
* [How to add a bookmark in Chrome](https://support.google.com/chrome/answer/188842?co=GENIE.Platform%3DDesktop&hl=en)
* [Discourse New User Guide](https://meta.discourse.org/t/discourse-new-user-guide/96331)

### 1.4.5 Activity 2 - Reply to a Topic in a Public Category

*Estimated time: 10 min*

#### 1.4.5.1 Instructions

1. Navigate back to the C-MOOR Academy Discussion Forum front page.
2. Click on the “Breakroom” Category. This is an informal category where we can chat about non-science related topics.
3. Look for a topic called “Enduring a Snow Storm” Click on the topic.
4. Read the topic and then click “Reply”.
5. The same box will pop up as you used to to post a topic earlier.
6. Write a reply to the topic.
7. When you are finished click “Reply” to post.

#### 1.4.5.2 Resources

* [C-MOOR Academy Discussion Forum](https://help.c-moor.org)
* [Discourse New User Guide](https://meta.discourse.org/t/discourse-new-user-guide/96331)

#### 1.4.5.3 Grading Criteria

* Submit URL to your Private Category Topic on Canvas

#### 1.4.5.4 Footnotes

**Contributions and Affiliations**

* Katherine Cox, Johns Hopkins University
* Valeriya Gaysinskaya, Johns Hopkins University
* Frederick Tan, Johns Hopkins University
* Stephanie Coffman, Clovis Community College

Last Revised: January 2025

# 2 Scientific Literature

## 2.1 Lecture - What’s in Your XYZ



[Slides: What’s in Your XYZ?](https://docs.google.com/presentation/d/1ph3LFw6i_mtv6ZJXssTf0-im7PhgV4FRJslDG0ICCws/edit?usp=sharing)

## 2.2 Pre-lab - Scientific Literature

### 2.2.1 Purpose

Obtain a high level overview of metagenomics by reading R.D. Sleator, C. Shortall, and C. Hill. Metagenomics. Letters in Applied Microbiology. 2008 Nov;47(5):361-6. [(pubmed.gov/19146522)](http://pubmed.gov/19146522)

### 2.2.2 Learning Objectives

* Read a review paper that summarizes the field of metagenomics.
* Broadly understand the scope of the review and the gaps in the field.

### 2.2.3 Introduction

The vast majority of all micro-organisms on Earch remain uncultured [(K.G. Lloyd et al, 2019)](https://journals.asm.org/doi/10.1128/msystems.00055-18). Additionally, in complex environments like soil and water, most micro-organisms remain unidentified [(M. Delgado-Baquerizo, 2019)](https://doi.org/10.1038/s41396-019-0405-0). The field of metagenomics is a culture-independend approach which aims to remedy these gaps in knowledge [(J. Handelsman, 2004)](https://pubmed.ncbi.nlm.nih.gov/15590779/). Metagenomics is the study of genomic (sequencing) data obtained directly from environmental (and other, e.g. clinical) samples and provides new meaningful information on the diversity and function of microorganisms.

### 2.2.4 Activity

*Estimated time: 50 min*

#### 2.2.4.1 Instructions

Read the review paper “Metagenomics” by Sleator, Shortall, and Hill, 2008 Lett Appl Microbiol and answer the following questions.

#### 2.2.4.2 Questions

1. What is one thing you learned or find interesting in the paper?
2. Define a term that is new to you (e.g. metagenome, microbiome, 16S rRNA).
3. Ask a question about the review paper.

#### 2.2.4.3 Grading Criteria

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

### 2.2.5 Footnotes

**Resources**

[Google Doc](https://docs.google.com/document/d/1-ruTySaAnSE-_5d6_LTdre4UmmUAx6TxMrTYBW-f3jQ/edit?usp=sharing)

**Contributions and Affiliations**

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

## 2.3 Discussion - Scientific Literature

### 2.3.1 Activity

*Estimated time: 25 min*

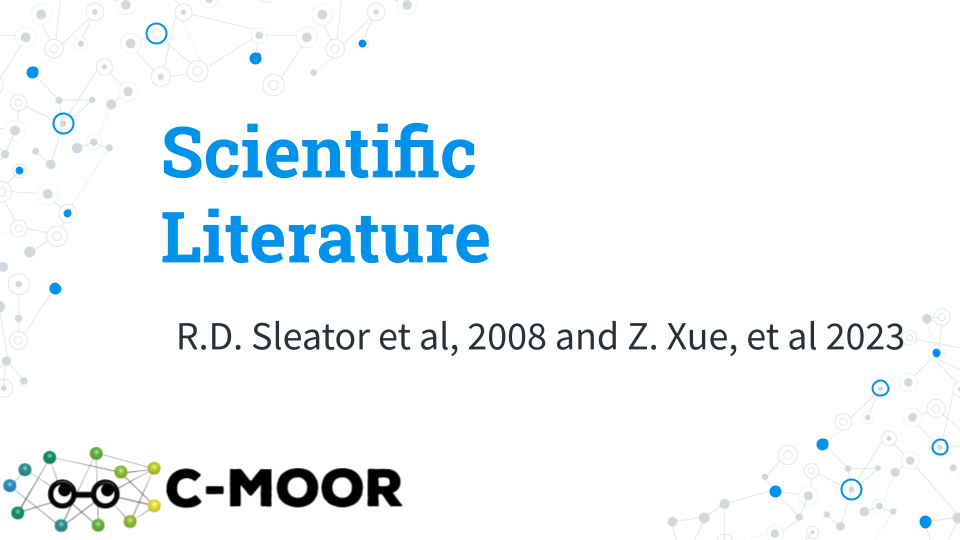
#### 2.3.1.1 Instructions

1. Form groups of four

* Add names to the “Microbial Mysteries Groups” sheet <https://docs.google.com/spreadsheets/d/11eoJgm9mehxGWWzh8IZYDCDmnCmSyshopPYHewvpC8c/edit?usp=sharing>

1. Pair up into groups (10 min)
2. Discuss – Each group member briefly describes answers to prelab assignment
3. Summarize – Identify best answer and add to slidedeck <https://docs.google.com/presentation/d/1nuA_gnL09_BPZVNJy5seL-HWejhMFLqILsmgmQztYKo/edit?usp=sharing>
4. Share group discussion (2 min each group)

## 2.4 Lecture - Scientific Literature



[Slides: Scientific Literature](https://docs.google.com/presentation/d/1zbjroITjBYmu-oxFT0qmOCx9LeEcSoaYrp_sWuy1OLM/edit?usp=sharing)

## 2.5 Activity - Scientific Literature

### 2.5.1 Purpose

Examine research on metagenomic diversity by reading Xue, *et al*. Metagenome sequencing and 103 microbial genomes from ballast water and sediments. Scientific Data.2023 Aug 10;10(1):536. [(pubmed.gov/37563185)](https://pubmed.ncbi.nlm.nih.gov/37563185/)

### 2.5.2 Learning Objectives

* Understand the purpose and experimental setup of the paper
* Understand the presented evidence (Figures and Tables) of the paper

### 2.5.3 Introduction

Understanding microbial composition and diversity in different environments is critical for assessing the benefits and threats of the bacterial community in that environment. In the publication by [Xue, et al. 2023](https://pubmed.ncbi.nlm.nih.gov/37563185/), the authors study microbial diversity in the ballast-tank water from two ships, with the idea that such a unique and isolated water environment may select for specific microbes. Luckily in their research they don’t find bacterium *Vibrio cholerae*, but that is exactly what they would find in the ballast water of cargo ships if they did the analysis during the cholera pandemic(s) of the 1800s.

### 2.5.4 Activity

*Estimated time: 90 min*

#### 2.5.4.1 Instructions

Based on the study by Xue, et al.2023, answer the following questions. The main text of the paper and the supplement can be found below, in the ‘Resources’ section of this assignment.

#### 2.5.4.2 Overview of the Paper (in class)

Determine the main objectives and purpose of the paper. Read the Abstract and the introduction with your group.

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?

#### 2.5.4.3 Methods (in class and homework)

1. Discuss how many and what samples were used for this study? Are there any replicates?
2. Discuss some methods used in this paper.
3. Discuss steps authors used to ensure their data is available to the public.

#### 2.5.4.4 Figures (in class and Homework)

**Methods.** How did the researchers test their hypothesis? Explain in your own words the methods in each figure.

| Figure | Methods |
| --- | --- |
| Fig. 1B | in class |
| Fig. 1C | in class |
| Fig. 2 | homework |
| Fig. 3A | homework |
| Fig. 3B | homework |
| Fig. 3C | homework |

### 2.5.5 Results (in class and Homework)

**Results.** What are the main findings from each figure?

| Figure | Main Findings |
| --- | --- |
| Fig. 1B | in class |
| Fig. 1C | in class |
| Fig. 2 | homework |
| Fig. 3A | homework |
| Fig. 3B | homework |
| Fig. 3C | homework |

#### 2.5.5.1 Conclusions (Homework)

1. Read the discussion section. What were the main conclusions the authors made in this study?
2. Do the figures agree with their conclusion?

#### 2.5.5.2 Future Directions (Homework)

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?

1. 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?
* Do you think there are risks associated with such studies?

### 2.5.6 Grading Criteria

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

### 2.5.7 Footnotes

**Resources**

[Google Doc](https://docs.google.com/document/d/1kKnvMGq8jBfwKzC7W5YEJ7CtTgFNahAaDfV3LNknznM/edit?usp=sharing)

**Contributions and Affiliations**

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

## 2.6 Presentation - Scientific Literature

### 2.6.1 Activity

*Estimated time: 25 min*

#### 2.6.1.1 Instructions

1. Open the “Scientific Literature Presentation” slidedeck [here](https://docs.google.com/presentation/d/1wwfiTSgm0ialrYlBFnA21zdZsw7mVA58YPMiTnDGCrc/edit?usp=sharing)
2. For your assigned figure from Xue, et al., 2023, create two slides to present methods and results
3. Add bullet points on key details you understand (Notice) and questions you have (Wonder)
4. Search for and insert at least one additional image that relates to your figure
5. Update slide title to summarize your main takeaway
6. Present at next class (5 min each group)

### 2.6.2 Footnotes

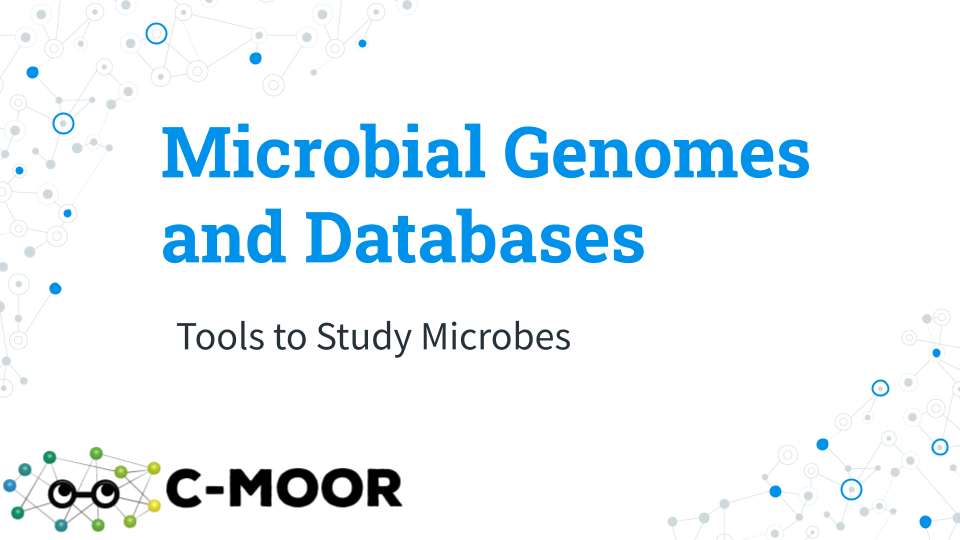
**Contributions and Affiliations**

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

Last Revised: January 2025

# 3 Microbial Genomes

## 3.1 Lecture - Microbial Genome



[Slides: Microbial Genomes](https://docs.google.com/presentation/d/1bnFhIIu6ZXSCjlrz5qWc7BZNz-7VfYWkKgbz3kL-ZNU/edit?usp=sharing)

## 3.2 Pre-lab - Microbial Genomes

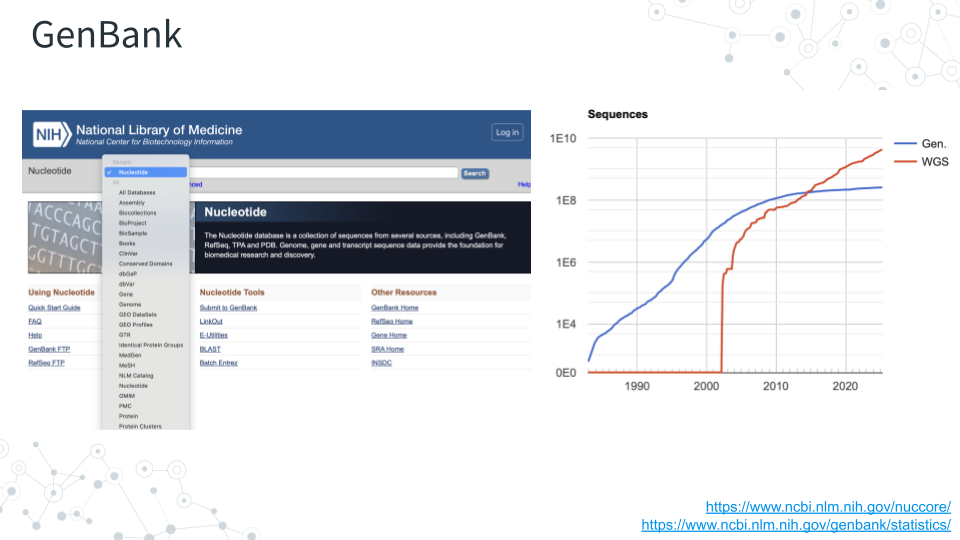
### 3.2.1 Purpose

Impress all the information that is freely available about well studied (and not so well studied) bacterial species. Start with E. coli as one of the best and longest studied before going into recently discovered bacteria (with less information).

### 3.2.2 Learning Objectives

1. GenBank – Explore sequence database of all publicly available DNA sequences
2. Sequence Browser – Observe genome organization using graphical representation
3. Bacterial database BV-BRC - Explore bacterial genes and their function
4. Taxonomy Browser – Identify relationships between taxa
5. Lifemap - Tree of Live viewer to visualize relationships between taxa

### 3.2.3 Activity 1 – GenBank



#### 3.2.3.1 Activity 1 - Part I

*Estimated time: 5 min*

#### 3.2.3.2 Instructions

Navigate to **GenBank** <https://www.ncbi.nlm.nih.gov/genbank> and enter ‘E. coli strain K-12’ in the search bar (which should default to **Nucleotide**). This will give you a list of almost 300,000 E. coli sequences. Go down the list and click on the entry that corresponds to the complete genome of the substrain MG1655, which should be on page 1! Click on the selection and answer the following questions.

#### 3.2.3.3 Questions

| **1. What is the size of this *E. coli* genome in bp and Kbp**? |
| --- |
|  |

| **2. Is the *E. coli* genome linear or circular**? |
| --- |
|  |

| **3. What is the ACCESSION number?** |
| --- |
|  |

#### 3.2.3.4 Activity 1 - Part II

*Estimated time: 5 min*

#### 3.2.3.5 Instructions

Back in GenBank <https://www.ncbi.nlm.nih.gov/genbank>, under the **Nucleotide** search tab, instead of typing ‘E. coli strain K-12’, type in the ACCESSION number you found (NZ\_CP169634).

#### 3.2.3.6 Questions

| **1. How many records do you observe after an ACCESSION number entry?** |
| --- |
|  |

| **2. What can you conclude about what is the ACCESSION number?** |
| --- |
|  |

### 3.2.4 Activity 2 – Genomes, genes, and other databases (BV-BRC)

#### 3.2.4.1 Activity 2 - Part I

*Estimated time: 5 min*

#### 3.2.4.2 Instructions

After entering your accession number, the on the right of the page, under **Related Information**, click on Assembly link to explore genome assembly information.

#### 3.2.4.3 Questions

| **1. How many chromosomes does E. coli have?** |
| --- |
|  |

| **2. What is the genome coverage of this sequenced genome?** |
| --- |
|  |

| **3. How many genes were annotated for this genome?** |
| --- |
|  |

#### 3.2.4.4 Activity 2 - Part II

*Estimated time: 10 min*

#### 3.2.4.5 Instructions

Go back and on the top of the page and click on **Graphics** to explore genome browser.

1. Hover along one of many green vertical sticks.

* You can also zoom into a smaller and smaller genomic region for higher resolution.

1. Hover along one of many red vertical sticks.

* You can also zoom into a smaller and smaller genomic region for higher resolution.

#### 3.2.4.6 Questions

| **1. What do the ‘green sticks’ represent?** |
| --- |
|  |

| **2. What do the ‘red sticks’ represent?** |
| --- |
|  |

| **3. Record 5 genes you found present in E. coli**. |
| --- |
| Gene 1 |
| Gene 2 |
| Gene 3 |
| Gene 4 |
| Gene 5 |
|  |

#### 3.2.4.7 Activity 2 - Part III

*Estimated time: 10 min*

#### 3.2.4.8 Instructions

To learn more about the genes of interest and their function scientists often use specialized databases. One such bacterial database is BV-BRC <https://www.bv-brc.org>. Use BV-BRC to find information about some of the E. coli genes. If there are 0 results then indicate No results found.

**For the 3 genes below, and one gene of your choice from activity above**, in the BV-BRC **Search** space,

**a)** from a dropdown menu select **“Pathways”** ,

**b)** type in the gene name and **click enter**. This will result in a lot of entries for different organisms.

**c)** To retrieve information specifically for *E. coli (Escherichia coli)*, in the **Keyword** space type “Escherichia”.

**d)** Check one of the boxes corresponding to *E. coli* strains and enter below which E. coli strain **(Genome Name)** you selected. If no E. coli entry is present, select and record another bacterial Genome Name to learn about your gene’s function.

#### 3.2.4.9 Questions

| **1. Record Genome Name associated with the following genes**. |  |
| --- | --- |
| **Gene ID** | **Genome Name** |
| Gene 1: ampC | Escherichia coli 07798 |
| Gene 2: mgtA |  |
| Gene 3: cdd |  |
| Gene 4: Your gene |  |
|  |  |

| **2. For the 3 genes from the activity above, record gene Product**. |  |
| --- | --- |
| **Gene ID** | **Gene Product** |
| Gene 1: ampC | Beta-lactamase |
| Gene 2: |  |
| Gene 3: |  |
|  |  |

| **3. For the 3 genes from the activity above, record Pathway Name (relates to function)**. |  |
| --- | --- |
| **Gene ID** | **Pathway Name** |
| Gene 1: ampC | beta-Lactam resistance |
| Gene 2: |  |
| Gene 3: |  |
|  |  |

**Note**, For some well characterized genes you can additionally obtain more detailed information about the gene/protein function.

* To do so, from the ‘green’ menu on the right, select **FEATURE** option, to learn from the Special Properties section about a special property of your gene.

### 3.2.5 Activity 3 - Taxonomy and tree of life

*Estimated time: 30 min*

#### 3.2.5.1 Activity 3 - Part I

#### 3.2.5.2 Instructions

1. Back in NCBI enter the *E. coli* accession number again. Under **Related Information** on the right, click on **Taxonomy** and then on the provided link for the *E. coli*.
2. Find Lineage information. Full Lineage information contains 7 core taxonomy ranks: Kingdom, Phylum, Class, Order, Family, Genus and Species, plus any additional classification ranks. To just get the 7 core lineage names, click on Lineage link for the abbreviated Lineage, or, simply hover over lineage names.

#### 3.2.5.3 Questions

| 1. **Record 7 core taxonomy ranks below**. |  |
| --- | --- |
| Kingdom: |  |
| Phylum: |  |
| Class: |  |
| Order: |  |
| Family: |  |
| Genus: |  |
| Species: |  |
|  |  |

#### 3.2.5.4 Activity 3 - Part II

#### 3.2.5.5 Instructions

1. As with the BV-BRC database above, we can use another database called **Lifemap** to visually explore the *E. coli* in the context of the tree of life. Go to <https://lifemap-ncbi.univ-lyon1.fr>, type E. coli and click species tab.
2. On the tree map, the yellow tag will indicate *E. coli*. Use plus and minus tabs to zoom in and out and visualize E. coli relative to other organisms on the map.

#### 3.2.5.6 Questions

| **1. How many Domains of life are there?**. |
| --- |
|  |

| **2. Zoom into and find nodes for the E. coli Genus, Family, Order, Class and Phylum. Do they match what you found in the NCBI Taxonomy Browser for Lineage?**. |
| --- |
|  |

| **3. What are some other members of the Genus (Escherichia) to which E. coli belongs?**. |
| --- |
|  |

| **4. What are some other members of the Family (Enterobacteriaceae) to which E. coli belongs?**. |
| --- |
|  |

| **5. What are some other members of the Order (Enterobacterales) to which E. coli belongs?**. |
| --- |
|  |

| **6. What are some other members of the Class (Gammaproteobacteria) to which E. coli belongs?**. |
| --- |
|  |

| **7. What are some other members of the Phylum (Pseudomonadota) to which E. coli belongs?**. |
| --- |
|  |

### 3.2.6 Grading Criteria

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

### 3.2.7 Footnotes

**Resources**

* Google Doc

**Contributions and Affiliations**

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

Last Revised: February 2025

## 3.3 Discussion - Microbial Genomes

### 3.3.1 Activity

*Estimated time: 25 min*

#### 3.3.1.1 Instructions

1. Form new groups of four [here](https://docs.google.com/spreadsheets/d/11eoJgm9mehxGWWzh8IZYDCDmnCmSyshopPYHewvpC8c/edit?usp=sharing).
2. Pair up into groups (10 min).
3. Discuss – Each group member briefly describes answers to prelab assignment.
4. Summarize – Identify best answer and add to slidedeck [here](https://docs.google.com/presentation/d/1PxqLt-MDFCTSBWm0-j-mY268b3YUiPMRqYMtcJ5JDL4/edit?usp=sharing)
5. Share group discussion (2 min each group)

### 3.3.2 Footnotes

**Contributions and Affiliations**

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

Last Revised: February 2025

## 3.4 Project - Microbial Genomes

### 3.4.1 Purpose

Explore information about new bacterial MAGs from Zue Z et al, Nature Scientific Data 2023 <https://pubmed.ncbi.nlm.nih.gov/37563185/>.

### 3.4.2 Learning Objectives

1. Explore and better understand MAGs and contigs by following up one MAG and one contig in NCBI.
2. Utilize knowledge learned by navigating MAG information using previously learned tools such as GenBank, Sequence Browser, BV-BRC, Taxonomy Browser and Lifemap.
3. Go deeper and see what you can find on the organism from your MAG of choice and a gene from a MAG of choice using any tools available to you (Pubmed, Google, BV-BRC, other databases).

### 3.4.3 Activity 1 – MAGs and Taxonomy

*Estimated time: [40] min*

#### 3.4.3.1 Activity 1 - Part I - Explore a MAG

#### 3.4.3.2 Instructions

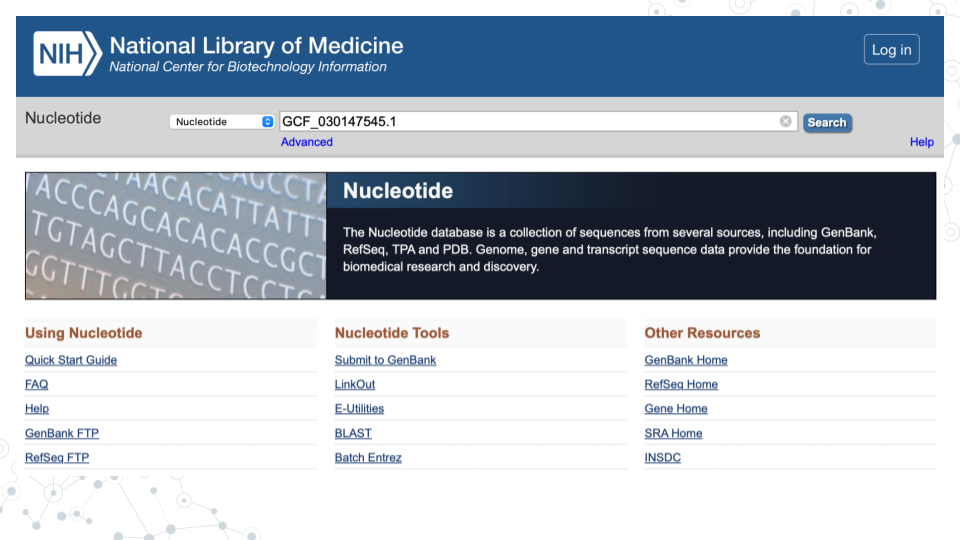
* In the research study by Zue Z et al, Nature Scientific Data 2023 <https://pubmed.ncbi.nlm.nih.gov/37563185/>, 17 of 103 uncovered MAGs from ballast water or sediment were of very high quality and completeness. You will follow up on one of them!
* **Choose one of these 17 MAGs** (see 17 GenBank IDs below) to follow up in this activity.
* Use GenBank <https://www.ncbi.nlm.nih.gov/nucleotide/> to answer questions.

| GenBank IDs for 17 quality MAGs |
| --- |
| 1. GCF\_030147545.1 |
| 2. GCF\_030148435.1 |
| 3. GCF\_030149225.1 |
| 4. GCF\_030148195.1 |
| 5. GCF\_030148855.1 |
| 6. GCF\_030148245.1 |
| 7. GCF\_030149045.1 |
| 8. GCF\_030148385.1 |
| 9. GCF\_030147715.1 |
| 10. GCF\_030149085.1 |
| 11. GCF\_030149145.1 |
| 12. GCF\_030148125.1 |
| 13. GCF\_030149425.1 |
| 14. GCF\_030149235.1 |
| 15. GCF\_030148515.1 |
| 16. GCF\_030147875.1 |
| 17. GCF\_030149465.1 |

#### 3.4.3.3 Questions

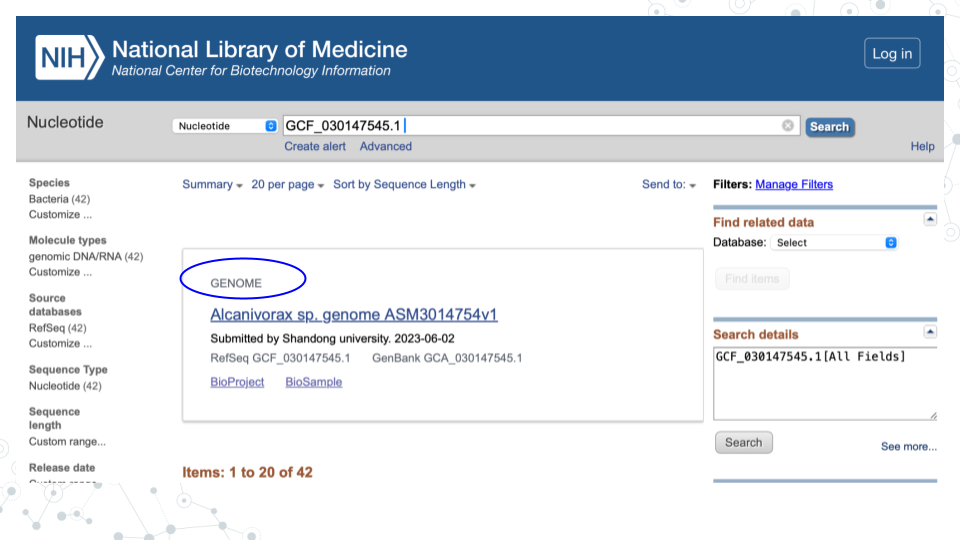
| **1. Record the GenBank ID number for one of 17 high quality MAGs for follow up below**. |
| --- |
|  |

In the GenBank <https://www.ncbi.nlm.nih.gov/nucleotide/>, under the Nucleotide search tab **enter** the GenBank number of your MAG (e.g. GCF\_030147545.1) and click **Search**.



| **2. Record below the GENOME name for the MAG assembly associated with the GenBank ID you entered.** |
| --- |
|  |

For the 1st MAG (GCF\_030147545.1), the GENOME name is “Alcanivorax sp. genome ASM3014754v1”:



Click on the GENOME name of your MAG to **explore genome assembly summary** information.

| **3. What is the Taxon of your MAG?** |
| --- |
|  |

| **4. What is the genome size of your MAG?** |
| --- |
|  |

| **5. How many contigs contributed to your MAG assembly?** |
| --- |
|  |

| **6. How many genes were annotated?** |
| --- |
|  |

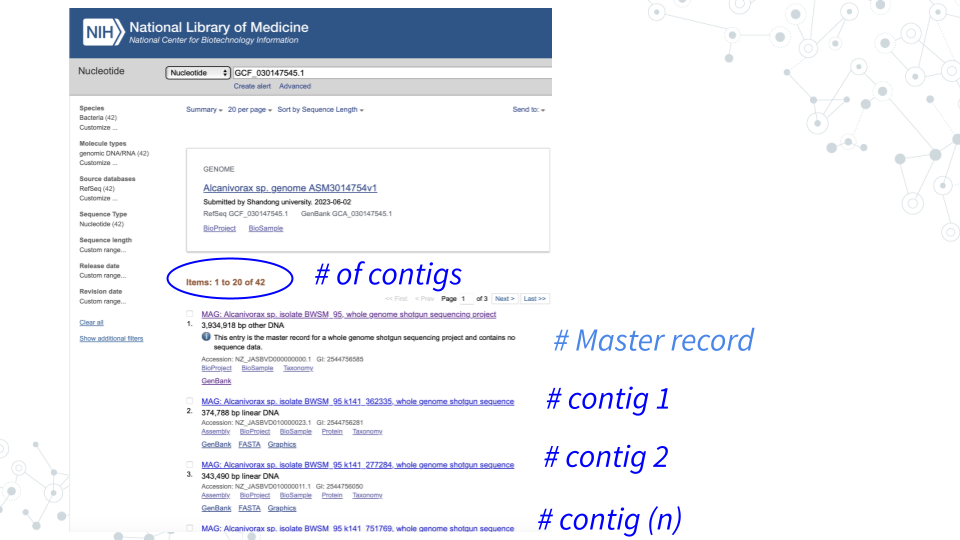
| **7. How comparable is your MAG genome to the E. coli genome from the microbial-genomes-pre-lab, in terms of genome size and number of genes?** |
| --- |
|  |

| **8. As you can see, MAGs are made up of Contigs. Based on your lecture and reading material, and this exercise, in your own words define MAGs and Contigs below.** |
| --- |
| MAGs: |
| Contigs: |
|  |

#### 3.4.3.4 Activity 1 - Part II - Explore a Contig

#### 3.4.3.5 Instructions

* MAGs are made up of Contigs. To see which contigs make up your MAG, go back to GenBank <https://www.ncbi.nlm.nih.gov/nucleotide/>, and search for your MAG using its GenBank ID (e.g.GCF\_030147545.1).
* You will see that MAG GCF\_030147545.1 is composed of 42 contigs:



| **1. Each contig has its own accession number. Choose a contig of reasonably large size (> 75 kb), click on the contig, then find and record below the Contig’s ID (Accession number/GenBank ID) for further examination.** |
| --- |
| Tip: You can sort the entries by length via’ Sort by Sequence Length’ on top! |
|  |

1. For the contig you chose in activity above, under **Related Information** on the right, click on **Taxonomy** and **then click again on the provided link**.
2. Find Lineage information. Full Lineage information contains 7 core taxonomy ranks: Kingdom, Phylum, Class, Order, Family, Genus and Species, plus any additional classification ranks. To just get the 7 core lineage names, click on Lineage link for the abbreviated Lineage, or, simply hover over lineage names.

| **2. Record 7 core taxonomy ranks for your Contig.** |  |
| --- | --- |
| Kingdom: |  |
| Phylum: |  |
| Class: |  |
| Order: |  |
| Family: |  |
| Genus: |  |
| Species: |  |
|  |  |

#### 3.4.3.6 Activity 1 - Part III - Visualize contig in a tree of life

#### 3.4.3.7 Introduction

1. Use Lifemap to visually explore the contig taxonomy in the context of the tree of life. Go to <https://lifemap-ncbi.univ-lyon1.fr> and **enter the lowest taxonomy rank observed for your contig** (most likely the species or genus level, but can also correspond to order or family).
2. On the tree map, use plus and minus tabs to zoom in and out and visualize your Contig entry relative to other organisms on the map. Zoom in and find nodes corresponding to the higher taxonomic ranks. For example, if your contig corresponds to genus level classification, you will not be able to identify species level information, but you will be able to identify the corresponding Family, Order, Class and Phylum.

| **1. Record 7 core taxonomy ranks for your Contig.** |  |
| --- | --- |
| Kingdom: |  |
| Phylum: |  |
| Class: |  |
| Order: |  |
| Family: |  |
| Genus: |  |
| Species: |  |
|  |  |

| **2. What are some other members of the Genus to which your Contig belongs?** |
| --- |
|  |

| **3. What are some other members of the Family to which your Contig belongs?** |
| --- |
|  |

| **4. What are some other members of the Order to which your Contig belongs?** |
| --- |
|  |

| **5. What are some other members of the Class to which your Contig belongs?** |
| --- |
|  |

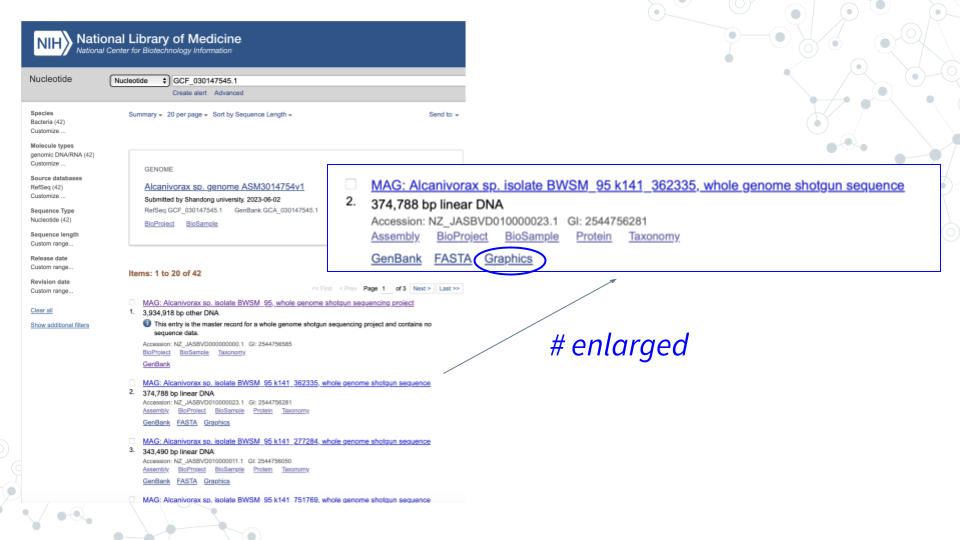
| **6. What are some other members of the Phylum to which your Contig belongs?** |
| --- |
|  |

### 3.4.4 Activity 2 – Genomes, Genes, and Databases

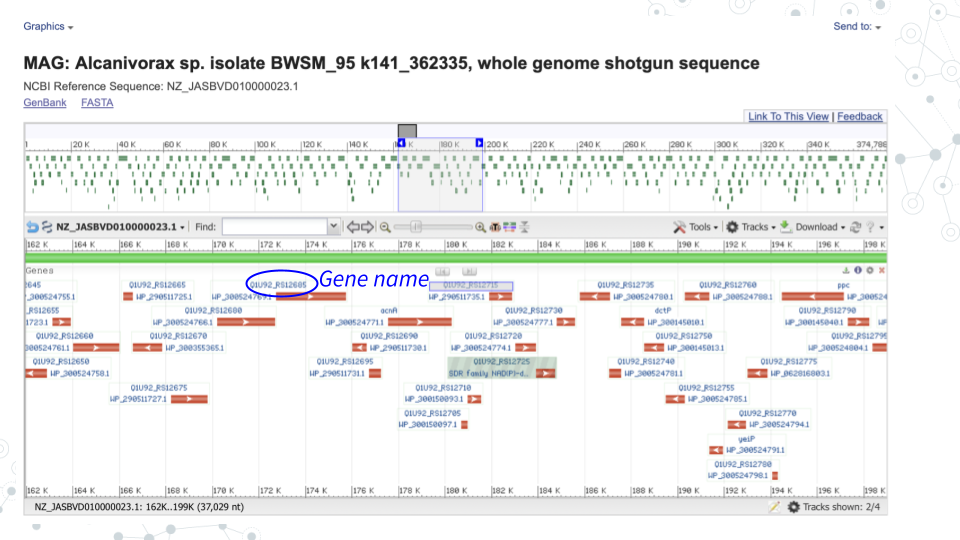
*Estimated time: [20] min*

#### 3.4.4.1 Instructions

1. In GenBank <https://www.ncbi.nlm.nih.gov/nucleotide/>, for the contig you chose in activity 1 click on **Graphics** to explore the genome browser and the genes.



1. **Select genes of interest**. A lot of genes found in the Contigs will have no ‘familiar’ short symbol and instead have a long alphabetical-and-numerical- name. Such genes are either uncharacterized, hypothetical or have functional or structural similarity to known genes/proteins, but have not been confirmed. However, some Contig genes will be annotated with a ‘familiar’ short gene symbol, matching known genes. Please use the genes with the short symbols for this activity, since the ‘other’ genes will not be found in the databases.



1. For the 3 genes of choice, use the BV-BRC [bv-brc.org](https://www.bv-brc.org) to find more information on your genes.

#### 3.4.4.2 Questions

| \*\*1. Record 5 genes of interest for your Contig. |  |
| --- | --- |
| Gene1: |  |
| Gene2: |  |
| Gene3: |  |
| Gene4: |  |
| Gene5: |  |

For the 3 genes above, in the [bv-brc.org](https://www.bv-brc.org) **Search** space, a) from a dropdown menu select **“Pathways”** , b) type in the gene name and click enter. This will result in a lot of entries for different organisms. c) check one of the boxes corresponding to the gene of interest and enter below which species/strain (**Genome Name**) you selected.

| \*\*2. Record Genome Name associated with your 3 genes of interest in bv-brc database. |  |
| --- | --- |
| **Gene** | **Genome Name** |
| Gene1: |  |
| Gene2: |  |
| Gene3: |  |

| \*\*3. Record gene Product associated with your genes of interest in bv-brc database. |  |
| --- | --- |
| **Gene** | **Product** |
| Gene1: |  |
| Gene2: |  |
| Gene3: |  |

| \*\*4. Record Pathway Name associated with your genes of interest in bv-brc database. |  |
| --- | --- |
| **Gene** | **Pathway Name** |
| Gene1: |  |
| Gene2: |  |
| Gene3: |  |

### 3.4.5 Activity 3 – Go Deeper!

*Estimated time: [30] min*

#### 3.4.5.1 Instructions

Use any tools available at your disposal to follow up on your MAG and gene of interest from activities 1 and 2 above. Some suggested tools include PubMed, Google, [MBGD](https://mbgd.nibb.ac.jp), BV-BRC, [BacDive](https://bacdive.dsmz.de).

1. For the **taxa** you identified for your chosen MAG in activity 1, what can you learn about this organism (species or genus for example) in 15 minutes using any tools at your disposal?

| \*\*a. What did you learn? |
| --- |
|  |
|  |

| \*\*b. What tools did you use? |
| --- |
|  |
|  |

1. For one of the **genes** you identified for your chosen contig in activity 2, what can you learn about this gene in 15 minutes using any tools at your disposal?

| \*\*a. What did you learn? |
| --- |
|  |
|  |

| \*\*b. What tools did you use? |
| --- |
|  |
|  |

### 3.4.6 Grading Criteria

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

### 3.4.7 Footnotes

**Resources**

* Google Doc

**Contributions and Affiliations** - Valeriya Gaysinskaya, Johns Hopkins University - Gauri Paul, Clovis Community College - Frederick Tan, Johns Hopkins University

Last Revised: May 2025

## 3.5 Presentation - Microbial Genomes

### 3.5.1 Activity

*Estimated time: 25 min*

#### 3.5.1.1 Instructions

1. Open the “Microbial Genomes Presentation” [slidedeck](https://docs.google.com/presentation/d/17hFixGrtYtM0EzEOcU1aAi0CVRu36OqtNIXtbuiu0Os/edit?usp=sharing)
2. Create at least two slides (feel free to make more!) to present your findings. E.g.,

* Describe the MAG you explored (Activity 1) and some of its genes (Activity 2)
* What did you discover about your organism and gene (Activity 3)?

1. Iterate your slides being sure to:

* Add bullet points on key details you understand (Notice) and questions you have (Wonder)
* Include one or more plots or images that support your points
* Create slide titles that summarize your main takeaway

1. Present at next class (5 min each group)

### 3.5.2 Footnotes

**Contributions and Affiliations**

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

Last Revised: February 2025

# 4 Taxonomy Profiling

## 4.1 Activity - Taxonomy Profiling Spreadsheet

### 4.1.1 Purpose

First hands-on experience with real data! Compare kraken2 output for [Zymo Gut Microbiome Standard](https://www.zymoresearch.com/products/zymobiomics-gut-microbiome-standard?srsltid=AfmBOoqP_zq131c2GTidPCM0j6yA3JFcGQ0haUNu1jAJI9RQ9qsXLYSF) and [Zymo Human Fecal Reference](https://files.zymoresearch.com/protocols/d6323-zymobiomics_fecal_reference_protocol.pdf). Introduce concepts of taxa and relationships, begin forming data analysis goals like comparing how many species, most abundant species, etc. See accompanying [slides](http://docs.google.com/presentation/d/16lpgWFU6jzh-e7HuwXLHmUFpsnE8NreMzL-nTn8cJVk).

### 4.1.2 Learning Objectives

1. Explore taxonomy with Kraken 2 taxonomic assignment output.
2. Compare and contrast taxonomy between Zymo Gut Microbiome Standard and Zymo Human Fecal Reference.

### 4.1.3 Introduction

Metagenomics is the direct analysis of the genomes through genome sequencing of an environmental sample (soil, water, gut, etc). The purpose of the taxonomic classification of metagenomic sequences is to catalogue, classify and identify the species inhabiting a given environment. In the process, new species may get identified! After sampling, DNA extraction, DNA sequencing and genome assembly, genome annotation is used to assign taxonomy to the sequenced sample DNA. Here is where the Kraken 2 tool comes in; Kraken 2 is a taxonomic classification tool which assigns taxonomy to sequencing reads.

### 4.1.4 Activity 1 – Explore Zymo Gut Standard Metagenomic Diversity

*Estimated time: 25 min*

#### 4.1.4.1 Instructions

Perform the activity below and answer the embedded **questions**.

1. Access tax-data-gut.tsv and open with Google Sheets [here](http://drive.google.com/file/d/1vL6adVIrqxpONbae8rUsneK3tbdCpmR-)
2. Identify what information is provided in columns of the tax-data-gut taxonomy file.

* Col A = Counts
* Cols B-H correspond to taxonomic ranks k(Kingdom), p(Phylum), c(Class), o(Order), f(Family), g(Genus) and s(Species)
* Each row corresponds to a different taxa. There are 153 taxa that were classified for this sample.

1. Create a header row and enter column information.

#### 4.1.4.2 Questions

**1. Evaluate what proportion of data was taxonomically classified.**

* Insert a new column A; we will use this temporary column for calculations, so you can name this column “Calculations”.
* In e.g. cell A2, calculate the sum of all reads observed in the gut std sample.

| 1A. How many total counts are there? |
| --- |
|  |

* In e.g. cell A3, determine the percentage of unclassified reads.

| 1B. What percentage of reads are unclassified? |
| --- |
|  |

* In e.g. cell A4, determine the percentage of classified reads.

| 1C. What percentage of reads are classified? |
| --- |
|  |

**2. Identify abundant taxa (those at >1%).**

1. Select columns B through I
2. In the Data menu, select “Sort range by column B (Z to A)”
3. Insert a new column C; we will use this temporary column for calculations; you can name this column “% abundance”.
4. In new column C, calculate % abundance for each row by dividing each count value by the total number of reads and multiplying by 100.
5. Quantify abundant taxa.

| 2A. How many abundant taxa (at >1%) do you observe? |
| --- |
|  |

1. List abundant taxa you identified in a table below.

* To consolidate the different abundant taxa, in e.g. new column D, copy the lower taxonomic rank identified for the abundant (at >1%) taxa.
* Then, enter the results into a table below.

| 3A. What abundant taxa do you observe? |  |
| --- | --- |
| **% abundance** | **Taxonomy** |
| 20.1 | s\_Faecalibacterium\_prausnitzii |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

**4. Compare your results with the expected taxa and abundance for** [**Zymo gut standard documentation**](https://www.zymoresearch.com/products/zymobiomics-gut-microbiome-standard?srsltid=AfmBOor0X27Jf1gfXVmyGu5nZq3M6fx6OJXdEc0t6rqSRBPww2qeY-Yd)**?**

* Note, the Kraken2 output does not distinguish different *E. coli* strains, so just combine them all into a single *E. coli group*!

| 4A. How do your results overall compare with the expected taxa and % abundance from Zymo gut standard? |
| --- |
|  |

| 5. Calculate ‘Low abundance’ for < 1% abundant taxa by adding together taxa at <1%. What percentage of reads are classified in a low abundance taxa? |
| --- |
|  |

| 6. Create a barplot of % abundance for your 12 abundant taxa via Insert Chart. Paste your barplot of % abundance for the 12 most abundant taxa. |
| --- |
|  |

### 4.1.5 Activity 2 (OPTIONAL) – Compare with Zymo Fecal Reference

*Estimated time: 20 min*

#### 4.1.5.1 Instructions

Perform the optional activity below and answer the embedded **questions**.

In this activity, repeat steps of the Activity 1 above, but now using [tax\_data\_fecal.tsv](http://drive.google.com/file/d/1CLQw9yqoqWl5caLm-ZmiHpLNtUo_Zo4s) dataset corresponding to Zymo fecal reference. The tax\_data\_fecal.tsv dataset comes from a real human fecal sample, in contrast to the tax\_data\_gut.tsv sample you explored in the Activity 1, which corresponds to cultured and pooled known species combined at specific proportions to make up a predictable standard population.

* Perform Activity 1 exercises using tax\_data\_fecal data, then, use questions below to compare the two datasets.
* See D6323 Zymo Fecal Microbiome References documentation (pg. 4) in the Resources section below.

#### 4.1.5.2 Questions

| **1. Which dataset is classified better, gut or fecal??** |
| --- |
|  |

| **2. Are there any abundant taxa (at >1%) in common between the gut standard and fecal reference?** |
| --- |
|  |

| **3. In your opinion, does the gut standard mimic the fecal reference well or not?** |
| --- |
|  |

### 4.1.6 Grading Criteria

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

### 4.1.7 Footnotes

**Resources**

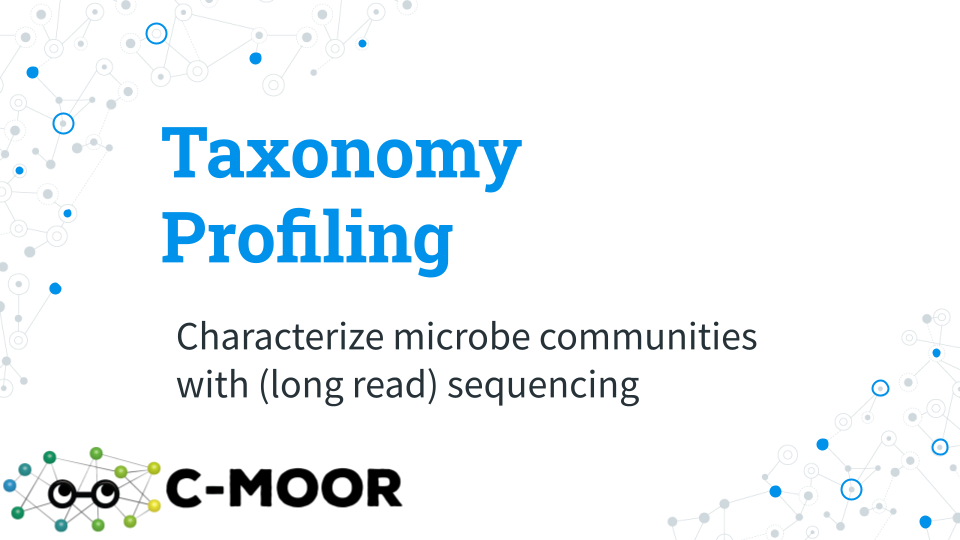
* Google Doc
* [D6331 Zymo Gut Microbiome Standard documentation](https://www.zymoresearch.com/products/zymobiomics-gut-microbiome-standard?srsltid=AfmBOor0X27Jf1gfXVmyGu5nZq3M6fx6OJXdEc0t6rqSRBPww2qeY-Yd)
* [D6323 Zymo Fecal Microbiome References documentation](https://files.zymoresearch.com/protocols/d6323-zymobiomics_fecal_reference_protocol.pdf?_gl=1*cych1b*_gcl_au*MzEzNzgzNjc0LjE3MzQ5NTk3NzY)

**Contributions and Affiliations**

* Valeriya Gaysinskaya, Johns Hopkins University
* Gauri Paul, Clovis Community College
* Frederick Tan, Johns Hopkins University

Last Revised: January 2025

## 4.2 Lecture - Taxonomy Profiling



[Slides: Taxonomy Profiling](https://docs.google.com/presentation/d/10P4ktKWSrRpM1YAWYW31tLly_jf1rcfgCMbzZ9Tx9xY/edit?usp=sharing)

## 4.3 Prelab - Taxonomy Profiling

### 4.3.1 Purpose

To use a variety of Galaxy tools to perform Quality Control (QC), taxonomy profiling, and visualization of a metagenomics sample.

### 4.3.2 Learning Objectives

Use Galaxy tools to:

1. Perform Quality Control (QC) on your raw data by checking the quality of your raw reads
2. Assign taxonomy labels to your reads
3. Visualize the classified metagenome

In this exercise, using Galaxy, you will be:

1. Using NanoPlot tool to examine the quality of your sequencing reads
2. Running a workflow to perform taxonomy profiling and visualization in a single step

### 4.3.3 Introduction

To find out which microorganisms are present in the sample, it is important to have high-quality DNA sequences. To ensure high-quality sequence input, QC (and in many cases also read trimming and filtering) are routinely performed on raw sequences. The reads can then be used to determine which species, genera, families, and other taxonomic ranks are present in the sample. To assign taxonomy, we can compare the reads of the sample to a reference database, i.e. sequences of known microorganisms stored in a database, using Kraken2, which is a k-mer based taxonomic assignment tool. We can then use a tool like Krona to interactively visualize and explore the composition of a metagenome.

### 4.3.4 Activity 1 – QC Reads

*Estimated time: 50 min*

#### 4.3.4.1 Activity 1 - Part I - Import dataset into Galaxy

*Estimated time: 15 min*

#### 4.3.4.2 Instructions

1. Import dataset into Galaxy.
2. Open the zymo-gut-standard public history <https://usegalaxy.org/u/valerie-g/h/zymo-gut-standard-d6331-subset-1>
3. Click on **Import this history**, select Copy only the active, non-deleted datasets and then Copy History.
4. Confirm Zymo\_Gut\_Standard\_D6331\_subset exists in your history by clicking on the Home button “Galaxy” on top left ().
5. Click on Zymo\_Gut\_Standard\_D6331\_subset to explore content.

#### 4.3.4.3 Questions

| 1. What is the size of this downloaded dataset subset? |
| --- |
|  |

| 2. What is the format/extension of the downloaded file? |
| --- |
|  |

| 3. Click on the Display (eyeball) icon and describe what you see in the 4 lines of the fastq file? |
| --- |
| Line 1: |
| Line 2: |
| Line 3: |
| Line 4: |

#### 4.3.4.4 Activity 1 - Part II - Run Nanoplot in Galaxy to assess sequence quality

*Estimated time: 15 min*

#### 4.3.4.5 Instructions

1. Run Nanoplot in Galaxy.
2. Explore NanoPlot tool parameters - click on the Tools icon on the left of the page. Then, in the search bar enter ‘NanoPlot’ and select the **NanoPlot** tool. Explore NanoPlot functionality via examining **Tool Parameters**.
3. Run NanoPlot using default settings. Under Tool Parameters, **check the following settings**:

* Under **files** there are 4 options to select a fastq dataset: Single dataset, Multiple datasets, Dataset collection or option ‘…’ which is Browse or Upload Datasets. Browse to select your fastq dataset. **Note**, Galaxy tool may pre-select the correct dataset already for you so just make sure that the file is correct.
* Click on Run Tool and wait ~5-10 minutes as the NanoPlot job is scheduled, run, and complete.

#### 4.3.4.6 Questions

| 1. Under **Type of file(s) to work on**, check to see which input files are compatible with NanoPlot and name 2 file extension options listed. |
| --- |
| File extension name 1: |
| File extension name 2: |

| 2. Click to expand *Options for filtering or transforming input prior to plotting* and name 3 options you could use to filter your sequencing data. |
| --- |
| 1. |
| 2. |
| 3. |

| 3. **Run Nanoplot** using default tool settings and record how many output files you obtained after running NanoPlot and list their names. |
| --- |
| # of output files |
| Names of output files |

#### 4.3.4.7 Activity 1 - Part III - View NanoPlot Results in Galaxy

*Estimated time: 15 min*

#### 4.3.4.8 Instructions

1. View and examine NanoPlot Results in Galaxy.

* By clicking on the Display icon (eyeball) next to the NanoPlot output.

#### 4.3.4.9 Questions

1. Click on the Display icon (eyeball) next to the NanoPlot output files to view results.

| A. How many bases were sequenced? |
| --- |
|  |

| B. Why is mean read length longer than the median read length? - Hint: think skewness <https://wikipedia.org/wiki/Skewness> |
| --- |
|  |

| C. Record Reads >Q20 metric value. Given that Q20 quality (Phred) score corresponds to read accuracy of 99% (or 1 in 100 errors), do you think this dataset is of a good sequence quality? |
| --- |
|  |

1. Click on the Display icon (eyeball) next to the NanoPlot output HTML report.

| A. Scroll down to view the ‘Weighted histogram of read lengths’ histogram. From this plot estimate the range of read lengths obtained |
| --- |
|  |

| B. Scroll down to view the ‘Yield by length’ cumulative plot which shows sequencing yield based on read length. From this plot do shorter (10kb or less) or longer sequences produce more data? |
| --- |
|  |

### 4.3.5 Activity 2 – Taxonomy Profiling in Galaxy

*Estimated time: 50 min*

#### 4.3.5.1 Activity 2 - Part I - Run ‘Taxonomy Profiling’ workflow in Galaxy

*Estimated time: 15 min*

#### 4.3.5.2 Instructions

1. Run ‘Taxonomy Profiling’ public workflow.
2. Open the taxonomy-profiling public workflow <https://usegalaxy.org/u/cutsort/w/taxonomy-profiling> and click on **Run**.
3. Browse to select your fastq dataset by clicking on the ‘…’ tab.
4. Under **kraken\_database** select ’Prebuilt Refseq indexes: PlusPF(Standard plus protozoa and fungi)(Version:2022-06-07 - Downloaded: 2022-09-04T165121Z).
5. Click **Run Workflow** with the following parameters:
6. Wait ~15-30 minutes as the Kraken2, KrakenTools, and Krona jobs are scheduled, run, and complete.
7. Examine select aspects of Kraken2 tool.
8. While the Taxonomy Workflow is running, click on the Tools icon on the left of the page.
9. Then, in the search bar enter ‘Kraken2’’ and select the **Kraken2** tool.

#### 4.3.5.3 Questions

| A. How does Galaxy describe the Kraken2 tool in its descriptor on top of the page? |
| --- |
|  |

| B. Scroll down to **Help/What it does** section, and in your own words paraphrase the paragraph describing how Kraken 2 works. |
| --- |
|  |

| C. Record how many output files you obtained from the Taxonomy Workflow and list their names. |
| --- |
|  |

#### 4.3.5.4 Activity 2 - Part II - View Kraken2 results

*Estimated time: 15 min*

#### 4.3.5.5 Instructions

1. Examine **converted\_kraken\_report**.

* Click on the Display icon (eyeball) next to the output file with **converted\_kraken\_report**. This report should look familiar from week 1 taxonomy-profiling-spreadsheet activity.

#### 4.3.5.6 Questions

1. Open and scroll through the **converted\_kraken\_report** report to answer the following questions:

| A. How many Unclassified reads are there? |
| --- |
|  |

| B. How many Kingdoms are there and what are they? |
| --- |
|  |

| C. How many Phyla are there and what are they? |
| --- |
|  |

| D. Using the total *number\_of\_reads* you obtained from the **NanoStats** (NanoPlot) metrics, and the value of *Unclassified* reads from the **converted\_kraken\_report**, calculate % unclassified and % classified taxa. |
| --- |
|  |

| E. Click on the **converted\_kraken\_report** entry and just below the ‘Add Tags’ to see the number of lines and columns in the file. The number of lines corresponds to the number of taxa detected. Excluding Unclassified subset, how many taxa were identified? |
| --- |
|  |

1. Examine **kraken2\_with\_pluspf\_database\_output\_report** by clicking on the Display icon (eyeball) next to the output file.

This output report is an extended version of the converted\_kraken\_report. The output contains 6 columns. See info for select column headers below:

* Column 1: Percentage (%) of a given taxon
* Column 2: # of reads per given taxon
* Column 4: A rank code, indicating (U)nclassified, (R)oot, (D)omain, (K)ingdom, (P)hylum, (C)lass, (O)rder, (F)amily, (G)enus, or (S)pecies. Note, that in this extended file, some rank codes will have numbers associated with them; Ignore this aspect of the document for the moment.
* Column 6: Identified taxa/scientific name.

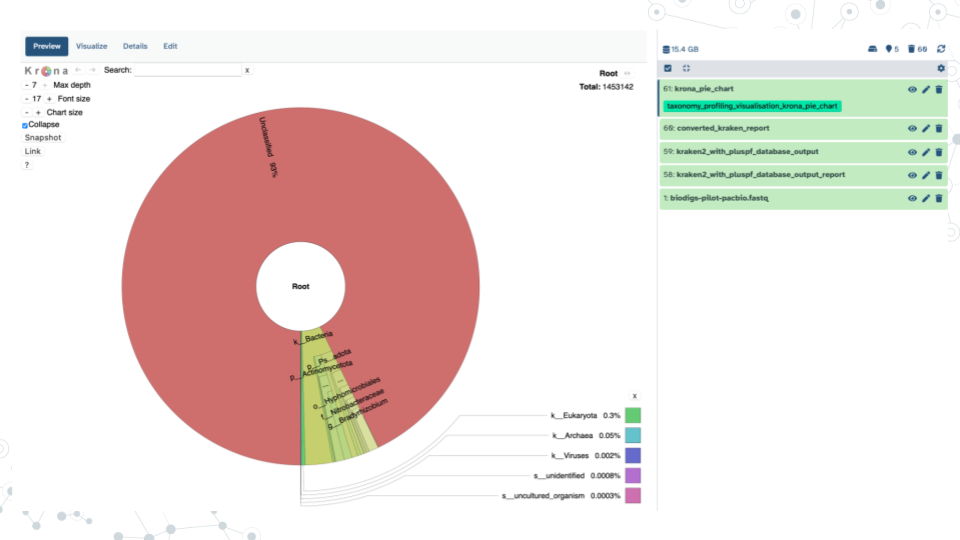
| A. What is the percentage of Unclassified taxa listed? Does it match what you calculated in section 2-2.1? |
| --- |
|  |

| B. What is the percentage of Classified taxa listed? Does it match what you calculated in section 2-2.1? |
| --- |
|  |

| C. Find and record the 3 most abundant Phyla (p\_) by percentages. As rows are not sorted by abundance, you may find it helpful to search using for “P” using -F to highlight rows corresponding to phyla. |
| --- |
|  |

1. View Krona Results

Krona pie chart is one of the outputs of the Taxonomy workflow, and it is an interactive visualization tool for exploring the composition of metagenomes. Click on Display icon (eyeball) next to the **Krona\_pie\_chart** dataset to explore Krona pie chart results.



| A. What are the 2 main phyla you observe? |
| --- |
|  |

| B. What appears to be the more diverse phyla of the two and why? |
| --- |
|  |

| C. Examine how many reads and percent of reads classified as phylum Firmicutes. |
| --- |
|  |

| D. What is one of the most abundant Orders in phylum FIrmicutes based on number and % of reads? |
| --- |
|  |

| E. Within the most abundant Order from your answer above, record one Family member, one genus member and one species member; also include percent abundances. |
| --- |
|  |

| F. Some potentially pathogenic bacterial species are often present in a healthy gut, but at relatively low amounts. What % of bacteria is represented by *C. difficile* (*Clostridiodes difficile*) in this gut standard sample? How does it compare to the expected proportion in [ZymoBIOMICS® Gut Microbiome Standard](https://files.zymoresearch.com/datasheets/ds1712_zymobiomics_gut_microbiome_standard_data_sheet.pdf?_gl=1*1njkhlx*_gcl_au*MzE1NTQwNzEuMTczODI1OTA5OA..)? |
| --- |
|  |

### 4.3.6 Grading Criteria

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

### 4.3.7 Footnotes

**Resources**

* Google Doc
* Species composition in the Gut Microbiome Standard dataset: [ZymoBIOMICS® Gut Microbiome Standard](https://files.zymoresearch.com/datasheets/ds1712_zymobiomics_gut_microbiome_standard_data_sheet.pdf?_gl=1*1njkhlx*_gcl_au*MzE1NTQwNzEuMTczODI1OTA5OA..)
* If interested reading more about Kraken 2, see [Kraken 2 publication](https://pubmed.gov/31779668)

**Contributions and Affiliations**

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

Last Revised: May 2025

## 4.4 Discussion - Taxonomy Profiling

### 4.4.1 Activity

*Estimated time: 25 min*

#### 4.4.1.1 Instructions

1. Form new groups of four

* <https://docs.google.com/spreadsheets/d/11eoJgm9mehxGWWzh8IZYDCDmnCmSyshopPYHewvpC8c/edit?usp=sharing>

1. Pair up into groups (10 min)
2. Discuss – Each group member briefly describes answers to prelab assignment
3. Summarize – Identify best answer and add to slidedeck

* <https://docs.google.com/presentation/d/16cRBFzezun6MGXcJEm5uP2KercIccO-dFvVx58_ClNo/edit?usp=sharing>

1. Share group discussion (2 min each group)

### 4.4.2 Footnotes

**Contributions and Affiliations**

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

Last Revised: February 2025

## 4.5 Project - Taxonomy Profiling

### 4.5.1 Purpose

To use a variety of Galaxy tools to perform Quality Control (QC), sequence quality filtering, taxonomy profiling, and visualization of a metagenomics soil sample sequenced with long-read Nanopore technology.

### 4.5.2 Learning Objectives

In this exercise, using Galaxy tools you will be 1. Performing QC and quality filtering of your soil metagenomics data with NanoPlot and fastp tools. 2. Running a workflow to perform taxonomy profiling and visualization of a soil metagenome.

Throughout these objectives you will be comparing soil and gut metagenomes.

### 4.5.3 Introduction

The total time for a Galaxy step to complete depends on and will increase based on multiple factors such as large input file size, long queue when many other people are analyzing data, the complexity of the job itself, or an error. See table below for the minimum time a step will take for this assignment – be sure to start early as when Galaxy is busy each step can take 2-to-10 times longer to complete.

Note, that you can save time by 1) submitting multiple jobs that use the same input (NanoPlot and fastp) and 2) submitting a job like the taxonomy workflow that uses the fastp output as input as soon as the output appears in your history even before the fastp job finishes.

**Table of approximate minimum times for a job to be completed on Galaxy using specified tools.**

| Nanoplot | fastp | taxonomy workflow |
| --- | --- | --- |
| 15 min | 15 min | 30 min |

### 4.5.4 Activity 1 – QC

*Estimated time: 50 min*

#### 4.5.4.1 Activity 1 - Part I - Import data and run NanoPlot

#### 4.5.4.2 Instructions

1. Import dataset into Galaxy.
2. Open the nanopore-soil-pilot public history <https://usegalaxy.org/u/valerie-g/h/nanopore-soil-pilot-1>
3. Click on Import this history, select Copy only the active, non-deleted datasets and then Copy History.
4. Confirm Nanopore-soil-pilot-subset exists in your history by clicking on the Home button on top left ().
5. Run **NanoPlot** tool in Galaxy to assess sequence quality using **default settings**.
6. Click on the Tools icon. Then, in the search bar enter ‘NanoPlot’ and select the NanoPlot tool.
7. Under **files** browse to select your Nanopore-soil-pilot-subset fastq dataset.
8. Click on **Run Tool** and wait ~10 minutes as the NanoPlot job is scheduled, run, and complete.

#### 4.5.4.3 Questions

| 1. Click on the Display icon (eyeball) next to the NanoPlot output NanoStats report and record: |
| --- |
| Read mean length (mean read length): |
| Read Mean quality (mean\_qual): |
| Proportion of reads with quality > Q20 (Reads > Q20): |

**2. Compare NanoPlot results.**

|  | Nanopore soil pilot (this activity) | Zymo gut standard (taxonomy profiling pre-lab) |
| --- | --- | --- |
| mean read length: |  |  |
| mean\_qual: |  |  |
| Reads > Q20: |  |  |

| 3. Which dataset has better sequence quality, Zymo-gut-standard (taxonomy profiling pre-lab) or Nanopore-soil pilot (taxonomy profiling project)? Why? |
| --- |
|  |

#### 4.5.4.4 Activity 1 - Part II - Quality filtering with fastp

#### 4.5.4.5 Instructions

Although the majority of bases in the soil dataset are of high quality ( >Q20, or 1 in 100 base error), we can filter out very low quality reads to further improve dataset quality. In this activity run fastp tool to filter out reads with a high proportion of low quality bases (<Q15) using default settings.

1. Click on the Tools icon. Then, in the search bar enter **fastp** and select the fastp tool.
2. Click on **Run Tool** and wait ~10 minutes as the fastp job is scheduled, run, and complete.

#### 4.5.4.6 Questions

**1. Compare your dataset before and after filtering using fastp: HTML report output.**

|  | **Before** | **After** |
| --- | --- | --- |
| Mean Length |  |  |
| total reads |  |  |
| total bases |  |  |
| Q20 bases (%): |  |  |
| Q30 bases (%): |  |  |

| 2. Compare BEFORE and AFTER quality plots from activity above. What key quality improvement can you observe after fastp quality filtering. Hint 1: look at 5’ end, 3’ end and the middle? Hint 2: Pay attention to the y-axis |
| --- |
|  |

### 4.5.5 Activity 2 – Taxonomy Profiling

*Estimated time: 50 min (~35 min computing)*

#### 4.5.5.1 Activity 2 - Part I - Run Taxonomy Profiling Workflow

#### 4.5.5.2 Instructions

1. Run ‘Taxonomy Profiling’ workflow on your fastp-filtered data from Activity 1 and view results.
2. Open the taxonomy-profiling public workflow <https://usegalaxy.org/u/cutsort/w/taxonomy-profiling>
3. Click on Run
4. Browse to select your fastp-filtered fastq dataset “fastp on data1:Read 1 output” dataset by clicking on the ‘...’ tab.
5. Under **kraken\_database** select ’Prebuilt Refseq indexes: PlusPF(Standard plus protozoa and fungi)(Version:2022-06-07 - Downloaded: 2022-09-04T165121Z).
6. Click **Run Workflow**
7. Wait ~15-30 minutes as the Kraken2, KrakenTools, and Krona jobs are scheduled, run, and complete.
8. Click on the Display icon (eyeball) next to the output file with converted\_kraken\_report. Explore metagenomic diversity of soil by performing taxonomy profiling spreadsheet activity you did during week 1.
9. Click on **converted\_kraken\_report**, find the download button and **download** the report.
10. Change the extension of your taxonomy file from .tabular to .tsv.
11. Upload your taxonomy .tsv file to Google Drive and open with Google Sheets.
12. Create a header row and enter column Information.

* Col A = Counts
* Cols B-H correspond to taxonomic ranks k(Kingdom), p(Phylum), c(Class), o(Order), f(Family), g(Genus) and s(Species).
* Each row corresponds to a different taxa.

1. Evaluate what proportion of data was taxonomically classified.

* Insert a new column A; we will use this temporary column for calculations, so you can name this column “Calculations”.
* In e.g. cell A2, calculate the sum of all reads observed in the soil sample.

#### 4.5.5.3 Questions

| 1. How many total read counts are there? |
| --- |
|  |

| 2. Determine percentage of reads that are unclassified |
| --- |
|  |

| 3. What percentage of reads are classified? |
| --- |
|  |

**4. Identify the most abundant taxa (those at >0.1%).** - Remember, soil is one of the most diverse microbial environments with many more microbial species than in the gut. Therefore, abundant species can still be quite low abundance.

1. Select columns B through I
2. In the Data menu, select “Sort range by column B (Z to A)”
3. Insert a new column C; we will use this temporary column for calculations; you can name this column “% abundance”.
4. In new column C, calculate % abundance for each row by dividing each count value by the total number of reads and multiplying by 100.

| How many ‘abundant’ taxa (at > 0.1%) do you observe? |
| --- |
|  |

| 5. What are the taxonomic ranks of most abundant taxa? |
| --- |
|  |

| 6. What is the most abundant eukaryote observed and its read count? |
| --- |
|  |

| 7. What is the most abundant archaea observed and its read count? |
| --- |
|  |

| 8. What is the most abundant virus observed and its read count? |
| --- |
|  |

#### 4.5.5.4 Activity 2 - Part II - Analyze Kraken2 results

#### 4.5.5.5 Instructions

1. Click on the Display icon (eyeball) next to the output file with **kraken2\_with\_pluspf\_database\_output\_report**. This output report is an extended version of the converted\_kraken\_report. The output contains 6 columns. See info for select column headers below:

* Column 1: Percentage (%) of a given taxon
* Column 2: # of reads per given taxon
* Column 4: A rank code, indicating (U)nclassified, (R)oot, (D)omain, (K)ingdom, (P)hylum, (C)lass, (O)rder, (F)amily, (G)enus, or (S)pecies. Note, that in this extended file, some rank codes will have numbers associated with them; Ignore this aspect of the document for the moment.
* Column 6: Identified taxa/scientific name.

Of note: The benefit of kraken2\_with\_pluspf\_database\_output\_report is that it summarizes converted\_kraken\_report and calculates summary percentages for taxonomic ranks. For example, your converted\_kraken\_report has hundreds of lines for phylum Proteobacteria, while kraken2\_with\_pluspf\_database\_output\_report has 1 line summarizing the percent abundance of all Proteobacteria.

#### 4.5.5.6 Questions

| 1. What is the percentage of Unclassified taxa? Does it match your calculations in Activity 2 - Part I? |
| --- |
|  |

| 2. What percentage of bacteria is Proteobacteria, the most abundant Phyla observed? |
| --- |
|  |

| 3. What is the most abundant class observed and at what percentage? |
| --- |
|  |

#### 4.5.5.7 Activity 2 - Part III - Krona Pie Chart

#### 4.5.5.8 Instructions

Krona pie chart is one of the outputs of the Taxonomy workflow, and it is an interactive visualization tool for exploring the composition of metagenomes.

1. View Krona results: Click on the Display icon (eyeball) next to the output file named **krona\_pie\_chart**.
2. Double click on Bacteria kingdom (k\_Bacteria) to explore further.
3. Answer questions below

#### 4.5.5.9 Questions

| 1. What are the 2 main phyla you observe? |
| --- |
|  |

| 2. What appears to be the more diverse phyla of the two? |
| --- |
|  |

**3. Compare your taxonomy from soil, to the gut taxonomy results from your taxonomy-prelab** (Zymo-gut-standard [ZymoBIOMICS® Gut Microbiome Standard](https://files.zymoresearch.com/datasheets/ds1712_zymobiomics_gut_microbiome_standard_data_sheet.pdf?_gl=1*1njkhlx*_gcl_au*MzE1NTQwNzEuMTczODI1OTA5OA..).|

**3A. Fill out the comparison table below**

|  | Nanopore soil pilot | Zymo gut standard |
| --- | --- | --- |
| What are 2 most abundant phyla |  |  |
| What are 2 most abundant species |  |  |
| % Classified taxa |  |  |
| % Unclassified taxa |  |  |

| 3B. Discuss taxonomy diversity between soil and gut, providing 3 points: |
| --- |
| 1) |
| 2) |
| 3) |

### 4.5.6 Grading Criteria

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

### 4.5.7 Footnotes

**Resources** - Google Doc - Species composition in the Gut Microbiome Standard dataset: [ZymoBIOMICS® Gut Microbiome Standard](https://files.zymoresearch.com/datasheets/ds1712_zymobiomics_gut_microbiome_standard_data_sheet.pdf?_gl=1*1njkhlx*_gcl_au*MzE1NTQwNzEuMTczODI1OTA5OA..)

**Contributions and Affiliations**

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

Last Revised: May 2025

## 4.6 Presentation - Taxonomy Profiling

### 4.6.1 Activity

*Estimated time: 25 min*

### 4.6.2 Instructions

1. Open the “Taxonomy Profiling Presentation” slidedeck <https://docs.google.com/presentation/d/1qsGKQ3M-etSpH1JFN270YfYgtYKxCsE-ou2B7gw25Gc/edit?usp=sharing>
2. Create at least two slides (feel free to make more!) to present your findings e.g.

* Describe the sample, drawing inspiration from Fig 2 (taxonomic distribution) and Fig 3a (read abundance) in the Xue et al 2023 ballast water study
* Propose several steps for a follow-up analysis to compare this soil sample with one of the studies described in the Possible Dataset Activity, drawing inspiration from your submissions as well as those of your classmates in the Discussion Forum

1. Iterate your slides being sure to

* Add bullet points on key details you understand (Notice) and questions you have (Wonder)
* Include one or more plots or images that support your points
* Create slide titles that summarize your main takeaway

1. Present at next class (5 min each group)

### 4.6.3 Footnotes

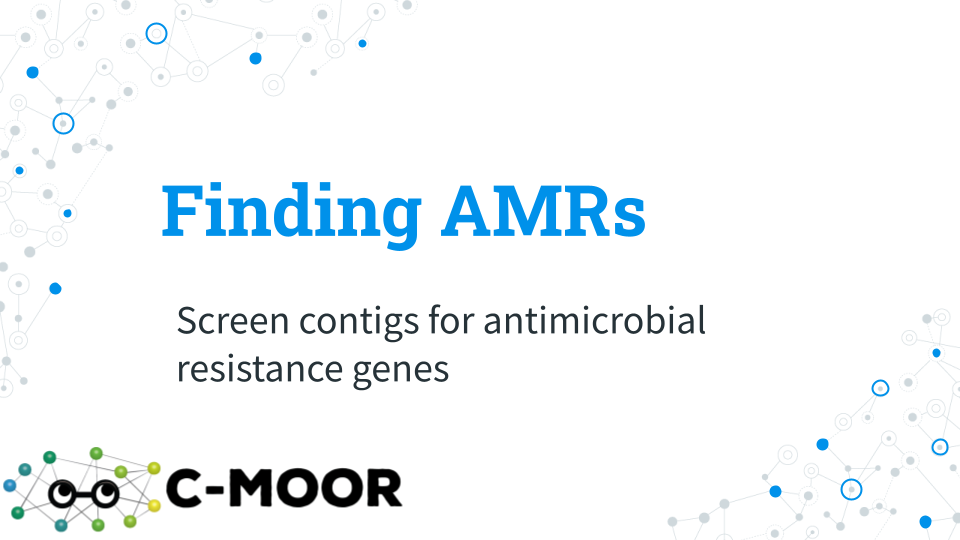
**Contributions and Affiliations**

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

Last Revised: February 2025

# 5 Finding AMRs

## 5.1 Lecture - Finding AMRs



[Slides: Finding AMRs](https://docs.google.com/presentation/d/165OHha9IYOctuyzg1CG0LwGxNnbC85JyJfMZT6xnYHw/edit?usp=sharing)

## 5.2 Pre-lab - Finding AMRs

### 5.2.1 Purpose

To use a Galaxy tools to perform de novo genome assembly of sequencing reads into contigs, visualize the contigs and find antimicrobial resistance (AMR) genes.

### 5.2.2 Learning Objectives

Use Galaxy tools to:

1. Use Flye tool to perform de novo genome assembly of long reads into ‘contigs’
2. Visualize contigs with Bandage Image tool
3. Use ABRicate tool to find AMRs in contig assemblies

### 5.2.3 Introduction

Genome assembly is the process of reconstructing genomes from DNA sequence reads. Accurate and continuous genome assembly from sequenced fragments, even very long fragments, is challenging. Flye, is a long-read assembly algorithm that aims to produce highly contiguous genome assemblies and overcome some of the assembly challenges, like repetitive DNA sequences. Using Flye we will hope to be able to reconstruct bacterial genomes and plasmids and enable detection of important genes like AMRs. Ideally, after assembly we want to get back circular contigs as that would typically indicate an entire microbial genome(s) or plasmid(s).

Antimicrobial resistance is the ability of microbes to evade one or more antibiotics, leading to multidrug resistance and ability to survive and even thrive in the presence of antibiotics. Detecting and studying antibiotic-resistant pathogens is therefore extremely important to human health. However, the environmental reservoirs of resistance determinants are poorly understood. Certainly the indiscriminate and sometimes inappropriate use of antibiotics by humans (e.g. in the hospitals, in food production) has contributed to the emergence of resistant bacterial strains, but there are many other ways microbes can acquire AMR. For example, the environment like soil is emerging as a key reservoir of these antibiotic resistance genes. For more information on AMRs see the following review articles: [10.1038/nrmicro2312](https://doi.org/10.1038/nrmicro2312) and [10.3390/antibiotics13121112](https://doi.org/10.3390/antibiotics13121112).

**Table of approximate minimum times for a job to be completed on Galaxy using specified tools.**

* Note, these times apply only to the specific input file we will be using in this activity, the Zymo\_Gut\_Standard\_D6331\_subset that is ~340MB, and will take longer (or much longer) for larger (or much larger) input files.

| Flye | Bandage Image | ABRicate |
| --- | --- | --- |
| 5 min | < 5 min | < 5 min |

### 5.2.4 Activity 1 – Flye assembly

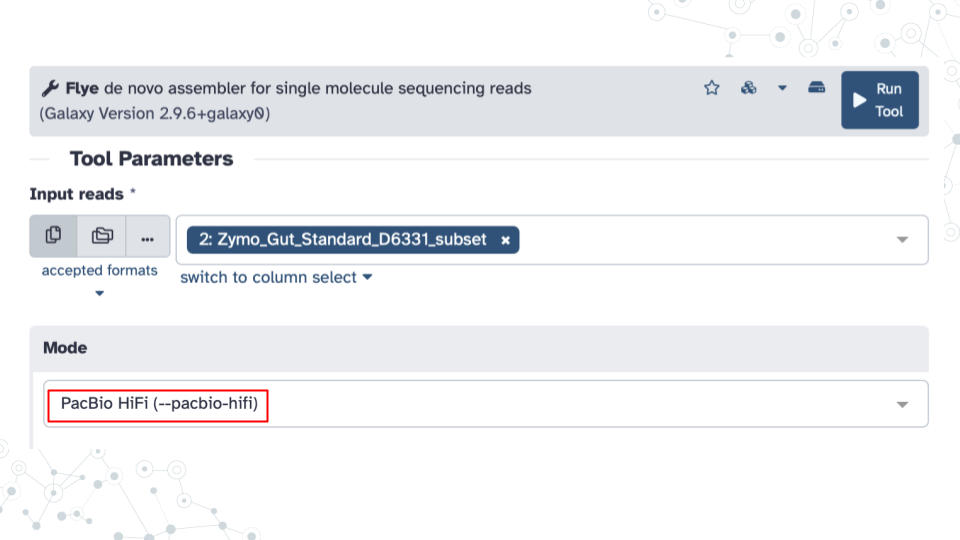
*Estimated time: 45 min*

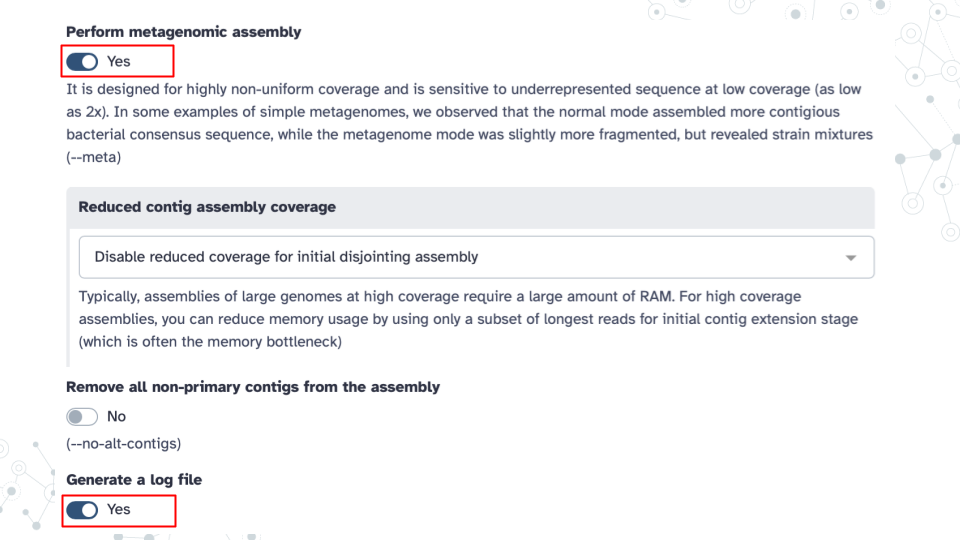
#### 5.2.4.1 Instructions

**1. Run Flye in Galaxy** - using Zymo Gut Standard D6331 subset to assemble gut microbial genomes.

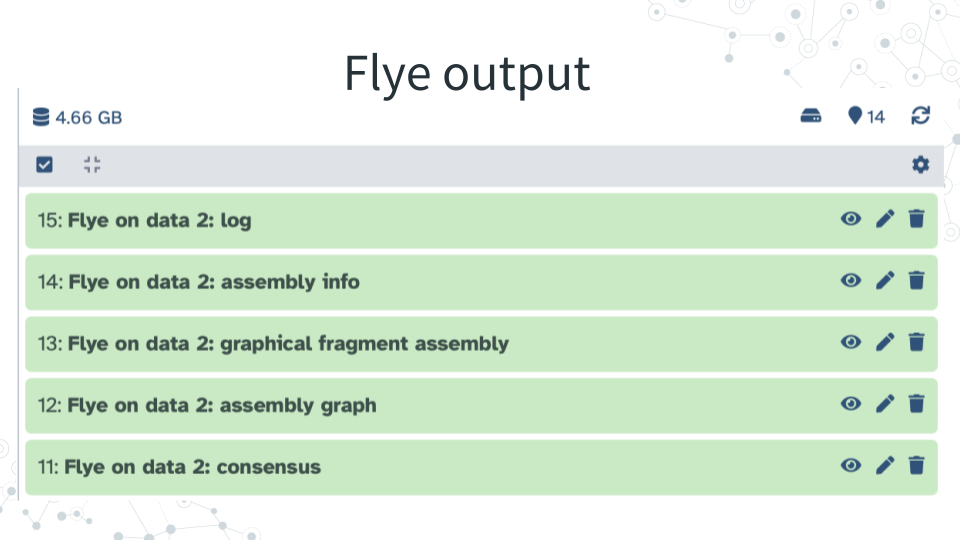
1. Obtain .fastq file from Zymo\_Gut\_Standard\_D6331\_subset: <https://usegalaxy.org/u/valerie-g/h/zymo-gut-standard-d6331-subset-1>
2. Name your new history **“Finding gut AMRs”**.
3. Run Flye tool to assess sequence quality using the following Tool Parameters:

* Under **Input Reads**: select your Zymo\_Gut\_Standard\_D6331 **.fastq** dataset.
* Under Mode: select PacBio HiFi (--pacbio-hifi) option, since the sequences were obtained using PacBio HiFi sequencing technology.
* Under **Perform metagenomic assembly**: select Yes
* Under **Generate a log file**: select Yes





**2. View Flye results** - Explore Flye output files and answer questions below.



#### 5.2.4.2 Questions

**1. Explore Flye tool purpose and output.**

| A. In your own words describe the purpose of Flye based on the Introduction section, and from the Purpose section of Flye tool description. |
| --- |
|  |

| B. How many Flye output files did you get back, and what are they? |
| --- |
|  |

| C. What are the file extensions (formats) for the following Flye output files? |
| --- |
| consensus |
| graphical fragment assembly |

**2. Explore Flye log report file.**

| A. Based on your Flye output log report file: At the very bottom of the very long file find how many bases were assembled. What is the Total length? |
| --- |
|  |

| B. Based on your Flye output log report file: At the very bottom of the file find the length of the longest assembled fragment? Look for Largest frg? |
| --- |
|  |

| C. What proportion of input was assembled into contigs? |
| --- |
| *At the very top of the log file you will find that the input number of bases was 177,760,975 (Look for Total read length). Compare to the total length after assembly from your answer to question 2A above* |
|  |
|  |

**3. Explore Flye assembly info file.**

| A. Based on your Flye output assembly info file sorted by contig length (high to low, in base pairs, bp) - What is the longest `contig size? |
| --- |
|  |

| B. Based on your Flye output assembly info file sorted by contig length (high to low, in base pairs, bp) - What is the shortest contig size? |
| --- |
|  |

**4. Explore Flye output consensus file (in FASTA format).**

| A. What is the beginning line of the FASTA format? |
| --- |
|  |

| B. How does FASTA sequence format differ from FASTQ sequence format? |
| --- |
| *See this link for a quick summary comparison of the 2 formats [https://compgenomr.github.io/book/fasta-and-fastq-formats.html](https://compgenomr.github.io/book/fasta-and-fastq-formats.html)* |
|  |

**5. Test your general understanding of genome assembly.**

| A. Summarize your experience with assembling a genome with Flye. |
| --- |
| *E.g., were you surprised at the percentage of assembled input, at the length of the largest contig, at the abundance of linear contigs, or anything else?* |
|  |

| B. What would you want the ideal genome assembly tool to do? |
| --- |
|  |

### 5.2.5 Activity 2 – Visualizing contigs

*Estimated time: 15 min*

#### 5.2.5.1 Instructions

Run **Bandage Image** tool in Galaxy on your Flye: **graphical fragment assembly** output (in gfa1 format) using default parameters.

#### 5.2.5.2 Questions

| 1. Paste the resulting image below. |
| --- |
|  |

| 2. Describe contig profile based on the Bandage Image. |
| --- |
|  |

| 3. Do you expect to obtain more contigs, larger or circular contigs with more sequencing reads? why? |
| --- |
|  |

### 5.2.6 Activity 3 – Finding AMRs

*Estimated time: 30 min*

#### 5.2.6.1 Instructions

1. Run **ABRicate** tool in Galaxy using Flye **consensus** as input using the following Tool Parameters:

* Under **Input Reads**: select your Flye: consensus output in FASTA format
* IMPORTANT: Under **Advanced Option**s: select NCBI Bacterial Antimicrobial Resistance Reference Gene Database as your database option; the default‘resfinder’ may not work well.

1. Explore **ABRicate report file**.

**Abricate output report** has the following information:

| Column | Description |
| --- | --- |
| FILE | The filename this hit came from |
| SEQUENCE | The sequence in the filename |
| START | Start coordinate in the sequence |
| END | End coordinate in the sequence |
| GENE | ABR gene name |
| COVERAGE | What proportion of the gene is in our sequence |
| COVERAGE\_MAP | A visual represenation of coverage map (gaps or no gaps) |
| GAPS | Was there any gaps in the alignment - possible pseudogene? |
| %COVERAGE | Proportion of gene covered |
| %IDENTITY | Proportion of exact nucleotide matches |
| DATABASE | The database this sequence comes from |
| ACCESSION | The genomic source of the sequence |

1. Answer questions below.

#### 5.2.6.2 Questions

**1. Explore Abricate output report.**

| A. How many AMR genes were detected? This is the number of rows in your file |
| --- |
|  |

| B. How many DIFFERENT AMR genes were detected and what are their GENE names? |
| --- |
|  |

| C. What are the different AMR genes resistant to? What is their RESISTANCE? |
| --- |
|  |

| D. How many DIFFERENT contigs had AMRs? |
| --- |
|  |

**2. Research an AMR gene.**

| A. Research and write a small paragraph report on one of the AMR genes. |
| --- |
| *Use any search tools for your research, but we encourage you to use PubMed [https://pubmed.ncbi.nlm.nih.gov/](https://pubmed.ncbi.nlm.nih.gov/) where you can find many scientific articles on the topic if you search for e.g. your AMR gene name, or resistance name or using a sentence as input. Talk about anything of interest, e.g., which microbes have the AMR of interest, what is the substance to which the gene shows resistance to, where could the resistance to this substance come from, what are possible health implications, etc.* |
|  |

| B. Ask one question you want to know about AMRs? |
| --- |
|  |

### 5.2.7 Grading Criteria

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

### 5.2.8 Footnotes

**Resources** - Google Doc

**Contributions and Affiliations**

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

Last Revised: May 2025

## 5.3 Discussion - Finding AMRs

### 5.3.1 Activity

*Estimated time: 25 min*

#### 5.3.1.1 Instructions

1. Form new groups of four

* <https://docs.google.com/spreadsheets/d/11eoJgm9mehxGWWzh8IZYDCDmnCmSyshopPYHewvpC8c/edit?usp=sharing>

1. Pair up into groups (10 min)
2. Discuss – Each group member briefly describes answers to prelab assignment
3. Summarize – Identify best answer and add to slidedeck

* <https://docs.google.com/presentation/d/1ZC8yjuk8V5pt3yh1p4_jkhTw0Pu4mHViAmNF-UPWAHs/edit?usp=sharing>

1. Share group discussion (2 min each group)

## 5.4 Project - Finding AMRs

### 5.4.1 Purpose

To explore soil metagenomics using Galaxy tools for de novo genome assembly (Flye), assembled contig visualization (Bandage), finding antimicrobial resistance (AMR) genes (ABRicate) and binning contigs into larger groups/bins (MetaBAT2). You will also explore information about the taxonomic classifier GTDB-Tk. This will allow you to compare genomic assemblies and AMRs between soil and Zymo gut standard.

### 5.4.2 Learning Objectives

1. Use **Galaxy** tools to
2. Use **Flye** tool to perform de novo genome assembly of long reads into ‘contigs’
3. Visualize the contigs with **Bandage** visualization tool
4. Use **ABRicate** tool to find AMRs in contig assemblies
5. Use **MetaBAT2** to bin contigs into larger MAGs
6. Learn about **GTDB-Tk** taxonomy classifier tool

### 5.4.3 Introduction

In this activity we will practice de novo genome assembly with Flye, using soil metagenomes from Nanopore sequencing . You will have an opportunity to compare and contrast assembled contigs and antimicrobial resistance profiles of soil and gut and think about the differences and similarities in microbial diversity of the two environments.

In this project you will also learn about MetaBAT2 and GTDB-Tk. Let’s start now. The most up to date long read metagenomics workflow includes contig assembly (e.g. using Flye tool), followed by contig binning into larger metagenome-assembled genomes (MAGs) (with e.g. MetaBAT2 tool) and finally MAG classification (e.g. with GTDB-Tk). MetaBAT2 is an algorithm that bins (or groups) sequence fragments (contigs) into larger MAGs or draft genomes. Subsequently, MAGs can be taxonomically classified by GTDB-Tk.

You will not be executing GTDB-Tk in this activity to stay within a reasonable activity time frame. However, for your project work, you will have a chance to test GTDB-Tk on your genome assemblies and bins. A note to your future self working on a project - as a rule, after genome assembly, if your contigs are 500 kb or above, they will be considered large enough to be passed to GTDB-Tk without the need for binning. Contigs of < 500 kb will be binned and passed on to GTDB-Tk for taxonomy classification as bins.

**Table of approximate minimum times for a job to be completed on Galaxy using specified tools.**

* Note, these times apply only to the specific input file we will be using in this activity, the nanopore-soil-subset that is 5.4 GB

| Flye | Bandage | ABRicate | MetaBAT2 |
| --- | --- | --- | --- |
| 5 hours | < 5 min | < 10 min | < 10 min |

### 5.4.4 Activity 1 – Genome assembly with Flye

*Estimated time: 30 min (activity time DOES NOT include the Flye run time on Galaxy)*

#### 5.4.4.1 Instructions

1. Run Flye in Galaxy on quality filtered (with fastp tool) nanopore soil pilot subset [nanopore-soil-subset-filtered](https://usegalaxy.org/u/valerie-g/h/nanopore-soil-subset-filtered) to de novo assemble soil microbial genomes.
2. Obtain .fastq file from a subset of Nanopore-sequencing soil study

* [https://usegalaxy.org/u/valerie-g/h/nanopore-soil-subset-filtered](%5Bhttps://usegalaxy.org/u/valerie-g/h/nanopore-soil-subset-filtered)

1. Name your new history **“Finding soil AMRs”**
2. Run **Flye** tool to assess sequence quality using the following **Tool Parameters**

* Under **Input Reads**: select your nano pore-soil-subset-filtered **.fastq** dataset.
* Under **Mode**: select --nano-raw option, since the sequences were obtained using Nanopore sequencing technology.
* Under **Perform metagenomic assembly**: select Yes
* Under **Generate a log file**: select Yes

1. Explore Flye output **assembly info** file which is sorted by length (in base pairs, bp) of the contig (high to low).

#### 5.4.4.2 Questions

| 1. How many contigs were assembled? |
| --- |
| *Note: Since each contig is represented by a separate row (or line) in the assembly info file, simply clicking on the assembly info file and recording the number of lines listed in the file will correspond to the number of contigs* |
|  |

| 2. What is the longest contig size? |
| --- |
|  |

| 3. What percent of input was assembled into contigs? |
| --- |
| - *Note 1: Based on the log file, the input going into flye assembly was 6,103,654,873 bases.* |
| - *Note 2: Based on the log file, the output going of flye assembly was 154,251,885 bases.* |
|  |

| 4. Why do you think only a small fraction of reads was assembled into contigs? |
| --- |
|  |

**5. Compare soil assembly to the Zymo gut standard assembly provided the following observations:**

* For this activity you can consult back to your Prelab: Finding AMRs.
* The number of contigs assembled from **339.9 MB** of sequencing data from the Zymo gut standard D6331 subset was **265**, much smaller than for the filtered soil sample of 5.4 GB! Yet, the largest contig size for the Zymo gut standard was **2,158,044** (almost 9 times larger) and circular (while the largest contig for the soil sample was linear).

| 5A. Why do you think the number of contigs in the soil sample was so much higher than the number of contigs in the Zymo gut standard? |
| --- |
| - *Hint 1 - it is NOT because of the difference in the size of the sequencing file.* |
| - *Hint 2 - Think about possible differences in the microbial diversity of the two samples.* |
|  |

| 5B. Why do you think it was possible to assemble a much larger and circular contig with the Zymo gut standard sample compared to the soil sample? |
| --- |
|  |

### 5.4.5 Activity 2 – Contig visualization with Bandage

*Estimated time: 15 min*

#### 5.4.5.1 Instructions

1. Run **Bandage** Image tool in Galaxy to visualize contigs.

* Run **Bandage Image** tool in Galaxy using your **Flye: graphical fragment assembly file** (in gfa1 format) as input, using default parameters.

#### 5.4.5.2 Questions

| 1. Paste the resulting image below. |
| --- |
|  |

| 2. Describe contig profile based on the Bandage Image result. |
| --- |
|  |

### 5.4.6 Activity 3 – Finding AMRs

*Estimated time: 30 min*

#### 5.4.6.1 Instructions

1. Run **ABRicate** tool in Galaxy using Flye consensus as input using the following **Tool Parameters**:

* Under **Input Reads**: select your Flye: consensus output in FASTA format
* IMPORTANT: Under **Advanced Options**: select NCBI Bacterial Antimicrobial Resistance Reference Gene Database as your database option; the default ‘resfinder’ may not work well.

1. Explore **ABRicate report** file.

* Note, Abricate output report has the following information:

| Column | Description |
| --- | --- |
| FILE | The filename this hit came from |
| SEQUENCE | The sequence in the filename |
| START | Start coordinate in the sequence |
| END | End coordinate in the sequence |
| GENE | ABR gene name |
| COVERAGE | What proportion of the gene is in our sequence |
| COVERAGE\_MAP | A visual represenation of coverage map (gaps or no gaps) |
| GAPS | Was there any gaps in the alignment - possible pseudogene? |
| %COVERAGE | Proportion of gene covered |
| %IDENTITY | Proportion of exact nucleotide matches |
| DATABASE | The database this sequence comes from |
| ACCESSION | The genomic source of the sequence |

#### 5.4.6.2 Questions

| 1. How many AMR genes were detected? This is the number of rows in your file. |
| --- |
|  |

| 2. How many DIFFERENT AMR genes were detected and what are their GENE names? |
| --- |
|  |

| 3. What are the different AMR genes resistant to? What is their RESISTANCE? |
| --- |
|  |

| 4. How many DIFFERENT contigs had AMRs? |
| --- |
|  |

**5. Research and write a small paragraph report on one of the AMR genes you found.**

* Use any search tools for your research, but we encourage you to use PubMed <https://pubmed.ncbi.nlm.nih.gov/> where you can find many scientific articles on the topic if you search for e.g. your AMR gene name, or resistance name or using a sentence as input.
* Talk about anything of interest, e.g., which microbes have the AMR of interest, what is the substance to which the gene shows resistance to, where could the resistance to this substance come from, what are possible health implications, etc.

| 5A. Report on one of the AMR genes you found. |
| --- |
|  |

| 5B. Why do you think the AMRs you found in soil differ from the AMRs you found in the gut (from your pre-lab)? |
| --- |
|  |

### 5.4.7 Activity 4 – Bin contigs with MetaBAT2

*Estimated time: 20 min*

#### 5.4.7.1 Instructions

1. In Galaxy, find and click on **MetaBAT2** tool and explore tool parameters.
2. Run **MetaBAT2** tool in Galaxy using Flye consensus as input using the **following Tool Parameters**:

* Under **Fasta file containing contigs**: select your Flye: consensus output in FASTA format
* Under **Output options**: from the **Extra outputs** dropdown menu select: Process log file.

#### 5.4.7.2 Questions

**1. Explore MetaBAT2 tool and parameters.**

| 1A. What is the function of MetaBAT2 tool based on Galaxy tool description on top? |
| --- |
|  |

| 1B. Under Tool Parameters for MetaBAT2, find and record below the Minimum size of a contig for binning (a value given in basepairs, bp). |
| --- |
|  |

| 1C. Under Tool Parameters and Output options for MetaBAT2, find and record below the Minimum size of a bin as the output. |
| --- |
|  |

**2. Explore MetaBAT2 tool output.**

| 2A. Open MetaBAT2 Process log output file and record how many bins were formed from contigs. |
| --- |
|  |

| 2B. Open MetaBAT2 Process log output file and record how many bases in total were used to form bins. |
| --- |
|  |

| 2C. What percent of contig bases formed bins (given that 154,251,885 bases were in Flye output)? |
| --- |
|  |

| 2D. Based on percent of contigs that formed bins (from activity 4-2.1 above) did metaBAT2 do a good job of binning the contigs? |
| --- |
|  |

**3. Examine MetaBAT2 Bin sequences output, which is a DATASET COLLECTION, where each collection is a separate bin.**

| 3A. Click on the MetaBAT2 Bin sequences output. How many bins (or Galaxy ‘folders’) are there? |
| --- |
|  |

| 3B. Click on the MetaBAT2 Bin sequences output and then on bin 1. Without ‘eyeballing’ the fasta file, note how many sequences (contigs) were included in bin 1. |
| --- |
|  |

| 3C. Click on the MetaBAT2 Bin sequences output and then on bin 2. Without ‘eyeballing’ the fasta file, note how many sequences (contigs) were included in bin 2? |
| --- |
|  |

### 5.4.8 Activity 5 – Read about GTDBtk tool in Galaxy

*Estimated time: 15 min*

#### 5.4.8.1 Instructions

1. In Galaxy, find and click on **GTDB-Tk Classify genomes.**

#### 5.4.8.2 Questions

| 1. Read the GTDB-Tk Classify genomes tool’s “What it does” part and summarize what GTDB-Tk does. |
| --- |
|  |

| 2. What is the ideal input sequence for GTDB-Tk classification: 1) raw sequences, 2) contigs or 3) large contigs(>500kb) and MAGs? |
| --- |
|  |

**3. Visit** [**https://gtdb.ecogenomic.org/stats/r220**](https://gtdb.ecogenomic.org/stats/r220) **to explore the database and database statistics.**

| 3A. How many bacterial species are present in GTDB database Release 220? |
| --- |
|  |

| 3B. Scroll through website . Although it has a lot of complex information, what is one thing you found interesting about GTDB-Tk content? |
| --- |
|  |

### 5.4.9 Grading Criteria

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

### 5.4.10 Footnotes

**Resources**

* Google Doc

**Contributions and Affiliations**

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

Last Revised: June 2025

## 5.5 Presentation - Finding AMRs

### 5.5.1 Activity

*Estimated time: 25 min*

#### 5.5.1.1 Instructions

1. Open the “Finding AMRs Presentation” slidedeck

* <https://docs.google.com/presentation/d/1AXZkQ0CyIasnkwyz0uvPsFic9eFrnPDQX_DbyffNmxQ/edit?usp=sharing>

1. Create slides covering at least two of these topics
2. Summarize the metagenome assembly statistics for the BioDIGS soil sample (Results)
3. Report on one of the AMR genes you found (Discussion)
4. Describe GTDB and the associated GTDB-Tk software toolkit (Methods)
5. Iterate your slides being sure to
6. Add bullet points on key details you understand (Notice) and questions you have (Wonder)
7. Include one or more plots or images that support your points
8. Create slide titles that summarize your main takeaway
9. Present at next class (5 min each group)

#### 5.5.1.2 Footnotes

**Contributions and Affiliations**

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

Last Revised: February 2025

# 6 Possible Datasets

## 6.1 Lecture - Possible Datasets



[Slides: Possible Datasets](https://docs.google.com/presentation/d/1VxwSmAY8BUs3EfVcxPm3I8kNYJWjVqoHJrGOX3X3sog/edit?usp=sharing)

## 6.2 Activity - Possible Datasets

### 6.2.1 Activity

*Estimated time: 50 min*

#### 6.2.1.1 Instructions

1. Skim three abstracts
2. Pick one and answer the following questions
3. Notice – What about this abstract most interests you?
4. Dataset – Summarize at a high level where the samples came from, how many there are, and what technology was used for sequencing.
5. Wonder – Two or three questions you would like to ask using this (and any other) datasets.
6. Post your answers by replying to the “Project Work: Possible Datasets” topic in the Discussion Forum

| Possible Datasets (Long-read PacBio) |  |  |  |
| --- | --- | --- | --- |
| **Soil** |  |  |  |
|  | Antarctic | PRJNA1126331 | <https://pubmed.gov/39300163/> |
|  | Biocrust | PRJNA691698 | <https://pubmed.gov/34795375/> |
| **Water** |  |  |  |
|  | Fresh Water | PRJNA924152 | <https://pubmed.gov/36823661/> |
|  | Ocean Water | PRJNA853328 | <https://pubmed.gov/36448813/> |
| **Human Gut** |  |  |  |
|  | Vegan/Omnivore | PRJNA750084 | <http://pubmed.gov/36289209> |
|  | Infant Nutrition | PRJNA1139951 | <http://pubmed.gov/31022095> |
| **and More!** |  |  |  |
|  | Lamb Gut | PRJNA595610 | <http://pubmed.gov/34980911> |
|  | Deadwood | PRJNA603240 | <http://pubmed.gov/39627869> |
|  | Cheese | PRJNA778418 | <http://pubmed.gov/9948695> |
|  | Whey | PRJNA454439, PRJNA477604 | <http://pubmed.gov/6593500> |

### 6.2.2 Grading Criteria

* Submit URL to your reply on Canvas

### 6.2.3 Footnotes

**Contributions and Affiliations**

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

Last Revised: May 2025

# 7 Kickstart Project Work

## 7.1 Lecture - Kickstart Project Work



[Slides: Possible Datasets](https://docs.google.com/presentation/d/1j1u4frdYt18nVmjMs6ZPRoP1KfXxognVUYG-2itH6kE/edit?usp=sharing)

## 7.2 Activity - Kickstart Project Work

### 7.2.1 Learning Objectives

1. Import an example SRA sequencing data file by accession id (e.g. SRR)
2. Perform classification of nanopore-sequenced soil MAGs with GTDB-Tk
3. Import an SRA dataset you would like to analyze for your research project
4. Use a new Galaxy tool that we have not worked through in class

### 7.2.2 Activity 1 - Import an SRA file into Galaxy

#### 7.2.2.1 Instructions

1. Upload sequencing reads for accession id SRR29980924 in Galaxy using **Faster Download and Extract Reads in FASTQ** tool .
2. Make your Galaxy history sharable.

#### 7.2.2.2 Questions

| 1. Share your Galaxy history with SRR29980924fastq data you downloaded with Faster Download and Extract Reads in FASTQ tool. |
| --- |
| Galaxy history Link: |

| 2. Find and record the following features of your downloaded dataset below. |
| --- |
| File size: |
| File format: |
| Job Runtime (Wall Clock): |
| First line of the sequencing file: |

| 3. Inspect SRR29980924 entry in NCBI <https://www.ncbi.nlm.nih.gov/sra>. Obtain the following sequencing information. |
| --- |
| Organism: |
| Instrument: |
| Strategy: |
| Run: |

### 7.2.3 Activity 2 – Taxonomically classify nanopore-soil-subset MAGs with GTDB-Tk in Galaxy

#### 7.2.3.1 Instructions

Using **GTDB-Tk Classify genomes** tool, classify the nanopore-soil MAGs you obtained from **MetaBAT2** binning of contigs during Project: Finding AMRs, Activity 4. Note, MetaBAT2 bins are a collection, so try using **Dataset collection** (instead of multiple individual files) as input.

#### 7.2.3.2 Questions

| 1. Record summary of MAG classification of Bacteria below. |
| --- |
|  |

| 2. Record summary of MAG classification of Archaea below. |
| --- |
|  |

| 3. Share your Galaxy history to GTDBtk classification below. |
| --- |
|  |

### 7.2.4 Activity 3 – Import an SRA dataset of interest for your research project

#### 7.2.4.1 Instructions

Upload sequencing reads from a dataset of interest using the **Faster Download and Extract Reads in FASTQ** tool. Refer back to the table in the **Activity: Possible Datasets** for BioProjects with known long read metagenomics datasets.

#### 7.2.4.2 Questions

| 1. Record the following features of your downloaded dataset below. |
| --- |
| Accession ID: |
| File size: |
| Job Runtime (Wall Clock): |

| 2. Inspect your SRA entry in NCBI <https://www.ncbi.nlm.nih.gov/sra>. Obtain the following sequencing information. |
| --- |
| Organism: |
| Instrument: |
| Strategy: |
| Run: |

### 7.2.5 Activity 4 – Use a new Galaxy tool that may be useful for your project

#### 7.2.5.1 Instructions

Find one new Galaxy tool that you think may be useful for your research project. A few possibilities to get you started include:

* Prokka [Galaxy tutorial](http://training.galaxyproject.org/training-material/topics/genome-annotation/tutorials/annotation-with-prokka/tutorial.html)
* Bakta [Galaxy Tutorial](https://training.galaxyproject.org/training-material/topics/genome-annotation/tutorials/bacterial-genome-annotation/tutorial.html)
* antiSMASH [Galaxy tutorial](<http://training.galaxyproject.org/training-material/topics/genome-annotation/tutorials/secondary-metabolite-discovery/tutorial.html>
* CheckM2 [Galaxy tutorial](https://training.galaxyproject.org/training-material/topics/genome-annotation/tutorials/bacterial-genome-quality-control/tutorial.html)
* Tools mentioned in tutorials like [GTN: Bacterial Genome Annotation](https://training.galaxyproject.org/training-material/topics/genome-annotation/tutorials/bacterial-genome-annotation/tutorial.html)
* Tools used in workflows like <https://nf-co.re/mag> (not all tools will be available on Galaxy)

#### 7.2.5.2 Questions

| 1. Briefly describe the tool e.g. how it works, what it takes as input, and what it produces as output. |
| --- |
|  |

| 2. Describe briefly any problems you had running the tool and troubleshooting steps you tried. |
| --- |
|  |

| 3. If you were able to get the tool to run, describe the results and whether they provide additional information for your project. |
| --- |
|  |

### 7.2.6 Grading Criteria

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

### 7.2.7 Footnotes

**Resources**

* Google Doc

**Contributions and Affiliations**

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

Last Revised: June, 2025

## 7.3 Project Work Organizer

### 7.3.1 Purpose

**Kickstart Project Work** - The purpose of kickstart project work is to gain practice in scientific exploration and research, communication, and work on your scientific poster! A few ideas to get you started can be found in the **Possible Datasets** Lecture and Acvitivity below. **Your project work will culminate in Poster making, sharing and presentation.** a. Each group will present their capstone project during an in-class poster presentation.

**Project Work phase** contains many elements to scaffold your research as you progress towards a final poster presentation. These include round table updates, written check-ins, advisory meetings, and peer reviews. During the Project Work phase, you will be documenting, summarizing and updating relevant documenents which will fall under 3 categories:

1. Round Table Data & Troubleshooting\*\* - Reviewed and discussed as a community In-Class   
2. Written Check-Ins\*\* - Group-specific write-ups checked by instructor(s)  
3. Advisory Meetings\*\* - Group-specific in-person meetings between individual groups and instructor(s)

**Project Work Organizer** - Research lab meetings are great forums to receive community analysis and feedback. Many classes going forward will be spent discussing your project progress, alternating between data results and troubleshooting methods. Refer to this organizer to see what we’ll discuss on a given day and find the appropriate slidedeck.

Use this organizer to find instructions and track documents.

### 7.3.2 Round Table Data & Troubleshooting

Research lab meetings are great forums to receive community analysis and feedback. In-class round table data and troubleshooting sessions and slidedecks will be used to report on and get feedback on your ongoing research. Consider these in-class project work updates your research lab meetings!

#### 7.3.2.1 Instructions

**1. Track and share your work and progress using Slidedecks**

1. Instructor will create a single Slidedeck - **Round Table Data** for all groups
2. Instructor will create a single Slidedeck - **Round Table Troubleshooting** for all groups
3. Instructor will create a single Slidedeck - **Poster** for all groups
4. Instructor will create a single Slidedeck - **Final Presentation** for all groups

**2. Use the following (sample) project Work organizer for your In-class Project Work updates.**

| Course Week | Day 1 | Day 2 |
| --- | --- | --- |
| 7 | Round Table Data |  |
| 8 | Round Table Troubleshooting |  |
| 9 | Slidedeck - Round Table Data | Round Table Troubleshooting |
| 10 | Slidedeck - Round Table Data | Round Table Troubleshooting |
| 11 | Slidedeck Poster Review | Poster Review |
| 12 | Science Talks |  |
| 13 | Final Presentations |  |

### 7.3.3 Written check-ins

Written check-ins encompass a written document for groups-instructor interaction, and will also be used for peer reviews.

#### 7.3.3.1 Instructions

**1. Track and share your work and progress using test document**

For each group, instructor will create a text document **Doc** which will include the following sections:

**a. Written Check-in (Doc)** - Progress will be assessed through several written check-ins. These documents will allow you to summarize the work you’ve done, receive feedback from colleagues, and provide another opportunity to ask for help. Your check-in should be concise but complete (full sentences) and include the following information:

* New progress since last submission (15 pts)
* How you’ve addressed prior feedback (15 pts)
* Poster organization (10 pts)
* Struggles you are encountering and questions you would like advice on (10 pts)

**b. Written Check-in (Poster)** - The second written check-in will focus on starting your poster using material you developed for your Round Table updates. Posters should be 30 inches wide and we have provided a template with example two and three column formats. You can make your own design by selecting the Blank layout. Kickstart your poster by adding the following details:

* Title and authors (10 pts)
* Hypothesis along with diagram or figure (10 pts)
* Dataset table with sample metadata including accession ID, collection method, etc. (10 pts)
* Methods in the form of details and figure (10 pts)
* Results (10 pts)

**c. Peer Review** - Each person will work individually to peer review two separate projects. Review the poster associated with each project and provide feedback in group’s Check-in document on the following:

* Aspects of the project that you find fascinating
* Questions you have about the approach or results
* Suggestions on next steps, poster organization, etc.

### 7.3.4 Advisory Meetings

**Advisory Meeting** - Each group will schedule a meeting with the instructors to discuss their progress. These meetings will take place outside of class and should be scheduled before the deadline. A link to the when2meet schedule can be found in the Group Information section.

### 7.3.5 Sample project work organizer

With your instructor, during your Project Work, fill out the Project Work organizer below.

| Group | Check-in | Poster | Meeting #1 | Meeting # 2 |
| --- | --- | --- | --- | --- |
| Group A | Doc | Slides | Date 1 | Date 2 |
| Group B | Doc | Slides | Date 1 | Date 2 |
| Group C | Doc | Slides | Date 1 | Date 2 |
| Group D | Doc | Slides | Date 1 | Date 2 |

### 7.3.6 Footnotes

**Contributions and Affiliations**

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

Last Revised: June 2025

# 8 Galaxy Tutorials

## 8.1 A short introduction to Galaxy

### 8.1.1 Introduction

Galaxy <galaxyproject.org> is a free, open-source system for analyzing data, authoring workflows, training and education, publishing tools, managing infrastructure, and more. Among the notable features:

* [Graphical user interface](https://training.galaxyproject.org/training-material/topics/introduction/tutorials/galaxy-intro-101/tutorial.html) (GUI) for interactively running tools
* [Toolshed](https://toolshed.g2.bx.psu.edu/) with 10,000 tools ready to run
* Full featured [workflow](https://training.galaxyproject.org/training-material/topics/galaxy-interface/tutorials/workflow-editor/tutorial.html) functionality
* Terabytes of the latest, curated [reference data](https://galaxyproject.org/admin/reference-data-repo)
* Extensive training [tutorials](https://training.galaxyproject.org/) and infrastructure
* Large international [community](https://galaxyproject.org/community) of users and developers

### 8.1.2 Activity

*Estimated time: 40 min*

#### 8.1.2.1 Instructions

1. Review slides for “A short introduction to Galaxy”

* <training.galaxyproject.org/training-material/topics/introduction/tutorials/galaxy-intro-short/slides.html>

1. Complete the hands-on tutorial

* <training.galaxyproject.org/training-material/topics/introduction/tutorials/galaxy-intro-short/tutorial.html>

### 8.1.3 Grading Criteria

* Submit URL to your shared Galaxy history on Canvas

### 8.1.4 Footnotes

**Resources**

* Introduction to Galaxy Analyses [topic](https://training.galaxyproject.org/training-material/topics/introduction/)

**Contributions and Affiliations**

* Frederick Tan, Johns Hopkins University

Last Revised: January 2025

## 8.2 QC and Galaxy Workflows

### 8.2.1 Introduction

In the “A short introduction to Galaxy” activity you learned how to upload a file, use a tool, view results, view histories, extract and run a workflow, and share a history. You will now practice using these skills to do one of the first things when encountering a new sequencing dataset – quality control (QC). After assessing the quality of both short and long reads, gain more practice with workflows by creating and editing a new workflow.

### 8.2.2 Activity 1 – Quality Control (QC)

*Estimated time: 90 min*

#### 8.2.2.1 Instructions

1. Complete the “Quality Control” hands-on tutorial: [training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/quality-control/tutorial](https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/quality-control/tutorial.html)

* **NOTE**: Do all hands-on steps in a single Galaxy history

### 8.2.3 Activity 2 – Creating, Editing and Importing Galaxy Workflows

*Estimated time: 30 min*

#### 8.2.3.1 Instructions

1. Complete the “Creating, Editing and Importing Galaxy Workflows” hands-on tutorial: [training.galaxyproject.org/training-material/topics/galaxy-interface/tutorials/workflow-editor/tutorial](https://training.galaxyproject.org/training-material/topics/galaxy-interface/tutorials/workflow-editor/tutorial.html)

### 8.2.4 Grading Criteria

* Submit URL to your shared Galaxy “Quality Control” history on Canvas

### 8.2.5 Footnotes

**Resources**

* Using Galaxy and Managing your Data [topic](https://training.galaxyproject.org/training-material/topics/galaxy-interface/)
* Introduction to Galaxy and Sequence analysis [pathway](https://training.galaxyproject.org/training-material/learning-pathways/intro-to-galaxy-and-genomics.html)

**Contributions and Affiliations**

* Frederick Tan, Johns Hopkins University

Last Revised: February 2025

## 8.3 Galaxy data collections

### 8.3.1 Introduction

Collections in Galaxy can be useful when you have to deal with large numbers of datasets. Examples of when you might use collections include when you have many DNA sequencing samples (e.g. .fastq files) or when you have many genome assembly contigs (e.g. .fasta files). Dataset collections allow you to not only organize these datasets but they also allow you to manipulate them in a single batch operation.

### 8.3.2 Activity

*Estimated time: 45 min*

#### 8.3.2.1 Instructions

1. Review the “Datasets versus collections” FAQ

* [training.galaxyproject.org/training-material/faqs/galaxy/histories\_datasets\_vs\_collections.html](http://training.galaxyproject.org/training-material/faqs/galaxy/histories_datasets_vs_collections.html)

1. Complete the “Using dataset collections” hands-on tutorial

* [training.galaxyproject.org/training-material/topics/galaxy-interface/tutorials/collections/tutorial.html](http://training.galaxyproject.org/training-material/topics/galaxy-interface/tutorials/collections/tutorial.html)

### 8.3.3 Grading Criteria

* Submit URL to your shared Galaxy history on Canvas

### 8.3.4 Footnotes

**Resources**

* Using Galaxy and Managing your Data [topic]

**Contributions and Affiliations**

* Frederick Tan, Johns Hopkins University

Last Revised: January 2025

# 9 SciServer Tutorials

## 9.1 Meet R

### 9.1.1 Introduction

R is a popular programming language for statistical computing and data visualization [wikipedia](<https://wikipedia.org/wiki/R_(programming_language)>. Notable features include the [RStudio IDE](https://posit.co/products/open-source/rstudio) built to help you be more productive with R (and Python); the [tidyverse](https://tidyverse.org/packages) collection of packages that aims to make data science faster, easier, and more fun; and importantly the [Bioconductor](https://bioconductor.org) project with its active support community and robust genomics tooling. Now that you completed “test-driveR” and got a taste of analyzing a real world genomics dataset using R it’s time to learn more about how R stores and operates on data using concepts such as variables, vectors, and data.frames.

### 9.1.2 Activity

*Estimated time: 25 min*

#### 9.1.2.1 Instructions

1. Resume your C-MOOR LearnR container
2. Open sciserver.org in a web browser and log in to your account.
3. Click “Home” in the top menu to return to the home page.
4. Scroll down to the second set of boxes and click “Compute”.
5. Resume your C-MOOR LearnR container by clicking on its name.
6. Complete your next LearnR tutorial
7. Click on “Meet R!”. The tutorial will open in a new tab.
8. Complete the tutorial and answer the questions below.

#### 9.1.2.2 Questions

| 1. Variables – Define what is a variable, give a line of R code that stores information in a variable, and explain how one would view the data stored in a variable. |
| --- |
| Define variable: |
| Code: |
| Explain: |

| 2. Vectors – Provide two lines of R code, one which stores a vector of numbers in a variable, and one which performs a mathematical operation on each element of that variable (e.g. multiplication). |
| --- |
| Code 1: |
| Code 2: |

| 3. Plotting – Copy and paste as text your code from the final code block in the tutorial. Also insert a screenshot of your resulting plot. |
| --- |
| Code text: |
| Plot screenshot: |

### 9.1.3 Grading Criteria

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

### 9.1.4 Footnotes

**Resources**

* Google Doc
* [R cheat sheet](https://github.com/C-MOOR/cure-rnaseq/blob/master/tutorials/Rcheatsheet.md)

**Contributions and Affiliations**

* Katherine Cox, Johns Hopkins University
* Frederick Tan, Johns Hopkins University

Last Revised: February 2025

## 9.2 R for Data Science 1

### 9.2.1 Introduction

[R for Data Science (2e)](https://r4ds.hadley.nz) covers how to do data science with R using the tidyverse collection of packages. There are many possible reasons for the continued popularity of tidyverse, but one stand out principle the package authors embrace is that programs should be “easy to use by humans. Computer efficiency is a secondary concern” (see this and more at the [tidy tools manifesto](https://tidyverse.tidyverse.org/articles/manifesto.html#design-for-humans0). Here you will learn more about what kind of data can be organized into data.frames and possible ways to summarize, explore, and visualize datasets.

### 9.2.2 Activity

*Estimated time: 50 min*

#### 9.2.2.1 Instructions

1. Complete the “r4ds” LearnR tutorial.

* In addition to the “OCS – Global Diets” section being optional, you similarly do not need to complete all of the Exercises in 3.3.1.

1. Answer the questions below.

#### 9.2.2.2 Questions

| 1. data.frames – Data frames organize data into a 2-dimensional table of rows and columns like a spreadsheet with some special restrictions. For the mpg data.frame, what do rows represent? What do columns represent? |
| --- |
|  |

| 2. Functions – Most R commands are functions (e.g. mean()) which take some input (a number, a vector, a dataframe…) and produce some output (a summary, a plot, a new vector…). Describe one function you learned about that you anticipate being useful in the future. |
| --- |
|  |

| 3. Common Problems – What are some common problems you ran into? Refer to Part 2. Troubleshooting if you’d like some inspiration for how to put your experience into words. |
| --- |
|  |

| 4. [*optional*] OCS - Global Diets Plotting – Which foods show a difference based on sex? |
| --- |
|  |

### 9.2.3 Grading Criteria

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

### 9.2.4 Footnotes

**Resources**

* Google Doc
* [R cheat sheet](https://github.com/C-MOOR/cure-rnaseq/blob/master/tutorials/Rcheatsheet.md)

**Contributions and Affiliations**

* Katherine Cox, Johns Hopkins University
* Frederick Tan, Johns Hopkins University

Last Revised: February 2025

## 9.3 test-driveR

### 9.3.1 Introduction

SciServer [sciserver.org](https://www.sciserver.org) is an online platform for doing scientific data analysis. It is used by scientists studying astronomy, biology, oceanography, and more, and is free as long as you are using it for scientific research. Using SciServer means you do not need a fancy computer or need to install any special programs on your computer, you can just log in with your internet browser to start doing research. For this course, we have set up SciServer with customized collections of programs as well as the data that we’ll be analyzing. Once you sign up for SciServer and are added to the group for this course, you will be able to access these tools and begin your data analysis journey!

### 9.3.2 Activity

*Estimated time: 25 min*

#### 9.3.2.1 Instructions

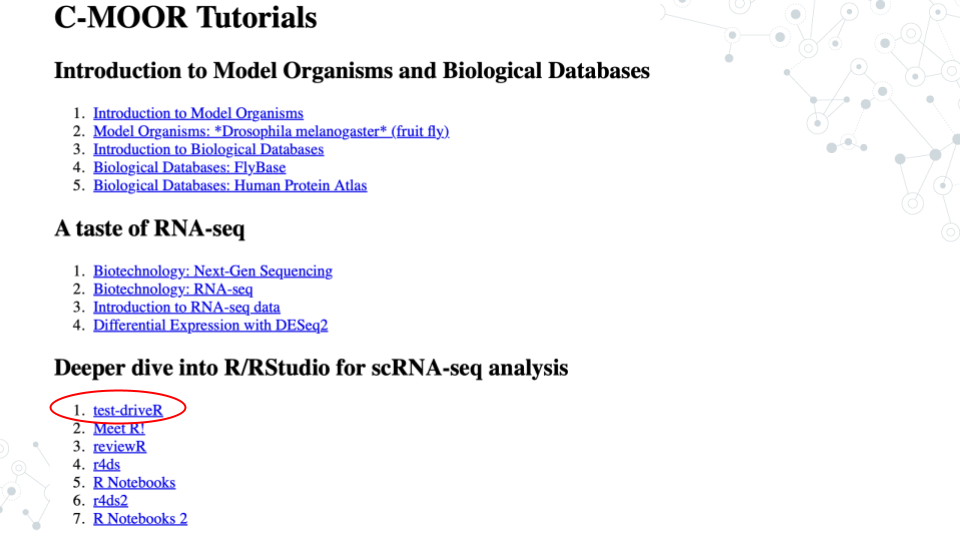
1. Accept invitation to join class SciServer Group
2. Navigate to <https://apps.sciserver.org/dashboard/groups>
3. Click “Accept Invitation”
4. Start up a “C-MOOR LearnR” compute container
5. Navigate to <https://apps.sciserver.org/compute>
6. Click “Create container”
7. Give your container a name (eg. my LearnR)
8. In the “Compute Image” drop-down menu, select “C-MOOR LearnR”
9. Under “Data Volumes”, check the box next to “C-MOOR Data”

 f. Scroll down and click “Create”. This may take a moment.

1. Start your C-MOOR LearnR container by clicking on its name

* NOTE: If you see “Error: C-MOOR data volume not mounted!” you most likely forgot to check the box next to “C-MOOR Data” when you created the container.

1. Complete your first LearnR tutorial
2. Click on “test-driveR”. The tutorial will open in a new tab.

 Complete the tutorial and answer the questions below.

#### 9.3.2.2 Questions

Use the blocks provided at the end of the tutorial to create variations of the code provided in the first three exercises (Sort Data, Extract Data, and Plot Data). When you are done, submit the code and resulting output that you found most interesting. Briefly describe why you found it most interesting.

| 1. Code – Copy and paste your actual code as text. |
| --- |
|  |

| 2. Output – Copy and paste as text if simple, otherwise use a screenshot. |
| --- |
|  |

| 3. Description. |
| --- |
|  |

### 9.3.3 Grading Criteria

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

### 9.3.4 Footnotes

**Resources**

* Google Doc
* [R cheat sheet](https://github.com/C-MOOR/cure-rnaseq/blob/master/tutorials/Rcheatsheet.md)

**Contributions and Affiliations**

* Katherine Cox, Johns Hopkins University
* Frederick Tan, Johns Hopkins University

Last Revised: February 2025

## 9.4 R for Data Science 2

### 9.4.1 Introduction

[R for Data Science (2e)](https://r4ds.hadley.nz) describes the importance of data visualization by saying that “a good visualization will show you things you did not expect or raise new questions about the data”. Here you will learn more about the [ggplot2](https://ggplot2.tidyverse.org) system for making graphs which is an elegant and versatile complement to what is available through software like Google Sheets. Increasingly more software packages use the ggplot2 system such as the Bioconductor [phyloseq](https://bioconductor.org/packages/phyloseq) package used for 16S rDNA analysis.

### 9.4.2 Activity

*Estimated time: 25 min*

#### 9.4.2.1 Instructions

1. Start the “r4ds2” LearnR tutorial.

* Focus on the first half of the tutorial, up to and including “4. Structure of a ggplot() command”

1. Bug fixes
2. For the Chapter 3, Quiz 2 on geom\_bar() aesthetics, you must check “size”
3. For the Chapter 4 exercises, you will need to add this code to the top of each code block

measles <- filter(us\_contagious\_diseases, disease=="Measles") measles\_MD <- filter(measles, state=="Maryland") measles\_VA <- filter(measles, state=="Virginia")

#### 9.4.2.2 Questions

There are seven examples of broken ggplot code at the end of “4. Structure of a ggplot() command”. Fix at least three of them and for one of them explain what was wrong and how you were able to figure it out. Remember that you must add the above three lines to the top of each code block.

| 1. What was the error? |  |
| --- | --- |
|  |  |

| 2. How did you figure it out? |  |
| --- | --- |
|  |  |

### 9.4.3 Grading Criteria

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

### 9.4.4 Footnotes

**Resources**

* Google Doc
* [R cheat sheet](https://github.com/C-MOOR/cure-rnaseq/blob/master/tutorials/Rcheatsheet.md)
* [ggplot2 cheat sheet](https://rstudio.github.io/cheatsheets/data-visualization.pdf)

**Contributions and Affiliations**

* Katherine Cox, Johns Hopkins University
* Frederick Tan, Johns Hopkins University

Last Revised: February 2025

# 10 Scientific Posters

In this section we will go over how to create a scientific poster. We chose a poster as our final deliverable in C-MOOR because they:

* **Facilitate participation in research symposiums**: Students can use their poster to apply to present at research symposiums or conferences with minimal additional effort beyond what they do in class. If your institution does not have a research day, consider starting one using these posters as a foundation as we did at Clovis Community College! Your school’s library is another great point of contact in getting a poster session or research day set up. Some of the events students have participated in using their C-MOOR projects include:
  + The Clovis Community College Research Day
  + The Notre Dame of Maryland University Research Day
  + The GRADS-4C Conference (2025)



* **Serve as a physical representation of student achievement**: If printed, once finished with the class, students can choose to take their poster with them or leave it with the class to be hung up around the classroom or the hallways. This gives their research project more visibility, serves as an advertisement for other students who might be interested in taking the class, and showcases what students have accomplished.
* **Can easily be shared digitally**: A presentation can’t be repeated twice without a recording, and each recording can take up a lot of digital storage. A poster on the other hand, is a great compressed product that can be posted on a student’s (or the institution’s) social media and help get their work seen by others. Posters are even small enough that they can be attached to an email and sent to prospective supervisors and collaborators interested in a quick summary of their work. See our [Look at This!](https://help.c-moor.org/c/look-at-this/) category on the C-MOOR Academy Discussion Forum to see all the posters students have created and shared so far!
* **Posters can be used to measure student achievement and understanding**: We have previously used these posters to evaluate student learning, even years after students have left the classroom. Check out our [poster on posters!](https://drive.google.com/file/d/1kK6FBBLbHiAAsTUgxwzHRlK1qCo__eCy/view?usp=sharing)

## 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 Making Group Poster

### 10.3.1 Preparing to make a poster

An important part of scientific research is presenting your findings. Poster is a powerful visual way to communicate new and exciting findings, share ideas and get feedback. No wonder Poster sessions are an integral part of any conference or symposium. Throughout research project work, students will work with their groups to put together a scientific research poster.

### 10.3.2 Part 1 - Choose a Template

*Estimated time: 5 minutes*

1. With your group, open a suggested [poster template](https://docs.google.com/presentation/d/1chz02nzYklAEjOtHrPsJZccGy1sFUZvqdlfcUyfj3I4/edit?usp=sharing) to use for your poster.
2. Discuss with your group how you will divide up the work and exchange important information (e.g. phone numbers, email). Consider the following sections:

* Abstract/Introduction
* Methods
* Results
* Conclusions/Discussion
* References
* Acknowledgments

Notice these posters are sized at 36 x 42 inches, but there are many other commonly used sizes. Refer to your instructor on what poster size to use; not every poster will fit every stand when printed (if applicable). Notice that the posters in this slidedeck come in two-column, three-column, and mixed formats. Think about your figures and tables and select a format that makes the most sense for them. There are also some example posters in the last two slides that we will use to examine the anatomy of a research poster further in the next part.

### 10.3.3 Part 2 - Make an Academic Research Poster

*Estimated time: —*

1. Complete the following components of your research poster with your group. You might not do them in this order, but these are the components you are being graded on. For more details on each of these sections and their role in a scientific paper, see the Scientific Literature Lab.

#### 10.3.3.1 The header: title, authors, and affiliations

Since you completed this work as a team of scientists, all of your team members are considered authors. List your team members in alphabetical order. The author’s affiliation is the university, college, research institution or company that the work was conducted at. An example affiliation for Clovis Community College would be: “Department of Biology, Clovis Community College, California, United States.” If multiple institutions exist, you will need to use superscript to denote who has what affiliation (see the example poster).

We will include the C-MOOR logo, the logo of your institution(s), and the most immediately relevant funding source (if applicable) in the header.

**Check the header of your poster:**

* ☐  All authors are listed
* ☐  All authors’ home institutions are referenced
* ☐  The title of the poster is focused on your specific project topic (ex. genes, variables, model organism)
* The title of the poster describes the main result of your research
* The title of the poster does not overstate the findings or significance of your research
* ☐  The C-MOOR logo is included
* ☐  The logos of your institutions are included

#### 10.3.3.2 Abstract/Introduction

An abstract or introduction is a concise summary of your work. An effective abstract/introduction will inform the reader of the relevant background to the research, scientific hypothesis being tested, the purpose of the study, the main methods, and the most important results and conclusions. You may notice that not all posters online contain and abstract; different fields of science have slightly different requirements and formats for abstracts.

Abstract is a kind of introduction because it provides context for your work by describing what is already known in the field, as well as an unknown that your research is addressing. The latter is often called the gap in knowledge.

Abstract is often difficult to write. When writing your abstract/introduction, it may be helpful to start writing at the end. Identify your scientific question, your hypothesis and the gap of knowledge first. Then brainstorm what you will need to tell your readers in terms of context and background.

**Your abstract/introcuction should do the following:**

Regarding background information

* ☐  Include relevant background information such as:
* A problem or question at hand,
* Model organism,
* Gene of interest and its function,
* Important associations (e.g. between genotype and phenotype),
* Phenotypes (typical, disease-state),
* Biological processes

Regarding your research:

* ☐  Why your research is novel
* ☐  Why your research matters (ex. any relevant disease states)
* ☐  The aim and/or hypotheses of your research
* ☐  Summarize your methodology
* e.g. Our hypothesis is that the healthy gut microbiome has more microbial diversity than the gut of an individual with a celiac disease.
* e.g. We hypothesize that maternal antibiotic treatment correlates with higher antibicrobial resistance in infant microbiome.
* ☐  State your main results without interpretation
* e.g. We found that gene X was differentially abundant between the eye and all other tissues
* e.g. We found an association between age and the presence of Y bacteria
* ☐  Clearly stated aim, hypothesis, and/or objectives of the research

Additionally, check that

* ☐  The abstract text should be no longer than ~200 words
* ☐  Include a diagram figure or table relevant to background and/or conclusion
* ☐  Include appropriate in-line citations are included

#### 10.3.3.3 Materials and Methods

The materials and methods section will detail your analysis of the data. Don’t provide any of your results, just the methods. If you did not generate the data yourself from raw samples, you will simply cite the paper that made them instead of detailing their construction. We will also need to list any programs we used and provide credit to their creators.

Some other things you might include would be what type of analysis you decided to do (which parts of the body parts you analyzed, sets of genes, what p-value you used, etc.).

**Check your poster’s materials and methods section:**

* ☐  You have listed the source of your data (ex. the published study your data came from) and added the appropriate in-line citation
* ☐  You have listed all R packages and/or websites you used (ex. DESeq2, ClustrProfiler, Phyloseq, FlyBase, Human Protein Atlas, CellxGene) and added the appropriate in-line citation
* ☐  If relevant, list what p-value you tested for significance at (usually 0.05)
* ☐  Another student from your class would be able to replicate your research based on what you wrote

#### 10.3.3.4 Results

The Results section is where you will detail your data primarily through **figures and tables**, though sometimes written text is included. Begin by creating your tables and figures.

Figure body:

* Has high resolution images, well spaced and labeled parts and text.
* Place the figures and tables in order of how you want to present them and name them such as **Figure 1, Figure 2, Table 1, Table 2, etc**.

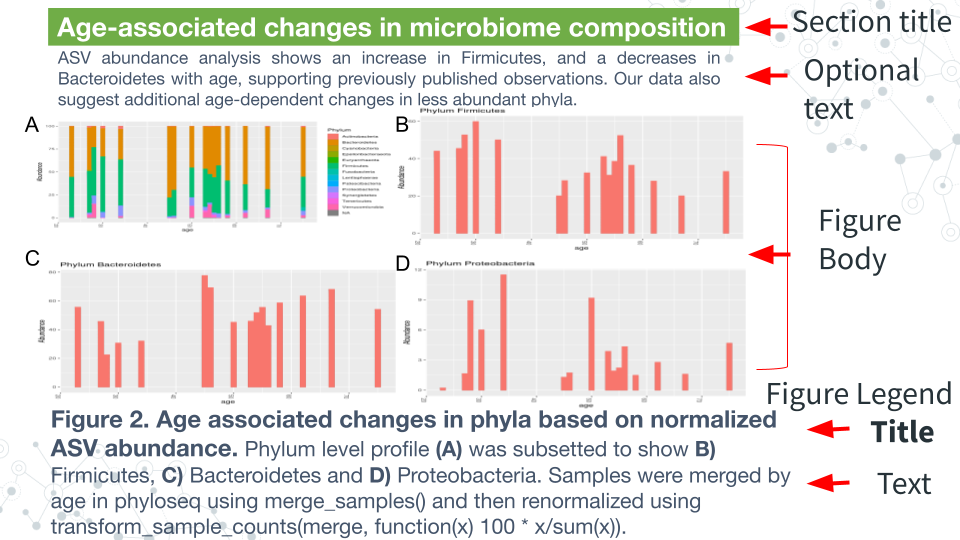
Figure Legend:

* All figures and tables should have their legends *underneath*
* The legend should be in a smaller font than the main text on the poster.
* A legend should includes a **figure title**, which is a declarative statement summarizes findings. This is the takeaway from the figure/table.
* A legend also includes **figure text** which includes detail on the context and/or methodology used to generate the figure or table to give context to the audience on what they are looking at.

1. Identifies (and can briefly describe parts)
2. Includes brief description of of methods necessary to understand figure and
3. Unless this can be combined with 1) or 2) should include relevant statistics (stats).

Each Figure and Table can have additional (optional) text:

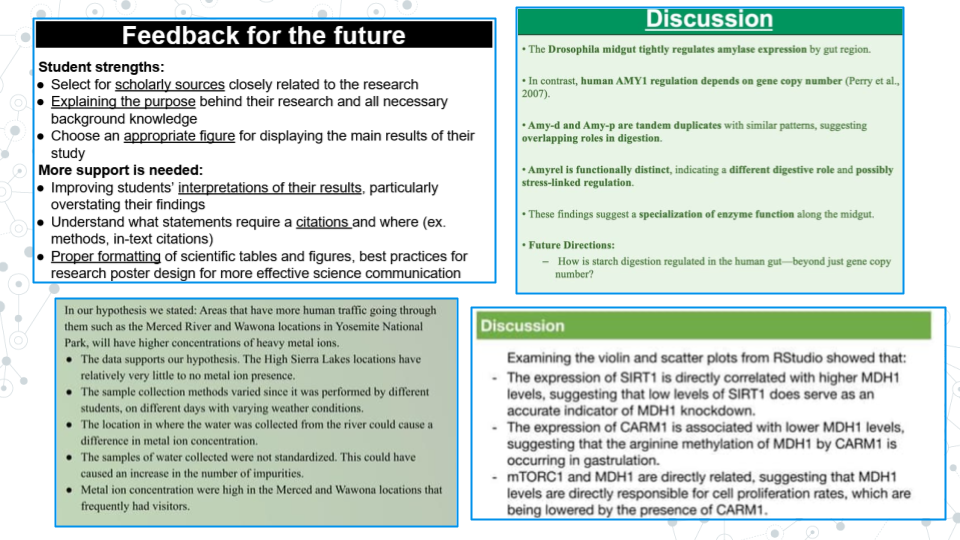
* Section Title - Summarizes section or figure, similar or broader than figure legend title. Attention Grabber.
* (Optional) text - describes the findings of the figure upfront to engage.



**Check your poster’s results section:**

* ☐  You have a minimum of 2 figures on the poster
* ☐  All figures have a labeled x and y-axis
* ☐  All figures have a legend underneath them
* ☐  All tables have a legend above them
* ☐  The legends have 1 sentence summarizing the figure or table’s main idea
* ☐  The legends have accompanying text detailing the methodology and or other context of the figure/table.

#### 10.3.3.5 Conclusions/Discussion



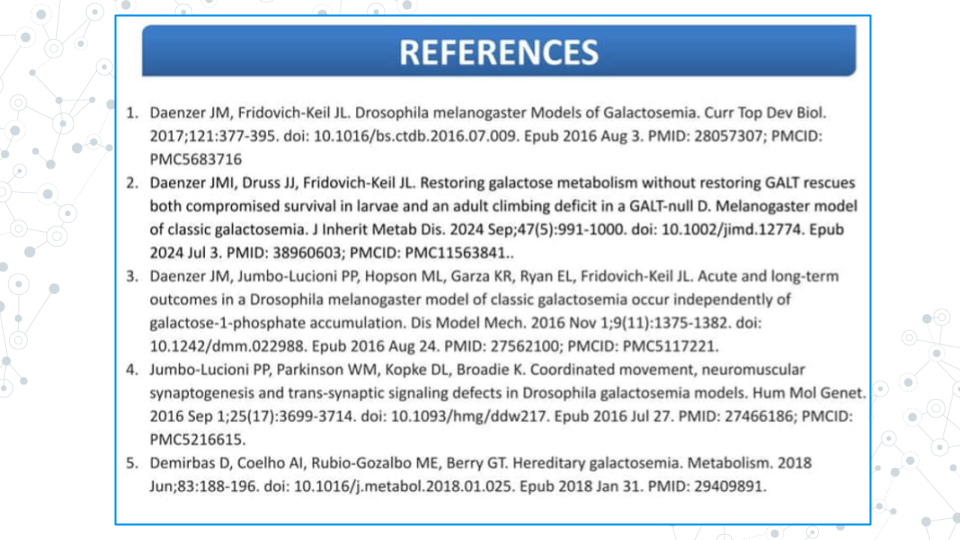
The discussion section of the paper is your chance to analyze and interpret your results. For your lab report, make sure your discussion section includes all of the following:

* What do your results mean?
* How do they fit into the bigger picture?
* If any experiments did not give expected results, hypothesize why that might have been the case and propose alternate experiments that could confirm or clarify your results.
* Include at least one sentence of future work that you would do if you had more time or what students in upcoming semesters could do to continue to answer your questions.

**Check your poster’s discussion section:**

* ☐  1 sentence for each major result that relates to your hypothesis
* ☐  1-2 sentences how these results matter based on the problem you described in your introduction
* ☐  1 sentence describing future work someone else could do as a logical next step in addressing the problem you described in your introduction
* ☐  If your results were unexpected, hypothesize why that might be the case

#### 10.3.3.6 References



All the references that you cite on your poster must be present in a References section including the following sections: Introduction, Methods, and Discussion. To save space on our posters, we will number our references (ex. 1-5) and use the numbers as citations throughout the text of your poster. You may have a lot of references; it is okay to put them in tiny text if you have to in order to make them fit.

There are many different ways to format the reference section. We will make ours in alphabetical order by the first author’s last name. All of your sources must be scientific journals and should use the following format:

Authors (year) “Title.” Journal Name, vol. #, page #s, DOI

Online article that is also in print:

Haussecker D., Huang Y., Lau A., Parameswaran P., Fire A. Z. and M. A. Kay (2010) “Human tRNA-derived small RNAs in the global regulation of RNA silencing.” RNA, Vol. 16, page 637-695, <doi:10.1261/rna.2000810>

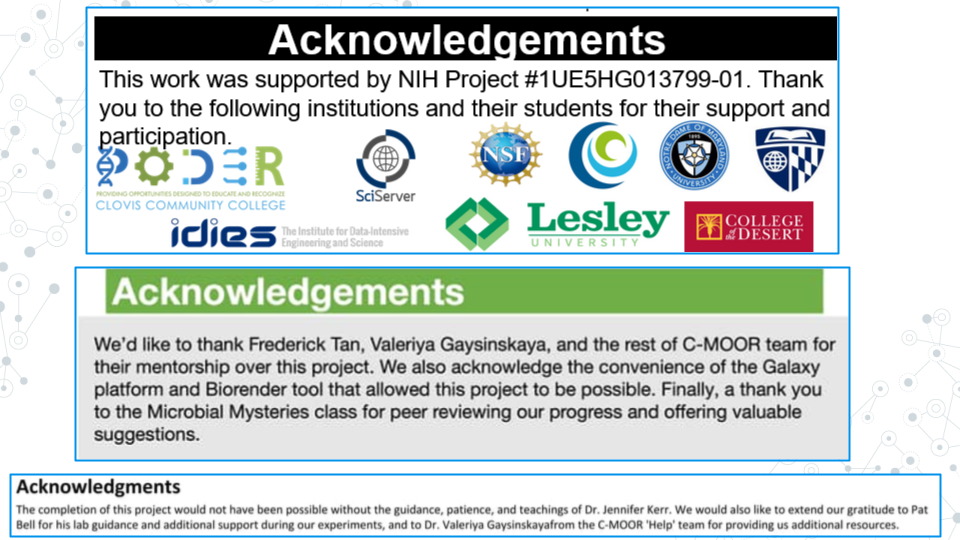
Online article only:

Marianes, A. and A. C. Spradling (2013) “Physiological and stem cell compartmentalization within the Drosophila midgut.” eLife, <doi:10.7554/eLife.00886>

**Check your references section**:

* ☐  All in-text citations in the introduction are included in the references
* ☐  All in-text citations in the methods are included in the references
* ☐  All in-text citations in the discussion are included in the references
* ☐  All references are from scientific journals
* ☐  All text that requires an in-text citation has one

#### 10.3.3.7 Acknowledgements



The acknowledgements section is where you give thanks to the people, organizations, and institutions that have supported you in your research. If relevant, include the grant ID # of your funding source. Institutions and organizations that have contributed to your research - but you do not belong to - can be thanked here.

**Check your acknowledgements section:**

* ☐  Thank any funding source and the grant # are listed here
* ☐  Thank your research mentor (if applicable)
* ☐  Thank those who have provided you with technical support (if applicable)

### 10.3.4 Part 3 - Proofread and Add Final Touches

*Estimated time: 30 min to an hour*

1. Each group member should re-read the poster from beginning to end and fix any typos or grammatical errors.
2. Check the alignment of figures, text boxes, titles, etc.
3. Add some finishing touches. You can play with the color, the font, add additional images if it’s relevant.

### 10.3.5 Part 4 - Canvas Discussion

*Estimated time: 30 min*

You will turn in your poster to be graded as a group in a Canvas Assignment and post it to a Canvas Discussion to be viewed by the class.

1. Convert your poster to a pdf.
2. Have one member of your group turn in the pdf of your poster to the Graded Canvas Assignment. This assignment is already set up so that if one group member turns it in, it will show as submitted for all students in the group. This is where your instructor will grade you poster as a group.
3. Have one member of your group post a pdf of your poster in the Canvas C-MOOR Poster Discussion.
   1. With your poster, introduce your group members and copy and paste your abstract into the post.
   2. Insert your pdf into the post and edit the link so that it automatically shows the inline preview. This will make it easier for students to view your poster.
4. As an individual, read through the other posters from different groups.
5. Post comments

### 10.3.6 Grading Criteria

Everyone in the group earns the same grade, so it is important to work together.

| Points | Category |
| --- | --- |
| 2 points | Title, Authors and Affiliations |
| 3 points | Abstract |
| 4 points | Introduction 2 - relevant background 2 - drosophila as a model system |
| 5 points | Methods |
| 5 points | Results |
| 4 points | Discussion |
| 5 points | References ( 2 primary sources, 5 sources total) |
| 2 points | Poster Organization |

### 10.3.7 Footnotes

#### 10.3.7.1 Resources

* [Google Doc]

#### 10.3.7.2 Contributions and Affiliations

* Stephanie R. Coffman, Clovis Community College
* Valeriya Gaysinskaya, Johns Hopkins University
* Frederick Tan, Johns Hopkins University
* Sayumi York, Notre Dame of Maryland University

Last Revised: June 2025

## 10.4 Activity - Share Your Poster

### 10.4.1 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.

#### 10.4.1.1 Activity 1 – Share Your Poster

*Estimated time: 20 min*

#### 10.4.1.2 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** |
|  |

#### 10.4.1.3 Questions

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

### 10.4.2 Activity 2 – Present Your Poster

*Estimated time: 40 min*

#### 10.4.2.1 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

#### 10.4.2.2 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) |

### 10.4.3 Grading Criteria

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

### 10.4.4 Footnotes

**Resources**

* [Google Doc]

**Contributions and Affiliations**

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

Last Revised: April 2025

# 11 Next Steps

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>

## 11.3 Grading Criteria

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

## 11.4 Footnotes

**Resources**

* Google Doc

**Contributions and Affiliations**

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

## 11.5 Update 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.

#### 11.5.0.1 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.

## 11.6 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.

## 11.7 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.



### 11.7.1 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!

### 11.7.2 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.

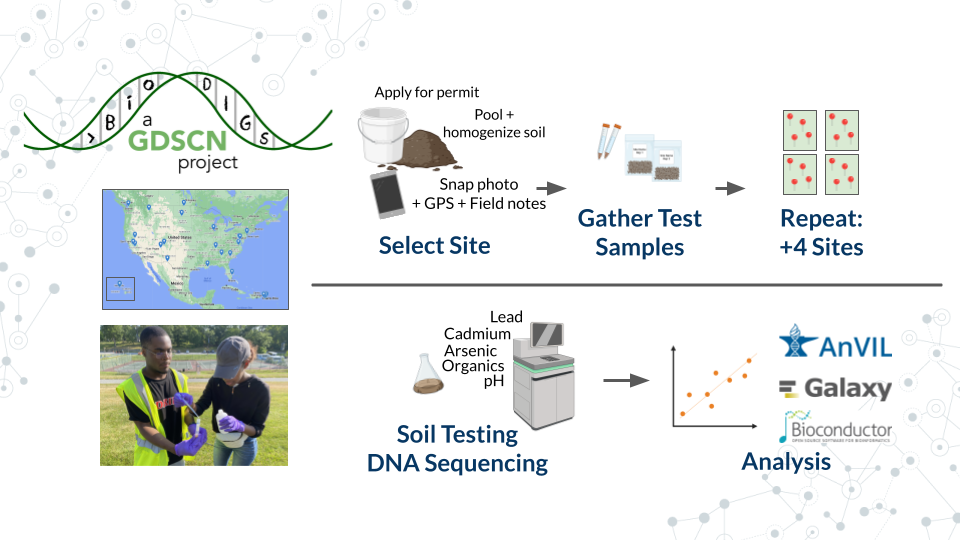
I previously completed a project as a part of an RNA-seq miniCURE and have some exposure to working with DESeq2, though I still have much to learn. My group and I profiled the gene expression of the genes XYZ in the Drosophila midugt and I presented our poster at our university’s research day poster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

## 11.8 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.

# About the Authors

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

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

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

# 12 C-MOOR Scholars

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>

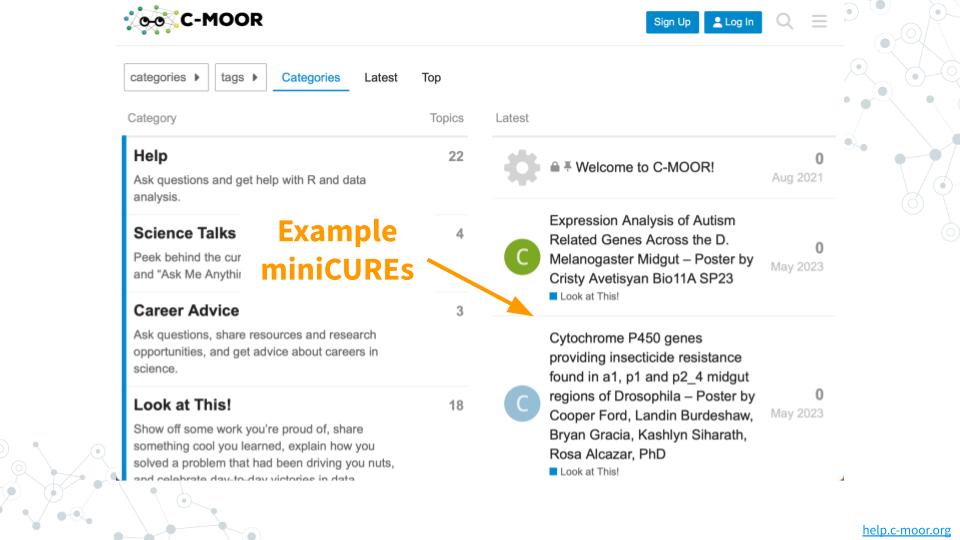
Meet the C-MOOR Scholars and learn how you can support them

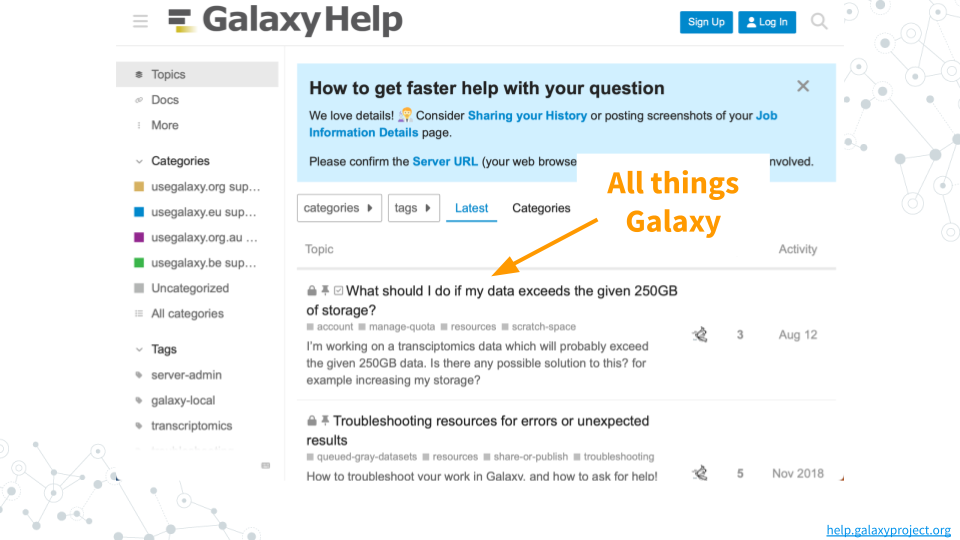
* <https://www.cloviscollege.edu/alumni-and-community/c-moor/c-moor-scholars.html>

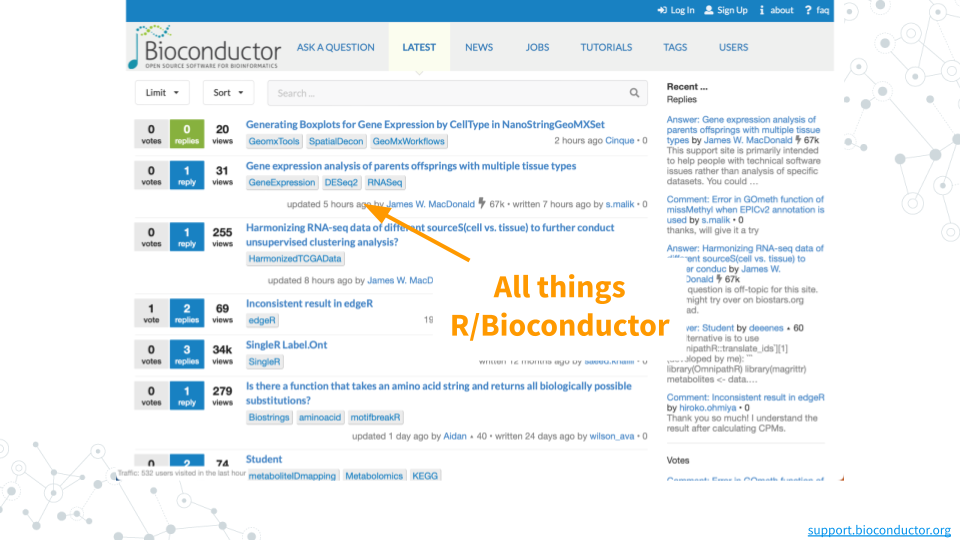
# 13 Online Community

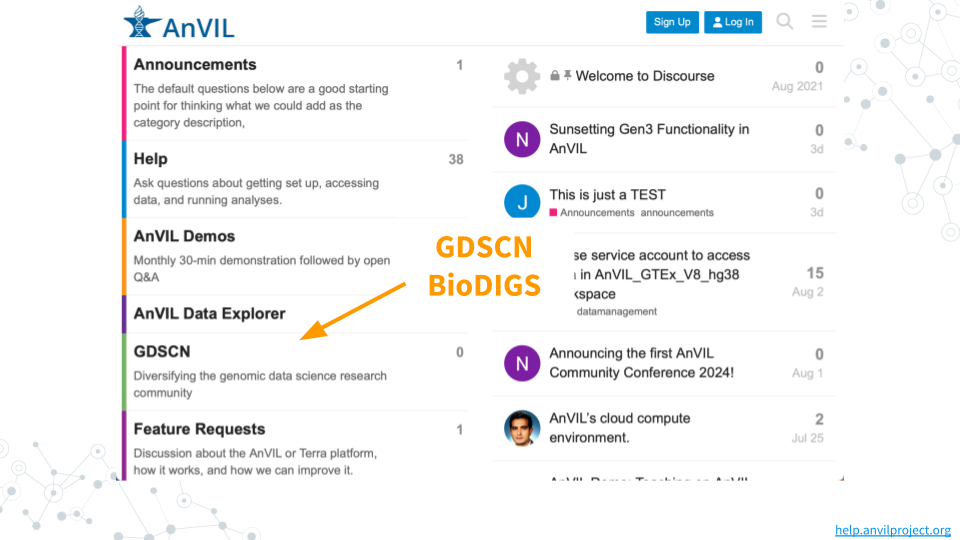
The genomic data science community is a remarkably open community that continually strives to increase opportunitites and broaden access to the scientific enterprise (e.g. [GDSCN 2022](https://pubmed.gov/35858750)). There are many opportunities to interact online with scientists that share a common research 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>









# References