Course Name

June 02, 2025

Table of Contents

# About

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



Figure 0.1: C-MOOR logo

### Audience and Prerequisites

What is the target audience?

**Prerequisites**:

* Prereq 1
* Prereq 2

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

### Format

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

### Learning Goals

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

### Core Competencies

This activity addresses the following core concepts and competencies:

**Vision and Change**

**Genetics**

**Bioinformatics**

Core concepts and competencies are taken from the following sources:

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

### C-MOOR Content Collection

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

# 1 Introduction

## 1.1 Lecture: 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 Lecture: 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 Activity: Taxonomy Profiling Spreadsheet

### 2.2.1 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).

### 2.2.2 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.

### 2.2.3 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.

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

*Estimated time: 25 min*

#### 2.2.4.1 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.

#### 2.2.4.2 Questions

1. Evaluate what proportion of data was taxonomically classified.
2. Insert a new column A; we will use this temporary column for calculations, so you can name this column “Calculations”.
3. In e.g. cell A2, calculate the sum of all reads observed in the gut std sample.

| **How many total counts are there?** |
| --- |
|  |

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

| **What percentage of reads are unclassified?** |
| --- |
|  |

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

| **What percentage of reads are classified?** |
| --- |
|  |

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

| **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.

**What abundant taxa do you observe?**

| **% abundance** | Taxonomy |
| --- | --- |
| 20.1 | s\_Faecalibacterium\_prausnitzii |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

1. 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*!

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

1. Calculate ‘Low abundance’ for < 1% abundant taxa by adding together taxa at <1%.

| **What percentage of reads are classified in a low abundance taxa?** |
| --- |
|  |

1. Create a barplot of % abundance for your 12 abundant taxa via Insert Chart.

| **Paste your barplot of % abundance for the 12 most abundant taxa** |
| --- |
|  |

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

*Estimated time: 20 min*

#### 2.2.5.1 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.

#### 2.2.5.2 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?** |
| --- |
|  |

### 2.2.6 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

### 2.2.7 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

## 2.3 Activity: A short introduction to Galaxy

### 2.3.1 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

### 2.3.2 Activity

*Estimated time: 40 min*

#### 2.3.2.1 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>

### 2.3.3 Grading Criteria

* Submit URL to your shared Galaxy history on Canvas

### 2.3.4 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

## 2.4 Homework: Scientific Literature Prelab

### 2.4.1 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)

### 2.4.2 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.

### 2.4.3 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.

### 2.4.4 Activity

*Estimated time: 50 min*

#### 2.4.4.1 Instructions

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

#### 2.4.4.2 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.

#### 2.4.4.3 Grading Criteria

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

### 2.4.5 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

## 2.5 Discussion: Scientific Literature Prelab

### 2.5.1 Activity

*Estimated time: 25 min*

#### 2.5.1.1 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)

## 2.6 Lecture: Scientific Literature



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

## 2.7 Homework: Scientific Literature Activity

### 2.7.1 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/)

### 2.7.2 Learning Objectives

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

### 2.7.3 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.

### 2.7.4 Activity

*Estimated time: 90 min*

#### 2.7.4.1 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.

#### 2.7.4.2 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?

#### 2.7.4.3 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.

#### 2.7.4.4 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 |

### 2.7.5 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 |

#### 2.7.5.1 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?

#### 2.7.5.2 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?

### 2.7.6 Grading Criteria

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

### 2.7.7 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

## 2.8 Presentation: Scientific Literature Activity

### 2.8.1 Activity

*Estimated time: 25 min*

#### 2.8.1.1 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)

### 2.8.2 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 Homework: Microbial Genomes Prelab

### 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 Homework: QC and Galaxy Workflows

### 3.3.1 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.

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

*Estimated time: 90 min*

#### 3.3.2.1 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

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

*Estimated time: 30 min*

#### 3.3.3.1 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)

### 3.3.4 Grading Criteria

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

### 3.3.5 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

## 3.4 Discussion: Microbial Genomes Prelab

### 3.4.1 Activity

*Estimated time: 25 min*

#### 3.4.1.1 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)

### 3.4.2 Footnotes

**Contributions and Affiliations**

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

Last Revised: February 2025

## 3.5 Homework: Microbial Genomes Project

### 3.5.1 Purpose

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

### 3.5.2 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).

### 3.5.3 Activity 1 – MAGs and Taxonomy

*Estimated time: [40] min*

#### 3.5.3.1 Activity 1 - Part I: Explore a MAG

#### 3.5.3.2 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 |

#### 3.5.3.3 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: |
|  |

#### 3.5.3.4 Activity 1 - Part II: Explore a Contig

#### 3.5.3.5 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: |  |
|  |  |

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

#### 3.5.3.7 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?** |
| --- |
|  |

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

*Estimated time: [20] min*

#### 3.5.4.1 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.

#### 3.5.4.2 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: |  |

### 3.5.5 Activity 3 – Go Deeper!

*Estimated time: [30] min*

#### 3.5.5.1 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? |
| --- |
|  |
|  |

### 3.5.6 Grading Criteria

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

### 3.5.7 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

## 3.6 Activity: test-driveR

### 3.6.1 Introduction

SciServer ([sciserver.org](https://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!

### 3.6.2 Activity

*Estimated time: 25 min*

#### 3.6.2.1 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”



1. Scroll down and click “Create”. This may take a moment.
2. 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.



1. Complete the tutorial and answer the questions below.

#### 3.6.2.2 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. Explanation.** |
| --- |
|  |

### 3.6.3 Grading Criteria

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

### 3.6.4 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

## 3.7 Presentation: Microbial Genomes Project

### 3.7.1 Activity

*Estimated time: 25 min*

#### 3.7.1.1 Instructions

1. **Open the “Microbial Genomes Presentation” slidedeck** [here](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.
3. Describe the MAG you explored (Activity 1) and some of its genes (Activity 2)
4. What did you discover about your organism and gene (Activity 3)?
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)

### 3.7.2 Footnotes

**Contributions and Affiliations**

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

Last Revised: February 2025

# 4 Taxonomy Profiling

## 4.1 Lecture: Taxonomy Profiling



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

## 4.2 Homework: Taxonomy Profiling Prelab

### 4.2.1 Purpose

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

### 4.2.2 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

### 4.2.3 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.

### 4.2.4 Activity 1 – QC Reads

*Estimated time: 50 min*

#### 4.2.4.1 Activity 1 - Part I: Import dataset into Galaxy

*Estimated time: 15 min*

#### 4.2.4.2 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.

#### 4.2.4.3 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: |

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

*Estimated time: 15 min*

#### 4.2.4.5 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.

#### 4.2.4.6 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 |

#### 4.2.4.7 Activity 1 - Part III: View NanoPlot Results in Galaxy

*Estimated time: 15 min*

#### 4.2.4.8 Instructions

1. View and examine NanoPlot Results in Galaxy.

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

#### 4.2.4.9 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? |
| --- |
|  |

### 4.2.5 Activity 2 – Taxonomy Profiling in Galaxy

*Estimated time: 50 min*

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

*Estimated time: 15 min*

#### 4.2.5.2 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.

#### 4.2.5.3 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. |
| --- |
|  |

#### 4.2.5.4 Activity 2 - Part II: View Kraken2 results

*Estimated time: 15 min*

#### 4.2.5.5 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.

#### 4.2.5.6 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..)? |
| --- |
|  |

### 4.2.6 Grading Criteria

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

### 4.2.7 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

## 4.3 Lecture: Possible Datasets



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

## 4.4 Homework: Possible Datasets Activity

### 4.4.1 Activity

*Estimated time: 50 min*

#### 4.4.1.1 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> |

### 4.4.2 Grading Criteria

* Submit URL to your reply on Canvas

### 4.4.3 Footnotes

**Contributions and Affiliations**

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

Last Revised: May 2022

## 4.5 Homework: Galaxy data collections

### 4.5.1 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.

### 4.5.2 Activity

*Estimated time: 45 min*

#### 4.5.2.1 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)

### 4.5.3 Grading Criteria

* Submit URL to your shared Galaxy history on Canvas

### 4.5.4 Footnotes

**Resources**

* Using Galaxy and Managing your Data [topic]

**Contributions and Affiliations**

* Frederick Tan, Johns Hopkins University

Last Revised: January 2025

## 4.6 Discussion: Taxonomy Profiling Prelab

### 4.6.1 Activity

*Estimated time: 25 min*

#### 4.6.1.1 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)

### 4.6.2 Footnotes

**Contributions and Affiliations**

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

Last Revised: February 2025

## 4.7 Homework: Taxonomy Profiling Project

### 4.7.1 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.

### 4.7.2 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.

### 4.7.3 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 |

### 4.7.4 Activity 1 – QC

*Estimated time: 50 min*

#### 4.7.4.1 Activity 1 - Part I: Import data and run NanoPlot

#### 4.7.4.2 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.

#### 4.7.4.3 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? |
| --- |
|  |

#### 4.7.4.4 Activity 1 - Part II: Quality filtering with fastp

#### 4.7.4.5 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.

#### 4.7.4.6 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 |
| --- |
|  |

### 4.7.5 Activity 2 – Taxonomy Profiling

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

#### 4.7.5.1 Activity 2 - Part I: Run Taxonomy Profiling Workflow

#### 4.7.5.2 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.

#### 4.7.5.3 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? |
| --- |
|  |

#### 4.7.5.4 Activity 2 - Part II: Analyze Kraken2 results

#### 4.7.5.5 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.

#### 4.7.5.6 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? |
| --- |
|  |

#### 4.7.5.7 Activity 2 - Part III: Krona Pie Chart

#### 4.7.5.8 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

#### 4.7.5.9 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) |

### 4.7.6 Grading Criteria

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

### 4.7.7 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

## 4.8 Activity: Meet R!

### 4.8.1 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.

### 4.8.2 Activity

*Estimated time: 25 min*

#### 4.8.2.1 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.

#### 4.8.2.2 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: |

### 4.8.3 Grading Criteria

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

### 4.8.4 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

## 4.9 Presentation: Taxonomy Profiling Project

### 4.9.1 Activity

*Estimated time: 25 min*

### 4.9.2 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)

### 4.9.3 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 Homework: Finding AMRs Prelab

### 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 Lecture: Scientific Posters



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

## 5.4 Homework: Scientific Posters Activity

### 5.4.1 Activity

*Estimated time: 50 min*

#### 5.4.1.1 Instructions

1. Review “Presentations Guidelines for Posters”.

* <https://ur.umbc.edu/urcad/resources/posters-guidelines>

1. Skim three posters from among the following (must be Biology if from UMBC Posters)
2. C-MOOR Posters – <https://help.c-moor.org/c/look-at-this/8>
3. UMBC Posters – <https://ur.umbc.edu/poster-presentation-examples>
4. PacBio Poster – <https://www.pacb.com/wp-content/uploads/PAG-Portik.pdf>
5. Pick one poster and address the following points.
6. **Notice** – What about this poster most interests you?
7. **Wonder** – Two or three questions you would ask the authors.
8. **Support** – Two or three suggestions on how the poster could be improved.
9. Post your answers by replying to the “Project Work: Scientific Posters” topic in the Discussion Forum

### 5.4.2 Grading Criteria

* Submit URL to your reply on Canvas

### 5.4.3 Footnotes

**Contributions and Affiliations**

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

Last Revised: February 2025

## 5.5 Homework: r4ds

### 5.5.1 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.

### 5.5.2 Activity

*Estimated time: 50 min*

#### 5.5.2.1 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.

#### 5.5.2.2 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? |
| --- |
|  |

### 5.5.3 Grading Criteria

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

### 5.5.4 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

## 5.6 Discussion: Finding AMRs Prelab

## 5.7 Homework: Finding AMRs Project

## 5.8 Activity: r4ds2

## 5.9 Presentation: Finding AMRs Project

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

# References