A Data Carpentry- Style Metagenomics Workshop

Claudia Zirión-Martínez[1], **Diego Garfias-Gallegos**[1], **Tania Vanessa Arellano-Fernandez**[3,6], **Aarón Espinosa-Jaime**[6], **Edder D Bustos-Díaz**[2], **José Abel Lovaco-Flores**[4,6], **Luis Gerardo Tejero-Gómez**[5], **J Abraham Avelar-Rivas**[3,4], and **Nelly Sélem-Mojica** [5]

1 Laboratorio de Genómica Ecológica y Evolutiva, Langebio, Cinvestav, México. **2** Laboratorio de Evolución de la Diversidad Metabólica, Langebio, Cinvestav, México. **3** Laboratorio de Sistemas Genéticos, Langebio, Cinvestav, México. **4** BetterLab - C3. Irapuato, México. **5** Centro de Ciencias Matemáticas. UNAM, México. **6** Escuela Nacional de Estudios Superiores, Unidad León, UNAM, México.

Summary

Metagenomic analyses aim to explore the genomic diversity of communities in specific habitats by processing their DNA sequencing data. This analysis is achieved with specialized bioinformatics tools, which often require previous coding experience. Furthermore, beginners can struggle to build a pipeline from raw data to valuable biological insights. The Carpentries hosts open lessons used worldwide to analyze specialized datasets for beginners, including a Data Carpentry curriculum for individuals working with genomics sequencing data. However, a lesson addressing the specific challenges associated with metagenomics data and analyses was missing. We created a complete Metagenomics curriculum in The Carpentries Incubator, adapting and expanding on the Data Carpentry genomics curriculum. The curriculum provides an introduction to programming, teaching learners to access and handle metagenomics data, and to run commands with the software needed for completing metagenomics analyses. Content and exercises have been improved based on experience gathered in teaching the curriculum in three 16-hour online workshops. We expect to continue to enhance this lesson, which we hope is helpful as a teaching resource for new instructors in the field, and as a guide for newcomers wishing to perform metagenomic analyses from scratch.

Statement of Need

Bioinformatic tools are now essential to our understanding of biological systems. Open lessons for general-purpose coding languages and specialized topics such as genomics, ecology, and even metagenomics are already available (Darling et al., 2019), (*Lessons,* n.d.), (Kruchten, 2020). Nevertheless, a complete guide for shotgun metagenomics, assuming no prior knowledge of coding and provided hardware and software solutions, was missing. We introduce a curriculum to fill this gap, which teaches the skills required to conduct a comprehensive metagenomics workflow with Bash and R programming in a pre-installed remote server.

Content and Learning Objectives

The curriculum is comprised of four lessons (Fig 1). It assumes no previous programming experience but expects that students understand basic concepts of molecular biology and microbiology. The first two lessons, introductions to project planning and organization and programming in Bash respectively, are adapted from equivalent lessons in the Data

Carpentry Genomics curriculum (Becker et al., 2019). The third part includes a brief introduction to R, and the fourth teaches a complete shotgun metagenomics workflow using public data (Okie et al., 2020).
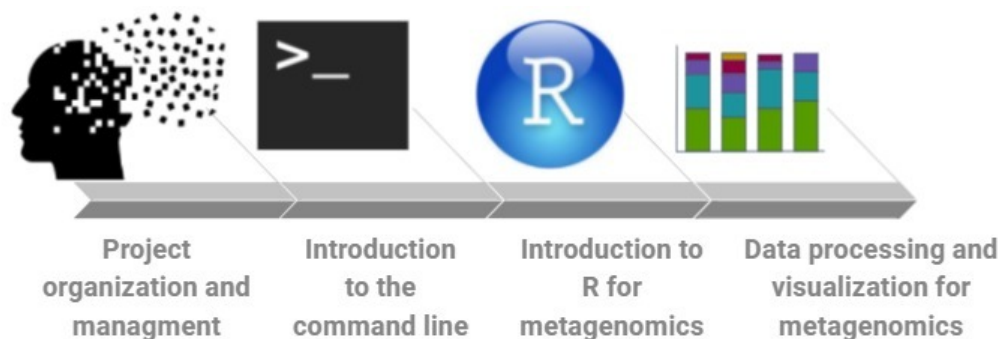

Figure 1: Fig 1.

Fig 1. The four lessons in the curriculum for a metagenomics workshop.

Learning objectives

Project organization and management for metagenomics - Plan, structure, organize, and document metagenomics data, metadata, and bioinformatics workflow. - Access public data on the NCBI sequence read archive (SRA) database.

Introduction to the command-line for metagenomics - Identify the benefits of the command line. - Navigate the file system, create, copy, move, and remove files and directories. - Work with files. - Combine commands and automate repetitive tasks. - Make an organized file structure for a bioinformatics project.

Introduction to R for metagenomics - Familiarize with RStudio and R functionality. - Distinguish the different data types. - Create and manipulate data frames. - Use and find help in R libraries.

Data processing and visualization for metagenomics - Explain the structure and contents of the data and metadata used in these lessons. - Assess the quality of sequencing data. - Trim and filter sequences based on their quality. - Perform a metagenomic assembly. - Obtain Metagenome-Assembled Genomes and check their quality. - Assign and visualize a taxonomy of reads and contigs. - Explore the diversity in a sample and calculate diversity estimates. - Discover more resources for metagenomics projects.

Instructional Design

The workshop includes 16 hours of content with live coding, formative assessment practice, and other supportive elements. During the development of the lessons, we considered three axes of teaching: cloud setup, standardized episodes, and teaching strategies.

1. Cloud setup: We set up remote machines to standardize the learning environment and teaching experience. This setup lowers entry barriers for students without experience in technical installations and provides enough computing power regardless of individual configuration. Students only need a computer with internet access and a terminal program installed. We also enlist the instructions for setting up the remote machines and an alternative installation guide to be used by people who prefer to follow the lessons on a different computer.

2. Standardized episodes: Following The Carpentries instructional design, lessons consist of multiple short *episodes* of content, each prefaced with the questions to be answered and the learning objectives for that section, and ending with concrete, clear messages of the learned content.
3. Teaching strategies
   - Live coding: Programming simultaneously with students gives a practical coding experience with examples of mistakes arising and being solved.
   - Exercises and discussion in small groups: Allows learners to solve problems with peers and incentivizes participation while applying what they learn to strengthen new knowledge and skills. Regular breaks for exercises provide learners and instructors with feedback about progress toward the learning objectives.
   - Content review: At the end and beginning of each session, we ask learners to revisit, list, and explain the content taught in past lessons. We use a collaborative document where simultaneous written and spoken review helps to reaffirm the content.

Teaching Experience

Our teaching team includes undergraduate and graduate students, postdocs, and professors, each with different perspectives that enrich the discussion of the lesson. We found it adequate to have two instructors for the online workshops, plus one helper for every five students. We are pleased to have recruited new helpers from the attendees of previous workshops, enlarging and further enriching our teaching community. To promote an interactive learning experience, we use a collaborative live document and small groups to do exercises and discussions. In this way, there are fewer barriers to student's participation, and they gain practical experience of the different solutions bioinformatic problems can have. We include bonus exercises around the lessons that are not meant to be solved by all the learners but are an opportunity for advanced participants to tackle more challenging tasks. They help to leverage the background and learning speed of the learners and provide opportunities to practice and reinforce the content learned. Another strength of our workshop structure is introducing programming languages followed by a practical example of their usage, so the metagenomic analysis grows while the coding skills learned become meaningful. The lessons are helpful even if the student's primary goal is to learn coding rather than metagenomics. We ask the attendees to fill out surveys before, during, and after the workshops to adapt our teaching strategies to the current and future learning groups. With the results of these surveys, we have been able to improve the curriculum's content, explanations, and exercises.

Story of the Project

The idea to create a lesson comprising all of the steps required to process and analyze metagenomics data arose when one of the development team wanted to learn how to understand metagenomics data but found it overwhelming even to decide where to start. As someone who already used and knew the advantages of the resources offered by The Carpentries, she recognized the absence of a curriculum about metagenomics on this platform. When she learned of The Carpentries Incubator, she reached out to a fellow student who had experience on the subject and started building the episodes. Other students in the same institute, who were also working on metagenomics, started contributing to the pipelines they were learning. Further team members were recruited as helpers after attending a workshop and later began collaborating to develop the episodes.

Becker, E. A., Teal, T., Michonneau, F., Sane, M., Reiter, T., Williams, J., Charbonneau, A., Kweskin, M., Anwar, M. Z., Cadzow, M., & Martinez, P. A. (2019). *Datacarpentry/genomics-workshop: Data carpentry: Genomics workshop overview, june 2019.* https://doi.org/10.5281/ZENODO.3260309

Darling, A. E., DeMaere, M. Z., Gaio, D., & Anantanawat, K. (2019). *Data for the gene school: Metagenomics workshop.* Zenodo. https://doi.org/10.5281/ZENODO.3585993

Kruchten, A. E. (2020). A curricular bioinformatics approach to teaching undergraduates to analyze metagenomic datasets using r. *Frontiers in Microbiology*, *11*, 2135. https://doi.org/10.3389/fmicb.2020.578600

Lessons. (n.d.). Retrieved September 12, 2021, from https://datacarpentry.org/lessons/

Okie, J. G., Poret-Peterson, A. T., Lee, Z. M., Richter, A., Alcaraz, L. D., Eguiarte, L. E., Siefert, J. L., Souza, V., Dupont, C. L., & Elser, J. J. (2020). Genomic adaptations in information processing underpin trophic strategy in a whole-ecosystem nutrient enrichment experiment. *eLife*, *9*, e49816. https://doi.org/10.7554/eLife.49816