Practical Computing Skills for Omics Data

PLNTPTH 6193 – Spring 2024 S2 – Independent Studies

Syllabus

Course Information

- **Course times and location:**
  - Synchronous Zoom meetings on Tue and Thu from 2:20 pm - 3:40 pm.
  - An optional 1-hour weekly recitation meeting on Mon – the time will be set during the first week of the course in consultation with the students.
- **Credit hours:** 2
- **Mode of delivery:** Distance Learning (DL; 100% online)

Instructor

- **Name:** Jelmer Poelstra (please just call me Jelmer)
- **Email:** poelstra.1@osu.edu (preferred mode of contact)
- **Phone Number:** 919-260-8253 (for emergencies)
- **Office location:** Selby Hall 018, Wooster Campus

- **Office hours:** Wed and Fri 10 - 11 am via Zoom or in my office. Reserve a spot via email. If these times don't work for you, we should be able to find another time ad hoc.

Course Description

As datasets have rapidly grown larger in biology, coding has been recognized as an increasingly important skill for biologists. This is especially true in "omics" research with data from e.g. genomics and transcriptomics, which usually cannot be analyzed on a desktop computer, where most software has a command-line interface, and where workflows can include many steps that need to be coordinated.

In this course, students will gain hands-on experience with a set of general and versatile tools for data-intensive research. The course will focus on foundational skills such as working in the Unix shell and writing shell scripts, installing software, and submitting jobs at a compute cluster (the Ohio Supercomputer Center), and building flexible, automated workflows. Additionally, the course will cover reproducibly organizing, documenting, and version-controlling research projects. Taken together, this course will allow students to reproduce their own research, and enable others to reproduce their research, with as little as a single command.

Learning Goals & Outcomes

This course is designed to provide students with foundational training in computing skills for reproducible research. At the end of this course, students will be able to start applying these skills and the associated tools in their own research and will have a firm understanding of how this can make their research more robust, reproducible, and efficient.

Upon successful completion, students should be able to:

- L1: Apply the Unix shell for a variety of omics data management and analysis tasks.
- L2: Work at a remote supercomputer (Ohio Supercomputer Center, OSC).
- L3: Writing and submitting shell scripts as batch jobs using Slurm.
- L4: Install and manage software using Conda and containers.
- L5: Use Git and GitHub for version control and research collaboration.
- L6: Use Markdown to document research and code.
- L7: Implement good project organization and data management in research projects.
- L8: Use a workflow system (Nextflow) to automate analysis pipelines.
- L9: Understand how the above-mentioned tools and techniques can make their research more reproducible and shareable.

How This Course Works

This course is divided into **weekly modules** that are released on Fridays. Zoom sessions will consist of a mixture of lecture, participatory live-coding ("code-along"), and exercises. During participatory live-coding, I will type, explain, and execute code, and you are expected to code along with me.

- **Synchronous Zoom sessions: required, twice per week** (will be recorded)
- **Homework**: **required**
    - Readings: There are readings for every week, and you are recommended to complete these before the first Zoom meeting of the week.
    - Exercises: Most weeks will include a set of exercises to be completed on your own time. These don't have to be submitted, but I may ask about them in class. You are recommended to do these after the second class of the week.
    - Surveys: There will be one or two brief student surveys.
- **Recitation: optional**
  During recitation (recap) sessions, we will discuss the exercises of the previous week. These will not be recorded to encourage student engagement.
- **Office hours**: **optional**
  To reserve a 30-minute office hour slot, send me an email.

Readings

The textbooks listed below are freely available through the OSU library, and PDFs will also be posted on Carmen. (But if you have the funds, consider buying a paper copy of one or both.)

- Allesina S, Wilmes M (2019). ***Computing Skills for Biologists***. Princeton UP.
  → Available online through the OSU library at https://library.ohio-state.edu/record=b8624007~S7.
- Buffalo V (2015). ***Bioinformatics Data Skills: Reproducible and Robust Research with Open Source Tools.*** O'Reilly Media, Inc.
  → Available online through the OSU library at https://library.ohio-state.edu/record=b9490023~S7.

Equipment and Software

- **Required:**
  - A computer, which can run any operating system (Windows/Mac/Linux).
  - Browser – A recently updated version of Chrome, Firefox, or Microsoft Edge.
- **Optional:**
  - External monitor – a secondary and preferably large monitor will be helpful but is not required (see below).

During the course, you will use the online resources of the Ohio Supercomputer Center (OSC, https://www.osc.edu/) through your browser. Prior to the start of the course, you will be asked to create an OSC account and will be granted access to the OSC Classroom Project associated with this course.

Because of our use of OSC, you can follow this course without other software and will not need a particularly powerful computer or a large amount of hard disk space. That said, if your computer is very slow, please contact me to make sure you will not run into problems.

Course websites

- All materials for this course will be made available to you using a **GitHub website (https://jelmerp.github.io/pracs-sp24/)** instead of through CarmenCanvas.
- You will not use CarmenCanvas to submit (final project) assignments either but will instead use your own GitHub repositories to do so, as you will learn during the course.
- CarmenCanvas will only be used to send course announcements and updates, and to share some files.

Additional Information

- During the synchronous Zoom sessions, we will often be doing hands-on coding in class ("participatory live coding"). To simultaneously see what I am doing and code along yourself, it will help to have a very large **monitor** or multiple monitors. If you don't have multiple monitors for a single device, you can connect to Zoom with two devices.
- **Zoom etiquette**:
    - Occasionally, you may be asked to *share your screen* and show your code, so try to have a setup where this is possible.
    - As much as possible, to the extent you are comfortable with it, please have your *camera on* during the Zoom sessions.
    - Please have your *microphone* on mute by default, but feel free to unmute yourself whenever you would like to say something.

Descriptions of Major Course Assignments

Final project

In your final project, you will be asked to combine the skills you have learned during this course into a small, well-documented, and reproducible data analysis project that uses code for all its steps.

You are encouraged to use a **dataset** from your own research or to find a publicly available dataset that is similar to what you will later encounter in your own research. Alternatively, I can provide a dataset and/or topic for you.

You will gradually build up your project during the latter part of the course:

1. **Project proposal** as a Git repository – due week 4
2. **Markdown draft** with a project outline and some documented code – due week 6
3. 10-minute **oral presentation**, describing the background of the project, the code and methods used, and the results – presented in class during week 7
4. **Final submission** – due Fri, Apr 29th

The project proposal, draft and final submission should be submitted through your GitHub repository with your final project (you will learn how to do this during the course!). The proposal and draft are due on Mondays at 11:59 PM, and the final submission is due on Apr 29th at 11:59 PM. You will give your oral presentation live during the final Zoom meeting of the course.

Your final project should be your own original work and collaborating with others is not permitted. Open-book and internet research is permitted and encouraged. You are prohibited in university courses from reusing previous work, that is, from turning in work from a past class

to your current class, even if you modify it. If you want to build on work you've explored in previous courses, please discuss the situation with the instructor.

Late submissions are only accepted without penalty in cases of emergency (illness or injury, family emergency) and students should communicate as soon as possible with the instructor when situations arise. Late submissions without communication with the instructor will result in a Fail grade for that submission. If a conflict arises for the day of the final project presentation, please communicate with the instructor directly.

Grading Scale

You will receive a final Satisfactory (S, credits awarded) or Unsatisfactory (U, no credits awarded) grade for the course. The following components will count towards your grade for a total of 50 points, where 35 or more points are needed to receive a Satisfactory Grade:

- Final project: proposal – 10 points
- Final project: draft – 10 points
- Final project: presentation – 10 points
- Final project: submission – 20 points

Course Schedule

"CSB" = *Computing Skills for Biologists*, "Buffalo" = *Bioinformatics Data Skills*.

MODULE	SESSION DATES	TOPIC	
		Required readings [+ optional]	*Assignments due*
1	Feb 27 & 29	Course intro & Shell I: Basics	
		CSB Ch. 0-0.1 & Ch. 1	
2	Mar 5 & 7	Project file organization & Markdown	
		Buffalo Ch. 2 [+ Ch. 3]	
SPRING BREAK			
3	Mar 19 & 21	Version control with Git and GitHub	
		CSB Ch. 2 [+ Buffalo Ch. 5]	
4	Mar 26 & 28	Shell II: Data tools, CLI software, & scripting	
		Buffalo Ch. 2 + 7 [+ Ch. 3]	*Project: proposal (3/25)*
5	Apr 2 & 4	Data & software management, OSC Slurm batch jobs	
		Buffalo Ch. 6 [+ Ch. 4]	
6	Apr 9 & 11	Reproducible workflows with Nextflow	
		(Will be assigned later)	*Project: draft (4/8)*
7	Apr 16 & 18	Recap & final project presentations	
		CSB Ch. 11, Buffalo Ch. 1	*Project: submission (4/29)*