TEACHING GENOMICS AT A PUI WITH BROWSER—ONLY ACTIVITIES

Link to this slide deck:

Plant and Animal Genomes

Resources and Programs for Undergraduate Education in Genomics

Bárbara D. Bitarello| Bryn Mawr College

January 2024

About me

- 2021-Present: Assistant professor at Bryn Mawr College (BMC), a small women's liberal arts college.
- Research: evolutionary &statistical genomics (humans and other primates).
- Bitarello Lab: currently 6 undergraduate researchers working on diverse projects in evolutionary & statistical genetics & phylogenetics
- Teaching:

- 100-level: Intro Bio
- 200-level: **Genomics (6h/week, 1/2 lab)**, Biostatistics with R,
- 300-level: Evolutionary Genetics & Genomics

Outline

1. Why browser-only?
2. Two projects/experiences from B216 (Genomics) that only require a browser

A. A soft-introduction to the command line and FASTQ files
B. The Genomics Education Partnership (GEP) and how I've adapted and contributed materials

Bonus: A quick mention about a third project involving R programming!

Why browser only?

Challenges for teaching:

A. getting all tools installed in a variety of OS and versions is often **frustrating** and **time-consuming**
B. campus computers: often **lack permissions** to get all the **required updates** and **installations** in a timely manner
C. some students use machines that **lack space or capability** for local installations (e.g. Chromebook)
D. the technical challenges intimidate students even more; the browser **keeps it familiar/simple**

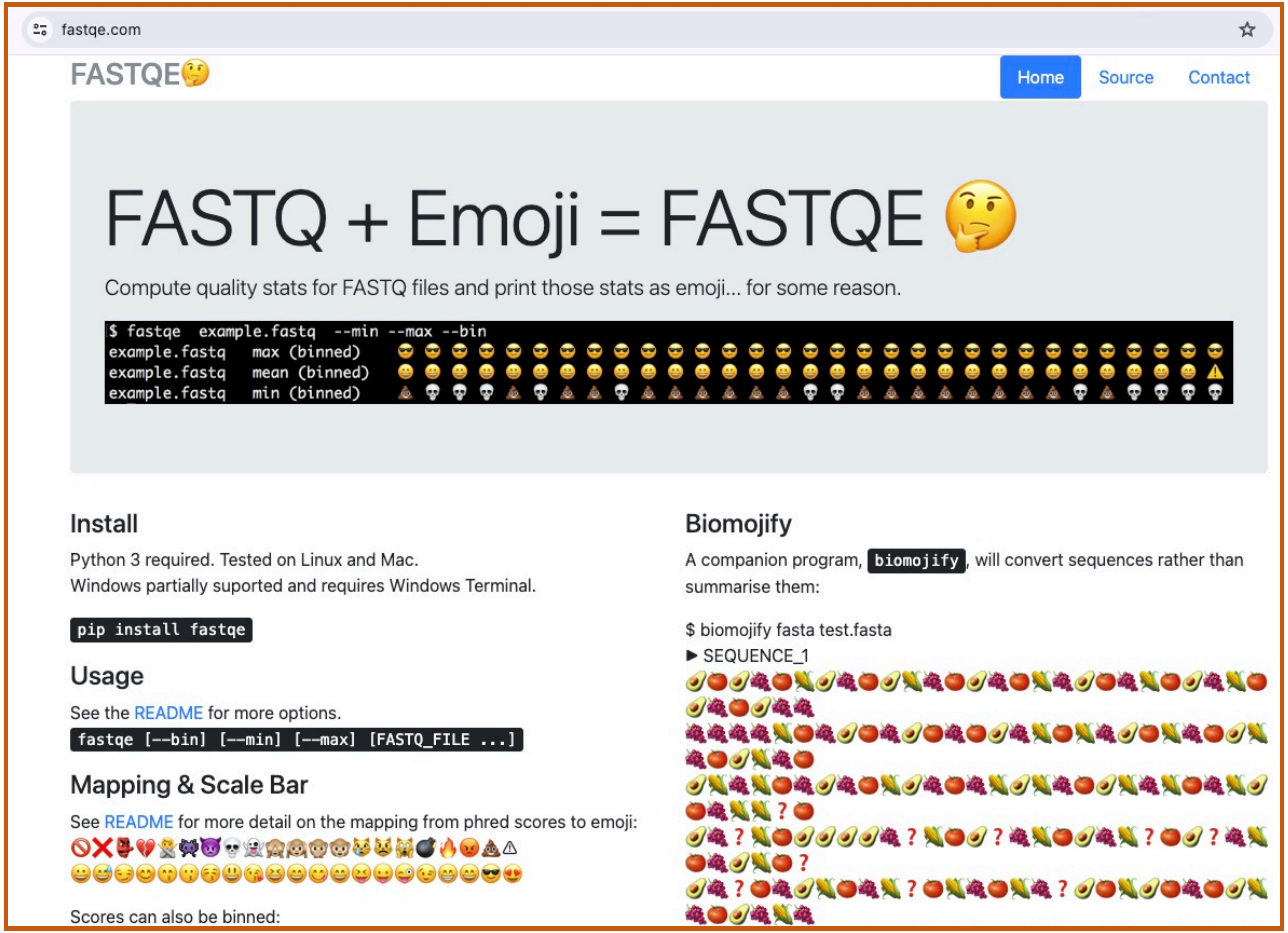
Browser-only activities bypass all of these hurdles!

Project 1: A soft introduction to the command-line and FASTQ files

Motivation

- Excellent materials from St. Jaquest et al. (2021), available in CourseSource
- Introduces the command-line and FASTQ files by using the FASTQE software
- I reached out to senior author Ray Enke about my adapted materials and here we are!

The FASTQE tool



FASTQ + Emoji = FASTQE 🤔

Compute quality stats for FASTQ files and print those stats as emoji... for some reason.

```
$ fastqe  example.fastq  --min --max --bin
example.fastq   max (binned)   😎 😎 😎 😎 😎 😎 😎 😎 😎 😎 😎 😎 😎 😎 😎 😎 😎 😎 😎 😎 😎 😎 😎 😎 😎 😎 😎 😎 😎 😎 😎
example.fastq   mean (binned)  😀 😀 😀 😀 😀 😀 😀 😀 😀 😀 😀 😀 😀 😀 😀 😀 😀 😀 😀 😀 😀 😀 😀 😀 😀 😀 😀 😀 😀 😀 ⚠️
example.fastq   min (binned)   💩 💀 💀 💀 💩 💀 💩 💩 💀 💩 💩 💩 💩 💩 💩 💀 💀 💩 💩 💩 💩 💩 💩 💩 💩 💀 💩 💀 💀 💀 💀
```

Install

Python 3 required. Tested on Linux and Mac.
Windows partially suported and requires Windows Terminal.

```
pip install fastqe
```

Usage

See the README for more options.

```
fastqe [--bin] [--min] [--max] [FASTQ_FILE ...]
```

Mapping & Scale Bar

See README for more detail on the mapping from phred scores to emoji:
🚫❌👺💔🙅👾👿💀👻🙈🙉🙊🐵😿😾🙀💣🔥😡💩⚠️
😀😅😏😊😙😗😚😃😘😆😄😋😄😝😛😜😉😁😄😎😍

Scores can also be binned:

Biomojify

A companion program, `biomojify`, will convert sequences rather than summarise them:

```
$ biomojify fasta test.fasta
▶ SEQUENCE_1
🥑🍅🥑🍇🍅🌽🥑🍇🍅🥑🌽🍇🍅🥑🍇🍅🌽🍇🥑🍅🍇🌽🍅🥑🍇🌽🍅
🥑🍇🍅🥑🍇🍇
🍇🍇🍇🍇🌽🍅🍇🥑🍅🍇🥑🍅🍇🍅🥑🍇🌽🍅🌽🍇🥑🍅🌽🍇🍅🥑🌽
🍇🍅🥑🌽🍇🍅
🥑🌽🍇🌽🍅🍇🥑🍇🍅🌽🥑🍇🍅🍇🌽🥑🌽🍇🍅🥑🌽🍇🌽🍅🍇🌽🥑
🍅🍇🌽🌽❓🍅
🥑🍇❓🌽🍅🥑🥑🥑🥑🍇❓🌽🍅🥑❓🍇🌽🍅🥑🍇🌽❓🍅🥑❓🍇🌽
🍅🍇🥑🌽🍅❓
🥑🍇❓🍅🍇🥑🌽🍅🍇🌽❓🍅🌽🍇🍅🌽🍇❓🥑🍅🌽🥑🍅🍇🍅🥑🌽
🍇🍅🥑🍇🌽🍇
```

Challenges in implementing the lesson

- FASTQE is a python package that needs to be installed, as well as its dependencies — and python installations are ~~always~~ often a nightmare!
- Proposed implementation using either a) local installation or b) Cyverse
- 2022: lost one entire class installing locally for each students and one student still could not get it to work; Cyverse did not work for any of them

Updated implementation

All is freely available:

- Shortened link I made for this presentation: http://tinyurl.com/33wkwjwt
- The Github repo with code and a slidedeck for students: https://github.com/bitarellolab/Genomics_Teaching

Learning goals

1) Have a soft introduction to the command line

- ls
- mkdir
- cd
- pwd
- more
- less
- wc
- pip
- conda

2) Have a soft introduction to FASTA and FASTQ files

3) Build an intuition around next-generation sequencing quality scores based on emojis

Structure

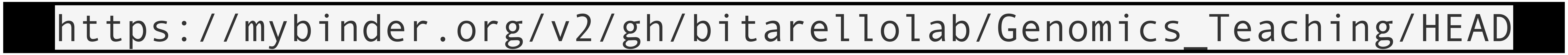
- Students follow the two tutorials through the mybinder.org link on their browser
  - Bash Basics tutorial
  - FASTQE tutorial

- Students hand in answer sheet at the end of class
- Later I wrote questions in a problem set where they had to go back to this and analyze different sequence files

Solution: using mybinder.org

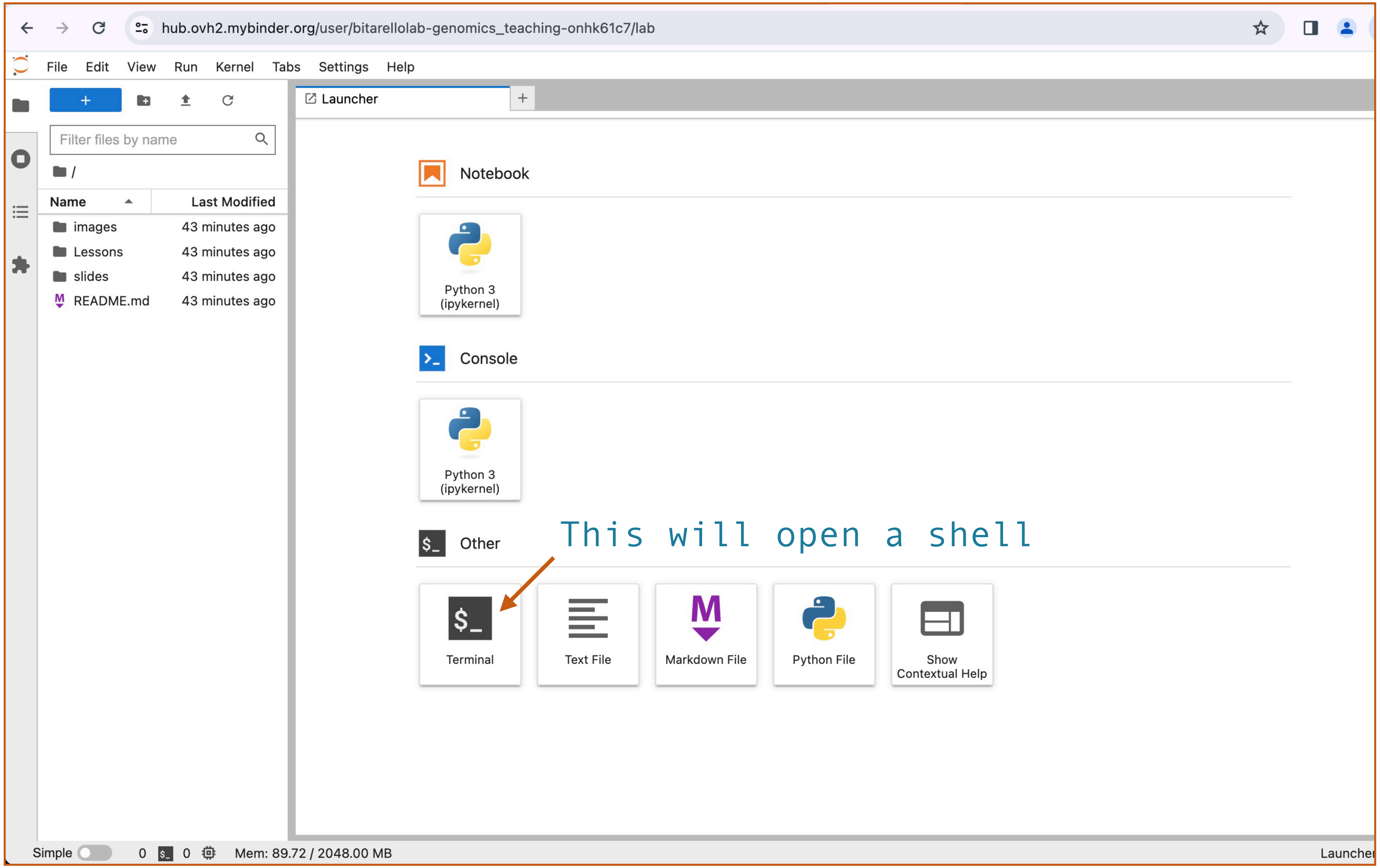
- Binder allows you to create **custom computing environments** that can be shared and used by many remote users.
- A Binder service is powered by BinderHub, an open-source tool.
- One such deployment lives at mybinder.org, and is free to use.

- To access the activity, go to this link:



```
https://mybinder.org/v2/gh/bitarellolab/Genomics_Teaching/HEAD
```

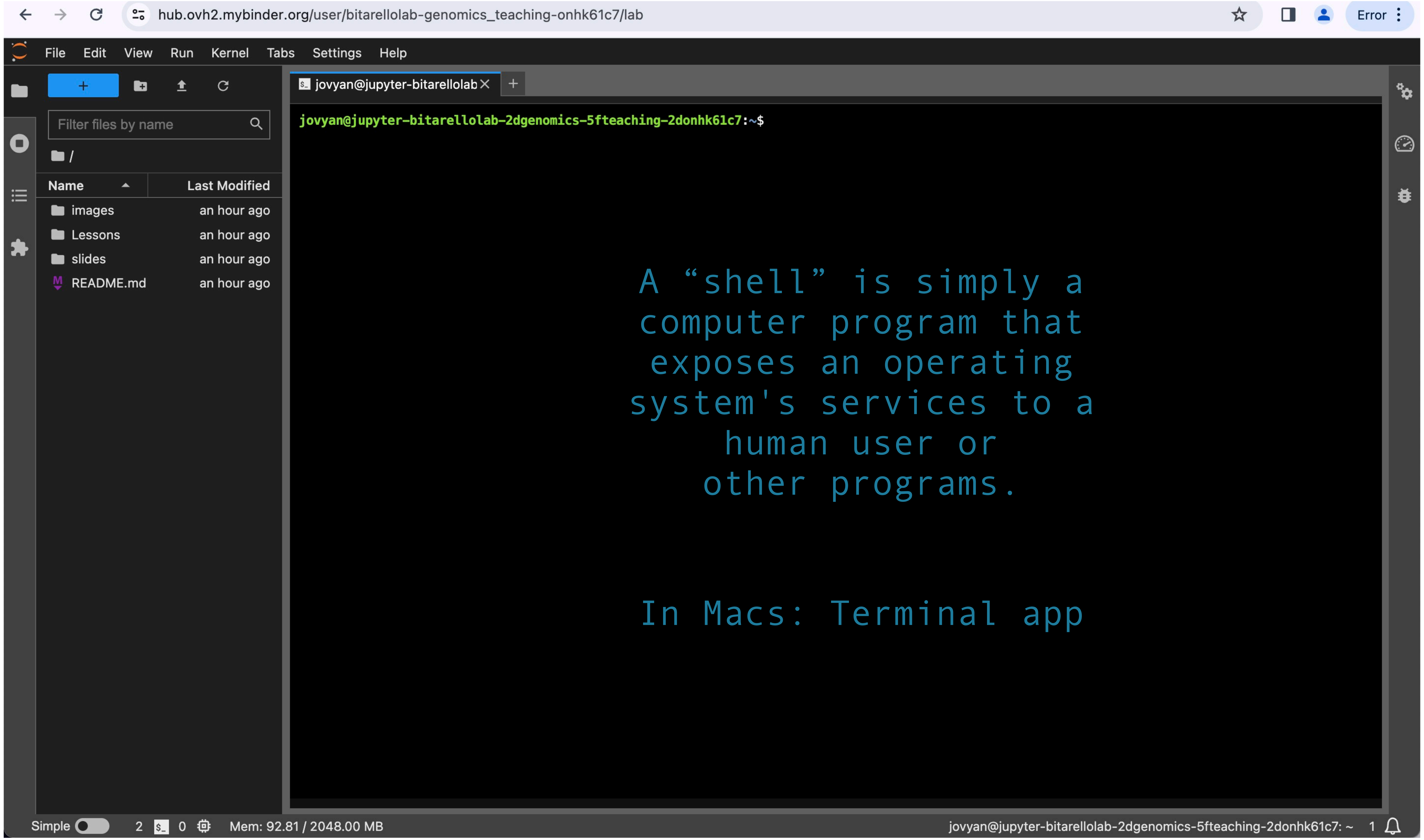
Landing page



Name	Last Modified
images	43 minutes ago
Lessons	43 minutes ago
slides	43 minutes ago
README.md	43 minutes ago

This will open a shell

A shell is now open



hub.ovh2.mybinder.org/user/bitarellolab-genomics_teaching-onhk61c7/lab

Name	Last Modified
images	an hour ago
Lessons	an hour ago
slides	an hour ago
README.md	an hour ago

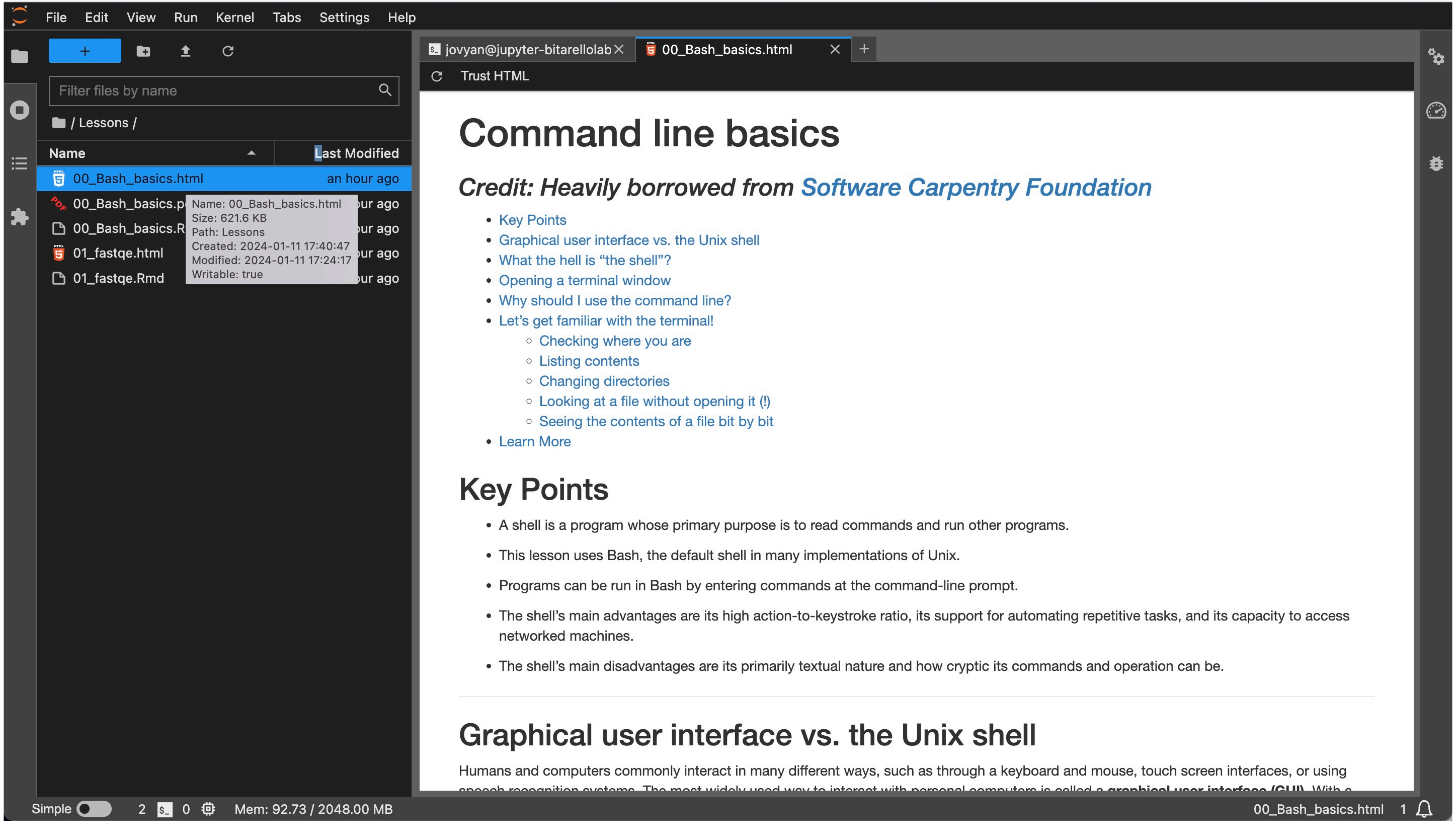
```
jovyan@jupyter-bitarellolab-2dgenomics-5fteaching-2donhk61c7:~$
```

A “shell” is simply a computer program that exposes an operating system's services to a human user or other programs.

In Macs: Terminal app

Simple
2
0
Mem: 92.81 / 2048.00 MB
jovyan@jupyter-bitarellolab-2dgenomics-5fteaching-2donhk61c7: ~
1

Soft introduction to the command line



Name	Last Modified
00_Bash_basics.html	an hour ago
00_Bash_basics.p	ur ago
00_Bash_basics.R	ur ago
01_fastqe.html	ur ago
01_fastqe.Rmd	ur ago

Name: 00_Bash_basics.html
Size: 621.6 KB
Path: Lessons
Created: 2024-01-11 17:40:47
Modified: 2024-01-11 17:24:17
Writable: true

Command line basics

Credit: Heavily borrowed from Software Carpentry Foundation

- Key Points
- Graphical user interface vs. the Unix shell
- What the hell is “the shell”?
- Opening a terminal window
- Why should I use the command line?
- Let’s get familiar with the terminal!
    - Checking where you are
    - Listing contents
    - Changing directories
    - Looking at a file without opening it (!)
    - Seeing the contents of a file bit by bit
- Learn More

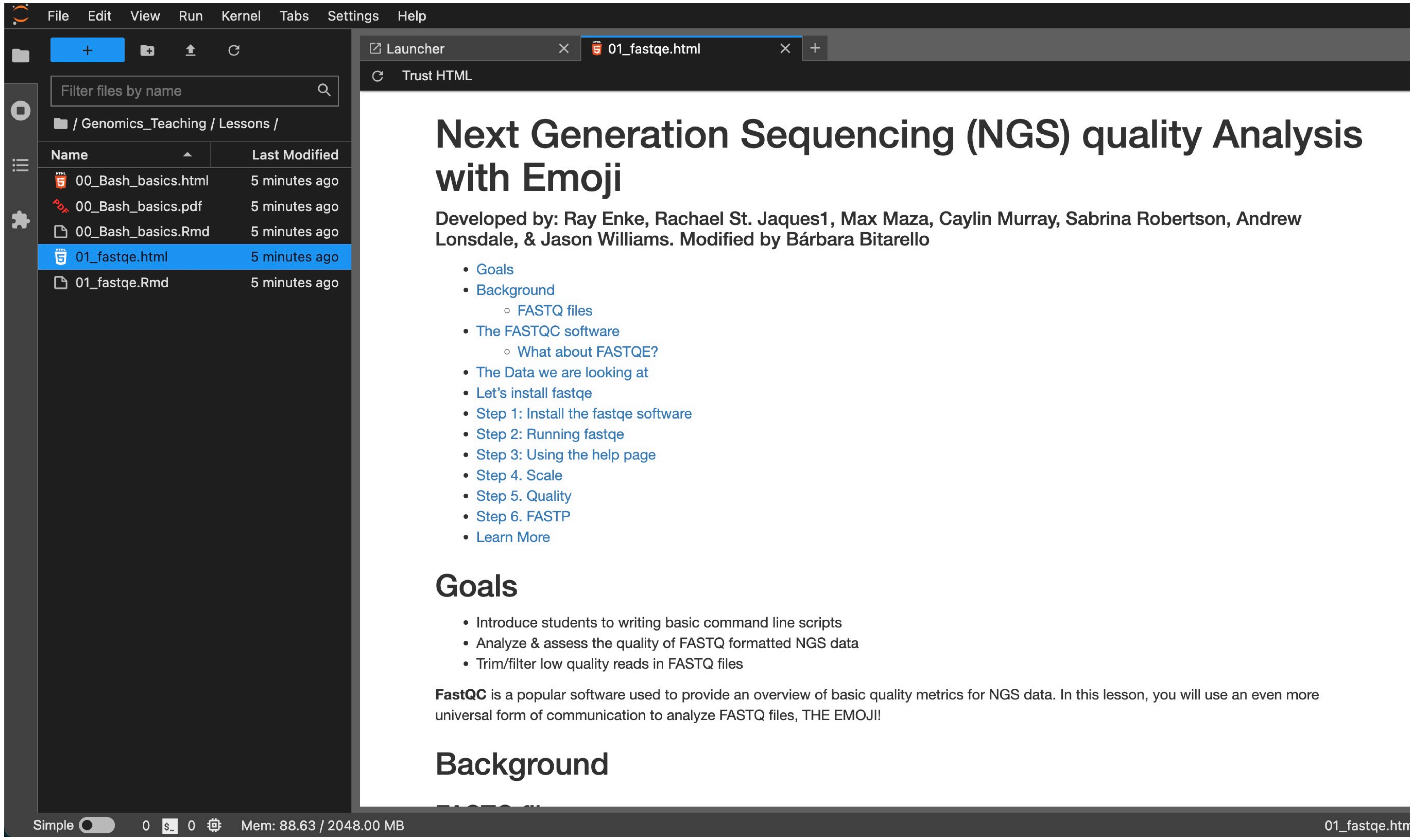
Key Points

- A shell is a program whose primary purpose is to read commands and run other programs.
- This lesson uses Bash, the default shell in many implementations of Unix.
- Programs can be run in Bash by entering commands at the command-line prompt.
- The shell’s main advantages are its high action-to-keystroke ratio, its support for automating repetitive tasks, and its capacity to access networked machines.
- The shell’s main disadvantages are its primarily textual nature and how cryptic its commands and operation can be.

Graphical user interface vs. the Unix shell

Humans and computers commonly interact in many different ways, such as through a keyboard and mouse, touch screen interfaces, or using speech recognition systems. The most widely used way to interact with personal computers is called a **graphical user interface (GUI)**. With a

Learning about FASTQ files with FASTQE



Name	Last Modified
00_Bash_basics.html	5 minutes ago
00_Bash_basics.pdf	5 minutes ago
00_Bash_basics.Rmd	5 minutes ago
01_fastqe.html	5 minutes ago
01_fastqe.Rmd	5 minutes ago

Next Generation Sequencing (NGS) quality Analysis with Emoji

Developed by: Ray Enke, Rachael St. Jaques1, Max Maza, Caylin Murray, Sabrina Robertson, Andrew Lonsdale, & Jason Williams. Modified by Bárbara Bitarello

Goals

- Introduce students to writing basic command line scripts
- Analyze & assess the quality of FASTQ formatted NGS data
- Trim/filter low quality reads in FASTQ files

FastQC is a popular software used to provide an overview of basic quality metrics for NGS data. In this lesson, you will use an even more universal form of communication to analyze FASTQ files, THE EMOJI!

Background

Where we're at

- Possibility: publish on QUBES/CourseSource to increase visibility
- Currently expanding/modifying the intro to command-line portion
- This works 99% of the time but mybinder.org is free and sometimes it gets busy…
- Currently working on some tricks to make the loading faster - it does but with this very unwieldy link. I've shortened it here so folks can access it:

http://tinyurl.com/33wkwjwt

Tl;dr

- This adapts St. Jacques et al. (2021) materials so that everything can be installed and run from a browser
- This implementation preserves the learning process of installing the packages while providing a uniform environment for all students
- Additionally: I was interested in expanding the intro to the command-line per se, as this course is a natural recruiting environment for new research students

References/links

The original publication describing the FASTQE activity

- St. Jacques RM, Maza WM, Robertson SD, Lonsdale A, Murray CS, Williams JJ, Enke RA. 2021. A fun introductory command line lesson: Next generation sequencing quality analysis with Emoji! CourseSource. https://doi.org/10.24918/cs.2021.17

The FASTQE tool:

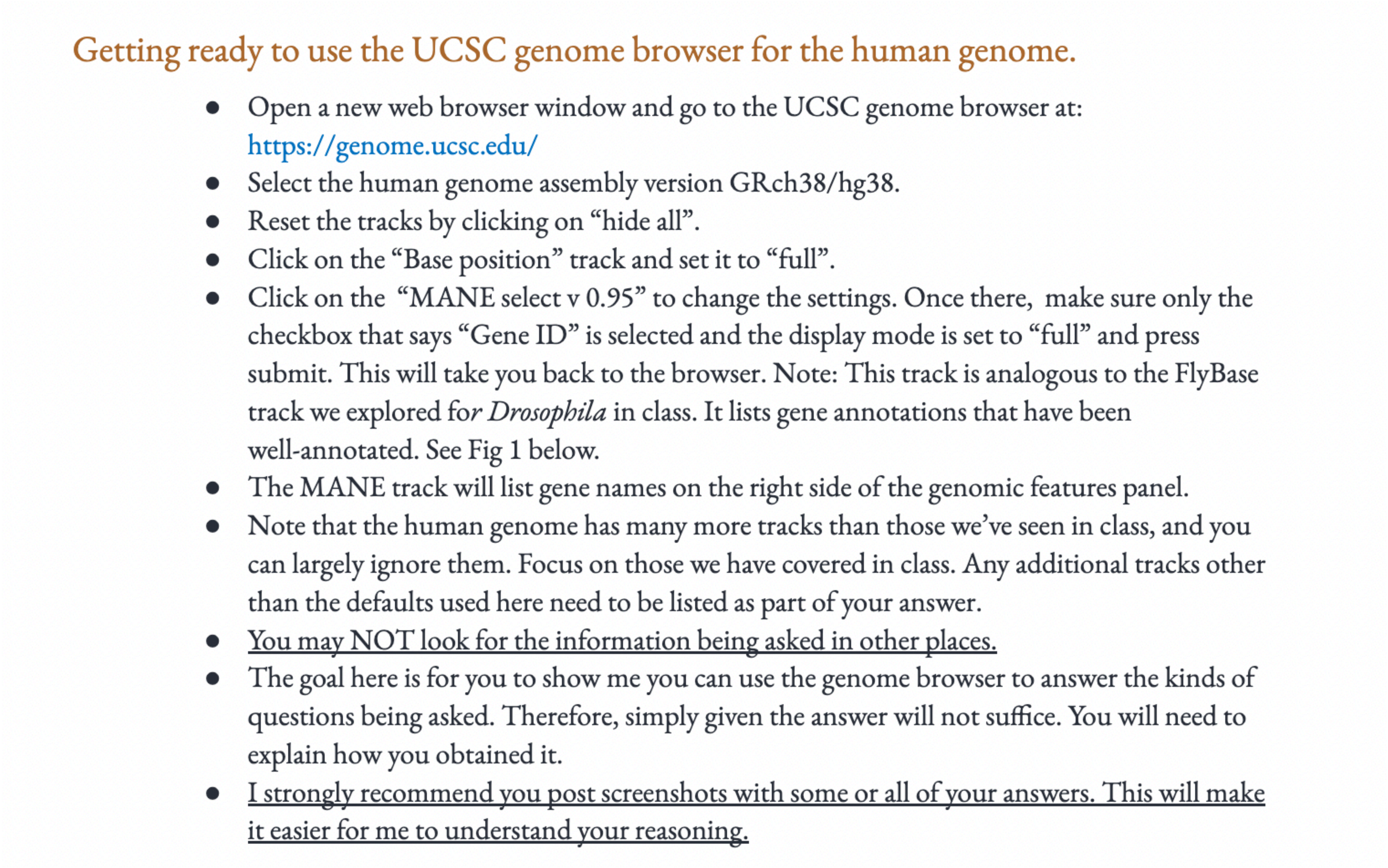
- Official Page: https://fastqe.com/
- Github: https://github.com/fastqe/fastqe

The command-line portion

- I took heavy inspiration from Intro to the command line: The Software Carpentry. https://swcarpentry.github.io/shell-novice/01-intro/index.html (Accessed March 22, 2023)

Project 2: The Genomics Education partnership (GEP)

Exam - take home, open book

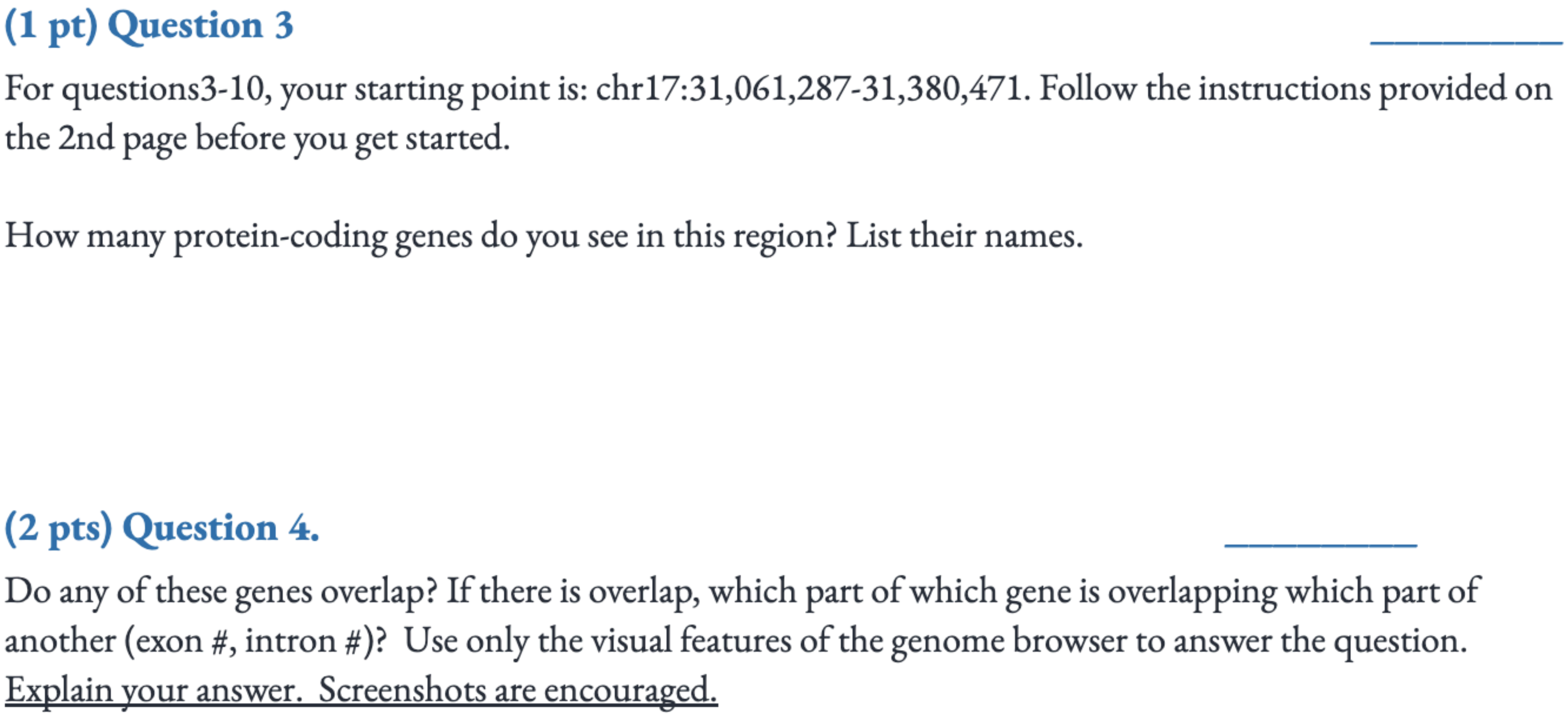


Getting ready to use the UCSC genome browser for the human genome.

- Open a new web browser window and go to the UCSC genome browser at: https://genome.ucsc.edu/
- Select the human genome assembly version GRch38/hg38.
- Reset the tracks by clicking on "hide all".
- Click on the "Base position" track and set it to "full".
- Click on the "MANE select v 0.95" to change the settings. Once there, make sure only the checkbox that says "Gene ID" is selected and the display mode is set to "full" and press submit. This will take you back to the browser. Note: This track is analogous to the FlyBase track we explored for *Drosophila* in class. It lists gene annotations that have been well-annotated. See Fig 1 below.
- The MANE track will list gene names on the right side of the genomic features panel.
- Note that the human genome has many more tracks than those we've seen in class, and you can largely ignore them. Focus on those we have covered in class. Any additional tracks other than the defaults used here need to be listed as part of your answer.
- <u>You may NOT look for the information being asked in other places.</u>
- The goal here is for you to show me you can use the genome browser to answer the kinds of questions being asked. Therefore, simply given the answer will not suffice. You will need to explain how you obtained it.
- <u>I strongly recommend you post screenshots with some or all of your answers. This will make it easier for me to understand your reasoning.</u>

ing on the same tracks used in labs for Drosophila

Exam - take home, open book



(1 pt) Question 3 __________

For questions3-10, your starting point is: chr17:31,061,287-31,380,471. Follow the instructions provided on the 2nd page before you get started.

How many protein-coding genes do you see in this region? List their names.

(2 pts) Question 4. __________

Do any of these genes overlap? If there is overlap, which part of which gene is overlapping which part of another (exon #, intron #)? Use only the visual features of the genome browser to answer the question. Explain your answer. Screenshots are encouraged.

ing on the same tracks used in labs for Drosophila

What I learned

- I recommend allowing enough time for them to finish the problem sets in class
- I often overestimated how much/how fast students could work through activities
- Students gave positive feedback on guided activities and exam questions following the format of incremental questions
- The good: having a deliverable made them come to class and take the activity seriously
- The bad: I don’t recommend giving students two modules on the same day and even less so one single problem set for two modules.

Next time

- I would like to use more materials next time and perhaps have a final project related to the activities from labs
- I want to give them at least one R activity
- Perhaps use more GEP materials for lecture-time active learning

Other projects

- Developing an R package with materials for B215: Biostatistics with R course. Currently not a package but many materials are freely available here: ADD LINK

Thank you!
Questions?

Email: bbitarello@brynmawr.edu (happy to share slides!)

Website: https://bitarellolab.digital.brynmawr.edu/

GitHub: https://github.com/bitarellolab

Twitter (X): @dudutchy