GENOME 569

Bioinformatics workflows for high-throughput sequencing experiments

What is a bioinformatics workflow?

A bioinformatics workflow is the collection of **scripts**, **programs**, and **procedures** you use to transform your data into a paper.

Goals for the course

Learn how to write reproducible workflows to analyze your data.

Learn how to use some awesome software tools.

Become a better programmer.

A bit about me

- **High-school**: worked as student software engineer for US Army
- **College**: software engineer for stock, futures, & foreign currency broker
- **After college**: software engineer for a “Business intelligence” company
- **Grad school**: Computer science (supercomputing, then bioinformatics)
- **Postdoc**: RNA biology, stem cell biology, genomics tech dev
- **Now**: single-cell genomics tech dev, developmental bio

How we’re gonna do this

I will spend about half of each class telling you about one or more new tools

You will spend the rest of the class trying it out

You will read a lot more about it after class

You will make use of it for the projects

What you will learn in this class

- UNIX scripts to automate tedious and computationally intensive jobs
- How to explore a dataset with R
- How to write an R package so others can do so on similar datasets
- How to make your data analysis and figure generation reproducible
- How to share all this code with the world

What you will NOT learn in this class

- Industrial-grade software engineering
- Computer science
- Computational biology
- Other kinds of biology
- Theory of any kind

How grading will work

- Show up to class
- Try to do the exercises in class
- Participate on slack. Help your classmates, and ask for help
- Try to do the programming homeworks.
- You get an A

The Classes

- Before each class, I will give you a lot of stuff to read, but it will be fun
- I will present some slides on a couple of really awesome programming tools each class.
- Then you will do some in class exercises with them
- When you have learned enough, you will be given a larger **programming project** that uses them.

The Projects

1. A pipeline to process raw single-cell RNA-seq reads

2. An electronic lab notebook that encapsulates your exploration of the dataset

RESEARCH ARTICLE

SINGLE-CELL GENOMICS

Comprehensive single-cell transcriptional profiling of a multicellular organism

Junyue Cao,[1,2*] Jonathan S. Packer,[1*] Vijay Ramani,[1†] Darren A. Cusanovich,[1†] Chau Huynh,[1] Riza Daza,[1] Xiaojie Qiu,[1,2] Choli Lee,[1] Scott N. Furlan,[3,4,5] Frank J. Steemers,[6] Andrew Adey,[7,8] Robert H. Waterston,[1‡] Cole Trapnell,[1‡] Jay Shendure[1,9‡]

Preliminary hassles

You will need a laptop in class.

You will need to be able to log into the GS cluster.

```
$ ssh <your-userid>@nexus2.gs.washington.edu
```

Let’s get started

Git

What is version control?

A *version control system* keeps track of changes to computer code over time

Modern VCS manage code contributions from many users, merge conflicting edits, allow changes to be compared, audited, and reverted

Almost all software companies and most open source projects use VCS.

Why do we use version control?

Share code easily with collaborators

Revert code if you make a mistake

Improves code clarity and quality

Keep a record of what you've done

Git

Git is a VCS (written by Linus Torvalds)

GitHub is a website that hosts Git repositories

We will use both in this class

Go create a GitHub account

(If you don’t already have one)

www.github.com

DM me your github ID on the class slack

Let’s use Git on the cluster

First, you need to generate an ssh key:

```
$ ssh-keygen -t rsa -b 4096 -C "your_email@example.com"

> Generating public/private rsa key pair.

> Enter passphrase (empty for no passphrase): [Just hit enter]
> Enter same passphrase again: [Just hit enter again]

$ cat ~/.ssh/id_rsa.pub
```

Copy that whole block of text (not the command prompt) into the clipboard

Add your ssh key to GitHub

(follow the instructions I am about to slack you)

https://help.github.com/en/github/authenticating-to-github/adding-a-new-ssh-key-to-your-github-account

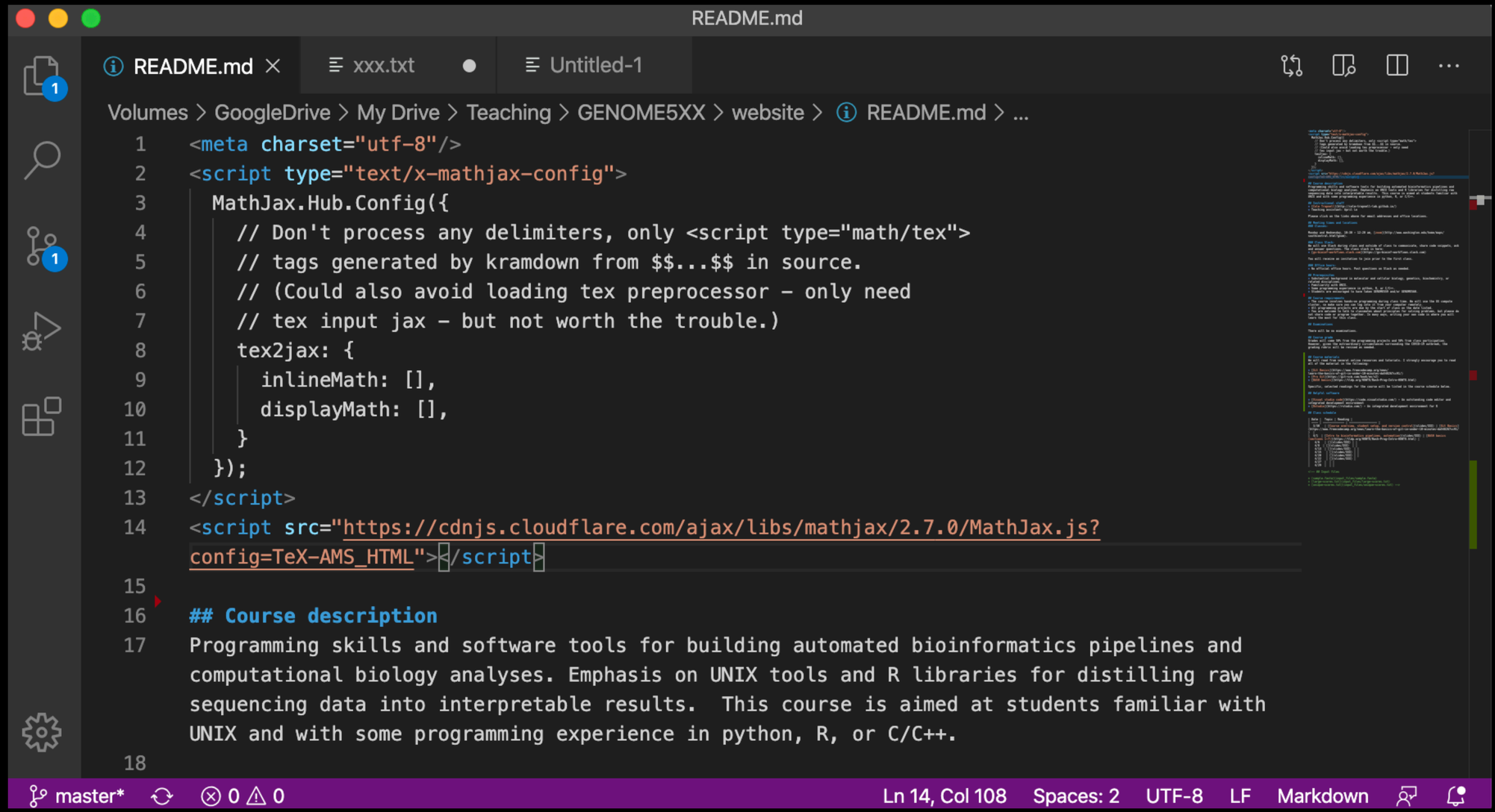
Clone the “hello world” repo

Click on this link: https://classroom.github.com/a/SEzXVNY0

```
$ git clone git@github.com:coletrapnell-teaching/hello-world-<your_gihub_id>.git
```

```
Cloning into 'hello-world'...
Warning: Permanently added the RSA host key for IP address '192.30.255.112' to the list of known hosts.
remote: Enumerating objects: 3, done.
remote: Counting objects: 100% (3/3), done.
remote: Total 3 (delta 0), reused 0 (delta 0), pack-reused 0
Receiving objects: 100% (3/3), done.
```

Visual Studio Code is a free code editor



Visual Studio Code is a free code editor

Download Visual Studio Code on your local computer

(optional) clone your GitHub Repo on your local computer as well

Open up README.md in the Hello World repo

Edit README.md and save it

Copy README.md back up to the cluster, overwriting the original

Push changes to “hello world”

```
$ cd hello-world-<your_gihub_id>
$ git stage README.md
$ git commit -m “Added my first checkin”
$ git push origin
```

THE END

For now