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

- 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

UNIX scripts to automate tedious and computationally intensive jobs

What you will NOT learn in this class

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

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

- 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
- dataset

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

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 to be able to log into the GS cluster.

You will need a laptop in class.

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



Let's get started



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 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

www.github.com

DM me your github ID on the class slack

(If you don't already have one)

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: <u>https://classroom.github.com/a/SEzXVNY0</u>

\$ 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

	•					RI					
ſ	(i) READ	ME.md $ imes$	≣ xxx.txt	•	≣ Untitled-	1					
	Volumes	s > GoogleD	rive > My Drive	e > Tea	aching > GENO	ME5XX					
\bigcirc	1	<meta cha<="" th=""/> <th>rset="utf-8"</th> <th>/></th> <th></th> <th></th>	rset="utf-8"	/>							
\succ	2	<pre><script type="text/x-mathjax-config"></pre></th></tr><tr><th></th><th>3</th><th>MathJax</th><th>.Hub.Config(</th><th>{</th><th></th><th></th></tr><tr><th>90</th><th>4</th><th>// Doi</th><th>n't process</th><th>any d</th><th>elimiters, o</th><th>only <s</th></tr><tr><th>٥(1)</th><th>5</th><th>// tag</th><th>gs generated</th><th>l by k</th><th>ramdown from</th><th>n \$\$</th></tr><tr><th></th><th>6</th><th>// (Co</th><th>ould also av</th><th>void l</th><th>oading tex p.</th><th>reproc</th></tr><tr><th>$\langle \rangle$</th><th>7</th><th>// te</th><th>x input jax</th><th>– but</th><th>not worth t</th><th>he tro</th></tr><tr><th><u>ਨ</u></th><th>8</th><th>tex2ja</th><th>ax: {</th><th></th><th></th><th></th></tr><tr><th>_</th><th>9</th><th>inl</th><th>ineMath: [],</th><th></th><th></th><th></th></tr><tr><th>ΗC .</th><th>10</th><th>dis</th><th>playMath: []</th><th>,</th><th></th><th></th></tr><tr><th>ш</th><th>11</th><th>}</th><th></th><th></th><th></th><th></th></tr><tr><th></th><th>12</th><th><pre>});</pre></th><th></th><th></th><th></th><th></th></tr><tr><th></th><th>13</th><th></script></pre>									
	14	<script s<="" th=""></script>									

EADME.md					
	ţ	Ţ			
> website > (i) README.md >					
		 Annual Annual Andrew Sterner Sterner and Sterner Sterner	 scoregi (gger"add;"tar"s Signature; reasons with a reasons wit	16. jul data papatana ani riskata function wata	
<pre>script type="math/tex"></pre>		 Destrong according to the basis shown for each Planet click on the basis shown for each With Destroy them and Socialization With Concerns. Basis and Machine And Specific and Basis and Machine And Specific and Basis and Machine And Specific action of the Antonyon Click and Antonyon Click and Specific actions. 	; addresses and afflice locations [cont](Mip://www.ardingles.ed	n fanns hages i	
\$\$ in source.		In walk are likely derived relative during and ansame quantums. The dataset during a regreterization on termination of the second from with remote an another termination and the second and an another termination is a structure during termination of the distribution of the termination of the second of the second second	is all class in communication, hence, hence, light handshear blass, stark, coat, riser to the Parel stars, not as Mark as medial.		
cessor – only need ouble.)		 biological de la la la compara de la compara de la compara en presente en presente en entre en la compara de la com	Alfalar Karlago, general L. et C.Con, alfantation and we attended and an experimental and an experimental processing of the second second second of the processing of the data of a data processing part and the data data processing part and take and processing part and take and processing part and take and processing part and take and participation of the second second and participation of the second second second and the second second second second second second and the second second second second second and the second second second second second second and the second second second second second second second and the second second second second second second second and the second second second second second second second second and the second second second second second second second and the second seco	alamistry, ar a the UL compute data. Alami, but plance de 'so alami pas atti.	
		Al Course prob France, add, cam bit. From the programma heating, game the advantationary threads young when add, to remain a course of the add read from same and addies research do. of the automatic in the following	a projects and DP, from class per lamon survening the CBCD-OF	riscipation. addresd, Un	
		 Child Amazari Malaya yina Ganara Childra Lange of the match of split to contact Childran The Bally Malaya (Subject Source and Bally and Similar Malaya (Subject Source) Bally Malaya (Subject Source) Bally Malaya (Subject Source) Bally Malaya (Subject Source) Similar Analysis (Subject Yorks and Malaya (Subject Source) Similar Analysis (Subjec	$\begin{array}{l} \begin{array}{l} x_{0}, y_{0}, y_{0$	uladila latas. cada adilar ani	
		************************************	rege when descriptions control () (define the maximum control () (define the house of physical control () and the second se	(Nor I (North Hild Research) (North Herican) (NORI Research	
		 Complex Partial States Complex Partial States Complex Partial States (Solidary Values Values) 			
jax/libs/mathjax/2.7.0/MathJax.js?					
uilding automated bioinformatics pipelines and UNIX tools and R libraries for distilling raw This course is aimed at students familiar wit n python, R, or C/C++.	h				
Ln 14, Col 108 Spaces: 2 UTF-8 LF	Ма	rkdown	ନ୍ଦି	Ļ	





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 push origin

\$ git commit -m "Added my first checkin"



For now