Genome Variant Analysis 2021

General Introduction to the Zoom edition

Background

- Doctorate from The Ohio State University
  - Epigenetic modifications and cell signaling in ovarian cancer
- Postdoctoral work with Jeffrey Barrick at UT Austin
  - Microbial evolution, and synthetic biology
- Self taught computational biologist
  - 9th year teaching this course

- Commonality:
  - Next-generation Sequencing

Disclaimers up front

- Spelling – Has never been a skill I possess, I'm sure there are errors left in the tutorials. Let me know if any interfere with learning.
- Typos – Will likely be your biggest problem in using the commands provided, and the biggest problem in your own work. Try to type the commands out to get practice, but remember you can copy paste.
- Names – Normally I apologize up front for not being able to remember peoples names. This year I apologize if I mispronounce your name, please do correct me if it bothers you.
- Royal "we" – Several of these tutorials were originally produced by former instructors or TA's and reference "we" or "your instructor(s). I thank them all for their help, and take responsibility for anything that goes wrong.

Zoom Disclaimers

- Please change your zoom name to start with a M or P depending on if you are using a Mac or PC.
- Feedback will be more important than ever, email or before/after class.
- Don’t be shy asking questions or wanting me to look at your screen.
- Dogs – I have dogs they are awesome and usually lazy during the day. Very possible at least one of them will decide to bark while I am talking. He’s very sorry he just can’t help it.
- Birds – I do NOT have a pet bird. I DO have a husky who does a pretty good bird call and an imitation of a screaming hyena. He’s also sorry.

Class Expectations

- Participant goals:
  1. Learn how to analyze my data, and have it fully analyzed by the end of the class.
  2. Maybe learn how to analyze NGS data in general.
- Teaching goals:
  1. Teach the fundamentals of NGS variant analysis.
  2. Provide context and exposure multiple types of data.
  3. Use example commands to familiarize you with variety of programs.
  4. Provide resources to enable you to do analysis you haven’t thought of yet.

Where to start?

- Many say “don’t know where to start” their data analysis once they have their data files.
- Ideally should have “started” weeks-months ago in planning experiments.
  - If you fall in this category don’t worry you are in a common situation.
  - Later presentations will explain differences in sequencing libraries and how to make the best choices going forward.

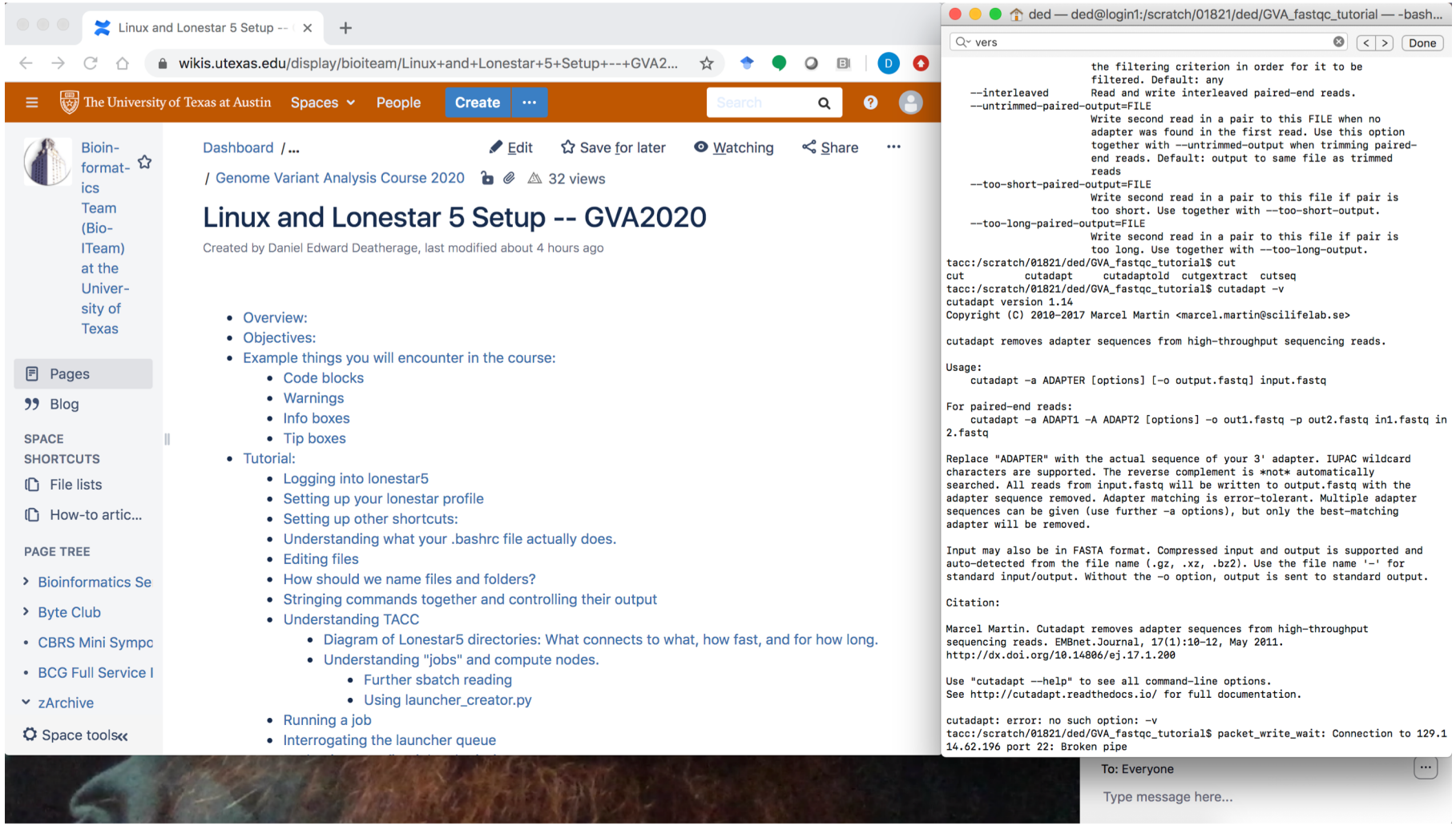
General notes about analysis

- There is no one “perfect” way to analyze next-gen data.
  - Results in many, nearly equivalent or redundant analysis types and tools.
  - Not all types of analysis are necessary, or even appropriate.
- Wet lab methods used to generate data may limit available/appropriate analysis methods.
  - Not-surprisingly, planning experiments from the beginning based on what answers/analysis you want can save time and money.

Introductions:

- Who are you?
- What lab are you in?
- What organism do you work with (can be as broad as virus, bacteria, plant, animal, human, etc)?
- 1-2 sentences on what type of variant analysis you are interested in
  - Type of reference genome
  - Mixed populations vs clonal isolates
  - Association vs causation studies

Set up recommendations



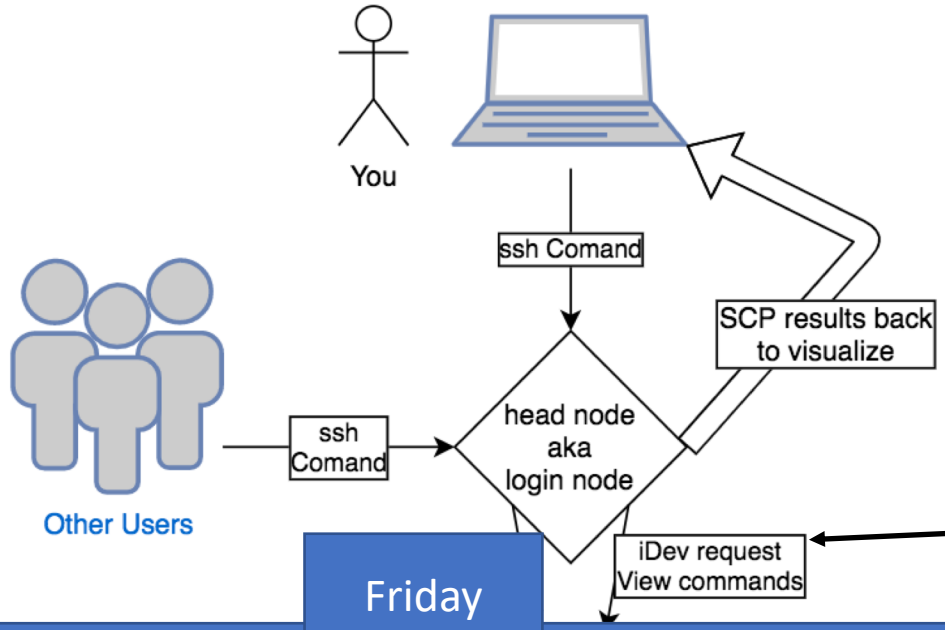
Computers computers computers

- Millions of reads, 100s of bp long, mapping to millions-billions of base long references.
- Mac/linux computers are your friend.
- Windows is getting better with powershell, but many analysis programs are only available for linux/mac.

- TACC is a time machine that lets you get stuff done much faster
  - This is where we will start the class.

TACC – where the analysis happens

- Where to start
  - Eventually end up with data, we all want to leverage other people's work to mimic their analysis as much as possible.
- Environment
  - These are variables stored on your computer, or TACC, that control how the computer behaves and what programs it has access to. Things like: HOME, WORK2, SCRATCH, PATH, .bashrc, .profile, modules, conda and a few others are discussed in the first tutorial.
- Programs – How do we access them
  - 3 most common/best ways from tutorial: TACC modules, conda system, direct downloads
  - Additional ways: repositories (github, docker, etc); other system managers: (pip, cpan, brew, etc)
  - BioITeam installed. (more on this later in the week)


You
ssh Comand
SCP results back to visualize
head node aka login node
ssh Comand
Other Users
iDev request
View commands
Won't use until tomorrow, but is discussed in tutorial
Friday

Tomorrow