

Genome Variant Analysis 2019

General Introduction

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
 - 7th year teaching this course
- Commonality:
 - Next-generation Sequencing

Disclaimers up front

- Spelling – Has never been a skill I possess, hopefully will only be noticeable if I write on board.
- 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 – I usually use this opportunity to apologize for my inability to remember people's names with a funny anecdote, with small class might not be an issue.
- Royal “we” – Several of these tutorials were originally produced by former instructors or TA's. I thank them all for their help, and take responsibility for anything that goes wrong.

Class Expectations

- Participant goals:
 1. Learn how to analyze my data, and have it fully analyzed by the end of the class.
 2. 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.

General notes about analysis

- There is no “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.

Computers computers computers

- Millions of reads, 100s of bp long, mapping to millions-billions of base long references.
- Windows is your enemy, linux/Mac is your friend.
- TACC is a time machine that lets you get stuff done much faster
 - This is where we will start the class.