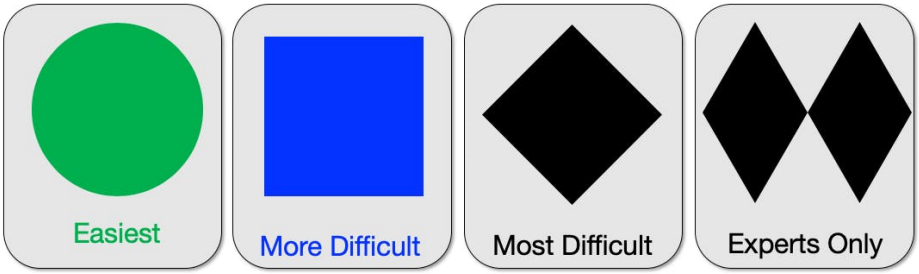
Overview of Final Project

Please note this is only part of the overview. As the semester continues, you will be provided with even more information*

I. **Overview**: Each student (or student group) will give an oral presentation describing their analysis of an NGS dataset. These presentations will be given by the student(s) on April 30th and May 2nd. All students are expected to be present for the entire session, provide feedback, and ask questions to each presenter. Student's will be asked to select from (3) distinct trails. The trail selected will guide the overall goal and analysis of the NGS dataset downloaded from GEO.

II. **Trails:** Ski trails are named and rated with different colors to signify to the skier the level of difficulty. For these trails, green signifies entry or beginner, blue signifies intermediate, while black signifies expert.

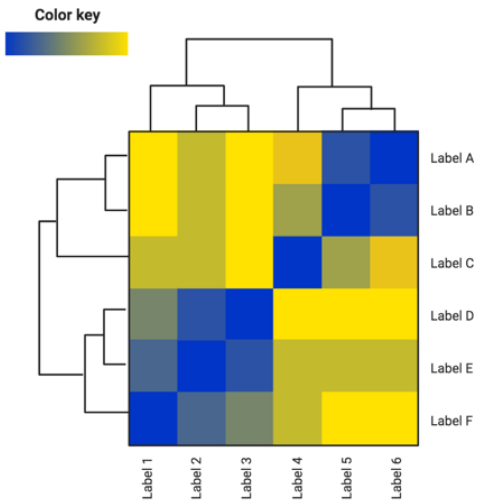


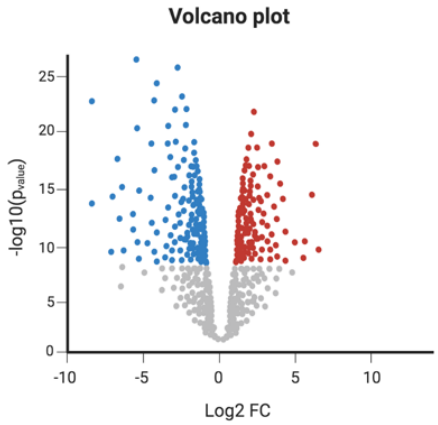
Similarly, the final project for this course will mirror these ratings with the understanding that within this class there are varying levels of expertise.

However, no matter the trail chosen, all students will be able to:

- ☑ Download NGS data from GEO
- ☑ Use FASTQC to assess fastq files, i.e. interpret data quality
- ☑ Use an adapter trimmer to remove adapters or low quality reads
- ☑ FASTQ → SAM → BAM → counts
- ☑ Generate basic visualizations i.e. heatmap, PCA, pathway analysis
- ☑ Interpret their results

Trail 1: Green Mountain Trail
"Replicate figure(s) in a primary research article and then change one parameter at the visualization stage"

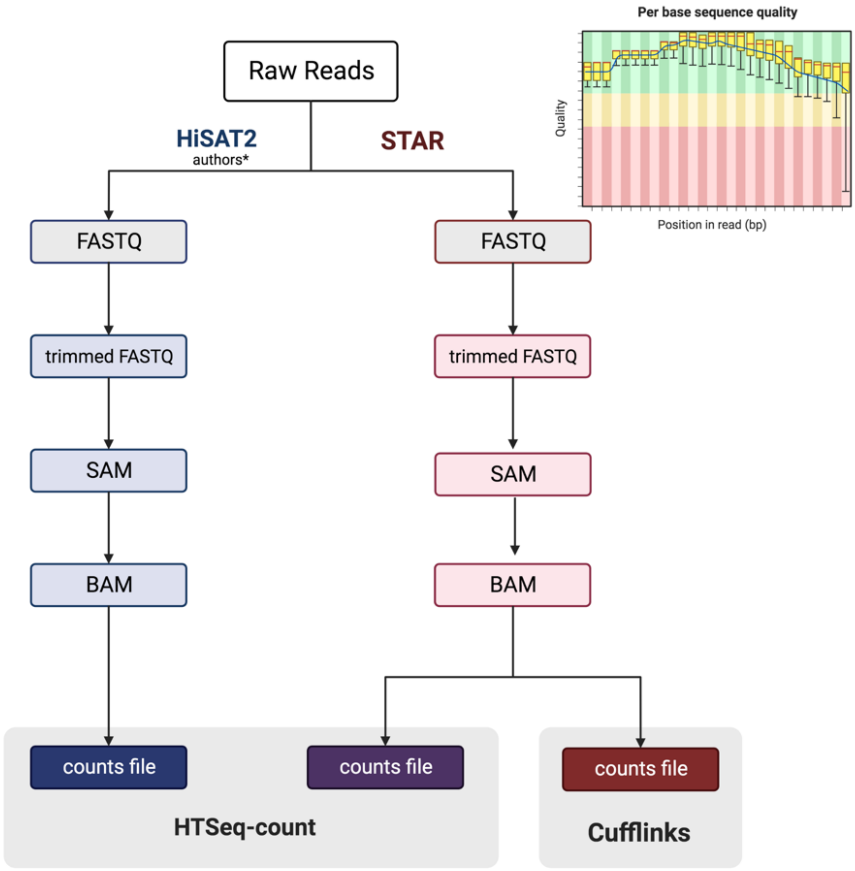




- Change the colors, scale, labels, and/or input

Trail 2: Blue Sky Trail

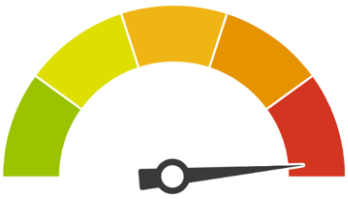
"Compare and Contrast select parameters within the NGS pipeline and describe how this impacts your final output"



- Requires an in-depth understanding of how the authors processed their dataset
- You will be required to give an in-depth overview of the entire analysis from raw reads to DE analysis + figures
- **We will not** cover all pipelines during class - so this does require ***independence***

Trail 3: Black Diamond Trail
"Process and Download an NGS dataset to test an original hypothesis"

- Graduate students will be **required** to select this option
- The final figure(s) you generate should not be found within the published dataset
- **Original, creating your own path, you are the trailblazer**



How do you Select A Trail?

What personal goal do you have?

- ☑ "I want to be confident downloading a dataset from GEO & replicating results" - Green Trail

AND

- ☑ "I want to added challenge"
  "I want to be able to understand the difference in using varying computational tools and when I would implement them"
  "I am thinking of bioinformatics as a future profession" - Blue Trail

- ☑ "I *want* to go to graduate school"
  "I'm in graduate school and I want to advance my research project" - Black DiamondTrail

III. **General:**

a. Undergraduate students will be allowed to work in pairs if they prefer. This is an individual assignment for graduate students unless granted permission.
b. Each student project will be allocated 15 minutes to present their findings and answer questions from the audience. The audience will be able to ask you questions ***during*** the presentation.
c. All team members must speak for the same amount of time and be ready to answer questions.

IV. **Presentation content & Grading Rubric**: This will be provided to students in March, and will posted on Blackboard and on the course website: (https://prodriguez19.github.io/Intro-to-rnaseq/)

V. **While selecting a primary research article, please consider the following:**

a. NGS data type selection:

Acceptable	**Unacceptable**
RNA-Seq	Single-cell RNA-Seq or ATAC-Seq
ChIP-Seq	Microarray
ATAC-Seq	Small RNA-Seq
Research-specific dataset **(permission required)**	Spatial Transcriptomic Dataset

b. Organism: I will provide the indexed genome for mm10, hg38, and hg19 for HISAT2 and STAR alignment*
c. Number of biological replicates available: at minimum 3-4 replicates per group is required (RNA-Seq), 3 or more for most other data types
d. Date of publication: 2000 to 2023
    i. Would prefer within the past 10-15 years but an older dataset will be considered if justified to the instructor

VI. **Timeline:**

	Selecting a dataset	**Download dataset**	**Index Genome**	**Alignment**
Estimated time to complete	1-2 weeks	24 hours	1hr – 3 days	3-7 days +
Comment		Per 5GB = 1.5 hrs = one sample	Depends on how large the genome is Dependent on alignment strategy	Dependent on the number of samples
Homework Assignment	~100 points Select dataset, and justify why dataset and trail were selected	~100 points FASTQC + interpretation		~100 points Alignment stats + interpretation Decision to be made on ***how*** to proceed based on interpretation
Due dates (tentative)	Feb 23rd	March 5th		March 26th

VII. **Important Disclosures**

- While in-class, we will be going through the basic steps of data processing using a dataset that is publicly available.
- This project requires that you use what you learned in-class and apply it to a different NGS dataset.
- Depending on the dataset, and your question question/aim, this will require independence.
- I will not know the quality of the dataset selected until about March. Therefore, depending on what we find we may need to pivot and change the intention of the final project goals.
- I am most familiar with advising on a human or mouse system. However, other organisms are completely fine to select. You will be in charge of understanding if for example "*…there are pathway analysis tools available for Drosophila…" or where to find the GTF file for bacteria.*
- We will hit many unforeseen hiccups. This is completely normal in the realm of bioinformatics! Be prepared to troubleshoot.
- I do not have control over how fast or slow your data will process on the VACC. The alignment step is the most COMPUTATIONAL HEAVY STEP of the ENTIRE pipeline. Please do not leave this for the last minute as the VACC does have multiple users!