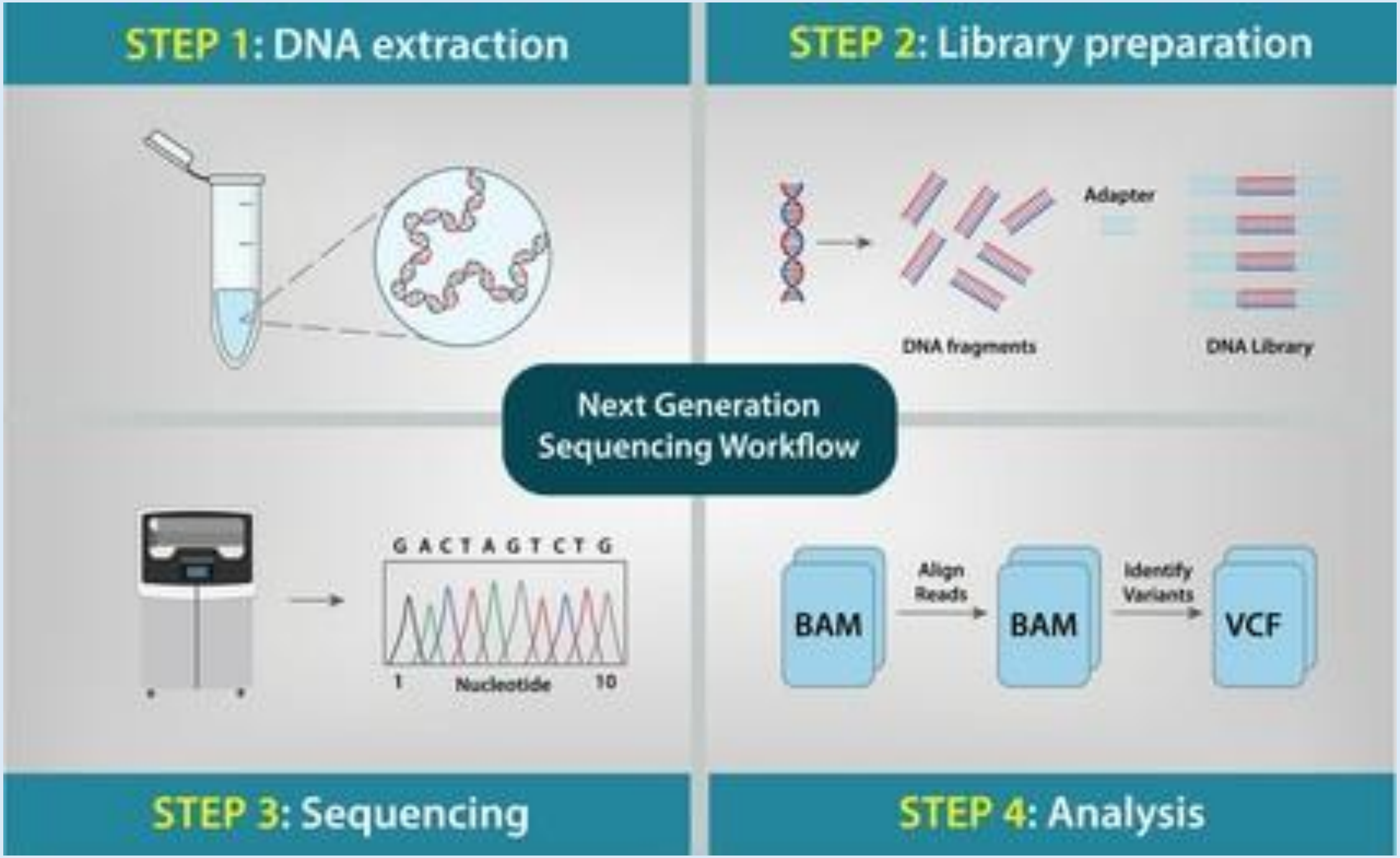


Final Project Discussion

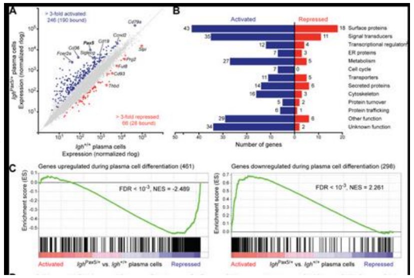
Princess Rodriguez, PhD

MMG 3320/5320
Spring 2025

Quick Overview of NGS technologies

Final Project Prompt

- You will analyze an NGS dataset of your choosing from "start" to "finish".
- You will begin by identifying your dataset.
- You will then download the data.
- You will then process it.
- You will then visualize it.
- You will then interpret and deliver your findings.



- Along the way you will perform QUALITY CONTROL

Final Project Delivery

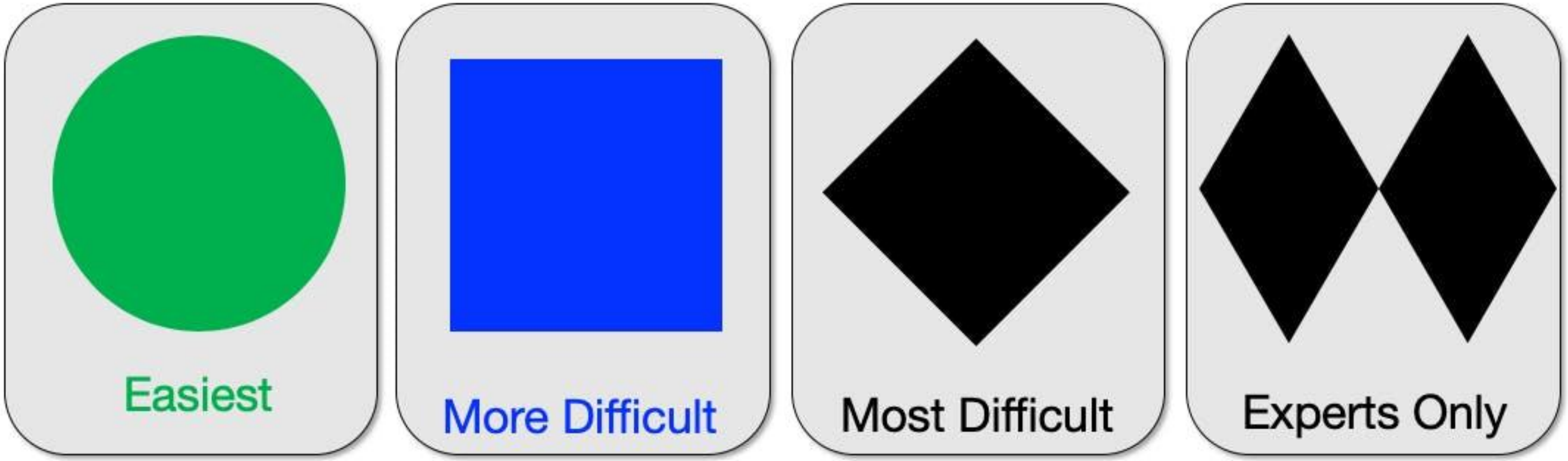
- **Undergraduate students** will be asked to submit a **written report** detailing their analysis and findings.

- **Graduate students** will deliver an **oral presentation** to communicate their results and interpretations.

- Everyone will submit a folder with their compiled analysis… more details to come!

- Everyone must be present for the final weeks' presentations.

Ski Trails



You will be asked to select a trail and a corresponding challenge.

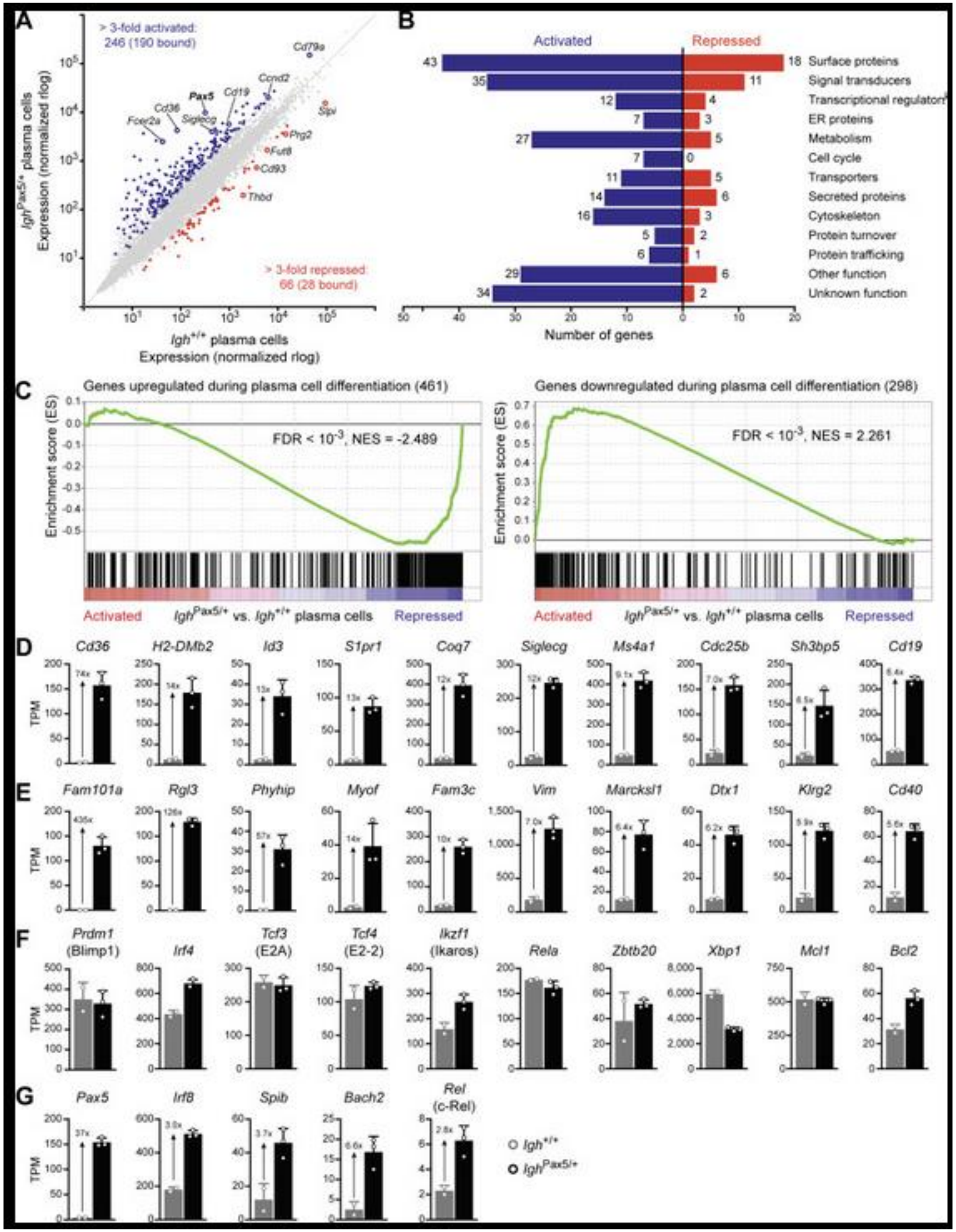
All challenge prompts below are *specific* to RNA-Seq. If you select a different kind of NGS dataset to analyze, I will generate a challenge prompt specific for your data type and trail.

Green Mountain Trail

Replicate a figure in a primary research article and then change one parameter at the visualization stage

Green Mountain Trail

Challenge 1: Adjusting the Threshold for Differential Gene Expression (DEG)

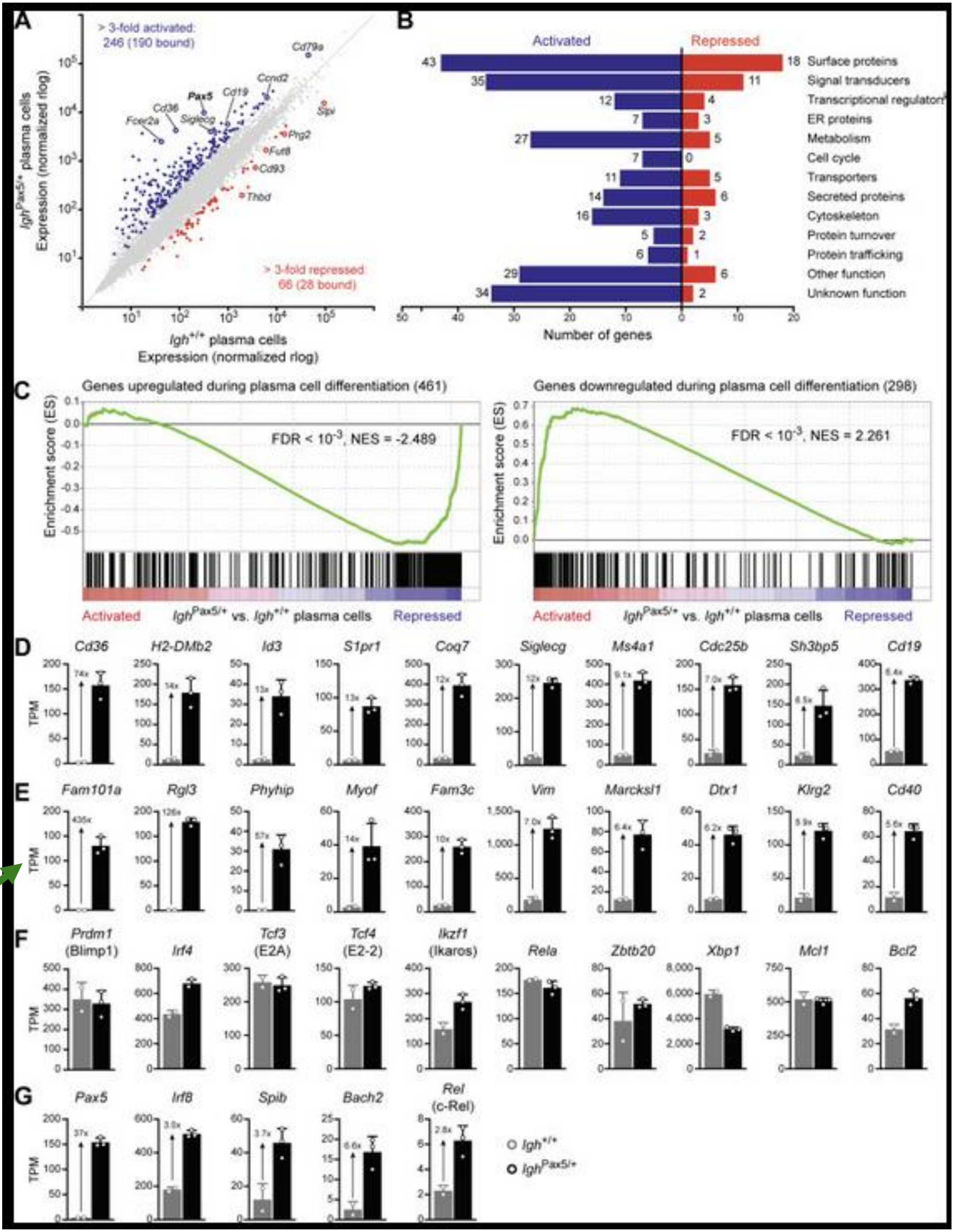


Green Mountain Trail

Challenge 2: Experimenting with Normalization Techniques

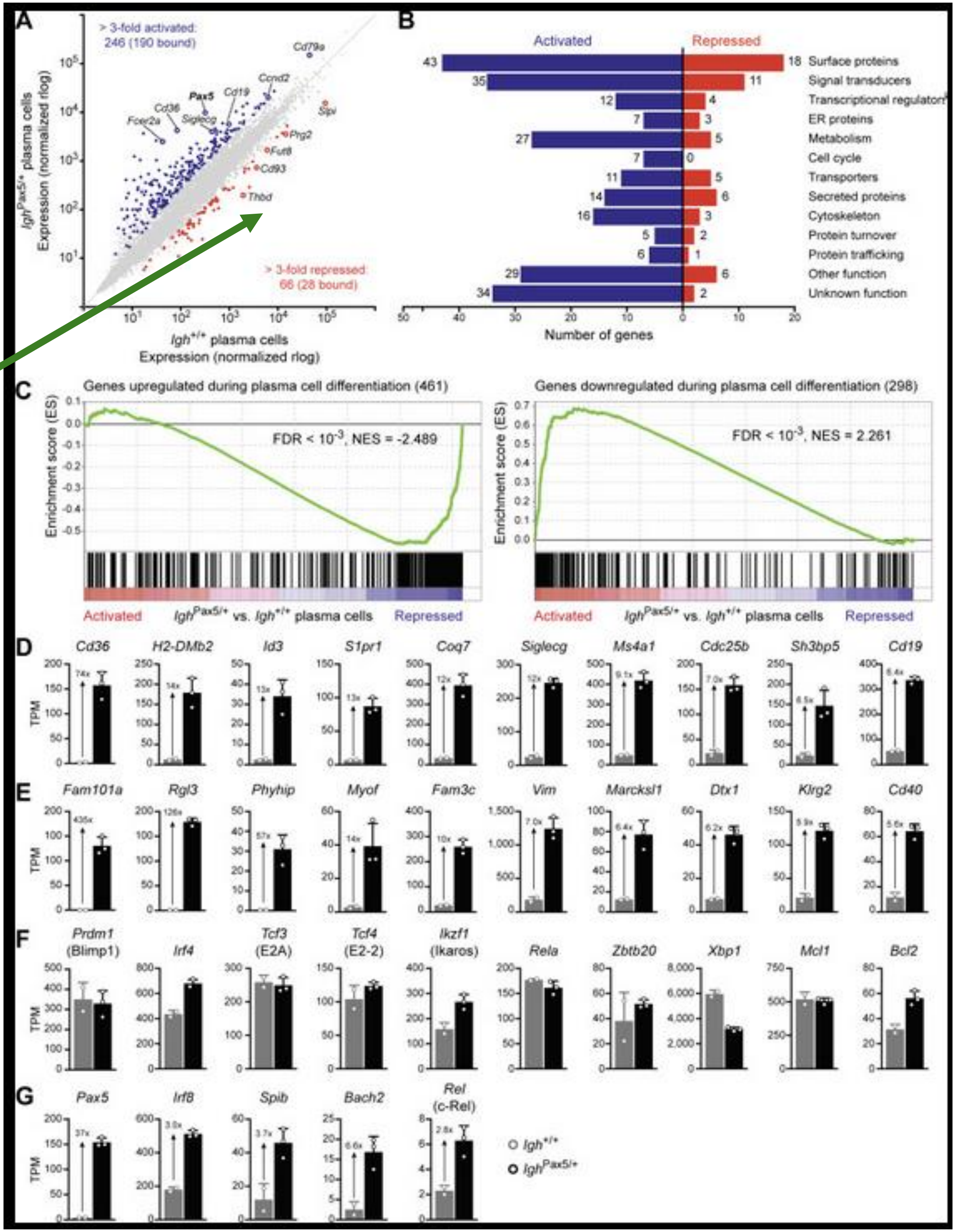
Compare visualizations of gene expression data normalized using two methods (e.g., **TPM, CPM, vs. DESeq2's variance-stabilizing transformation**). Assess how normalization affects downstream analyses like bar or box plots.

Do you know TPM means?



Green Mountain Trail

How would you label this MA plot?



Challenge 3: Changing Color Schemes/Labeling for Data Interpretation

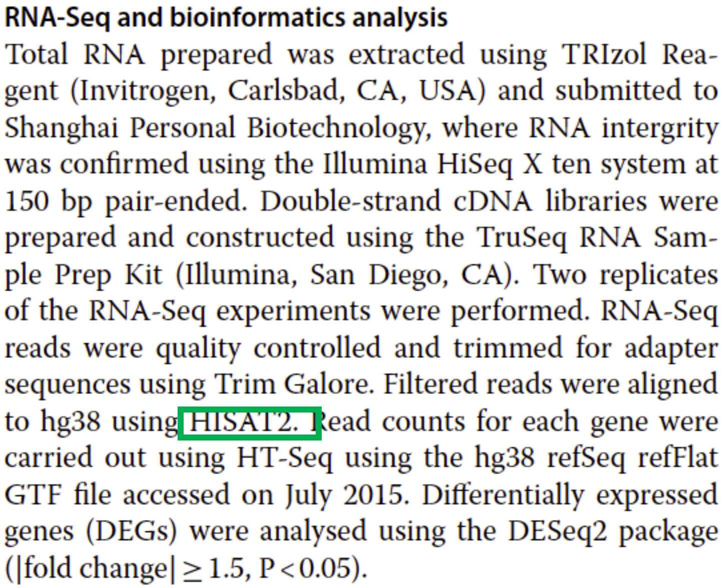
Adjust the color scale of a heatmap (e.g., changing from a red-green to a blue-yellow color scheme) and evaluate how the choice of visualization colors influences the clarity of expression trends and ease of data interpretation.

Blue Sky Trail

Compare and Contrast bioinformatic tools during the preprocessing stage and describe its impact on the data interpretation

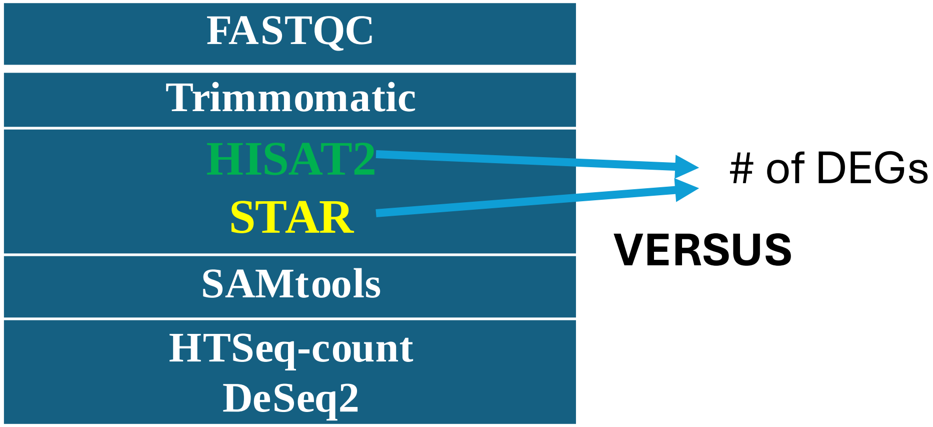
The bioinformatic pipeline we will learn in class is:

MMG3320	What it does…
FASTQC	Quality control FASTQC files
Trimmomatic	Trim adaptors and low quality reads
HISAT2 STAR	Alignment to Genome
SAMtools	SAM to BAM
HTSeq-count	Create counts files



RNA-Seq and bioinformatics analysis

Total RNA prepared was extracted using TRIzol Reagent (Invitrogen, Carlsbad, CA, USA) and submitted to Shanghai Personal Biotechnology, where RNA intergrity was confirmed using the Illumina HiSeq X ten system at 150 bp pair-ended. Double-strand cDNA libraries were prepared and constructed using the TruSeq RNA Sample Prep Kit (Illumina, San Diego, CA). Two replicates of the RNA-Seq experiments were performed. RNA-Seq reads were quality controlled and trimmed for adapter sequences using Trim Galore. Filtered reads were aligned to hg38 using HISAT2. Read counts for each gene were carried out using HT-Seq using the hg38 refSeq refFlat GTF file accessed on July 2015. Differentially expressed genes (DEGs) were analysed using the DESeq2 package (|fold change| ≥ 1.5, P < 0.05).



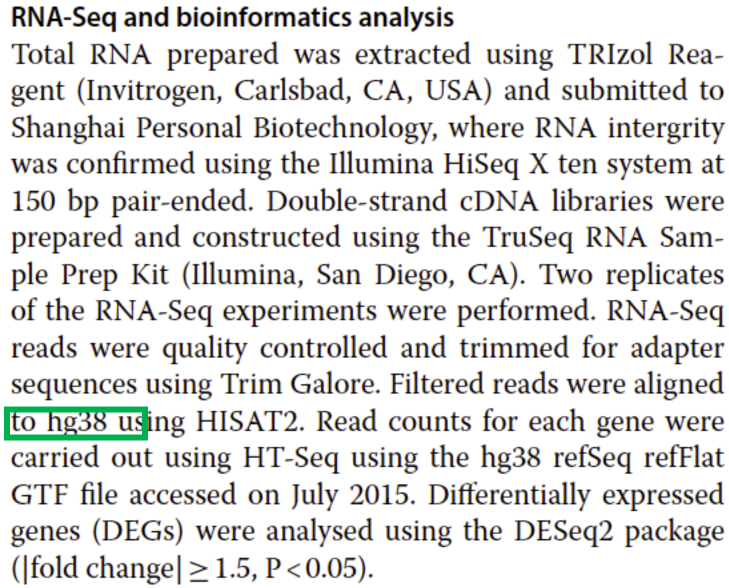
Information on the data processing can be found …?

GEO Gene Expression Omnibus

PubMed®

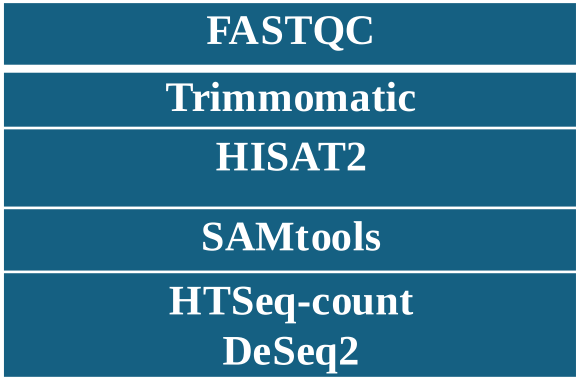
Challenge 1: Testing Different Alignment Tools

Align the RNA-Seq reads to the reference genome using two different aligners (e.g., HISAT2 vs. STAR). Compare metrics such as alignment rate, number of uniquely mapped reads, and runtime, and discuss how the choice of aligner might affect downstream analysis.



RNA-Seq and bioinformatics analysis

Total RNA prepared was extracted using TRIzol Reagent (Invitrogen, Carlsbad, CA, USA) and submitted to Shanghai Personal Biotechnology, where RNA intergrity was confirmed using the Illumina HiSeq X ten system at 150 bp pair-ended. Double-strand cDNA libraries were prepared and constructed using the TruSeq RNA Sample Prep Kit (Illumina, San Diego, CA). Two replicates of the RNA-Seq experiments were performed. RNA-Seq reads were quality controlled and trimmed for adapter sequences using Trim Galore. Filtered reads were aligned to hg38 using HISAT2. Read counts for each gene were carried out using HT-Seq using the hg38 refSeq refFlat GTF file accessed on July 2015. Differentially expressed genes (DEGs) were analysed using the DESeq2 package ($|\text{fold change}| \geq 1.5$, $P < 0.05$).



FASTQC

Trimmomatic

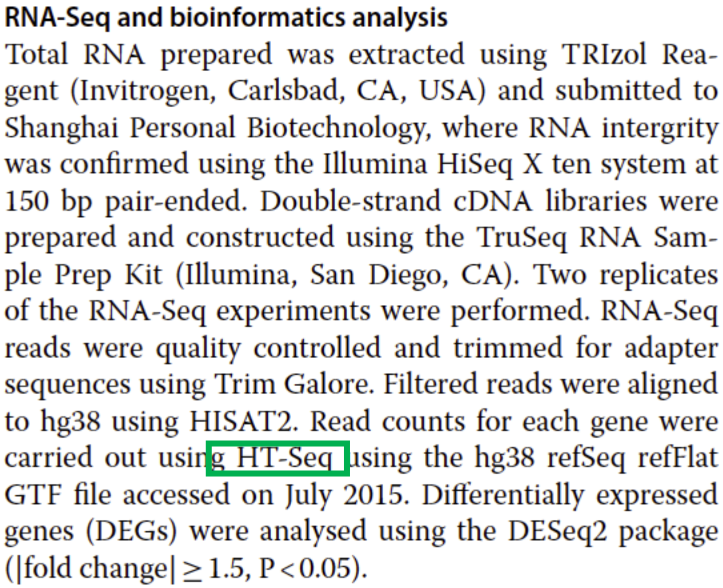
HISAT2

SAMtools

HTSeq-count
DeSeq2

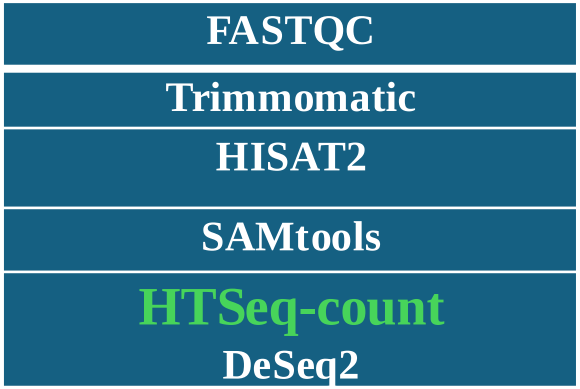
Challenge 2: Evaluating Reference Genome Versions

Map the RNA-Seq reads to two different versions of the reference genome (e.g., GRCh37 vs. GRCh38). Compare the alignment statistics and any differences in gene annotations. Discuss how the choice of reference genome might influence downstream results and biological interpretations.



RNA-Seq and bioinformatics analysis

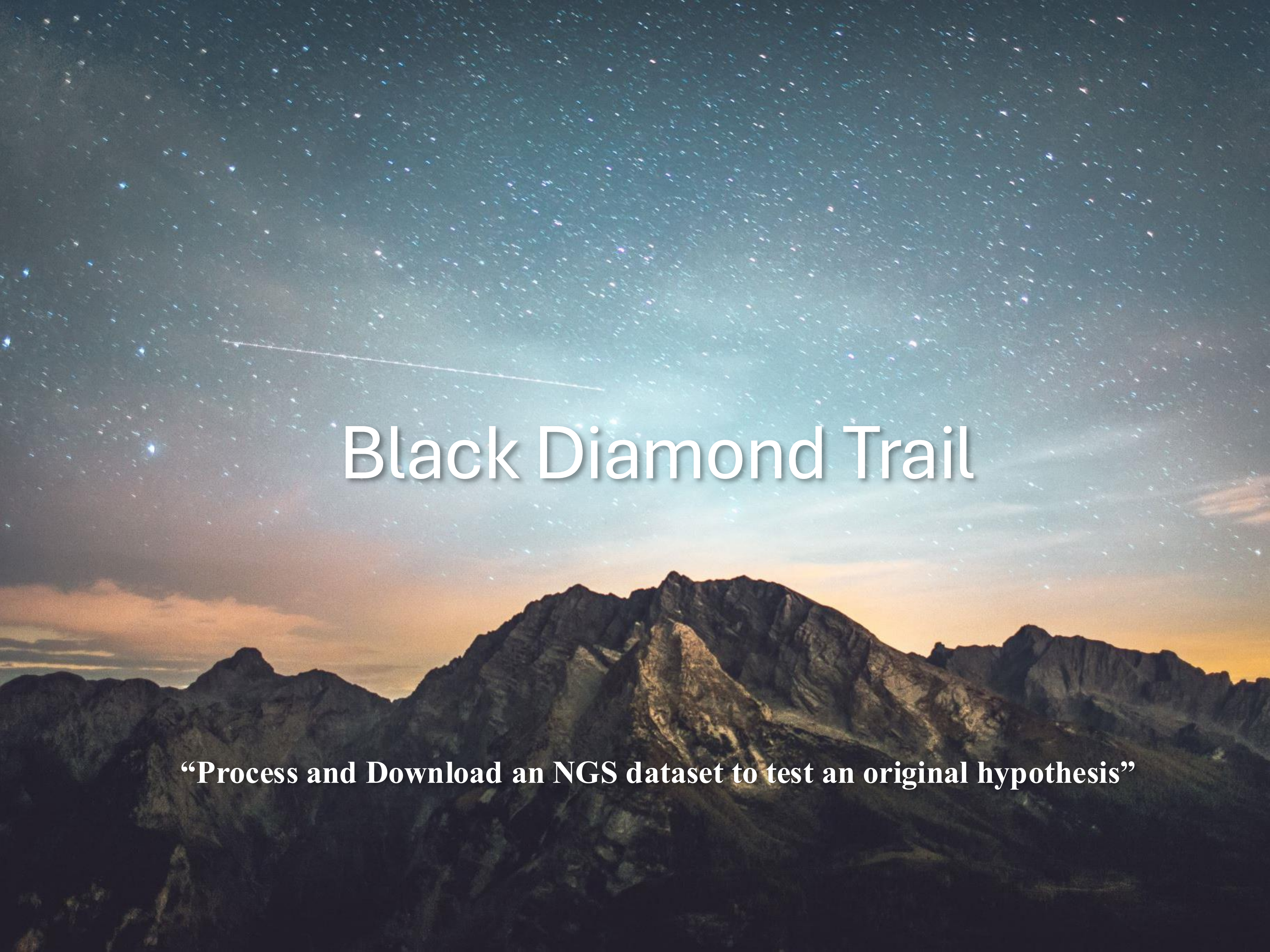
Total RNA prepared was extracted using TRIzol Reagent (Invitrogen, Carlsbad, CA, USA) and submitted to Shanghai Personal Biotechnology, where RNA intergrity was confirmed using the Illumina HiSeq X ten system at 150 bp pair-ended. Double-strand cDNA libraries were prepared and constructed using the TruSeq RNA Sample Prep Kit (Illumina, San Diego, CA). Two replicates of the RNA-Seq experiments were performed. RNA-Seq reads were quality controlled and trimmed for adapter sequences using Trim Galore. Filtered reads were aligned to hg38 using HISAT2. Read counts for each gene were carried out using HT-Seq using the hg38 refSeq refFlat GTF file accessed on July 2015. Differentially expressed genes (DEGs) were analysed using the DESeq2 package (|fold change| ≥ 1.5, $P < 0.05$).



Challenge 3: Comparing Count Generation Tools

Generate counts files using two different tools (e.g., HTSeq-count vs. featureCounts). Compare the total number of assigned reads, unassigned reads, and computational efficiency. Discuss how differences in counting strategies might influence downstream analyses such as differential expression.

The blue trail highlights the importance of tool selection during the preprocessing stage and its impact on the interpretation of RNA-Seq data.



Black Diamond Trail

“Process and Download an NGS dataset to test an original hypothesis”

Your overall approach will be different

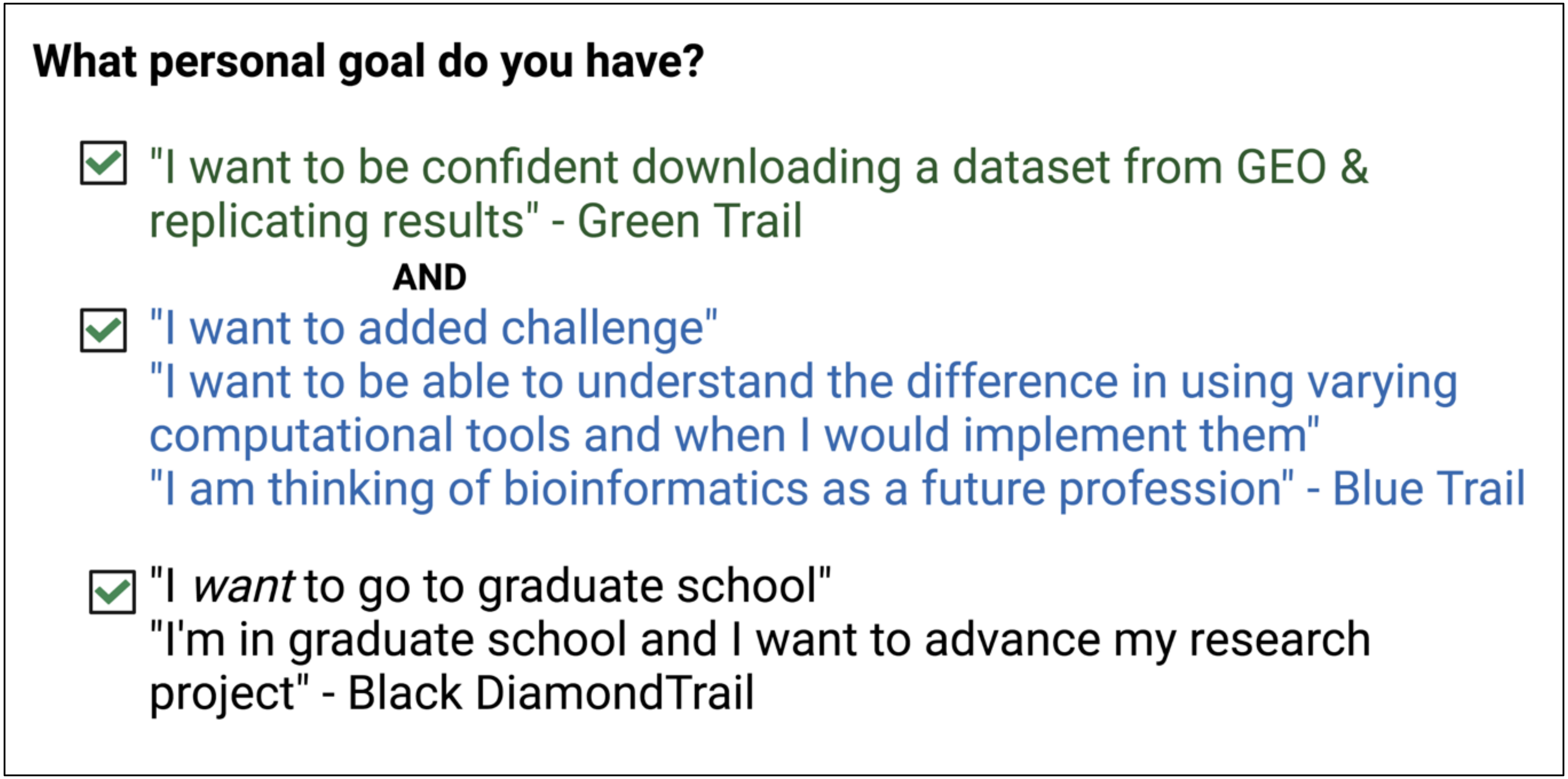
You are going in with a hypothesis and using the dataset to test this hypothesis.

“Compared to macrophages, I hypothesize that Dendritic cells activated with LPS will express an upregulation of glycolytic genes as opposed to genes required for oxidative phosphorylation.”

"Compared to macrophages, I hypothesize that Dendritic cells activated with LPS will express an upregulation of glycolytic genes as opposed to genes required for oxidative phosphorylation."

	Dataset	# of replicates
⇨	Macrophages (control)	3
⇨	Dendritic cells (control)	3
⇨	Macrophages + LPS	3
⇨	Dendritic cells + LPS	3
	~~Macrophages + Zeb1 KO~~	~~3~~
	~~Dendritic cells + Zeb1 KO~~	~~3~~
	~~Macrophages + Zeb1 KO~~ + LPS	~~3~~
	~~Dendritic cells + Zeb1 KO~~ + LPS	~~3~~

How do I 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

General

a. Undergraduate students will be allowed to work with a partner. This is an individual assignment for graduate students unless granted permission.

b. Graduate students must select either the blue or black trail.

c. Each graduate student will be allocated 15 minutes to present their findings and answer questions from the audience during the last week of class. All students are required to attend these sessions. The audience will be able to ask you questions ***during*** the presentation.

Selecting an NGS dataset

Acceptable	**Unacceptable**
RNA-Seq	Single-cell RNA-Seq
ChIP-Seq	Microarray
ATAC-Seq	Spatial Transcriptomics
Permission required: *Research-specific dataset Metagenomics WGS/WES	

Beware

	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 the number of samples Dependent on alignment strategy
Homework Assignment	~100 points Select dataset, and justify why dataset and trail were selected	~100 points FASTQC + interpretation		~150 points Alignment stats + interpretation Decision to be made on how to proceed based on interpretation
Due dates (tentative)	Mid Feb	Late Feb/ Early March		Early to Mid March

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.
- ***We both will not know the quality of the published dataset you 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!

Lessons from Last Year

- If you select black trail (*undergraduates*) but then see around April that your analysis is more aligned with green trail, this is 100% okay. But you must consult with me and tell me at least a week prior to your presentation that you will be ***changing trails***. There will be a major point deduction if your presentation and trail selected do not match!
- If you select the black trail, I expect an original hypothesis to be tested. Points will be deducted if this original hypothesis is not present or tested.
- I had multiple students throughout the years who opted to analyze a dataset “sitting in their lab.” Some of these students were wildly successful, others were not.