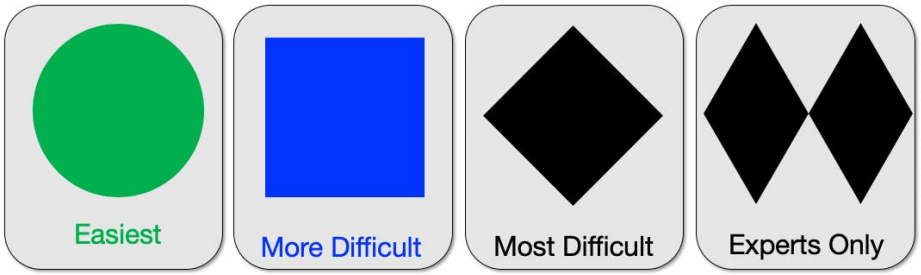
Overview of Final Project

I. **Overview**: Next-Generation Sequencing (NGS) refers to high-throughput technologies that enable rapid sequencing of DNA and RNA, revolutionizing genomics research. Unlike earlier Sanger sequencing methods, NGS platforms can process millions of fragments in parallel, providing an unprecedented scale of data. This makes NGS invaluable for applications such as genome assembly, transcriptome analysis, epigenetics, and metagenomics. With its high accuracy, cost-effectiveness, and versatility, NGS continues to drive discoveries in biology and medicine. In the coming weeks, students will analyze an NGS dataset of their choosing. Undergraduate students will be asked to submit a written report detailing their analysis and findings, while graduate students will deliver an oral presentation to communicate their results and interpretations. All students will select from three distinct tracks, each with a series of corresponding "challenges". The selected challenge prompt will guide the overall goal and analysis of the NGS dataset.

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

II. **Trails and Challenges:** 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

Challenge 1: Adjusting the Threshold for Differential Gene Expression (DEG)
Investigate how changing the log2 fold change threshold (e.g., from 1 to 0.5) impacts the number and biological interpretation of differentially expressed genes. Discuss the trade-off between sensitivity and specificity in DEG analysis.

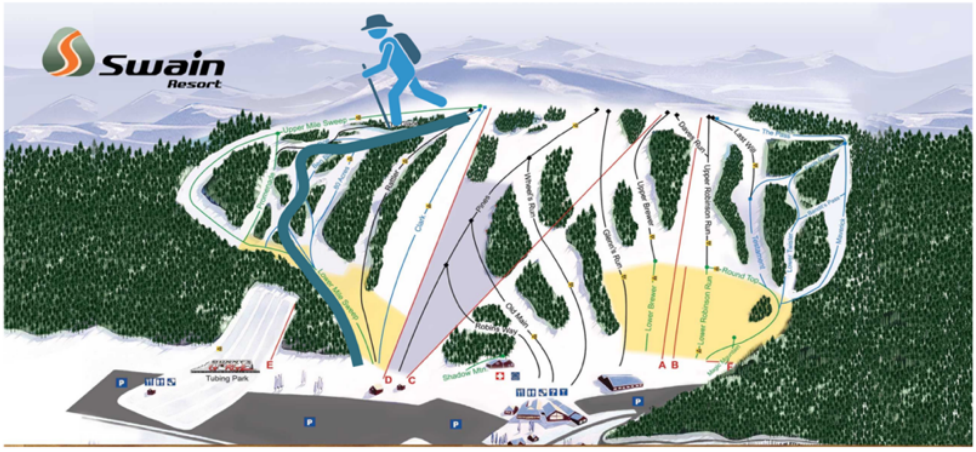
Challenge 2: Experimenting with Normalization Techniques
Compare visualizations of gene expression data normalized using two methods (e.g., TPM, CPM, vs. DESeq2's variance-stabilizing transformation). Assess how normalization affects downstream analyses like bar or box plots.

Challenge 3: Changing Color Schemes 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.

The green trail guides students in understanding how to make ethically responsible decisions when visualizing NGS data.

Trail 2: Blue Sky Trail

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



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.

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.

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.

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

Challenge 1: Creating Time-Series or Condition-Specific Plots
If your data includes multiple time points or conditions, create a figure (e.g., line plots or heatmaps) to visualize expression changes for key genes across these conditions. Highlight patterns or trends and discuss how they support or refute your biological hypothesis.

Challenge 2: Comparing Pathway Expression Across Groups
Use pathway analysis to identify key pathways enriched in a subset of your data. Create a customized plots (e.g. bar plots, dot plots, network graphs) to compare pathway activity between experimental groups not compared in the published work. Discuss how the visualization highlights the differences in pathway regulation.

Challenge 3: Annotating Single-Gene Expression Differences
Select a gene of interest from your dataset and create a violin plot or boxplot comparing its expression across conditions or groups. Customize the figure to include statistical annotations (e.g., p-values or fold changes) and explain why this gene is biologically significant.

For all black trail challenges you will be required to design a multi-panel figure that integrates multiple layers of analysis (e.g., a heatmap for expression patterns, a volcano plot for DEG results, and a GO enrichment bar chart). Explain how the combination of figures tells a cohesive story and enhances the overall interpretation of the data.

The black trail encourages students to think critically about data visualization while developing skills to create professional, publication-quality figures that clearly convey their original findings.

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 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.

IV. **Presentation content & Grading Rubric**: These will be provided to students in March and will also be posted on Brightspace and the course website.

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

a. NGS data type selection:

Acceptable	**Unacceptable**
RNA-Seq	Single-cell RNA-Seq
ChIP-Seq	Microarray
ATAC-Seq	Spatial Transcriptomic Dataset
Permission required: Research-specific dataset Metagenomics Variant analysis	

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 now
    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)	Mid Feb	Late Feb/ Early March		Early to Mid March

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.
- 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!

VIII. 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.