Assignment 3

Topics: Metagenomics, chromatin immunoprecipitation, and transcriptomics

Due date: 6 October 2025 at 11:59pm

Rules:

- You can work in group, but write your own answers
- You can use AI to help, but don't abuse it. Credit AI when used
- The objective of the assignment is to provide you with experience. Explain your work and observations. Don't just paste a screenshot of the result.
- You can contact me to ask for clarification

Credit: GPT-5 was used to aid the design of the assignment

Caution: Analyses in this assignment can take >30 minutes

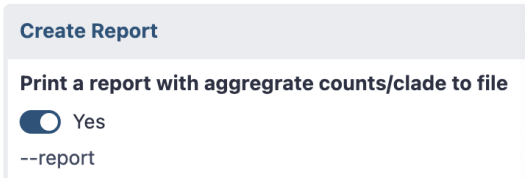
Part A. Hands-On Mini-Pipeline (Galaxy Platform)

We will again use the **Galaxy server** (https://usegalaxy.eu or https://usegalaxy.org) to analyze metagenomics data.

The data came from a 16S rRNA amplicon sequencing of gut microbiome of Thai population. Sequencing data in **FASTQ** format are available on **Sequence Read Archive** accession number **PRJDB5860**. To make the data more manageable, we are going to analyze only the following samples:

Samples from Bangkok	Samples from Ratchaburi
DRR095623	DRR095673
DRR095624	DRR095674
DRR095625	DRR095675

1. For each FASTQ file, run **Kraken2** to map taxonomy with the following parameter
    - Print a report with aggregate counts/clade to file



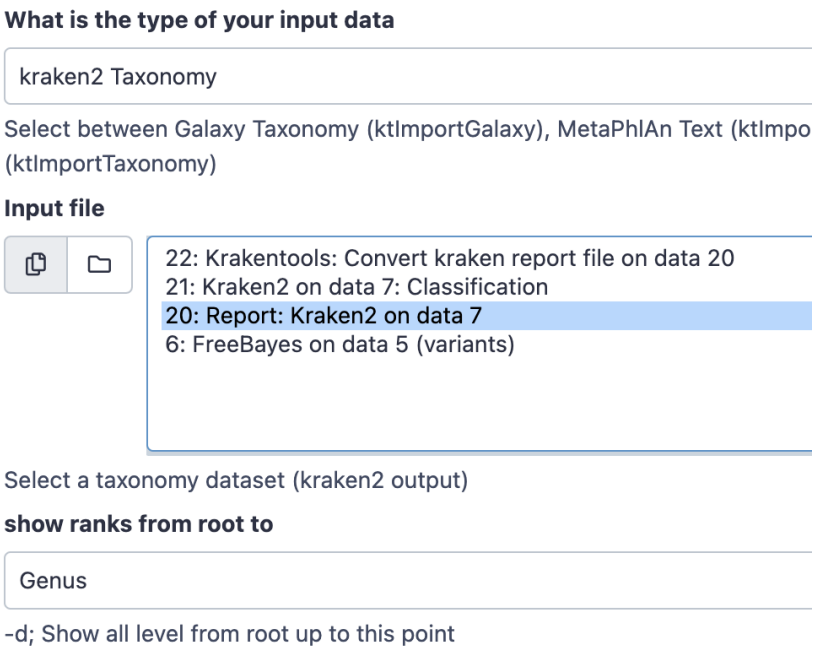
- Use **minikraken** database

2. Inspect the output from **Report on Kraken2**. There should be one tabular file (.tabular) per sample, which can be opened in MS Excel as shown below.

	A	B	C	D	E	F	G	H	I
1	1.24	115	115	U	0	unclassified			
2	98.76	9157	0	R	1	root			
3	98.76	9157	0	R1	131567	cellular organisms			
4	98.76	9157	87	D	2	Bacteria			
5	48.29	4477	5	P	1224	Proteobacteria			
6	48.08	4458	118	C	1236	Gammaproteobacteria			
7	46.14	4278	2991	O	91347	Enterobacterales			
8	13.43	1245	255	F	543	Enterobacteriaceae			
9	5.82	540	1	G	561	Escherichia			
10	5.78	536	500	S	562	Escherichia coli			

- Explain the information shown in columns A-F

3. Generate visualization of the taxonomic composition using **Krona pie chart**
   - Set **kraken2 Taxonomy** as the input and find your **Kraken2 Report** output
   - Select **Genus** for the depth of the visualization



4. Compare the visualization you get for samples from Bangkok and Ratchaburi
   - Is there any difference in gut microbiome compositions between the groups?
   - **Hint**: Krona pie chart is interactive, try clicking to explore the results

Part B. Literature Analysis (Metagenomics)

Study this landmark Human Microbiome study from 2012, https://www.nature.com/articles/nature11234, and answer the following questions:

- What was the experimental design of the study?
- What metagenomics techniques were used? What information did they provide?
- How was the variability of microbiome within an individual person evaluated?
- What were the key findings that impress you the most? Why?

Part C. Experimental Design Utilizing Transcriptomics and Epigenetics Analysis

Suppose you are studying **the changes in gene expression between healthy liver cells and liver cancer cells and identify epigenetics factor that regulate them**.

Propose an experimental design and computational analysis using techniques we have learned so far in this course to pinpoint differentially expressed genes (DEGs) and identify potential regulatory mechanisms that change their expression levels.

- Which omics techniques would you use? Why?
- Describe the data processing steps
- What analysis method(s) would you apply?
- How would you combine the results to pinpoint DEGs and epigenetics mechanisms that regulate them?

Part D. Critiquing LLM Responses

Do this after finishing Part C.

1. Feed the question from Part D into an LLM/AI of your choice:

"Design a sequencing-based experiment to identify genes whose expressions change between healthy liver cells and liver cancer cells as well as potential epigenetics mechanisms that regulate them. Justify the sequencing and computational analysis methods used. Do not include unrelated lab techniques."

2. What was the LLM's response?
3. Critique the response:
    - Which parts of the response are scientifically sound and useful?

- Which parts of the response are vague, inaccurate, or incomplete?
- Compare your own design from **Part C** with LLM's response.

4. Improving the prompt:
   - How would you improve the prompt to resolve the shortcomings in the LLM's response?
   - Try using your new prompt and comment on the results

Part E. Transcriptomic Differential Expression Analysis

We are going to identify differentially expressed genes across aerobic and anaerobic growth conditions of yeast (*S. cerevisiae*).

RNA-seq data is available at https://figshare.com/articles/dataset/Yeast_RNA-seq_data_and_transcriptome_for_kallisto-sleuth_demo_session/24182520. This dataset contains paired-end short-read data for 4 samples (8 files), as detailed below:

Condition	Replicate	Forward read	Reverse read
Aerobic growth	1	aerobic_r1_1.fq.gz	aerobic_r1_2.fq.gz
Aerobic growth	2	aerobic_r2_1.fq.gz	aerobic_r2_2.fq.gz
Anaerobic growth	1	anaerobic_r1_1.fq.gz	anaerobic_r1_2.fq.gz
Anaerobic growth	2	anaerobic_r2_1.fq.gz	anaerobic_r2_2.fq.gz

Study **Galaxy** tutorial at https://training.galaxyproject.org/training-material/topics/transcriptomics/tutorials/ref-based/tutorial.html and adapt it to analyze our yeast data.

For each tool, follow the parameter settings shown in the tutorial.

Tips: If you have difficulty passing the result from one analysis tool to the next tool, sometimes it might be quicker to export the result from **Galaxy**, do some formatting, and then upload back into **Galaxy**.

1. Run **FastQC** to check the quality of the sequencing data. Report what you found.
2. Run **Trimmomatic** to trim the sequencing reads (refer to guide from previous Assignment)
3. To run the alignment with **RNA STAR**, you will need to provide the reference genome (in FASTA format) and gene annotation (in GFF format). Obtain these files from

ENSEMBL's *S. cerevisiae* database version R64-1-1 (GCA_000146045.2): https://asia.ensembl.org/Saccharomyces_cerevisiae/Info/Index

   a. Look for the button to **Download DNA sequence (FASTA)**. Then select Saccharomyces_cerevisiae.R64-1-1.dna.toplevel.fa.gz
   b. Look for the button to **Download GFF**. Then select Saccharomyces_cerevisiae.R64-1-1.115.gff3.gz

4. Run **RNA STAR** to align reads in your samples to the reference genome
5. Run **featureCounts** to quantify read count per gene
6. Run **DESeq2** to perform differential expression analysis
   a. Attach graphics you receive from DESeq2 and write a caption for each graph to summarize the information shown and to discuss what you observed
   b. Summarize the differential expression result. How many genes significantly differ between the two growth conditions? What are the top up-regulated and down-regulated genes? Do their functions agree with the growth conditions or are they inconclusive?