Assignment 1

Topics: Sequence alignment and DNA sequencing

Due date: 3 September 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

Part A. Sequence Alignment

TP53 is one of the most studied human genes (https://www.nature.com/articles/d41586-017-07291-9). Let's use sequence alignment to find orthologs of this gene in other species and try to figure out when this gene emerged during evolution.

1. Retrieve human TP53 nucleotide sequence from **NCBI GenBank**.
2. Use **nucleotide BLAST** to identify homologous sequences in another species (choose one non-human primate and one non-mammalian species).
    - Report the top hits, percent identity, and alignment length.
    - Comment on how sequence similarity changes across species.
3. Translate your gene into its protein sequence (using ExPASy Translate or NCBI tools) and perform a **protein BLAST** search.
    - Compare protein-level vs DNA-level BLAST hits to the same species, which is more conserved.
4. Try BLASTX and tBLASTN.
    - Do you get any additional hits?

Part B. DNA Sequencing Technologies

Read the review article to learn more about DNA sequencing techniques:

- Heather JM, Chain B. *The sequence of sequencers: The history of sequencing DNA*. Genomics. 2016.

Below is a list of biological research questions. For each one, select the most suitable sequencing platform(s) and briefly justify your choice (2–3 sentences each).

You may combine multiple platforms and adjust the sequencing throughput and sequencing scope. There may be more than one correct answer:

- Detecting whether a newborn carries a single-gene mutation (e.g., in CFTR).
- Characterizing transcript isoforms and alternative splicing patterns in cancer.
- Studying taxonomic diversity of the human gut microbiome.
- Assembling the genome of a newly discovered plant species with no reference genome.
- Tracking tumor evolution by monitoring low-frequency mutations in circulating tumor DNA (liquid biopsy).
- Screening for unexpected structural genomic variants in a population.
- Detecting antimicrobial resistance genes in patients' blood and urine samples.

Part C. Applications of Sequencing as Molecular Assays

1. Find **a primary research article utilizing <u>at least two</u> sequencing-based assays** (e.g., not exome sequencing, ATAC-seq, ChIP-seq, RNA-seq, etc.).
2. Write a short summary (~300 words) about the technique:
    - What was the biological question?
    - How did the assays address the biological question?
    - What were the key findings?

Part D. Experimental Design utilizing Sequencing

Suppose you are interested in studying how **a specific mutation in a transcription factor affects gene regulation in cancer cells**.

Propose an experimental design using one or more sequencing-based assays.

- What assay(s) would you choose?
- What controls would you include?
- What outcomes would you expect, and how would they support/refute your hypothesis?

Part E. Critiquing LLM Responses

Do this after finishing Part D.

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

"Design an experiment to study how a specific mutation in a transcription factor affects gene regulation in cancer cells."

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 with your own design from Part D with LLM's response.
    - Reflect: What risks are there in relying solely on LLMs for scientific planning?