Assignment 4

Topics: Metagenomics, chromatin immunoprecipitation, and transcriptomics

Due date: 20 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

Part A. Hands-On Functional Enrichment Analysis

We will use **WebGestalt** (https://www.webgestalt.org/) to identify enriched biological functions and pathways in genes whose expression are affected in the temporal cortex of patients with major depressive disorder.

1. Obtain the list of affected genes by going to **MSigDB** (https://www.gsea-msigdb.org/gsea/msigdb/genesets.jsp)
    a. Look for C2 (curated gene sets): CGP (chemical and genetic perturbations)
    b. Within CGP, look for **ASTON_MAJOR_DEPRESSIVE_DISORDER_DN** (genes down-regulated in patients)
    c. What are the criteria for calling down-regulated genes in this set?
2. Given the gene list, describe how you would perform functional enrichment analysis in **WebGestalt**. Specifically, discuss whether it is a good idea to use Over-Representation, Gene Set Enrichment Analysis, or Network Topology-based method?
3. Let's perform Over-Representation analysis
    a. Use "Gene Ontology" (Biological Process, Cellular Component, and Molecular Function) to identify enriched terms
    b. Set the **Reference Set** to "Protein Coding Genes"
    c. Summarize your findings
4. Study the original paper Aston, C. *et al.* (https://www.nature.com/articles/4001565)
    a. Which transcriptomics platform was used?

    b. Given this knowledge, do you need to revise your Over-Representation analysis? Why and how?
    c. Revise your analysis in #3 as necessary
5. Compare your findings in #4 (or #3 if no re-analysis was necessary) to the results reported in the abstract of the original paper. Are they consistent?

Part B. Literature Analysis (Single-Cell Transcriptomics)

Study this recent paper from 2024, https://www.nature.com/articles/s41467-024-50478-8, **focusing on the results presented in Figure 1 and Figure 2**, and answer the following questions:

- How did the authors investigate tumor microenvironment using single-cell transcriptomics?
- How did the authors identify different cell types?
- Do the results shown in Figure 1b and 1c suggest that the cell types are clearly defined by transcriptomics data? Is there any potential ambiguity?
- How did the authors reconstruct cell type transition shown in Figure 2c?
    - What assumption or biological knowledge was used?
    - Which parts of the result were inferred computationally?

Part C. Experimental Design Utilizing Single-Cell Techniques

Suppose you are interested in the phenomenon of vasculogenic mimicry in cancer, where "some of the cancer cells gain endothelial-like properties and form vascular-like structures that connect to nearby blood vessel to acquire more nutrients and oxygen from the body" (https://pmc.ncbi.nlm.nih.gov/articles/PMC8071410/figure/cancers-13-01912-f002/).

If you can isolate these rare cell types and identify their unique molecular markers, you plan to engineer T-cell that can specially target these cells to both kill them and deliver other anti-tumor molecules to the vicinity of the tumor.

Design an experiment (combining single-cell and spatial techniques) that would allow you to identify these rare cell types. You may assume that you already have samples from tumor tissues of several patients that contain these cell types.

Hint: You may assume that these endothelial-like cancer cells must express some endothelial-specific gene markers, and that these endothelial-like cancer cells must be formed next to normal endothelial cells in the tumor tissue structure.

- How would you analyze single-cell transcriptomics data to identify these cells and subsequently their molecular markers?
- How would you use spatial techniques to assist the discovery of these cells?
- How would you use spatial techniques to validate the identified markers?

Part D. Using LLM to Aid Research Design

Do this after finishing Part C.

1. Ask an LLM/AI of your choice to help design the experiment in Part D. You are free to optimize your own prompts. Show your final prompt here.
2. Comment on your process of optimizing the prompt to obtain good LLM responses.
3. Compare your own design from **Part C** with LLM's response.