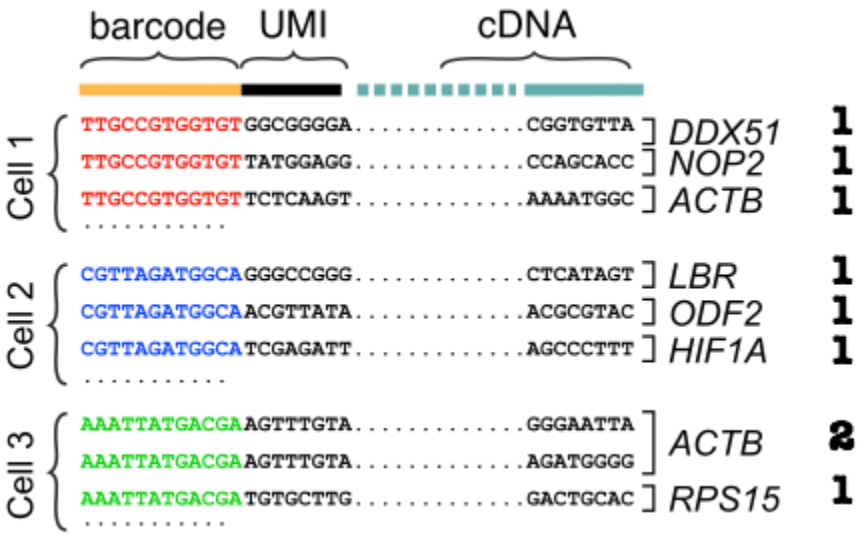
Problem set 5

Single-cell transcriptomics

In single-cell sequencing, in addition to the cDNA sequences, each read will also contain a barcode and an UMI portion at the beginning, as shown below.



Q1: Explain what barcode and UMI are and how they are interpreted.

In lecture and demo sessions, we learned that there are multiple quality filters that need to be applied to remove low-quality and unwanted barcodes.

Q2: Study https://www.nature.com/articles/s41588-022-01100-4. What kind of filters were used to select high-quality cells and samples? *Hint: The information may span across multiple paragraphs.*

Q3: During the demo, after we project the data onto 2D scatter plots (using either PCA or UMAP), why did we color the cells using some genes (CST3, NKG7, and PPBP)? How do these colorizations help us understand the data?

Q4: Pick one data integration or batch effect removal method (linear effect, mutual nearest neighbor, canonical correlation analysis) and explain how it works.

Q5: Explain how single-cell ATAC-seq data can be integrated with single-cell RNA-seq data.

The concept of RNA velocity (https://www.nature.com/articles/s41586-018-0414-6) helps us determine the ordering of cells during development. One key aspect is the comparison of the expression levels of unspliced and spliced forms of a gene. **Figure 1d** shows the expected trend of unspliced/spliced expression as the gene is up-regulated or down-regulated.

Q6: Explain the mechanisms that gave rise to the pattern of the graph in **Figure 1d**.

Q7: What are the differences between single-nucleus RNA-seq compared to single-cell RNA-seq? What are the pros and cons of each method?

Proteomics

Explore the proteomics dataset **PXD000674** deposited on PRIDE repository.

Q8: Fill in the following protocol information.

Is this a top-down or bottom-up proteomics experiment?	
If it is a bottom-up experiment, which enzyme was used to digest the proteins? If not, answer N/A.	
Is there a step to prevent Cysteine from forming di-sulfide bonds? If yes, what is the modification of Cysteine?	
How long was the liquid chromatography run?	
Which mass spectrometer was used? Which mass analyzer was used?	
Ware the mass spectra data acquired in data-independent or data-dependent mode?	
Would a large peptide with charge state of +6 be present in the data?	
Is this a label-free or labeled experiment?	

Q9: If you want to analyze this dataset using MaxQuant, which parameters would you set/change? Explain your decisions.

Q10: What are the pros and cons of top-down and bottom-up proteomics? When would you use one or the other technique?

Q11: Explain how peptide sequencing with mass spectrometry works.