Bi/BE/CS 183 2021-2022
Instructor: Lior Pachter
TAs: Tara Chari, Meichen Fang, Zitong (Jerry) Wang

Problem Set 10 - Final

You are not allowed to collaborate with others for this Final, though you may ask clarification questions on Piazza or at Office Hours. This Final is open note, which includes Lecture slides, your personal notes, and any resources provided in the notebook.

Submit your solutions as a single PDF file via Canvas by **5pm Wednesday March 16th**.

- If writing up problems by hand, please use a pen and not a pencil, as it is difficult to read scanned submission of pencil work. Typed solutions are preferred.
- For problems that require coding, Colab notebooks will be provided. Please copy and save the shared notebook and edit your own copy, which you should then submit by including a clickable link in your submitted homework. Prior to submission make sure that you code runs from beginning to end without any error reports.

Problem 1 (100 points)

In this problem you will process a single-cell dataset from the raw fastqs (sequencing reads of the cDNA library) to produce the cell x gene count matrix we usually work with.

Given this count matrix you will additionally investigate the impact of various normalization techniques and dimensionality reduction on clustering of the cells (i.e. looking for cell types). With the metadata for this dataset, the cell types and ages of the mice used, you will then compare the efficacy of logistic regression versus neural network based techniques for classifying the age of the mouse a cell came from. Together this will take you through common tasks performed on single-cell datasets and the effects of common data processing choices.

The Colab notebook is here.

Your edited version of the notebook *must be submitted* for this problem. Reminder to check that your notebook runs all the way through with the the `Runtime` → `Restart` and `Runtime` → `Run All` commands.