# Project Proposal: Uncovering Cellular Heterogeneity in Single-Cell RNA-Seq Data

Team Members: Alice Smith, Bob Johnson, Charlie Lee

Prepared by: Shiyu Wang (Andrew ID: shiyuwa2)

## Introduction

Understanding cellular heterogeneity is critical for revealing mechanisms underlying tissue development and disease progression. In this project, we propose to analyze a publicly available single-cell RNA-seq dataset—such as the Tabula Muris dataset—to investigate distinct cell populations and their gene expression patterns. Our analysis will help uncover latent structures within the data and ultimately facilitate the identification of novel cell subtypes.

## Related Works

Previous studies have successfully leveraged unsupervised learning to extract meaningful clusters from complex single-cell data. For example, the study by *A single-cell transcriptomic atlas characterizes ageing tissues in the mouse* (Nature, 2020) demonstrated the utility of clustering techniques in distinguishing cell types. Similarly, research on dimensionality reduction methods, such as Principal Component Analysis (PCA), has provided insights into the variability inherent in high-dimensional transcriptomic data (e.g., L. van der Maaten & G. Hinton, *Visualizing Data*, Journal of Machine Learning Research, 2008). These works motivate our combined approach of implementing both custom and standard machine learning methods to analyze single-cell gene expression profiles.

## Overview of Proposed Methods

To provide a comprehensive analysis of the dataset, we plan to implement and compare three distinct machine learning methods addressing different analytical tasks:

1. **Dimensionality Reduction (Unsupervised):** We will implement Principal Component Analysis (PCA) from scratch to reduce the high-dimensional gene expression data. This will allow us to visualize the primary sources of variance and assess reconstruction error as a metric for performance.

2. **Clustering (Unsupervised):** We plan to implement a custom k-means clustering algorithm. This method will be applied to the reduced data from PCA to identify distinct clusters corresponding to putative cell types. The quality of clustering will be evaluated using metrics such as the silhouette score and cluster purity (if cell type labels are partially available).

3. **Classification (Supervised):** To further validate our clustering findings, we will use an off-the-shelf package (e.g., scikit-learn) to implement a supervised learning method (e.g., logistic regression or Random Forest classifier). This classifier will predict known cell types using the gene expression data. Performance will be measured using standard metrics such as the F1 score, precision, and recall, with cross-validation to tune parameters.

# Evaluation Strategy

For each method, we will carefully document the parameter tuning process. In PCA, we will compare reconstruction errors across different numbers of principal components. For k-means clustering, the silhouette score will help determine the optimal number of clusters, and we will perform sensitivity analysis on initialization parameters. Finally, for classification, we will utilize k-fold cross-validation to optimize model hyperparameters and report metrics such as F1 score and confusion matrices to gauge prediction accuracy.

# Conclusion

This proposal outlines our approach to dissecting a complex single-cell RNA-seq dataset using both custom implementations and standard tools. By combining unsupervised and supervised learning tasks, we aim to provide a robust analysis of cellular heterogeneity and set the stage for future biological investigations.