Assignment Report

Advaith Kannan

210072

Introduction

In this assignment, the goal was to perform clustering on a given dataset using the Leiden clustering method, with different approaches to feature selection. The analysis was conducted in three main tasks:

- Task 1.1: Clustering with highly variable genes using ScanPy
- Task 1.2: Clustering with deviance-based feature selection.
- Task 3: Finding strategies to improve the produced clusters.

Task 1.1: Clustering with Highly Variable Genes

Objective:

- Perform clustering using the Leiden method.
- Preprocess the data with QC, normalization, feature selection (top 2000 highly variable genes), and PCA.
- Vary the resolution parameter to find the optimal clustering.
- Perform differential expression analysis to identify marker genes.
- Annotate clusters based on marker genes.

Approach:

1. Data Preprocessing

- Quality Control (QC): Removed low-quality cells and genes.
- Normalization: Applied normalization to ensure comparability across cells.
- Feature Selection: Selected the top 5000 highly variable genes to focus on the most informative features.
- Principal Component Analysis (PCA): Reduced dimensionality to capture the main variance in the data.

2. Clustering:

- Applied Leiden clustering with various resolution parameters.
- The quality of the clusters was assessed to produce the best results. (Hyperparameter tuning)

3. Differential Expression Analysis:

- Conducted differential expression analysis to identify marker genes for each cluster.

4. Manual Annotation

- The differentially expressed genes were searched on PanglaoDB. While some of the clusters were easily annotated (like dendritic cells and T-Cells), various others were rather ambiguous. Due to this ambiguity, these annotations have been omitted in the notebook submitted.

Results:

- The clustering results with highly variable genes were evaluated and found to be robust, providing meaningful biological insights.
- However, the Leiden Algorithm failed to cluster the various rare cell types visible.

Task 1.2: Clustering with Deviance-Based Feature Selection

Objective:

- Perform clustering using deviance-based feature selection.
- Compare clustering results with those obtained from Task 1.1.

Approach:

1. Feature Selection:

- Used deviance-based feature selection via the `pipeComp` package to identify relevant features.
- This method was applied to select features based on deviance metrics.

2. Clustering:

- Applied Leiden clustering to the dataset using the features selected through deviance-based methods.
- Compared the clustering results with those obtained in Task 1.1.

Results:

- The clustering results obtained with deviance-based feature selection did not match the quality of those from Task 1.1.
- The clusters were less distinct, making manual annotation a lot more ambiguous

Conclusion:

- The deviance-based feature selection did not perform as well as the highly variable gene-based feature selection in terms of clustering quality. The highly variable genes approach provided better cluster separation and biological interpretability.

Task 3: Improving the Clusters Formed

The objective of this task was to further improve upon the clustering step. While favourable results were not obtained, the following attempts were made to improve the clusters formed –

1. Using Individual Component Analysis (ICA) instead of Principal Component Analysis (PCA).
2. Using “phenotyping by accelerated refined community-partitioning” or PARC. (However, this was computationally expensive and impractical to run locally).

There were various other approaches ideated. These are some notable approaches which may be pursued if more time were allowed to accomplish this task –

1. Using multiple steps of clustering: In this scenario, first, Leiden clustering may be performed at a low resolution on the given dataset. Following this step, the under-clustered regions may be clustered even further to produce apt outputs. This may allow better clustering of rare cell types.
2. Making use of algorithms like scSID (https://doi.org/10.1016/j.csbj.2023.12.043) which identify rare cell types remarkably well. This too may be employed in an iterative manner after one round of Leiden Clustering.

Summary

Clusters were successfully formed using the Leiden clustering algorithm using ScanPy. The effects of deviance-based cell-clustering were compared to that produced using highly variable genes. Multiple approaches were attempted to produce improved clusters.

Feature Plots Obtained

Task – 1.1: