Evaluating Performance of Clustering Algorithms using Single-Cell RNA-Seq Data

Sarah Baalbaki, Sumitra Lele, Sarah Oladejo
Carnegie Mellon University- Computational Biology Department

Abstract

Efficient computational analyses of scRNA-seq data are required to extract inherent biological information from the huge quantity of information that single-cell RNA-Seq technology presents.

The integration of single-cell RNA-seq data with other machine learning algorithms has facilitated these methods of analysis.

Clustering is an upstream step that affects other downstream analysis, and it is therefore imperative to obtain accurate cluster assignment and number of cell types from scRNA-Seq data.

We plan to compare algorithms that Deep learning to assign the number of cell types.

Dataset

Our dataset is composed of scRNA-seq data from Smart-seq2 sequencing of FACS sorted cells of 20 organs from 7 mice.

The data contains gene expression information for individual cells within the given cell population.

We choose 5 tissue samples form this dataset, related to immunity:

- Spleen
- Thymus
- Marrow
- Brain Neurons
- Brain Microglia

We will cluster the genes to correctly identify different cell populations.

Data Preparation

Data Integration:

- Barcode for each cell type in annotations
- Integrated gene counts with cell annotations based on the barcodes
- Combined data across 5 tissues based on the genes

Data Preprocessing:

- Performed basic QC to remove mitochondrial genes
- Filtered cells with <200 genes
- We removed genes expressed in less than 3 cells

Data Normalization:

- Normalized based on log normalization

Algorithms

- SEURAT
- DESC

Methodology

DESC:

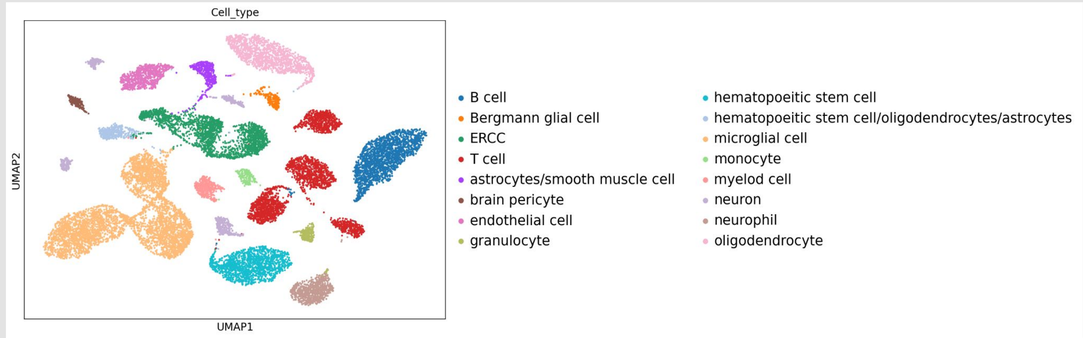
- implute counts dataset
- train the dataset
- predict cell types
- annotate based on expression of marker genes

SEURAT:

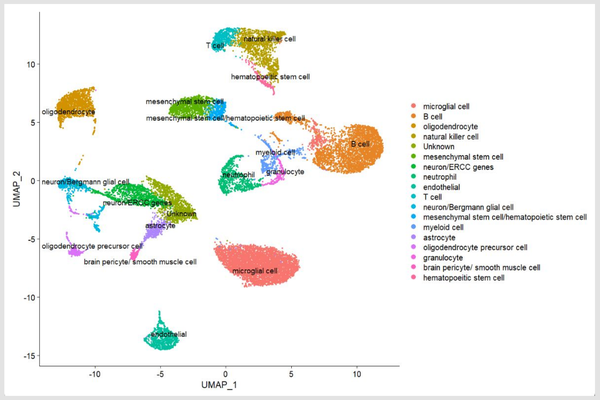
- find variable features
- scale the data
- run PCA
- find neighbors and find clusters
- plot the data in low dimensional space with UMAP

Results

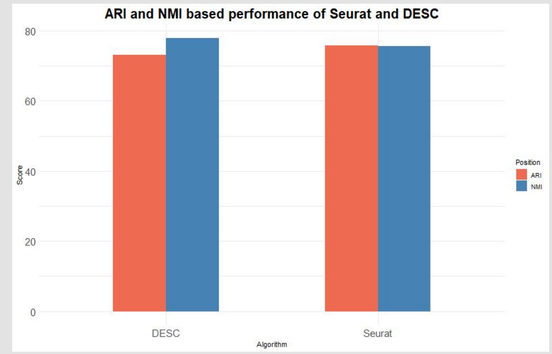
DESC:



SEURAT:



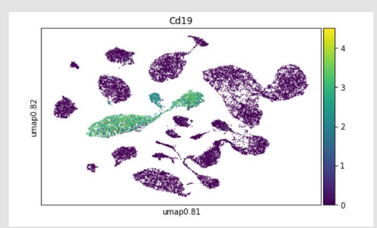
COMPARISONS:



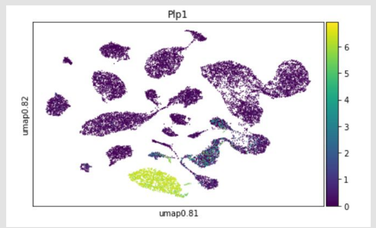
Benchmarking Seurat and DESC using rand index and normalized mutual information

MARKER GENES:

Cd19: B-Cells



Plp1: Hematopoietic stem cell



Conclusion and Future Work

- From our analysis, we see a comparable performance between DESC and SEURAT with both achieving a accuracy greater than 70%.
- However, both algorithms differ in their ability to delineate similar cell types such as stem cells (hematopoietic and mesenchymal) and immune cells. For example, DESC was able to cluster the ERCC-spike in transcripts more distinctly than SEURAT
- Benchmarking more algorithms and including more datasets will be crucial to draw a conclusion on the strengths and weaknesses of these clustering algorithms