Project 1 of Appied Data Science

Zihan Tang

October 2024

Abstract

This project implements data science methods to explore single-cell RNA sequencing (scRNA-seq) data. We apply clustering analysis and marker gene selection techniques to identify the regeneration-organizing cell (ROC) and the genes associated with this cell type. Different clustering methods and metrics are used to compare the results and visualize the data. Our findings indicate that Louvain and Leiden clustering outperform k-means. Moreover, t-test and Wilcoxon methods for identifying marker genes closely match the genes presented in the original paper, while the logistic regression method shows no matches.

1 Goal

In the paper "Identification of a Regeneration-Organizing Cell in the Xenopus Tail," the authors discovered the regeneration-organizing cell (ROC) by comparing regeneration-competent and regeneration-incompetent tadpoles.

This project aims to replicate the data exploration and analysis on scRNA-seq data to identify ROCs and determine the genes that differentiate these cells from others using marker gene methods. Finally, we compare the results obtained from different methods with those reported in the original paper.

2 Methods

2.1 Data Processing

The scRNA-seq data was processed using the following steps:

- Overview: Initial framework and inspection of the dataset.
- Filtering: Identification and retention of highly variable genes.
- Normalization and Log Transformation: Normalizing and log-transforming the data for downstream analysis.

2.2 Clustering Analysis

After preparing the data, we applied the following clustering methods to identify ROCs. The clusters were visualized using UMAP, with different clustering results shown in distinct colors.

- PCA + Louvain clustering
- PCA + Leiden clustering
- PCA + k-means clustering

2.2.1 PCA Analysis

We reduced the dimensionality of the data by selecting the top 31 principal components for further analysis.

2.2.2 Clustering Analysis

We used multiple clustering resolutions for Leiden and Louvain clustering methods and evaluated the clustering performance using metrics such as the Rand Index, Adjusted Rand Index (ARI), and Silhouette score.

2.3 Marker Selection and Gene Analysis

Clusters with a resolution of 0.5 were selected for marker gene identification. The following three marker selection methods were employed:

- Logistic regression
- T-test
- Wilcoxon test

3 Results

3.1 Data Processing

The single-cell RNA sequencing data used in this project was sourced from Xenopus laevis tadpoles at various stages following tail amputation. The dataset includes both regeneration-competent and regeneration-incompetent samples. After filtering, we retained 13,199 cells and 31,535 genes in a count matrix.

We first selected the data from timepoint 0 (the day of amputation) and calculated the mean and variance of each gene to obtain an overview of gene expression. Using the Fano factor as a metric, we filtered highly variable genes (those above the 65th percentile). After this filtering step, we retained 5,302 cells and 6,905 genes, which were then normalized and log-transformed for analysis. UMAP was implemented for visualization.

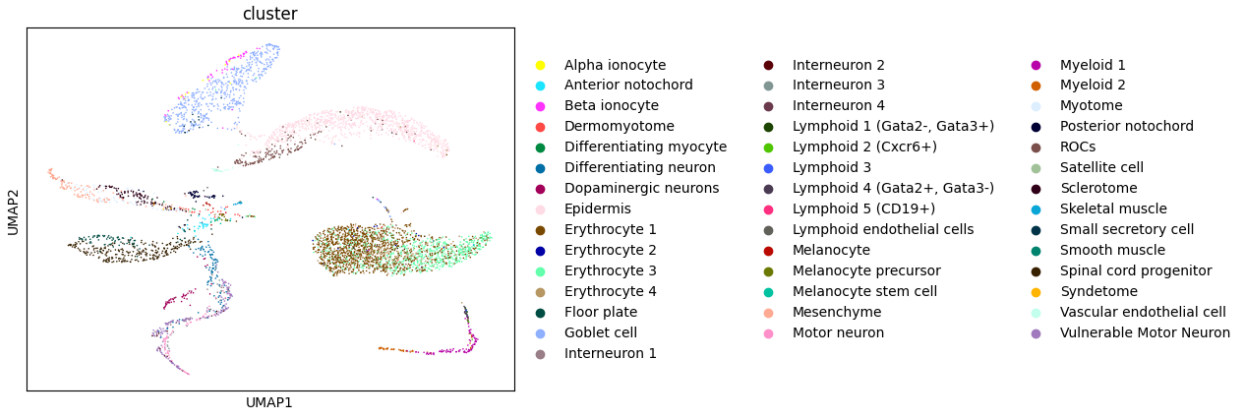

Figure 1: UMAP visualization of cell clusters identified in Xenopus tail regeneration.

3.2 Clustering Analysis

The PCA plot showed that different cell types overlapped, making it difficult to clearly separate ROCs from other cell types.

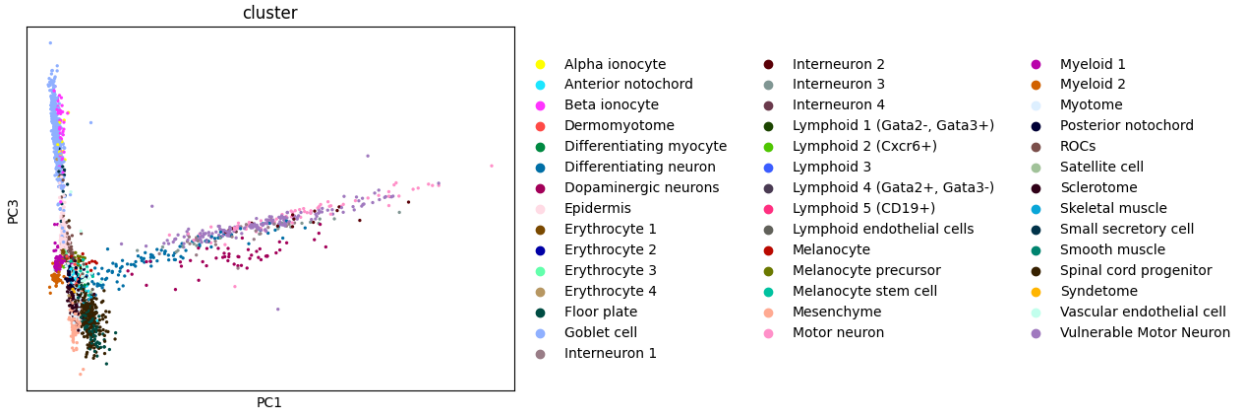

Figure 2: PCA visualization of cell clusters in Xenopus tail regeneration.

We found that the leidon and louvain clustering results are similar(in Figure 3), and much better than the kmeans. This is because when we use kmeans on the original data, the computation is so high; when we use kmeans on the principal component, it couldn't get ROCs correctly.

3.3 Gene Analysis

- Logistic regression: No marker genes matched those reported in the original paper.
- T-test: 32 marker genes matched those identified in the paper.
- Wilcoxon test: 32 marker genes matched those identified in the paper.

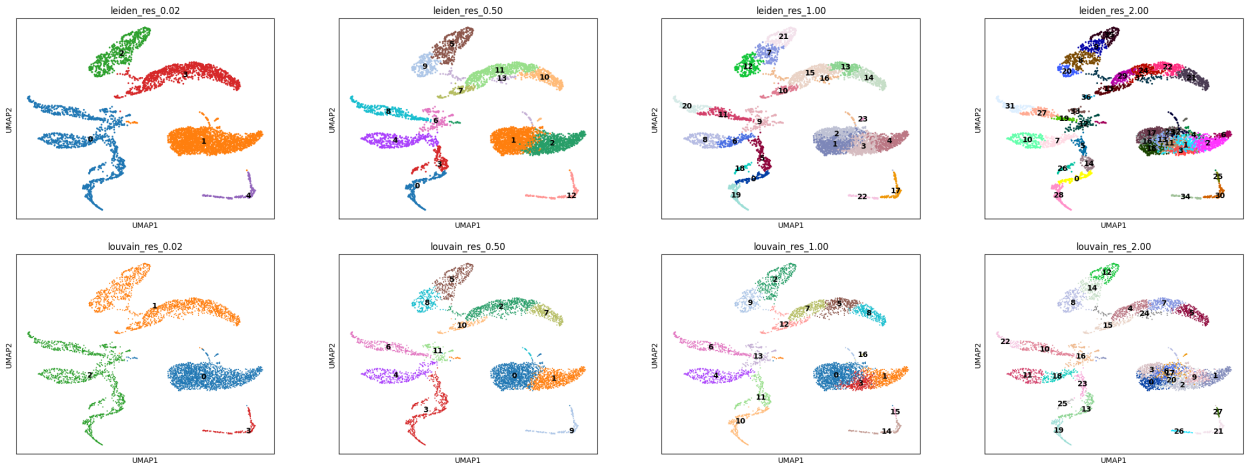

Figure 3: UMAP visualization of cell clusters from Leiden (top) and Louvain (bottom) clustering.

4 Conclusion

This project demonstrated that ROCs can be accurately identified using clustering methods and marker gene selection. Both Leiden and Louvain clustering performed well, while k-means struggled to identify ROCs. The t-test and Wilcoxon methods produced marker genes that closely matched those in the original paper, while logistic regression yielded no matches.

Future directions for exploration include:

- Testing alternative methods for selecting highly variable genes before clustering.
- Analyzing the resolution parameter in Leiden and Louvain clustering to optimize results.
- Refining marker gene selection to further distinguish ROCs from other cell types.

5 Code Availability

The code used for this analysis is available on a public GitHub repository: `https://github.com/Zihan-Hazel/Applied-Data-Science/tree/main`