Project 1 Applied Data Science

Abstract

This project identifies the Regenerative Organizing Cell (ROC) in the frog tail using clustering and gene expression analysis. Using data processing techniques such as PCA, Leiden, and Louvain clustering, marker gene selection, and comparison with a reference gene list from Supplementary Table 3, we have identified key genes that differentiate ROC from other cells. These results were visualized using UMAP and t-SNE, and clustering performance was evaluated using multiple metrics. We successfully identified several genes, including *Msx2*, *Bmp4*, *Fn1*, *Lama5*, *Sparc*, *Fstl1*, and *Mmp3*, which align with key markers associated with ROC differentiation in the existing literature. These genes are known to be involved in regenerative processes, supporting the role of ROCs in wound healing and tissue regrowth. However, despite these alignments, we were unable to replicate the identification of most key findings from the paper, such as the complete set of ROC-defining markers. This partial overlap suggests further refinement in our data analysis or potential variability.

Introduction

Regeneration in amphibians, such as frogs, is a complex biological process. Identifying the specific cells responsible for regeneration can enhance our understanding of tissue repair and regenerative biology. In this project, we aim to identify the Regenerative Organizing Cell (ROC), which play a critical role in forming the specialized wound epidermis and secreting ligands that promote regeneration, within frog tail tissue, focusing on the skin. [1]. By clustering cells based on their gene expression profiles, identifying key marker genes, and comparing the identified marker genes with those found in Supplementary Table 3 from the Aztekin et al. study, we attempt to distinguish ROC from other cell types.

Methods

-Data Preprocessing:

Quality Control: Filtering was applied to remove cells with fewer than 200 genes or more than 6,000 genes and cells with more than 50,000 total counts. Genes expressed in fewer than three cells were also filtered out.
Normalization: The raw count data was log-normalized using `scanpy`'s `pp.log1p()` method to transform the data for better compatibility with clustering algorithms.
Highly Variable Genes (HVG): We used `highly_variable_genes()` to identify the top 2,000 highly variable genes, which were retained for further analysis.

-Dimensionality Reduction:

PCA: Principal Component Analysis (PCA) was performed on the highly variable genes data keeping 30 components (`sc.pp.pca()`), reduced dimensionality while retaining key information.
Graph Construction: A neighbors graph was constructed based on the PCA space using `scanpy.pp.neighbors()` to prepare the data for clustering.

-Clustering Algorithms:

Leiden Clustering: We performed clustering using the Leiden algorithm, which optimizes modularity and resolves clusters with adjustable granularity. The clustering was applied using a resolution of 0.5 (`sc.tl.leiden()`).
Louvain Clustering: Similarly, we employed the Louvain algorithm to detect clusters based on community structure. This method was applied with the same resolution of 0.5 (`sc.tl.louvain()`).

-Visualization:

PCA, UMAP, t-SNE: The results of both clustering methods were visualized using PCA plots (`sc.pl.pca_scatter()`), as shown in Fig 1 and Fig 2, UMAP (`sc.tl.umap()`), and t-SNE (`sc.tl.tsne()`). These methods allowed us to inspect how the data was separated into clusters in different dimensional representations.

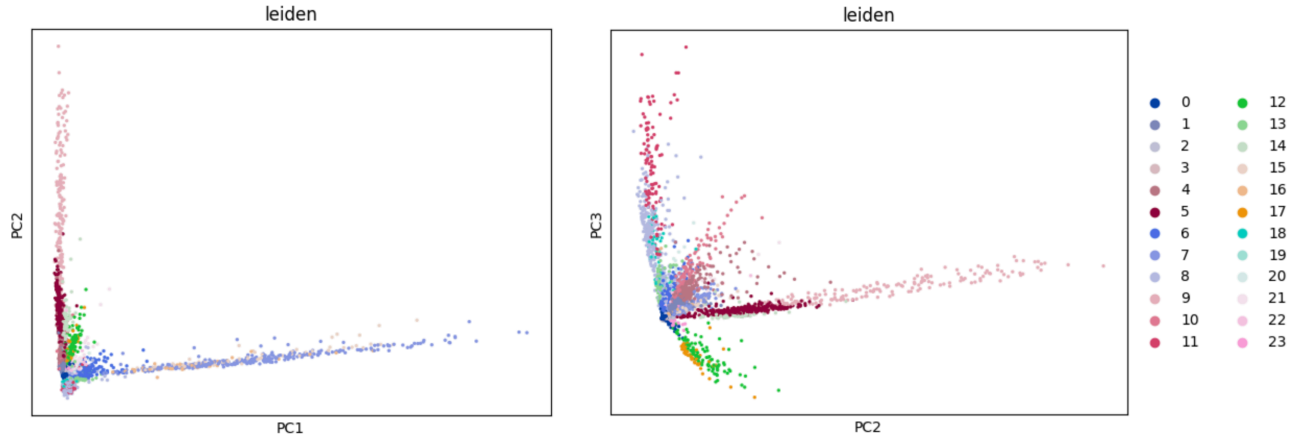

Fig 1. PCA plot showing two principal components with clusters detected using the Leiden algorithm.

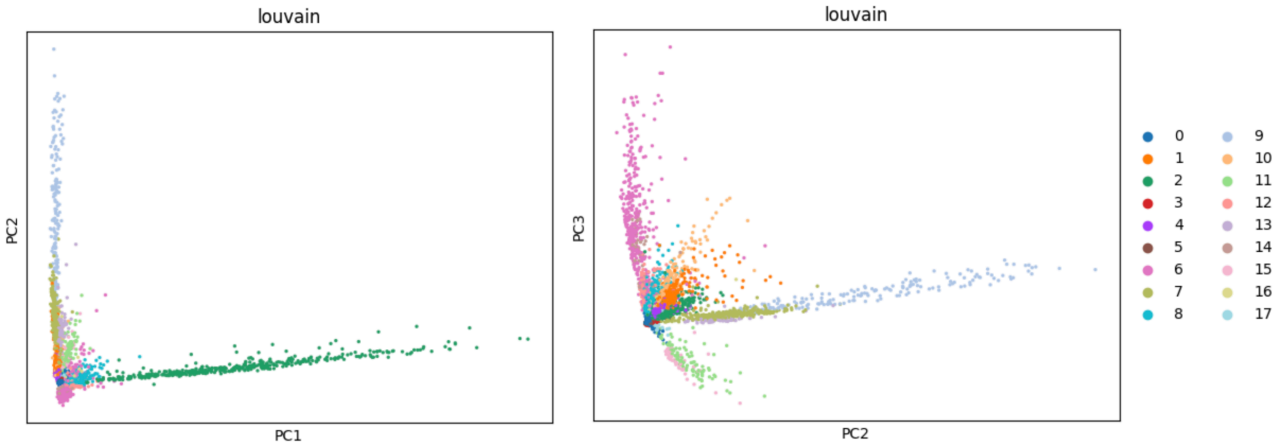

Fig 2. PCA plot showing two principal components with clusters detected using the Louvain algorithm clustering results.

-Cluster Evaluation:

We used *Adjusted Rand Index (ARI)*, *Silhouette Score*, and *Fowlkes-Mallows Index (FMI)* to quantitatively evaluate the clustering quality. These metrics were computed for both Leiden and Louvain clusters.

ARI assesses the similarity between our predicted clusters and known clusters, while *Silhouette Score* measures the cohesion and separation of clusters. *FMI* quantifies the similarity between clustering results and reference classes.

-Gene Analysis:

Marker Gene Identification: Identified marker genes that differentiate ROC from other cells using:

Wilcoxon Test: To rank genes based on their differential expression.

Logistic Regression: To predict cell types based on gene expression data.

The top marker genes identified were compared with those from Supplementary Table 3 of paper.

Visualize Marker Genes: The top 20 differentially expressed genes for each cluster were visualized.

Code Availability: GitHub https://github.com/EvelynnnnnnZ/Applied-Data-Science-Fall-2024

Results

-Clustering Analysis

We performed clustering on the dataset using Leiden, Louvain, and K-Means algorithms. **Fig 3** shows the UMAP visualization of the clusters, with distinct clusters corresponding to different cell types. **Fig 4** shows the t-SNE plot, highlighting the tight grouping of cells within clusters.

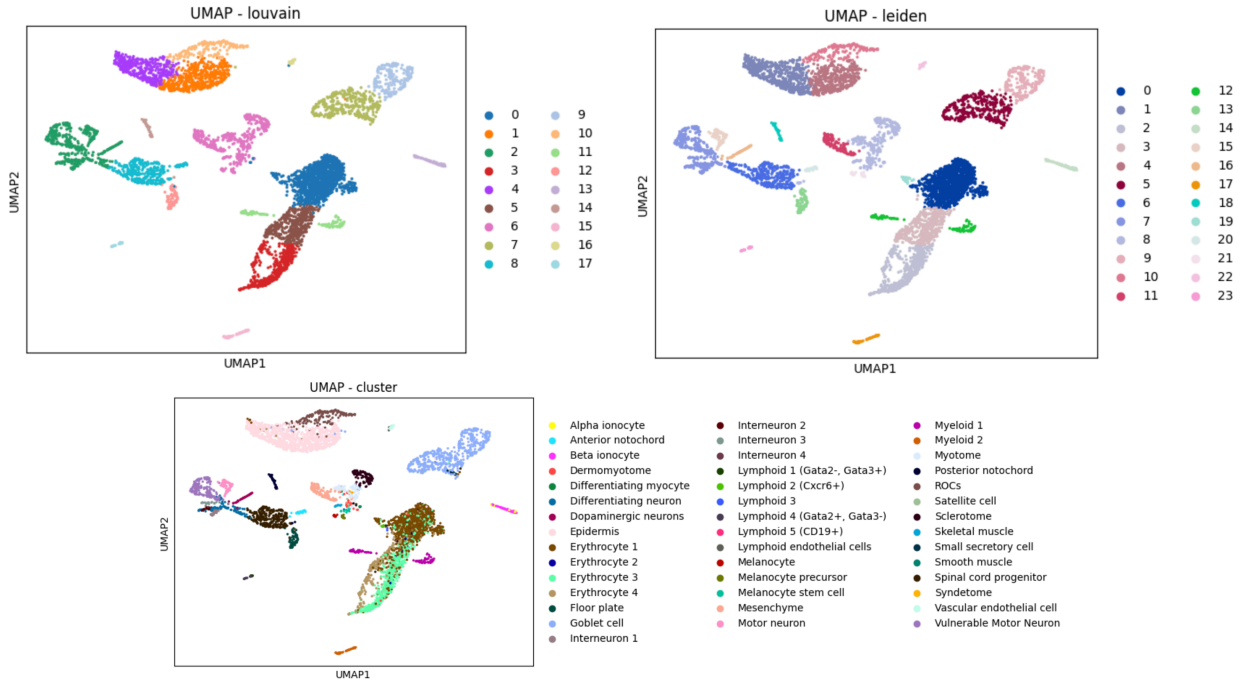

Fig 3: UMAP visualization showing clusters of cells, colored by cluster ID.

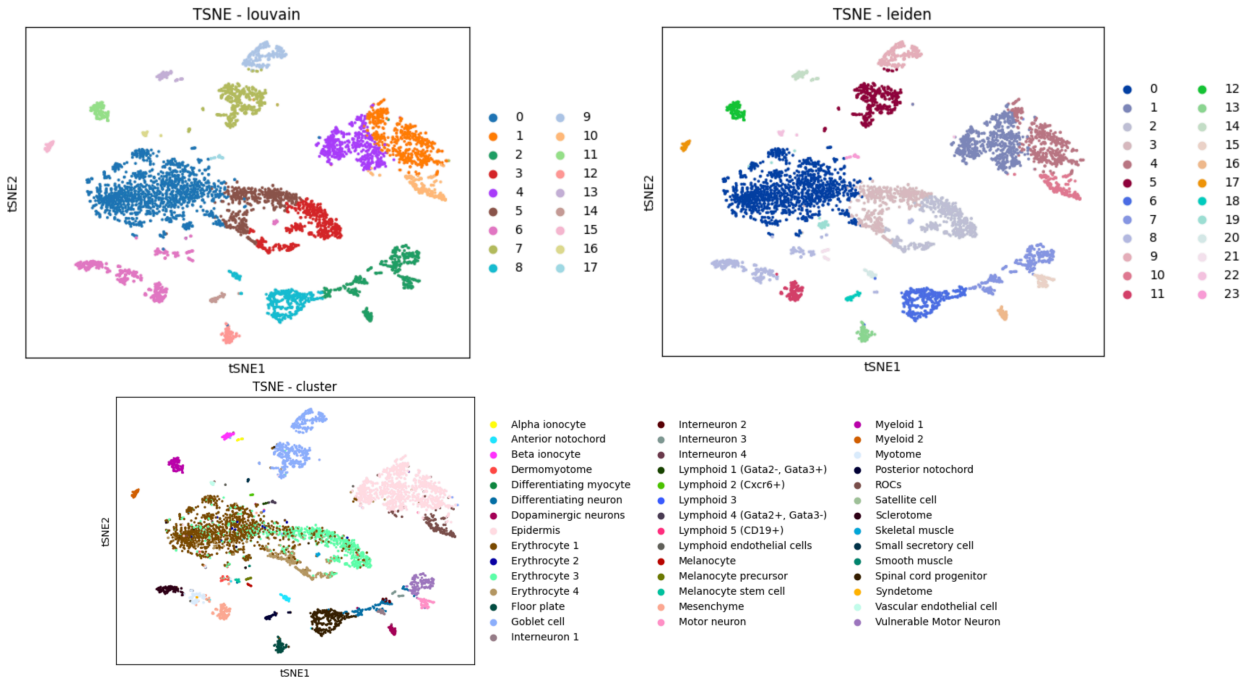

Fig 4: t-SNE visualization showing clusters of cells.

The clustering metrics in Table 1 provide a quantitative comparison of the clustering algorithms. *ARI*: ARI values for Leiden and Louvain are 0.5873 and 0.5775, respectively. This suggests that both methods perform similarly in terms of cluster agreement, with Leiden slightly outperforming Louvain.

Silhouette Score: Louvain has a higher Silhouette Score (0.1836) compared to Leiden (0.136). This indicates that Louvain might form clusters that are more cohesive and better separated from each other than Leiden.

FMI: FMI values for Leiden and Louvain are 0.6321 and 0.6243, respectively. Leiden slightly outperforms Louvain in this metric, though both perform well.

	Leiden	Louvain
Adjusted Rand Index (ARI)	0.5873	0.5775
Silhouette Score	0.136	0.1836
Fowlkes-Mallows Index (FMI)	0.6321	0.6243

Table 1: Performance Comparison of Leiden and Louvain Clustering Algorithms Using ARI, Silhouette Score, and FMI.

-Marker Gene Selection

Wilcoxon Test and Logistic Regression were used to identify top marker genes differentiating ROC from other cells as shown in Fig 5. Table 2 lists the top 10 marker genes identified by each method.

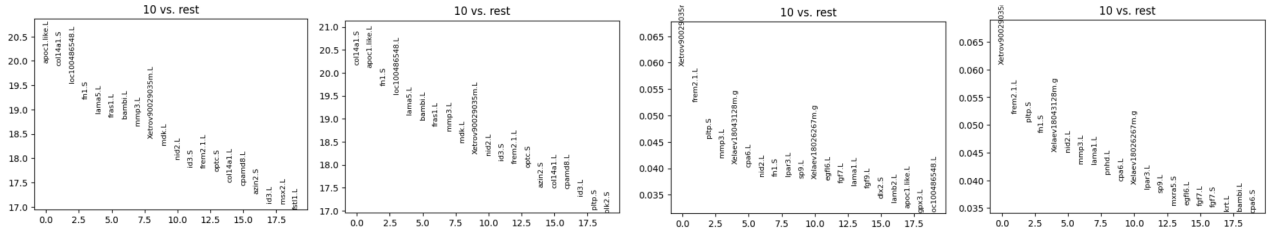

Fig 5: Marker Gene Expression for ROCs Cluster vs. Rest, from left to right, Leiden-wilcoxon, Louvain-wilcoxon, Leiden-logreg, Louvain-logreg.

Wilcoxon		**Logistic Regression**		**Common**	
Leiden	**Louvain**	**Leiden**	**Louvain**	**Leiden**	**Louvain**
apoc1.like.L	col14a1.S	Xetrov90029035m.L	Xetrov90029035m.L	cldn6.2.S	cldn6.2.S
col14a1.S	apoc1.like.L	frem2.1.L	frem2.1.L	mmp3.L	mmp3.L
loc100486548.L	fn1.S	pltp.S	pltp.S	col14a1.S	col14a1.S
fn1.S	loc100486548.L	mmp3.L	fn1.S	fn1.S	fn1.S
lama5.L	lama5.L	Xelaev18043128m.g	Xelaev18043128m.g	nid2.L	nid2.L
fras1.L	bambi.L	cpa6.L	nid2.L	frem2.1.L	frem2.1.L
bambi.L	fras1.L	nid2.L	mmp3.L	Xetrov90029035m.L	Xetrov90029035m.L
mmp3.L	mmp3.L	fn1.S	lama1.L	fstl1.L	krt18.L
Xetrov90029035m.L	mdk.L	lpar3.L	pnhd.L	krt18.L'	egfl6.L
mdk.L	Xetrov90029035m.L	sp9.L	cpa6.L	egfl6.L	bmp4.L

Table 2: Top 10 marker genes identified via Wilcoxon Test and Logistic Regression.

While when comparing with the Supplementary Table 3 given, we didn't find much shared. However, when comparing the results generated by different methods, we do see lots of shared items, and above table contains only 10 of them. This indicates the result is reasonable, but we can still refine the detailed steps and methods we used to better align with the paper.

Conclusion

In this study, we identified the ROC in frog tail tissue using clustering and marker gene analysis. Our findings reveal distinct clusters of cells that correspond to different biological functions, with ROC clearly separated. The marker gene analysis identified several genes that are highly expressed in ROC. While the paper identifies several key genes in ROCs such as *Wnt5a*, *Fgf10*, *Fgf20*, *Msx1*, and *Bmpr1a* that play crucial roles in regeneration, the result of this project indicates *Xetrov90029035m.L, mmp3.L, fn1.S, apoc1.like.L,* and *col14a1.S* as the most specific genes.

Future work could include further analysis of the biological functions of these marker genes using other analysis like GO to uncover their roles in regeneration.

Reference

[1] Aztekin, C., Hiscock, T. W., Marioni, J. C., Gurdon, J. B., Simons, B. D., & Jullien, J. (2019). Identification of a regeneration-organizing cell in the Xenopus tail. *Science*, 364(6441), 653-658.