## Abstract

In this study, i conducted comprehensive clustering analysis and gene marker identification on gene expression data. First, i performed data preprocessing, including data cleaning, removal of missing values, and filtering out inactive genes to ensure data quality and analytical accuracy. Subsequently, i selected 1000 highly variable genes and applied various normalization techniques, including log transformation and scaling, to optimize the dataset for further analysis.

I employed multiple clustering algorithms for analysis, including *PCA + Louvain* and *PCA + Leiden* methods (see Figure 1). To visualize the clustering results, we used UMAP and t-SNE techniques to reduce the dimensionality of the high-dimensional data and generated graphical representations of the clustering outcomes (see Figure 2). To evaluate the performance of different clustering algorithms, we calculated various clustering evaluation metrics such as RAND-index, silhouette score, and Adjusted Rand Index (ARI). Furthermore, to comprehensively assess the stability of the clustering results, i introduced additional evaluation metrics not covered in the course, such as the Fowlkes-Mallows Index (FMI), for an in-depth analysis.

In the gene marker selection and gene function analysis, i successfully identified that cluster 12 in the PCA + Leiden and cluster 9 in the PCA + Loouvain model corresponds to the Regenerative Organizing Cells (ROCs) label described in the original study. I further compared the identified gene markers with the results obtained from two different methods, Wilcoxon and t-test to verify the reliability and specificity of these genes. Additionally, i compared the identified gene sets with those listed in Supplementary Table 3 of the paper to confirm their overlap and consistency.

Eventually, I found that the most expressed gene in the Wilcoxon test, clustered 12 in the PCA + Leiden compared to the other classes, was the loc100486893.L gene while in the t test, the most expressed gene was Xelaev18045099m.g. As clustered 9 in the PCA + Louvain, the most expressed gene in the Wilcoxon test was the Xelaev18002241m.g gene while in the t test, the most expressed gene was hbd.s. However, in Table 3 of the Supplementary Material, wnt5a and loc100488523 were highly expressed in ROC cells



Key Words: **ROC Cells; Frog Xenopus Laevis;**


## Introduction

Regeneration is a remarkable biological phenomenon wherein certain organisms can restore damaged or lost tissues to their original state. Among vertebrates, the African clawed frog *Xenopus laevis* is known for its exceptional regenerative capabilities, particularly during its larval stage. The tail of *Xenopus* tadpoles, which comprises complex structures such as muscle, spinal cord, and vasculature, can fully regenerate following amputation. Understanding the cellular and molecular mechanisms underlying this process is crucial for advancing regenerative medicine.

In *Xenopus* tail regeneration, a novel cell type known as the Regeneration-Organizing Cell (ROC) has been identified as essential for initiating and orchestrating tissue repair and regeneration. These cells are characterized by the expression of specific markers such as **Lef-1** and **Tp-63** and are thought to act as a signaling hub, secreting key molecules that regulate the proliferation and differentiation of surrounding progenitor cells.

However, the precise role of ROC cells in *Xenopus* tail regeneration, as well as their molecular interactions with other cell types during the regenerative process, remains poorly understood. By elucidating the characteristics and functions of ROC cells, this study aims to provide deeper insights into the cellular dynamics of vertebrate tissue regeneration and uncover potential therapeutic targets for enhancing regenerative capacity in less regenerative organisms, including humans.

# Methods

## Code Availability
## https://github.com/Songyin0521/APDS-homework

## Data Preprocessing

During the data preprocessing stage, we first performed quality control and normalization of the raw gene expression data. Specifically, we filtered out low-quality cells with abnormal total gene expression(*NA*) and removed genes that were expressed in only a few cells or had very low expression levels across the dataset. Next, i identified highly variable genes (HVG), which capture the most significant variations in expression across cells, thereby improving the detection of subtle biological differences. We then normalized the gene expression values of each cell to ensure uniform total expression and applied log transformation and scaling to mitigate the effects of extreme values. This preprocessing workflow ensured that the dataset was balanced, reduced technical noise, and provided a robust foundation for downstream analyses.

## PCA Cluster& Visualization

In this section, we focused on dimensionality reduction and clustering analysis of the preprocessed gene expression data. We first applied Principal Component Analysis (PCA) to reduce the data's dimensionality, retaining the top 30 principal components, which capture the most variance in the dataset. This step helps to minimize noise and computational complexity, making the data more manageable for downstream clustering.

Subsequently, we constructed a k-nearest neighbor (kNN) graph based on the principal components, which serves as the foundation for clustering algorithms. We employed two popular clustering methods: Louvain and Leiden, both of which are widely used in single-cell RNA-seq analysis due to their efficiency and ability to detect distinct cell populations. By adjusting the resolution parameter, we controlled the granularity of the clustering, ensuring that meaningful subpopulations were identified.

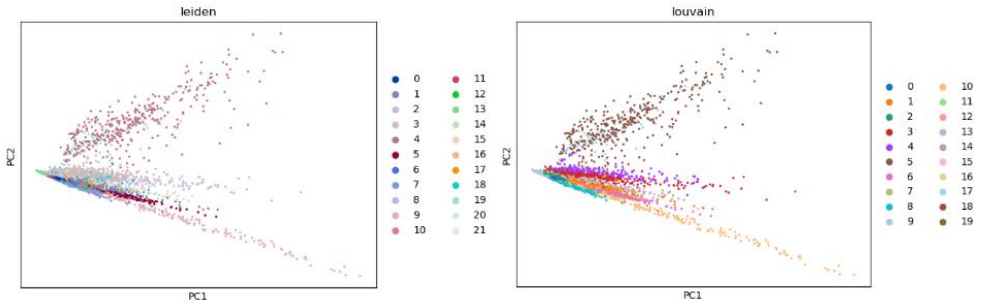


Figure 1 : PCA+ Leiden&Louvain Cluster

For visualization purposes, we utilized Uniform Manifold Approximation and Projection (UMAP) and t-distributed Stochastic Neighbor Embedding (t-SNE) techniques to project the high-dimensional data onto two-dimensional space. This visualization allowed us to clearly observe the clustering patterns and the relationships between different cell populations. The resulting plots provided an intuitive and comprehensive view of the clustering results, which were essential for identifying and interpreting distinct cell types in the dataset.

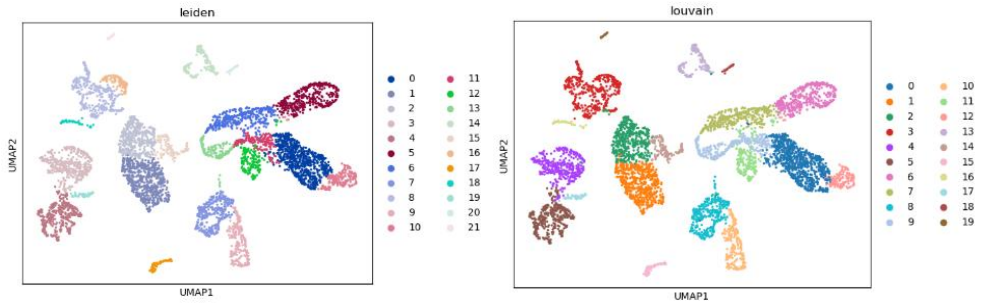


Figure 2 : UMAP& t-SNE in Leiden

Then we calculated various clustering evaluation metrics such as silhouette score, and Adjusted Rand Index (ARI) , Fowlkes-Mallows Index (FMI) to evaluate the difference between Louvain model and Leiden model.

Then,We employed statistical methods, the Wilcoxon test and t-test, to compare the expression levels of each gene between a target cluster and the remaining clusters.

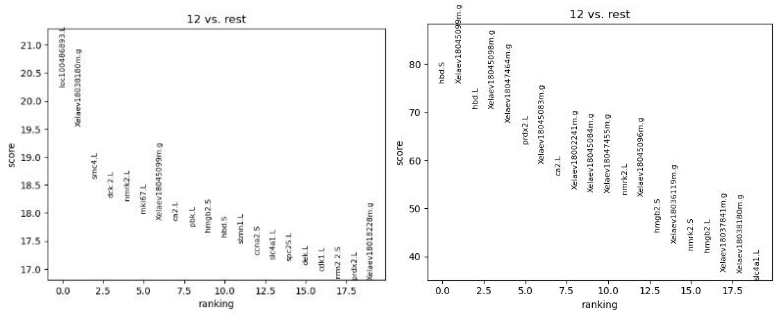


Figure 3: Cluster 12 specific genes in Leiden VS rest (Left:Wilcoxon  Right: T-test)

## Results

By compare the Plot of PCA+Leiden cluster to the cluster from Original paper, we can determine the cluster 12 in Leiden and cluster 9 In Louvain is ROC.

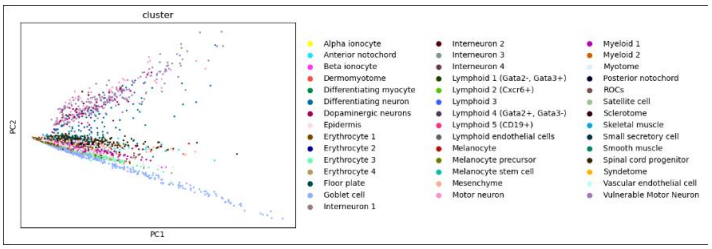


Figure 4: Cluster from Original paper (the brown cluster is ROC)

The result of Compute clustering metrics：

| Clustering metrics | Leidan | Louvain |
|---|---|---|
| Adjusted Rand Index (ARI) | 0.48360 | 0.48627 |
| Silhouette Score | 0.19341 | 0.20523 |
| Fowlkes-Mallows Index | 0.54329 | 0.54337 |

Table 1: 3 Clustering metrics for Leidan

| Wilcoxon | T test | Wilcoxon | T test |
|---|---|---|---|
| **Leidan** | | **Louvain** | |
| loc100486893.L | Xelaev18045099m.g | Xelaev18002241m.g | hbd.S |
| Xelaev18038180m.g | hbd.S | hbg2.L | hbd.L |
| smc4.L | hbd.L | ca2.L | Xelaev18045098m.g |
| dck.2.L | Xelaev18045098m.g | nmrk2.L | Xelaev18045099m.g |
| Xelaev18045099m.g | Xelaev18047464m.g | Xelaev18047455m.g | Xelaev18045083m.g |

Table 2: Top 5 specific genes in two cluster for Wilcoxon test and t-test

## Conclusion

Compared the specific genes(**Wn5t.a** and **loc100488523)** of ROC in Supplementary Material, our most specific genes in this two test is **loc100486893.L** , **Xelaev18045099m.g**, **Xelaev18002241m.g** and **hbd.S** respectively. Besides, we can find the gene called Xelaev180***** and hbd.S were expressed in large quantities, which may be the secret of **Wolverine frog——Super resilient.**