Identification of the Regenerative Organizing Cell in the Frog Tail Skin

Author: Estella Yu UNI: yy3644. **Due:** 1st, Oct, 2025

1. Abstract

In this study, we identified the *Regenerative Organizing Cell* (ROC) in *Xenopus* tail skin by analyzing single-cell RNA sequencing (scRNA-seq) data.

Multiple clustering algorithms (PCA + Louvain, PCA + Leiden, and PCA + kMeans) were applied to explore cell-type heterogeneity. Marker selection was performed using logistic regression and Wilcoxon rank-sum methods to identify signature genes distinguishing the ROC from other cells. These markers were compared with those listed in Supplementary Table 3 . Additionally, Gene Ontology (GO) enrichment revealed correlations in extracellular matrix organization and BMP signaling processes. Further analyses demonstrated that data denoising and batch integration improved clustering results. And data denoising influenced the gene marker results.

2. Introduction

Identifying the Regenerative Organizing Cell (ROC) is essential to understand how regeneration is initiated at the molecular level. And It could lead to potential applications in regenerative medicine, such as tissue repair and wound healing. The aim of this project was to locate and characterize the ROC population within the tail skin data. Using scRNA-seq, we performed clustering, marker identification, and Gene Ontology enrichment analyses. Additional preprocessing steps such as data denoising and batch integration were evaluated to assess their influence on clustering and marker selection.

3. Methods

3.1 Data Preprocessing and Clustering

The raw *AnnData* object(from'cleaned_processed_frogtail.h5ad') contained 13,199 cells and 26,166 genes across multiple developmental stages and experimental conditions(different batches). I used three clustering pipelines: **PCA + Louvain**, **PCA + Leiden**(state-of-the-art), **PCA + kMeans**. The Evaluation Metrics:

- **RAND Index** : agreement with known labels, the bigger the better
- **Silhouette Score** : cluster compactness and separation, the bigger the better
- **Adjusted Rand Index (ARI)** : correction of label matching, the bigger the better (primary index)
- **Calinski–Harabasz Index** : between and within cluster dispersion, the bigger the better

Calinski–Harabasz Index is not discussed in class, and it is defined as: $CH = \frac{tr(B_k)/(k-1)}{tr(W_k)/(n-k)}$. B_k and W_k denote dispersion between cluster and dispersion within cluster.

3.2 Marker Selection and Gene Analysis

Marker genes defining the ROCs were identified using two methods: **Logistic Regression**, **Wilcoxon Rank-Sum Test**. And GO enrichment analysis was also performed.

3.3 Data Denoising

To reduce technical noise, I used two smoothing techniques: **kNN-smoothing** and **iterative kNN-smoothing**

3.4 Batch Integration Over Time

Temporal batch effects across developmental stages were corrected using: **Combat** and **BBKNN**

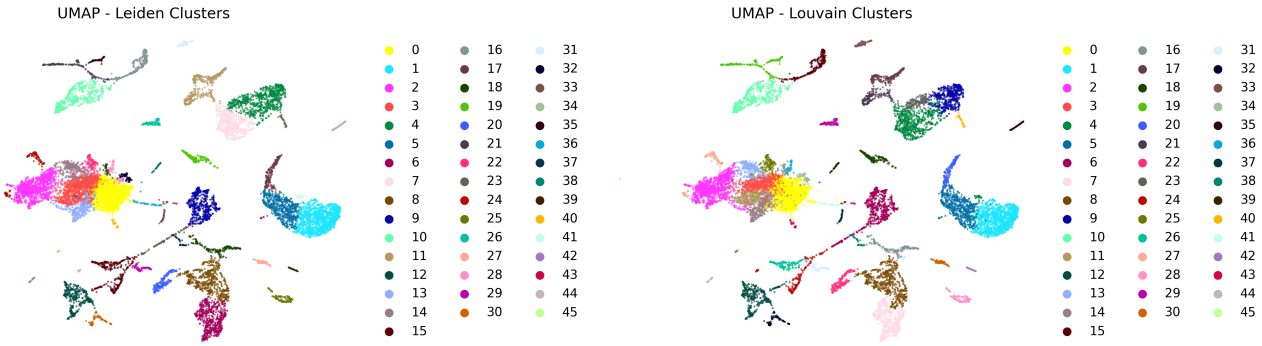
Code Availability

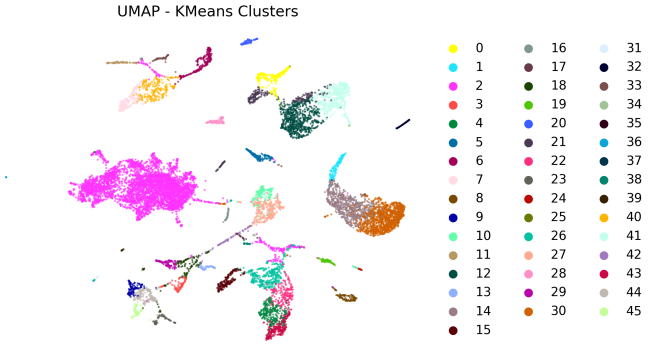
Colab Link: https://colab.research.google.com/drive/17uJ0pJc8RixlzKTRaiTYP0DLgIg8W8Ek?usp=sharing
Github Link: https://github.com/cccloves/Biological-Applied-DS#

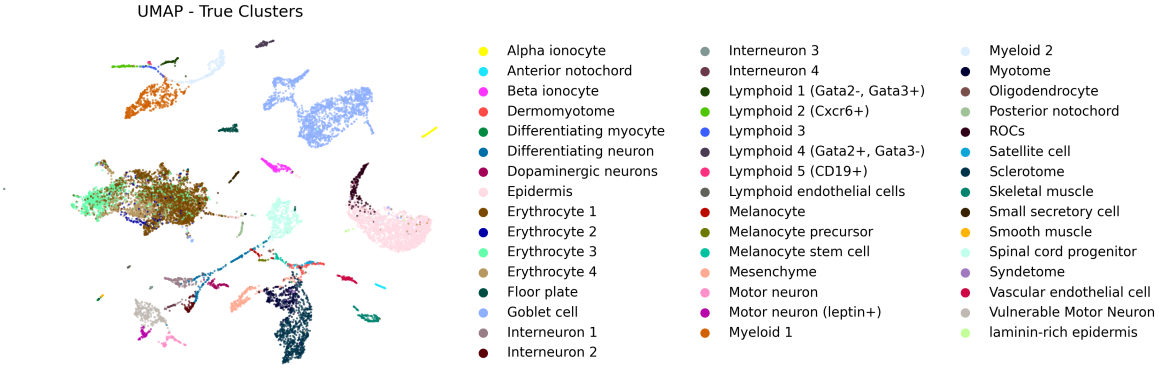
4. Results

Clustering Performance Summary

UMAP visualization was used for dimensionality reduction and cluster display. Louvain and Leiden produced highly similar results, both closely resembling the known biological clusters, while kMeans showed partial divergence.







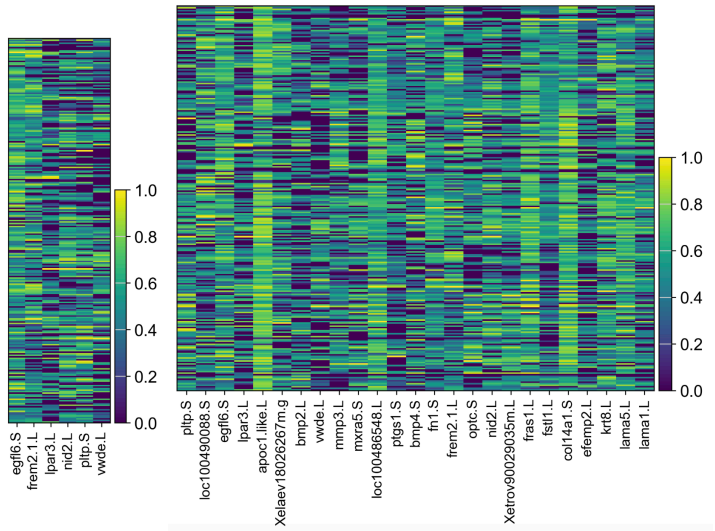
Metrics:

Method	Adjusted Rand Index	Rand Index	Silhouette Score	Calinski–Harabasz Index
Louvain	0.437	0.922	0.196	1023.231
Leiden	0.459	0.924	0.210	1004.550
kMeans	0.481	0.879	0.504	1859.406

Leiden achieved the high *Adjusted Rand Index* (0.459) and *Rand Index* (0.924), showing the best agreement with true biological labels while maintaining local connectivity. **Louvain** performed similarly to Leiden but with slightly lower separation quality. **kMeans** produced the highest *Silhouette Score* (0.504) and *Calinski–Harabasz Index* (1859.406), indicating more compact and well-separated clusters in PCA space.

Gene Expression Summary

The top 100 genes from each method were compared. **25 common marker genes** were shared across both methods: *bmp2*, *bmp4*, *egfl6*, *fn1*, *nid2*, and so on. Comparison with Supplementary Table 3 showed **six overlapping genes**: *egfl6, frem2, lpar3, nid2, pltp, vwde.* Below is the expression heatmap of 25 common marker genes and six overlapping genes:



GO enrichment analysis (performed using the Enrichr human database due to unabtainable *Xenopus* annotations) revealed that ROC marker genes were involved (Top overlap) in extracellular matrix organization and BMP signaling.

These terms suggest that ROC cells are linked to extracellular remodeling and developmental signaling, which is consistent with their regenerative role.

GO Term-Top 5 Overlap	Biological Meaning
extracellular matrix organization (GO:0030198)	Assembly and remodeling of extracellular matrix components such as collagen and elastin.
extracellular structure organization (GO:0043062)	Formation and maintenance of extracellular structural frameworks.
external encapsulating structure organization (GO:0045229)	Assembly of outer protective structures such as basal lamina.
muscle tissue morphogenesis (GO:0060415)	Development of muscle tissue architecture.
BMP signaling pathway involved in heart development (GO:0061312)	Role of BMP signaling in cardiac formation.

Denoising

After denoising, clustering structure became clearer and more biologically coherent. Clusters showed higher silhouette and CH indices, suggesting improved separability.

Marker selection after denoising produced greater overlap with the Supplementary Table 3 markers. After denoising, there are **27** common marker genes in both methods. In Supp Table 3, there are **19** overlap. It indicating that denoising enhances signal-to-noise ratio and improves biological interpretation.

Batch Integration

Batch integration(combat) improved clustering alignment across samples, reducing artificial separation caused by batch effects. The resulting clusters were more continuous and biologically meaningful (Figure 1, lower panel). However, the result of BBKNN + PCA + Leiden/Louvain is not good. The reason maybe that sometimes BBKNN “over-corrects,” mixing biologically distinct clusters simply because they come from different batches.

However, marker selection results remained largely unchanged, implying that batch correction affects cell grouping but not gene-level differences.

5. Conclusion

This analysis successfully identified the Regenerative Organizing Cell (ROC) in *Xenopus* tail skin and characterized its molecular signature. The ROC is defined by genes associated with extracellular structure formation and BMP-mediated developmental pathways.

Data denoising and batch integration significantly improved clustering and biological interpretability. And it demonstrates the importance of preprocessing in single-cell analysis. However, clustering or batch integeation do not influence the result of gene marker. The results align with published findings.