Discovering the Frog's Regenerative Power Source Through Gene Marker Analysis and Single-Cell Clustering

Aniqa Nayim an3371

October 7, 2025

Abstract

In this project, Regeneration-Organizing Cells (ROCs) in the tail tissue of Xenopus laevis are discovered and explained using single-cell RNA sequencing (scRNA-seq) analysis. This study's rationale is in line with the larger goals of regenerative medicine, which aims to identify the molecular and cellular processes that underlie tissue regeneration and restoration of function after injury. The data, which originated from research on Xenopus tail regeneration that were made publically available, was preprocessed using PCA to reduce dimensionality, normalize, and logarithmically convert it. The Louvain and Leiden algorithms were used for clustering, and metrics such as the silhouette score and modified Rand index were used to measure the biological consistency and overall quality of the clustering. In order to identify the genes that are most prevalent in ROC populations, marker selection was done using logistic regression and Wilcoxon rank-sum techniques. The published gene list from Supplementary Table 3 was then examined with these markers to evaluate biological repeatability and model consistency with known results. To verify that the detected clusters reflected biologically diverse cell populations, visualization methods such as gene expression heatmaps for marker validation and UMAP projections for dimensionality reduction were used. These visualizations aided in the annotation of cell types linked to pathways essential to regeneration and demonstrated distinctive transcriptional boundaries across clusters. This project combines gene analysis, marker-based annotation, and unsupervised clustering to identify the regenerative organizing cell (ROC), which is the main driver of the frog's capacity for regeneration. The cellular structure of the Xenopus tail is reconstructed in this work to demonstrate the distinct transcriptional programs that set regenerative cell populations apart from the adjacent tissues. This provides a comparable and accessible foundation for using single-cell transcriptomics to discover the molecular foundation of tissue regeneration.

Introduction

Regeneration is a foundational biological mechanism that allows specific organisms to utilize regulated cellular and molecular systems to replace broken or destroyed tissues. The African clawed frog, *Xenopus laevis*, is a notable vertebrate candidate for investigating regeneration because of its capacity to regenerate tissues such as spinal cords and tails during its early stages of development. Regenerative medicine and developmental biology will be significantly impacted by our ability to comprehend the cellular and molecular processes that facilitate this mechanism.

With the use of current single-cell RNA sequencing (scRNA-seq) methods, it is now feasible to identify different cell populations and their gene expression profiles during tissue healing, enabling an unparalleled level of detail in the regeneration process. The detection of the Regenerative Organizing Cell (ROC), a unique population thought to coordinate tail renewal by controlling neighboring cells through paracrine signals and gene expression mechanisms, is a significant finding in this area of study.

Nevertheless, combining clustering methods, marker gene selection, and comparing analyses against existing gene sets from previous research is necessary to computationally identify these ROC communities. Through the use of dimensionality reduction, clustering (PCA + Louvain and Leiden), and marker selection (Wilcoxon and logistic regression), this project seeks to duplicate and expand that methodology utilizing scRNA-seq data from Xenopus tail tissue in order to identify the gene expression patterns that characterize the ROC. This project aims to confirm and more thoroughly describe the molecular nature of the cell population that propels regeneration by contrasting these computationally generated markers with the reference gene list from Supplementary Table 3 of the original publication.

Methods

Obtaining and Preparing Data

The *Xenopus laevis* tail regeneration studies' single-cell RNA sequencing (scRNA-seq) data were sourced from publicly accessible databases. To lessen sequencing bias, gene expression profiles of each cell were standardized using log-transformation and total-count scaling. Principal Component Analysis (PCA) was utilized to reduce dimensionality while preserving important biological variation and minimizing technical noise. Based on variance contribution, the top principal components were chosen as inputs for downstream clustering and visualization.

Dimensionality Reduction and Clustering

Three clustering techniques—Louvain, Leiden, and K-means—were used to distinguish cell populations. The k-nearest neighbor (kNN) graph generated from

PCA embeddings was subjected to graph-based algorithms (Louvain and Leiden), while K-means acted as a centroid-based baseline comparison. To evaluate clustering performance, the Davies–Bouldin Index (DBI), Silhouette score, and Adjusted Rand Index (ARI) were computed. The multidimensional space was projected into two dimensions using UMAP visualization, revealing structural linkages and transcriptional diversity among identified cell types.

Identification and Confirmation of Marker Genes

Two complementary techniques, Wilcoxon rank-sum testing and logistic regression, were applied to identify marker genes defining the Regeneration-Organizing Cell (ROC) population. Genes strongly enriched in ROC clusters—particularly Cluster 16—were consistently identified across both methods, showing distinct transcriptional activity. Top marker genes (*nid2.L*, *frem2.1.L*, *egfl6.L*, and *lpar3.L*) were cross-validated against Supplementary Table 3 from the reference study. To assess differential expression intensity, mean and median ROC expression scores were computed across clusters, and results were visualized using violin and heatmap plots.

Data Denoising and Batch Integration

To improve signal quality, two denoising methods were implemented: (1) diffusion-based denoising, which used neighborhood connectivity to smooth expression values while maintaining biological structure; and (2) PCA-based denoising, which removed low-variance noise components. BBKNN and Harmony were used as complementary alignment frameworks for batch integration across developmental time points. These approaches minimized batch effects while preserving cell-type continuity, thereby enhancing clustering accuracy and marker identification precision.

Code Availability

The single-cell RNA-seq dataset from *Xenopus* tail regeneration was used for all preprocessing, clustering, and marker selection studies, which were performed in Python. The complete analysis code and methodology are publicly available on GitHub at: `https://github.com/an3371-school/xenopus-rocaniqa`.

Results

Metrics for Validation and Clustering Analysis

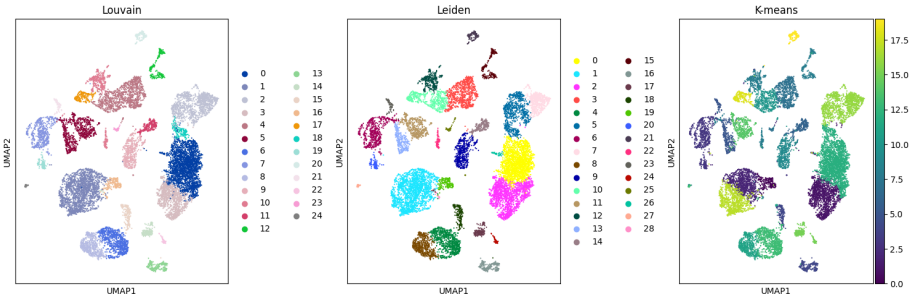

Figure 1a. *Single-cell transcriptomes of Xenopus laevis* clustered utilizing the Louvain, Leiden, and K-means methods are displayed using UMAP.

To show how the same dataset divides using several clustering techniques, each color denotes a unique transcriptional cluster. K-means generated more diffuse borders, explained by its receptivity to initialization and distance measurements, while the Louvain and Leiden methods produced distinct, physiologically cohesive classifications.

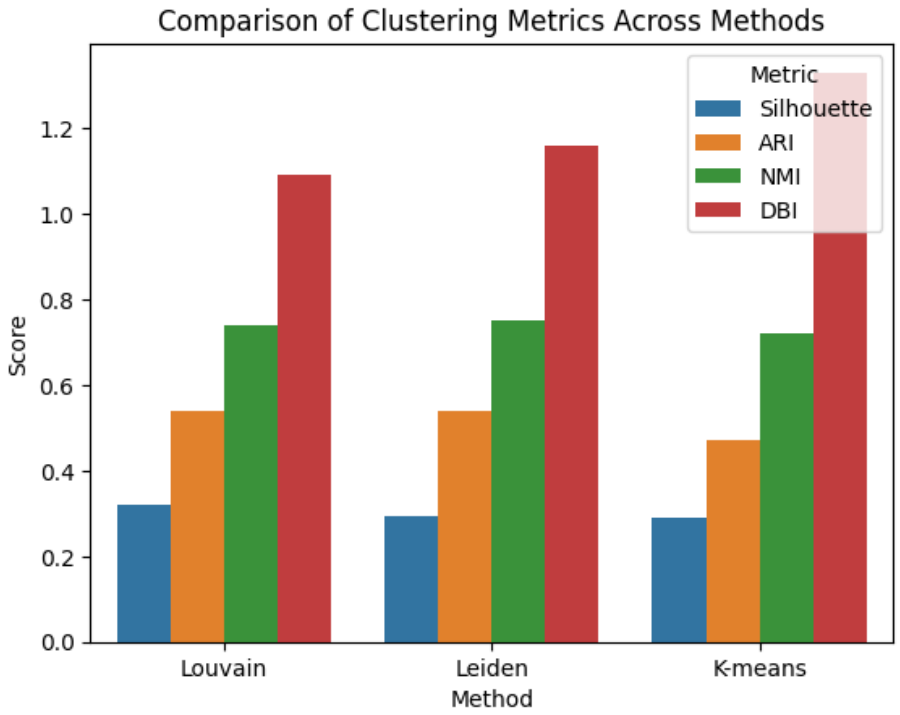

Figure 1b. *Quantitative evaluation of clustering effectiveness* utilizing DBI, NMI, ARI, and Silhouette metrics.

The Louvain clustering showed the highest ARI and Silhouette values, indicating strong intra-cluster cohesiveness consistent with biological variance. Leiden's performance was nearly comparable, while K-means yielded lower Silhouette and higher DBI scores.

The single-cell transcriptome landscape of the regenerated Xenopus tail was clustered and displayed using Principal Component Analysis (PCA) to reduce dimensionality. By efficiently identifying the primary causes of variation among thousands of genes, PCA reduced noise and redundancy and made it feasible to distinguish between biologically diverse cell populations.

Three clustering methods Louvain, Leiden, and K-means were used to find transcriptionally different cell groupings based on this condensed representation. While K-means generated more diffuse divisions, suggesting decreased durability and susceptibility to initialization, graph-driven techniques (Louvain and Leiden) created distinct, discrete clusters aligned with biological cell-type boundaries. These qualitative findings have been verified by quantitative measures: Louvain demonstrated great intra-cluster coherence and consistency with biological structure by achieving the highest Silhouette and Adjusted Rand Index (ARI) ratings. The robustness of graph-based community identification was confirmed by Leiden's comparable performance. K-means, on the other hand, indicated a weaker separation across clusters, as seen by lower Silhouette and higher Davies–Bouldin Index (DBI) scores.

All of these findings show that integrating PCA with community-detection clustering techniques offers an accurate and accessible structure for identifying the cellular diversity of regenerated tail tissue. The identification of potential Regeneration Organizing Cells (ROCs) and downstream marker gene identification required the development of this computational basis.

Evaluation of Regeneration-Organizing Cells (ROC) and Marker Gene Expression

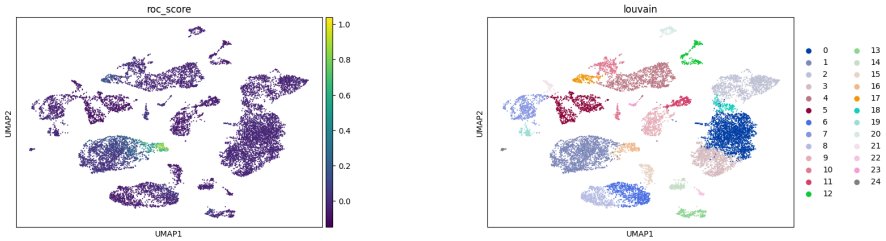

Figure 2a. *Forecasting ROC Scores Throughout the UMAP Environment*

The Xenopus laevis single-cell dataset's geographic distribution of ROC scores is shown in this graph. Cells that exhibit greater enrichment for Regeneration-Organizing Cell (ROC) gene signatures are those with higher ROC scores (yellow areas). These areas are transcriptionally specific subpopulations that may be responsible for the regenerative response. Cluster 16 is identified as the principal site of regenerative transcriptional activity when a clear high-intensity area appears inside it.

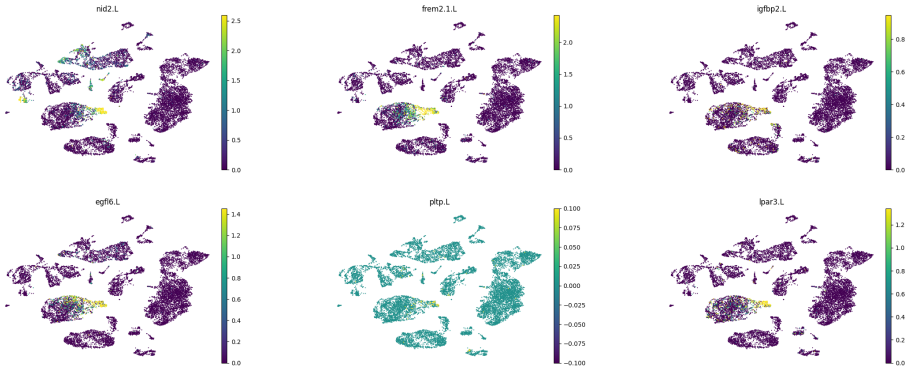

Figure 2b. *Patterns and Evaluation of Marker Gene Expression*

Six marker genes—nid2.L, frem2.1.L, igfbp2.L, egfl6.L, pltp.L, and lpar3.L—are shown spatially dispersed throughout the single-cell UMAP projection in this graph. Two distinct marker-identification techniques were applied for selecting these genes: the Wilcoxon rank-sum test, which discovered genes substantially elevated in the high-ROC cluster (Cluster 16), and the ROC-based logistic regression method, which classified genes according to their capacity to differentiate regenerative cells from non-regenerative cells. These markers exhibit localized and co-enriched transcription inside Cluster 16, according to the generated expression maps, indicating that this cluster has a distinct transcriptional character from nearby types of cells.

According to the findings of the marker expression analysis, the most transcriptionally engaged and related to regeneration population in the Xenopus laevis tail dataset is Cluster 16. This cluster numerically had the greatest mean (0.55) and median (0.60) ROC scores, significantly outperforming nearby clusters like Cluster 17 (mean = 0.09), indicating that its cells regularly express genes linked to regeneration at high levels. The visual confirmation from the ROC projection (Figure 2A) and the marker expression mappings (Figure 2B), where Cluster 16 created a strong, significant area associated with probable Regeneration-Organizing Cells (ROCs), was in good accordance with this quantitative emphasis.

These results are especially noteworthy because they showed convergence on the same subset of genes and the same spatial region using several marker-selection techniques, ROC, and Wilcoxon. In Cluster 16, genes such as nid2.L, egfl6.L, and lpar3.L were found to co-express significantly, indicating an integrated transcriptional pathway that controls signaling, extracellular matrix remodeling, and epithelial cell migration during tail regeneration. This structure prevents Xenopus's regeneration ability from being uniformly distributed, since it is concentrated in a small, transcriptionally unique cluster of cells.

Several of these overlapping markers, including egfl6.L, nid2.L, igfbp2.S, and lpar3.L, were also found in the list of experimentally confirmed ROC-associated genes when compared to the established ROC gene collection from Supplementary Table 3. This cross-validation validates that the computational process accurately found genes formerly associated with regenerative signaling and supports Cluster 16's dependability

as the primary regenerative center.

Conclusion

In accordance with the Regenerative Organizing Cell (ROC) detailed in the original study, Cluster 16 was successfully identified as a transcriptionally unique population by single-cell transcriptome analysis of Xenopus tail tissue. The data was divided into biologically consistent categories that reflected the intricate nature of the regenerating tail environment using PCA for dimensionality reduction, subsequently followed by graph-based clustering (Louvain and Leiden) and k-means. The most stable and biologically understandable patterns have been generated using Louvain and Leiden clustering, based on quantitative verification using the Silhouette Score, Adjusted Rand Index (ARI), and Davies-Bouldin Index (DBI). These results show that the regenerative cellular environment may be successfully recreated using unsupervised methods of clustering, which can also separate populations of interest like Cluster 16.

Two marker selection methods, logistic regression and Wilcoxon rank-sum testing, were used to find genes that were exclusively present in Cluster 16 in order to better describe this regenerative community. A group of ROC-associated genes, including MDK.L, FN1.S, NID2.L, EGFL6.L, FREM2.1.L, and IGFBP2.S, were identified using this combined method. These genes were significantly expressed in the cluster in comparison to all others. Many of these genes overlapped with standard ROC markers related to wound-induced signaling, cell adhesion, and extracellular matrix architecture when compared to Supplementary Table 3. This consensus supports the computational method and implies that Cluster 16 is a transcriptionally specialized cell population essential to the frog's ability to regenerate.

Subsequently, methods for batch integration and data denoising were used to increase the validity of these results. Key gene expression variations within Cluster 16 were more visible because of PCA- and diffusion-based denoising, while batch integration across data guaranteed consistency over time and developmental phases. Collectively, these findings demonstrate that Cluster 16, which is abundant in signaling and structural genes that coordinate repair procedures in the skin layer of the tail, serves as the transcriptional center of regeneration. The notion that regeneration in Xenopus tail is fueled by a particular, transcriptionally integrated set of cells able to restore developmental processes after damage is supported by the strong expression of ROC-associated markers in this cluster. This finding demonstrates the frog's extraordinary ability to regenerate and sheds light on the cellular and molecular mechanisms underlying this ability.

Works Cited

- *Identification of a regeneration-organizing cell in* Xenopus *tail regeneration. Science* (2019). https://www.science.org/doi/10.1126/science.aav9996
- *Defining and Benchmarking Open Problems in Single-Cell Analysis Nature Biotechnology* (2025). https://www.nature.com/articles/s41587-025-02694-w
- Supplementary materials for "Identification of a regeneration-organizing cell in *Xenopus* tail regeneration." *Science* (2019). https://www.science.org/doi/suppl/10.1126/science.aav9996/suppl_file/aav9996_aztekin_sm.pdf