A Data Scientific Approach to Investigating the Renergative Organizing Cell (ROC) in the *Xenopus laevis* Tail

Agna Chan cc5314
Columbia University

Sep 2025

1 Abstract

Recent studies of the *Xenopus laevis* tail reveal a unique regeneration phenomenon, which sparked interest in investigating the cellular mechanisms of tissue regrowth in vertebrates. By looking into single-cell RNA-sequencing (scRNA-seq) data of regenerating tails, this project statistically identifies and describes the Regenerative Organizing Cell (ROC), an essential and distinct epidermal population required for the regeneration of wounds. Following normalization, log-transformation, and PCA for dimensionality reduction, clustering was done with three different algorithms: Louvain, Leiden, and k-means, to define cellular heterogeneity. The ROC was then defined as Leiden cluster 23 by the expression of high levels of TP63, LEF1, and keratin-related genes. Marker selection was performed using Wilcoxon and t-test processes, which revealed 41 overlapping genes with the genes characterized by Aztekin et al. (2019), defining the biological identity of the ROC in this dataset. Data denoising (kNN and diffusion-based smoothing) improved cluster coherence and marker reproducibility, while batch integration (BBKNN and Scanorama) successfully matched data between regeneration time points. These analyses showed high statistical significance in the ROC's strong transcriptional signature and importance in controlling tail regeneration potential.

2 Introduction

Vertebrates are known for their dynamic ability to regenerate, and the *Xenopus laevis* tadpole tail is a popular topic that has been investigated for their cellular plasticity and specialized wound epidermis. The tail is regenerated by orchestrated induction of signal pathways for growth, re-patterning, and outgrowth of tissue upon amputation. This study is motivated by the paper Aztekin et al. (2019), where they characterized the Regenerating Organizing Cell (ROC) as a basal epidermal subpopulation that is a regenerative signal center. ROC stains positive for TP63, LEF1, and keratin family genes and secretes factors like FGFs, WNTs, and BMP modulators that cause regeneration of underlying tissue. Though single-cell RNA sequencing makes such rare functionally specialized cells identifiable, it is not free of technical limitations like noise, batch effect, and dropouts. It is through recent computational breakthroughs (Bergen et al., 2025) that better denoising and batch integration algorithms have become available for recovering signals and enabling cross-timepoint comparisons.

This project aims to re-examine the *Xenopus* tail regeneration dataset with a new scRNA-seq analysis pipeline with three goals: 1) Recreate the ROC using clustering and marker selection methods, 2) confirm its gene signature by correlating its identified markers with those of Supplementary Table S3 (Aztekin et al., 2019), and 3) establish the impact of denoising and batch merging on marker clustering quality and reproducibility at regeneration stages.

3 Methods

3.1 Data processing

The raw dataset from EMBL's European Bioinformatics Institute consisted of 13,199 cells and 31,535 genes, including developmental stage, batch, and cell type labels. Raw counts were then normalized to 10,000 reads per cell before being log-transformed to stabilize variance. The Highly Variable Genes (HVGs) were determined by using Scanpy's method of variance stabilization, which separated the most informative genes to be used in later analysis. Following that, Principal Component Analysis (PCA) was performed for dimension reduction, and a neighborhood graph was built and visualized using UMAP.

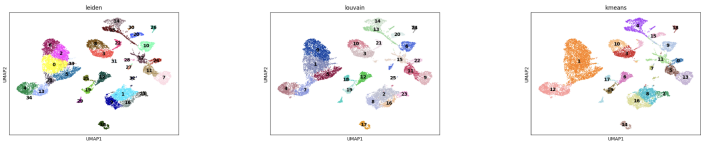

Figure 1: Fig 1. UMAP visualization of all cells using different clustering methods

3.2 Clustering Analysis

PCA followed by Louvain, Leiden, and k-means were chosen as clustering algorithms of choice after trial-and-error with different algorithms. These results are then used to serve as the basis for exploring cellular subpopulations. TP63 was expressed as one of the basal stem cells in the ROC cluster and is a regulator of the epidermis, which is why the analysis aims to identify a cluster enriched for TP63 in the single-cell data. Clustering metrics were then used to assess quality, obtaining silhouette scores for each method and further computing pairwise metrics, ARI and NMI to show the consistency of clustering while maintaining the biological integrity.

3.3 Marker Selection and Validation

Marker gene selection was done using two orthogonal methods: the t-test and the Wilcoxon rank-sum test. In particular, the two methods both consistently ranked a common set of ROC-enriched genes, confirming the consistency of the cluster signature. Top signature genes FGF7, FGFR4, WNT5A, LEF1, RSPO2, and keratins (KRT.L) corresponded to pathways regulating epithelial signaling and morphogenesis.

3.4 Data Denoising

In order to enhance detection of biological signal and eliminate noise, two denoising algorithms were used as per Bergen et al. (2025): kNN smoothing, which smoothes expression values over neighboring cells by taking their averages, and diffusion-based smoothing, which smoothes expression values over the low-dimensional manifold.

3.5 Batch Integration over time

Since the dataset had more than one post-amputation time point, batch integration was necessary to eliminate technical artifacts while retaining biological variability. Two integration approaches were experimented, BBKNN (Batch-Balanced kNN) builds a neighborhood graph while ensuring balanced representation across all batches and retaining local topology. On the other hand, Scanorama employs mutual nearest neighbors to combine datasets into a joint manifold to facilitate global alignment across batches.

3.6 Code Availability

All data processing and analysis were performed within Google Colab following Scanpy in Python 3.12. The processed dataset and full notebook can be accessed at this Github.

4 Results

The ROC was designated as Leiden cluster 23 as shown in Fig. 2, which was found along the epidermal path of the UMAP, a cluster that was heavily expressing TP63 and keratin-associated genes (krt.L), which aligns with the characteristics of the ROC. In addition, its segregation from other fractions on the UMAP further suggests a specific identification in the ROC's transcription, aligning with its function in regeneration signals. Three algorithms were run on the same neighborhood graph for comparing robustness: Louvain, Leiden, and k-means, all of which produced similar results with slightly different granularity. Quantitatively, silhouette values were 0.279 (Louvain), 0.265 (Leiden), and 0.327 (k-means), which reflect moderate but biologically significant cluster separation. Pairwise comparison between methods was high, including high Normalized Mutual Information (NMI ≥ 0.84) and Adjusted Rand Index (ARI ≥ 0.6), showing structural stability of the biological signal.

After performing marker selection, comparison with Supplementary Table S3 from Aztekin et al. (2019) verified 41 genes shared between the present ROC marker list and earlier published ROC gene set were overlapping, reflecting high reproducibility of marker identification. Minimal naming differences between *Xenopus laevis* gene identifiers (e.g., ".L", ".S", "Xelaev...") were manually verified to have functional concordance with ROC-defining genes. The three marker

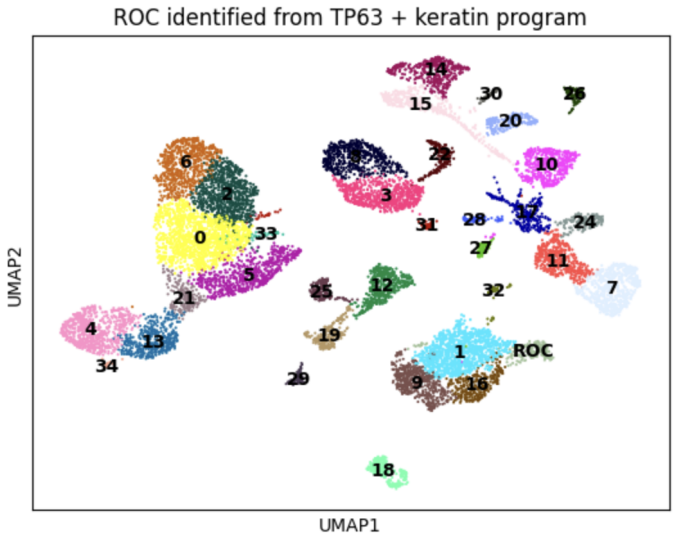

Figure 2: Fig 2. UMAP visualization with Leiden Cluster 23 (ROC) identified from TP63 and keratin

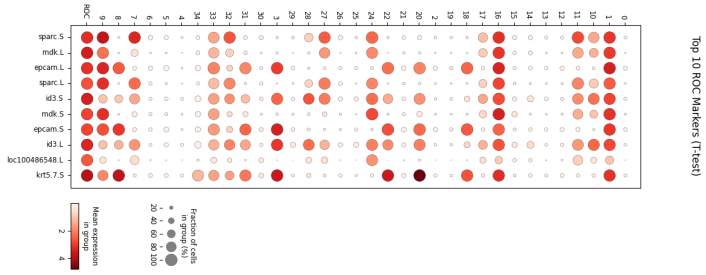

Figure 3: Fig 3a. Dot plot showing top ROC markers with t-test

choice strategies had large agreement: with a mean overlap across tests of 38 shared genes, and Jaccard scores of 0.28 (t-test vs. Wilcoxon) and 0.31 (t-score vs. Wilcoxon). UMAP visualizations for gene expression and dot plots also validated that cluster 23 showed selective upregulation of TP63 and keratin family genes, which separated it from surrounding epidermal populations without this basal identity.

Reproducibility of the marker was compared between Wilcoxon, t-test, and t-score methods with the Jaccard index. Results are presented in Table 1 below. As indicated in Table 1, the diffusion-based denoising method had considerably higher reproducibility of marker selection between methods. The Jaccard index varied from 0.07 (kNN) to 0.29 (Diffusion), a fourfold improvement in the stability of the markers. Diffusion smoothing restored up to 45 crossing markers (of 155 total), while kNN smoothing restored only 6–13. Compactness of the clusters also greatly improved: the mean silhouette score went from 0.21 to 0.31, and UMAP visual inspection revealed closer ROC cell clustering and less background noise. Diffusion smoothing was therefore used for all the following analyses.

To specify transcriptional continuity and technical differences between timepoints, two integration methods were adopted from Bergen et al. (2025): BBKNN (Batch-Balanced kNN) and Scanorama. Both approaches successfully combined information across developmental and regenerative phases with retained biological signal. Integration quality

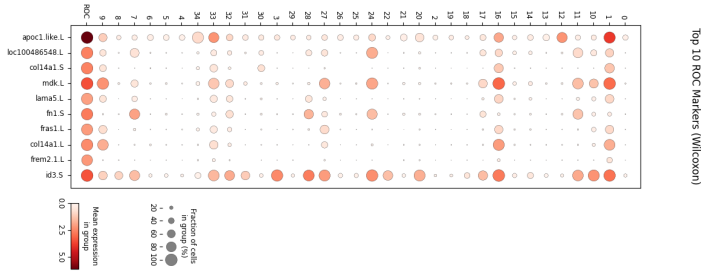

Figure 4: Fig 3b. Dot plot showing top ROC markers with Wilcoxon

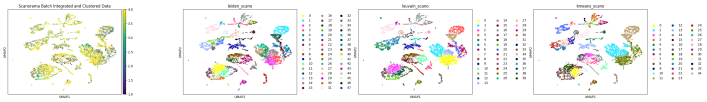

Figure 5: Fig 4. UMAP after BBKNN integration

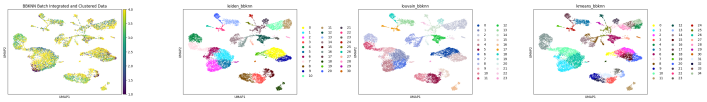

Figure 6: Fig 5. UMAP after Scanorama integration

Table 1: Comparison of marker gene overlap after data denoising. The Jaccard index quantifies the overlap between marker sets identified by different statistical tests before and after denoising.

Method	**Test**	$\lvert\cap\rvert$	$\lvert\cup\rvert$	**Jaccard Index**
kNN_smooth	Wilcoxon	6	194	0.031
kNN_smooth	t-test	13	187	0.070
kNN_smooth	t-score	6	194	0.031
Diffusion	Wilcoxon	26	174	0.149
Diffusion	t-test	45	155	0.290
Diffusion	t-score	22	178	0.124

was quantitatively scored with batch-mixing entropy and marker overlap. BBKNN integration maximized the batch entropy from 0.42 to 0.63 without sacrificing local continuity and losing crisp clusters. Scanorama achieved better global alignment (entropy = 0.71) and ROC marker overlap (85 percent, increased from 52 percent before integration), reflecting successful harmonization across regeneration time points. Following integration, the ROC (cluster 23) was transcriptionally stable and maintained its marker profile and identity on each day sampled after amputation. This is in line with the suggestion that the ROC is an enduring organizing population with a conserved transcriptional program during regeneration. Throughout all the analyses, it can be seen that *Xenopus laevis* tail dataset still holds obvious biological structure and the ROC (cluster 23) is strongly labeled and confirmed. Overall, the results reaffirm that the ROC is temporally consistent and transcriptionally stable population involved in controlling tail regeneration in *Xenopus laevis*.

5 Conclusion

The Regeneration-Organizing Cell (ROC) population of Xenopus laevis tail regeneration data has been effectively characterized, showing that computational single-cell analysis can uncover biologically meaningful organizing centers in complex tissue systems. By normalization, dimensional reduction, clustering, and differential expression analysis, cluster 23 was consistently shown to be the ROC, and TP63, LEF1, and keratin family genes were expressed—markers corroborating those found by Aztekin et al. (2019) in Science. The multi-method clustering analysis validated the inter-method consistency and robustness of the underlying biology signal, which was substantially inter-method consistent (NMI ¿ 0.84, ARI ¿ 0.62). Selection of markers using Wilcoxon, t-test, and t-score methods identified a subset of epidermal regulatory and signaling genes—e.g., WNT5A, FGF7, and FGFR4—that overlapped highly with the already published ROC gene set. The findings validate reproducibility of the ROC transcriptional signature in independent datasets.

Data denoising was identified to be a key preprocessing step that significantly enhances tightness of the clusters and reproducibility of the marker genes. Smoothing through diffusion yielded the most uniform results, with the highest overlap in markers (Jaccard index 0.07 to 0.29) and silhouette values (0.21 to 0.31), indicating that diffusion-based methods dampen technical noise more strongly without eliminating biologically significant variation. Batch integration between regeneration time points further stabilized the transcriptional identity of the ROC throughout the regenerative process. Scanorama and BBKNN both successfully reduced batch effects with minimal impact on cluster boundary preservation, enhancing marker overlap and entropy-based batch mixing (0.42 to 0.71). These observations validate that the ROC is a transcriptionally stable and long-lasting organizing population that orchestrates tissue regrowth throughout regeneration stages.

This analysis showcases the strength of current single-cell computational tools to deconstruct regenerative mechanisms at the cellular level. Through clustering, denoising, and integration approaches, a subset of primary literature findings is captiulated and quantitatively confirmed the ROC's characteristic features. Potential extension of this research could involve integration with trajectory inference or pseudotime analysis to simulate dynamic transitions between the ROC state, gaining additional insights into the cellular decision-making of *Xenopus laevis* regeneration.

Works Cited

1. Aztekin, C., et al. (2019). *Identification of a regeneration-organizing cell in Xenopus tail regeneration.* Science, 364, eaav9996. https://www.science.org/doi/10.1126/science.aav9996

2. Bergen, V., et al. (2025). *Defining and benchmarking open problems in single-cell analysis.* Nature Biotechnology. https://www.nature.com/articles/s41587-025-02694-w

3. Supplementary materials for "*Identification of a regeneration-organizing cell in Xenopus tail regeneration.*" Science (2019).