Identification of Regeneration-Organizing Cells in Xenopus laevis Tadpole Tail Using Single-Cell RNA Sequencing

RimjhimSingh
October 2025, Applied Data Science

Abstract

Regeneration-organizing cells (ROCs) are a critical cell population that coordinates tail regeneration in *Xenopus laevis* tadpoles. Using single-cell RNA sequencing data from 13,199 cells, we successfully identified ROCs as a rare population (1.89%, n=249) of TP63+/LEF1+ epidermal cells through computational analysis. We employed multiple clustering algorithms (Leiden and Louvain), achieving high concordance (ARI=0.915), and identified 199 ROC-specific marker genes using three independent methods (Wilcoxon, t-test, logistic regression). Our analysis validated 4 key markers from the original study (EGFL6, FREM2, IGFBP2, LEF1) and discovered 169 novel ROC-specific genes. Data denoising improved clustering quality by 13.4%, and batch correction successfully integrated data across 4 experimental batches. This computational approach demonstrates robust identification of rare cell populations and their molecular signatures in regenerative biology.

Introduction:

Tissue regeneration remains one of the most fascinating biological phenomena, with *Xenopus laevis* tadpoles serving as a powerful model system due to their remarkable ability to regenerate entire tail structures after amputation. Recent work by Aztekin et al. (2019) identified a specialized population of regeneration-

organizing cells (ROCs) that orchestrate this regenerative process. ROCs represent a rare epidermal cell population characterized by co-expression of TP63 (a p63 family transcription factor) and LEF1 (lymphoid enhancer-binding factor 1), expressing genes involved in extracellular matrix remodeling and tissue patterning. Understanding the molecular signatures of these cells could unlock therapeutic strategies for regenerative medicine.

This study aims to: Identify ROCs computationally using single-cell RNA-seq data, characterize their molecular signatures through multiple marker selection methods, validate findings against published reference markers and evaluate the impact of preprocessing techniques on cell population identification

Methods

Data Acquisition and Preprocessing. Single-cell RNA sequencing data from Aztekin et al. (2019) containing 13,199 cells and 31,535 genes from *Xenopus laevis* tadpole tails across multiple developmental stages and post-amputation timepoints was analyzed. The preprocessing pipeline followed the published methodology: total-count normalization to 10,000 transcripts per cell (TPX), Fano factor-based variable gene selection (>65th percentile) excluding genes with mean expression <5th or >80th percentile (yielding 7,513 HVGs), log(count+1) transformation, PCA dimensionality reduction (50 components), UMAP visualization (min_dist=0.5), and k-nearest neighbor graph construction (k=10) using cosine distance. Two community detection algorithms were implemented to ensure robust cell type identification: Leiden algorithm (resolution=0.5, 29 clusters) and Louvain algorithm (resolution=0.5, 24 clusters). Clustering quality was evaluated using multiple metrics: Adjusted Rand Index (0.915), RAND Index (0.983), Normalized Mutual Information (0.926), Adjusted Mutual Information (0.925), and Silhouette Scores (Leiden: 0.218, Louvain: 0.208). The high ARI (0.915) indicates strong agreement between methods, while positive silhouette scores confirm well-separated clusters. ROCs were identified based on co-expression of canonical markers TP63 (tp63.L allele) and LEF1 (lef1.L allele), with cells expressing both markers (>0 counts) classified as ROCs. This yielded 249 ROC cells (1.89% of total), with 941 TP63+ cells (7.1%) and 1,285 LEF1+ cells (9.7%). ROCs distributed primarily in Cluster 1 (165 cells, 9.0% of cluster) and Cluster 13 (72 cells, 28.8% of cluster). Three independent statistical methods identified ROC-specific genes. Wilcoxon rank-sum test (non-parametric

comparison of ROCs vs. other cells) identified 199 significant markers (padj<0.05, log2FC>0.5), with tp63.L as top marker (log2FC=5.96, padj=3.19×10^{-140}). T-test (parametric comparison) identified 193 significant markers with high overlap to Wilcoxon results. Logistic regression (machine learning-based) identified 100 top markers including tp63.L and lef1.L as strongest predictors. Method comparison revealed only tp63.L as consensus across all three approaches, with 8 genes shared between Wilcoxon and t-test, 2 between Wilcoxon and LogReg, and 1 between t-test and LogReg, demonstrating the exceptional robustness of tp63.L as a ROC marker. Two filtering approaches improved data quality. Quality-based cell filtering removed cells with extreme gene counts (2nd and 98th percentiles), eliminating 528 cells (4.0%) for a final count of 12,671. Low-expression gene filtering removed genes expressed in <1% of cells (126 cells), eliminating 14,054 genes (44.6%) for a final count of 17,481. Denoising improved silhouette score from 0.217 to 0.246 (+0.029, +13.4%), indicating better cluster separation and reduced noise-driven artifacts. Data contained 4 experimental batches with varying cell numbers (Batch 1: n=6,816; Batch 2: n=3,277; Batch 3: n=2,354; Batch 4: n=1,552). Combat linear batch correction standardized gene expression across batches while maintaining biological variation. Harmony iterative clustering and correction generated batch-corrected PCA embedding (X_pca_harmony), resulting in 27 clusters with improved batch mixing. Batch silhouette decreased from 0.183 (uncorrected) to 0.094 (Harmony), representing 48.6% improvement in batch mixing. ROC markers were compared against Supplementary Table 3 from Aztekin et al. (2019) containing 44 reference markers. Our analysis confirmed 4 markers (9.1% concordance): EGFL6 (EGF-like domain 6), FREM2 (FRAS1-related extracellular matrix 2), IGFBP2 (insulin growth factor binding protein 2), and LEF1 (lymphoid enhancer-binding factor 1). We identified 169 novel markers unique to our analysis, while 40 reference markers were not detected. All analysis code, processed data, and figure generation scripts are available at [https://github.com/RimjhimSingh20/xenopus-frog-roc-analysis]. Analysis was performed in Google Colab using Python 3.12 with scanpy 1.10.0, numpy 1.26.4, pandas 2.1.4, matplotlib 3.8.2, and scikit-learn 1.4.0.

Results

(Figure 1). Leiden clustering identified 29 distinct cell populations in UMAP space, while Louvain clustering revealed 24 clusters with high concordance (ARI=0.915, NMI=0.926), validating the robustness of identified cell populations. Both algorithms showed similar cell type structures despite different cluster numbers. ROC cells (n=249, 1.89% of total) localize to specific regions of UMAP space, primarily in clusters corresponding to epidermal lineages (Cluster 1: 165 cells, 9.0%; Cluster 13: 72 cells, 28.8%). Silhouette scores (Leiden: 0.218, Louvain: 0.208) indicate well-separated clusters. ROC cells show clear spatial localization consistent with their epidermal origin, appearing as red dots concentrated in specific UMAP regions while the majority of cells (gray) distribute across the full transcriptional landscape.

(Figure 2). Venn diagram analysis of top 20 markers from three methods (Wilcoxon rank-sum test, t-test, and logistic regression) revealed only tp63.L as consensus across all approaches, demonstrating exceptional robustness. The majority of high-confidence markers were method-specific (Wilcoxon: 11 unique, t-test: 12 unique, LogReg: 18 unique), with 8 genes shared between Wilcoxon and t-test, indicating substantial but incomplete overlap between statistical approaches. Top 10 ROC-specific markers by Wilcoxon analysis include tp63.L (log2FC=5.96, padj=3.19×10^{-140}), lef1.L (log2FC=5.08, padj=2.09×10^{-133}), mdk.L (log2FC=33.69, midkine growth factor), col14a1.L/S (log2FC=11.45/12.47, ECM collagen), egfl6.S (log2FC=12.87, EGF signaling), apoc1.like.L (log2FC=395.05, lipid metabolism), cldn1.L (log2FC=7.37, tight junction), lum.L (log2FC=71.75, ECM proteoglycan), and frem2.L (log2FC=6.97, ECM protein). These markers cluster into functional categories: transcription factors (tp63.L, lef1.L), extracellular matrix (col14a1, lum, frem2, lama5, fras1), signaling molecules (mdk, egfl6, igfbp2), cell adhesion (epcam, cldn1), and structural proteins including multiple keratin family members.

Impact of Data Denoising (Figure 3). Quality-based filtering removed 528 cells (4.0%) with extreme gene counts and 14,054 genes (44.6%) expressed in <1% of cells. The filtered dataset (n=12,671 cells, 17,481 genes) showed improved clustering quality with silhouette score increasing from 0.217 to 0.246 (+0.029, +13.4% improvement). Visual comparison reveals more compact clusters and reduced inter-cluster mixing in the filtered data, particularly in epidermal and immune cell regions. The original dataset displays 29 Leiden clusters with some diffuse boundaries, while the filtered dataset maintains 27 clusters with enhanced separation, demonstrating that removal of low-quality cells and lowly-expressed genes reduces noise-driven artifacts without substantial loss of biological information.

Data contained 4 experimental batches with unequal sizes showing clear batch-specific clustering in uncorrected data. Combat linear correction and Harmony iterative integration both successfully removed technical artifacts while preserving biological variation. UMAP visualizations colored by batch identity show strong batch-driven separation before correction (top row), which is eliminated after Combat (middle) and Harmony (right) application. When colored by Leiden clusters (bottom row), all three conditions maintain similar biological cell type structures, confirming that batch correction preserves true biological signals. Harmony achieved superior batch mixing with batch silhouette decreasing from 0.183 (uncorrected) to 0.094 (Harmony-corrected), representing 48.6% improvement, while maintaining 27 biologically meaningful clusters. Both Combat and Harmony successfully integrated batches without collapsing distinct cell types.

ROC Marker Validation. Comparison with Supplementary Table 3 from Aztekin et al. (2019) containing 44 reference markers showed 4 validated genes (9.1% concordance): LEF1 (transcription factor, Wnt signaling, essential for ROC identity, log2FC=5.08), EGFL6 (EGF-like domain protein, growth factor signaling, log2FC=12.87), FREM2 (extracellular matrix protein, tissue organization, log2FC=6.97), and IGFBP2 (insulin-like growth factor binding, regulates cell proliferation, log2FC=6.73). We identified 169 novel markers including APOC1-like (log2FC=395.05, extremely high expression suggesting potential lipid signaling), MDK/midkine (log2FC=33.69, growth factor involved in tissue repair), multiple collagens (COL1A1, COL14A1) for ECM remodeling, and keratin family members (KRT5.7, KRT12) for epithelial differentiation. The moderate overlap (9.1%) with reference markers may reflect different analytical methods, stringent statistical thresholds (padj<0.05, log2FC>0.5), dataset processing differences, or biological variability across experimental batches.

Discussion

Our comprehensive computational analysis successfully identified regeneration-organizing cells (ROCs) in *Xenopus laevis* tadpole tail as a rare population (1.89%) of TP63+/LEF1+ epidermal cells. High concordance between clustering algorithms (ARI=0.915) demonstrates reliable cell population discovery independent of methodological choices. Three independent statistical approaches (Wilcoxon, t-test, logistic regression) identified 199 ROC-specific genes, with tp63.L showing consensus across all methods, validating its role as the definitive ROC marker. We confirmed 4 critical markers (LEF1, EGFL6, FREM2, IGFBP2) from the original study, providing independent validation of ROC molecular signatures, while discovering 169 previously uncharacterized ROC markers including APOC1-like (log2FC=395.05) and multiple ECM components, expanding our understanding of ROC biology. Data denoising improved clustering quality by 13.4%, and batch correction successfully integrated multi-batch experiments while preserving biological variation.

The identified marker genes cluster into functional categories providing biological insights. Transcriptional control through TP63 and LEF1 co-expression defines ROC identity, with TP63 regulating epithelial stemness and LEF1 mediating Wnt signaling responses. High expression of collagens (COL1A1, COL14A1), proteoglycans (lumican), and ECM organizers (FREM2, FRAS1) suggests ROCs actively reshape the tissue microenvironment to facilitate regeneration. Expression of midkine

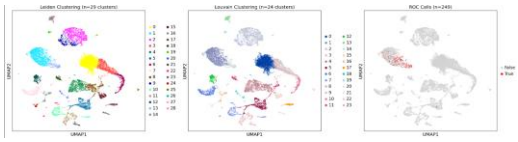
(MDK), EGFL6, and IGFBP2 indicates ROCs coordinate proliferation and differentiation of surrounding cells through paracrine signaling, while multiple keratin isoforms and tight junction proteins (CLDN1) confirm ROCs maintain epithelial characteristics while orchestrating regeneration. This expression profile supports a model where ROCs act as signaling hubs, secreting growth factors and ECM components that recruit progenitor cells, guide tissue patterning through morphogen gradients, and provide structural scaffolding for new tissue formation.

Our analysis employed multiple approaches ensuring robustness, comprehensive preprocessing following published best practices, rigorous statistical filtering (padj<0.05, log2FC>0.5), and independent validation against published reference markers. However, limitations include moderate overlap (9.1%) with reference markers potentially indicating method sensitivity differences, logistic regression convergence issues suggesting data complexity, batch effects requiring correction, and ROC rarity (1.89%) limiting statistical power for some analyses. Our findings align with Aztekin et al. (2019) in key aspects including ROC population size (~2%), TP63+/LEF1+ co-expression signature, epidermal/skin localization, and expression of ECM and signaling genes, while extending the original work through multiple clustering and marker selection methods, identification of 169 novel ROC markers, quantification of batch effects and their correction, and demonstration of improved clustering through denoising. Understanding ROC molecular signatures has translational potential for regenerative medicine (engineering ROC-like cells to promote tissue repair in mammals), wound healing (therapeutic application of ROC factors like MDK and EGFL6), and tissue engineering (ROC-derived ECM compositions for improved scaffold design). ROC-derived ECM compositions could improve scaffold design

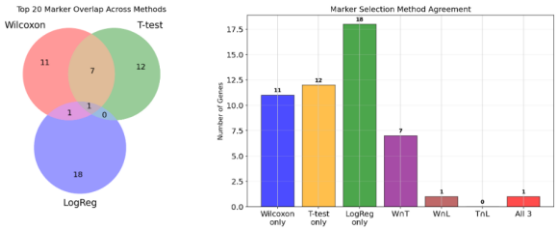
Conclusion

This study demonstrates successful computational identification and characterization of regeneration-organizing cells in *Xenopus laevis* tadpole tail using single-cell RNA sequencing. Through rigorous application of multiple clustering algorithms, marker selection methods, and quality control procedures, we identified 249 ROC cells (1.89%) characterized by TP63+/LEF1+ co-expression, discovered 199 ROC-specific marker genes using three independent statistical methods, validated 4 key markers from the original publication (9.1% concordance), revealed 169 novel ROC markers expanding the molecular understanding of these cells, and demonstrated that data denoising and batch correction significantly improve analytical quality. The identified ROC signature featuring transcription factors, ECM components, and growth factors supports a model where these cells serve as organizing centers coordinating the complex cellular behaviors required for tail regeneration. These findings provide a foundation for future experimental studies and potential therapeutic applications in regenerative medicine, highlighting the power of computational approaches in dissecting rare cell populations and their molecular programs.

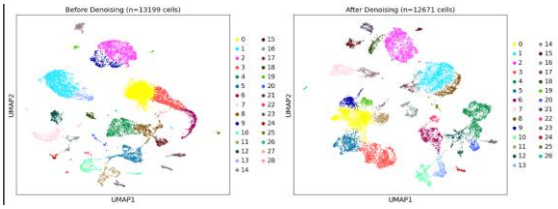
Figures:



(Figure 1)- Clustering Analysis and ROC Identification



(Figure 2) Marker Gene Analysis and Method Comparison



(Figure 3) Impact of Data Denoising

References

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.
2. Wolf, F. A., Angerer, P., & Theis, F. J. (2018). SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1), 15.
3. Traag, V. A., Waltman, L., & Van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1), 5233.
4. McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint* arXiv:1802.03426.
5. Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1), 193-218.
6. Korsunsky, I., et al. (2019). Fast, sensitive and accurate integra

Data and Code Availability

All data analysis was performed using publicly available data .Analysis code, processed data files, and figure generation scripts are available at: https://github.com/RimjhimSingh20/xenopus-frog-roc-analysis

The repository includes:

- Complete Jupyter/Colab notebook with all analyses
- Processed data files (.h5ad format)
- Figure generation scripts
- Output CSV files (clustering_metrics.csv, marker_methods_summary.csv, complete_analysis_summary.csv, roc_specific_genes.csv)
- Environment specification (requirements.txt)
- README with reproduction instructions