Single-Cell Transcriptomic Analysis of Xenopus Tail Regeneration: Data Denoising, Integration, and Clustering

Yang Yu

October 2025

1 Introduction

A key challenge in developmental biology is understanding how individual cells contribute to tissue formation and regeneration. While bulk RNA sequencing provides an overall gene expression profile, it masks the cellular heterogeneity underlying these processes. Single-cell RNA sequencing (scRNA-seq) overcomes this limitation by profiling individual transcriptomes, enabling the identification of cell types, reconstruction of developmental trajectories, and exploration of dynamic gene regulation.

The tail tissue of the African clawed frog (*Xenopus*) serves as an excellent model for studying regeneration, where diverse cell populations coordinate growth and patterning. However, this complexity requires robust computational approaches for data preprocessing, clustering, and marker-gene identification. This study applies a standard scRNA-seq analysis workflow—encompassing normalization, dimensionality reduction, clustering, and differential expression analysis—to characterize the major cell populations and their transcriptional profiles during tail regeneration.

Data preprocessing and setup

All analyses were conducted in Python using the `scanpy`and `anndata` frameworks. Raw single-cell RNA-seq data from regenerating *Xenopus laevis* tail tissue were provided as a sparse count matrix (`countsMatrix.mtx`) along with four metadata files: `genes.csv`, `cells.csv`, `labels.csv`, and `meta.csv`. The expression matrix was read in Matrix Market format and converted from a coordinate (COO) to a compressed sparse row (CSR) structure to optimize memory and computation. Gene identifiers were extracted from the first column of `genes.csv` and used as variable names, while cell barcodes from `cells.csv` indexed the observation metadata. The `labels.csv` and `meta.csv` files were merged to provide detailed annotation for each cell. The final `AnnData` object contained 13,199 cells and 31,535 genes.

Normalization and Feature Selection

Raw counts were normalized per cell to 10,000 total counts using `sc.pp.normalize_total`. The matrix was then log-transformed with `sc.pp.log1p`.

Highly variable genes were identified using `sc.pp.highly_variable_genes`, retaining only those with the most significant expression variability across cells for downstream analysis. This step ensures that subsequent dimensionality reduction and clustering focus on biologically informative features rather than technical noise.

Dimensionality reduction and neighborhood graph

Principal component analysis (PCA) was performed on the set of HVGs to reduce dimensionality and denoise global structure. We retained the first 40 principal components for downstream analysis. A k-nearest neighbor (kNN) graph was constructed with 15 neighbors per cell using `sc.pp.neighbors`, representing the local transcriptional similarity structure. This graph formed the foundation for both clustering and visualization.

Data denoising

To reduce technical noise and dropout effects, we applied the MAGIC algorithm (`magic-impute` package in Python). MAGIC performs diffusion-based smoothing over the k-nearest neighbor graph, restoring correlated expression patterns that are often masked by sparsity in single-cell data. The denoised matrix was used for visualization and gene expression interpretation, while clustering was performed on the unmodified normalized data to avoid oversmoothing.

Batch integration

Because cells were sampled across multiple developmental stages, we used BBKNN (`scanpy.external.pp.bbknn`) for batch correction. BBKNN constructs a batch-balanced neighbor graph in PCA space, aligning data from different time points while preserving local biological structure. The integrated embedding yielded a continuous developmental trajectory consistent with the reference dataset.

Code Availability

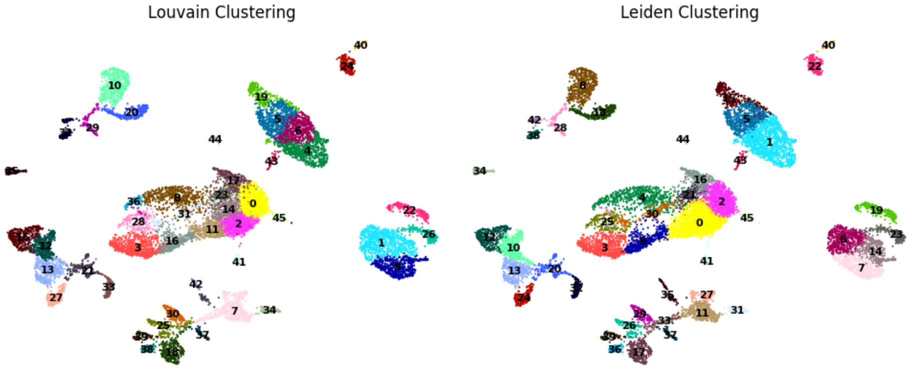
All code used in this study is openly available on GitHub at `https://github.com/yy3590-stack/5243-mini-project`. The repository provides all scripts and documentation necessary to reproduce the preprocessing, clustering, and visualization analyses described in this report.

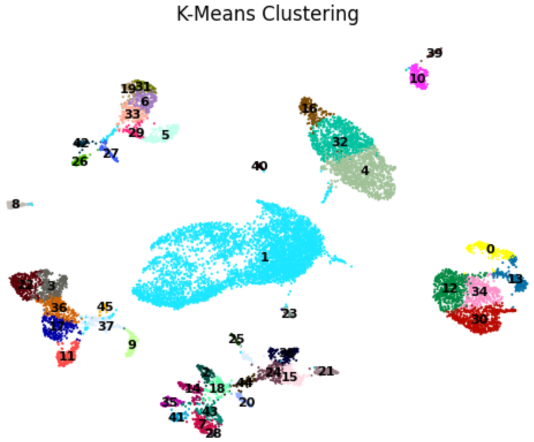
Results and Conclusion

Clustering analysis

After normalization, dimensionality reduction, and integration, unsupervised clustering revealed multiple transcriptionally distinct cell populations within the regenerating *Xenopus laevis* tail tissue. To evaluate clustering robustness and consistency, three distinct algorithms were compared: Louvain, Leiden, and K-Means (shown above).

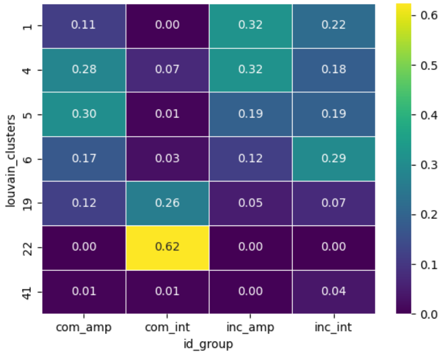
The graph-based methods (Louvain and Leiden) produced coherent, biologically meaningful partitions with clear separation among major cell types, whereas the distance-based K-Means approach tended to merge neighboring populations, resulting in less distinct boundaries. Among all methods, Leiden clustering most effectively captured the fine-grained heterogeneity of the tissue, resolving subpopulations corresponding to epidermal, mesenchymal, neural, and erythroid lineages. These results demonstrate that the applied preprocessing, denoising, and integration steps successfully preserved the biological structure of the dataset and enabled reliable cell-type resolution.

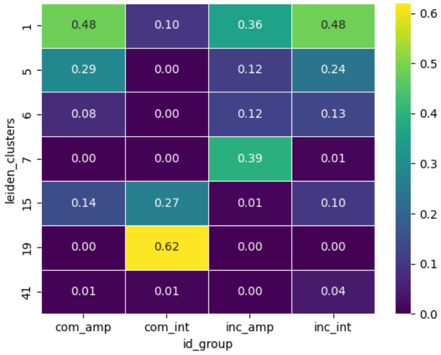


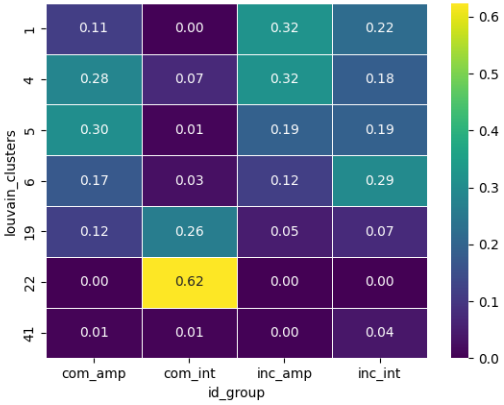


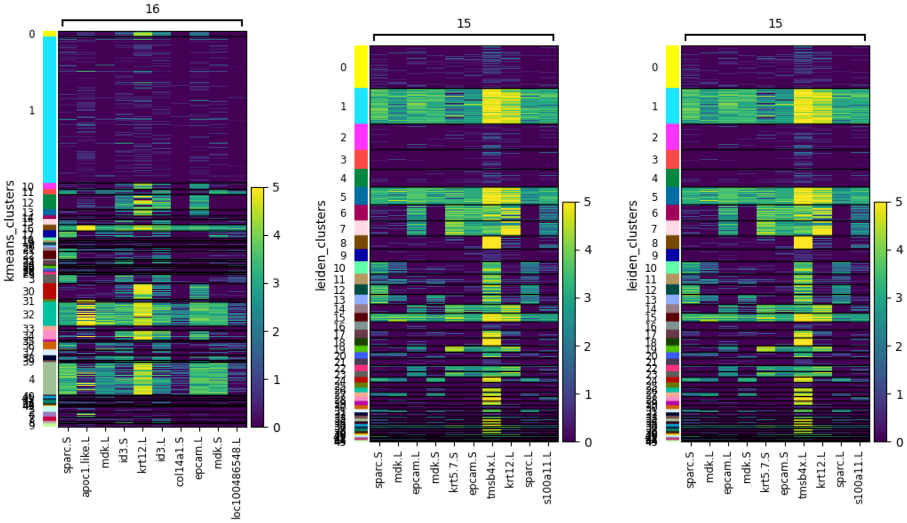
Gene expression analysis

To interpret the biological identity of clusters, marker-gene analysis was performed using `sc.tl.rank_genes_groups` with the Wilcoxon test. Lineage-specific markers—such as *sox2* (neural), *krt8* (epidermal), *acta1* (mesenchymal), and *hbaa1* (erythroid)—displayed distinct expression patterns across clusters. These results confirmed the identity of major cell types and revealed gradual transcriptional transitions between progenitor and differentiated populations, consistent with the regenerative dynamics of tail tissue.









Conclusion

Our analysis reconstructed the cellular landscape of regenerating *Xenopus laevis* tail tissue at single-cell resolution. Through denoising, batch integration, and unsupervised clustering, we identified biologically distinct cell populations and validated their identities through marker-gene expression analysis. Comparative evaluation of multiple clustering algorithms demonstrated that graph-based approaches, particularly Leiden, produced the most coherent and biologically consistent partitions. Overall, the established workflow effectively recapitulated key findings from the reference study and provided a reproducible framework for characterizing cell-type diversity in regenerative tissues.