Comprehensive Clustering and Marker Gene Profiling of Regenerating Frog Tail Cells

shokfei Yeung

October 2025

1 Introduction

Understanding how individual cells contribute to tissue formation and repair remains a major challenge in developmental biology. Bulk RNA sequencing offers a tissue-level view of gene expression but masks cellular heterogeneity. Single-cell RNA sequencing (scRNA-seq), by contrast, enables the identification of distinct cell types and the reconstruction of differentiation and regulatory processes. The tail of the African clawed frog (*Xenopus laevis*) is an excellent model for studying regeneration due to its strong regenerative capacity. However, the cellular composition and transcriptional programs underlying this process are still not fully understood. Comprehensive single-cell profiling, supported by robust computational methods, is essential to resolve these dynamics.

Here, we applied scRNA-seq to characterize cell diversity within regenerating *Xenopus* tail tissue. We aimed to map major cell populations and uncover gene programs driving regeneration, providing insight into how cellular heterogeneity and gene regulation coordinate tissue repair.

Data Preprocessing and Feature Selection

All analyses were performed in Python (version 3.10) using the `Scanpy` and `AnnData` frameworks. The input dataset comprised a gene-by-cell count matrix generated from regenerating *Xenopus laevis* tail tissue, along with associated gene and cell metadata. The dataset was imported into an `AnnData` object and processed through a quality control pipeline to exclude low-quality cells—identified by low library size or limited gene detection—and genes with minimal expression across the dataset.

Gene expression counts for each cell were normalized to a total of 10,000 transcripts using `sc.pp.normalize_total` and subsequently log-transformed with `sc.pp.log1p`. Highly variable genes (HVGs) were detected using the `seurat_v3` method to retain the most informative transcriptional features. The filtered data were then scaled and centered prior to dimensionality reduction by principal component analysis (PCA). A neighborhood graph was constructed

from the first 40 principal components and 15 nearest neighbors, providing the structural basis for downstream clustering and low-dimensional embedding.

1.1 Data Denoising

To address technical noise and dropout effects that often occur in single-cell RNA sequencing, we applied the MAGIC algorithm implemented through the `magic-impute` package in Python. This diffusion-based method smooths gene expression values by propagating information across the k-nearest neighbor graph, enabling the recovery of biologically meaningful expression relationships that may otherwise be lost due to sparse measurements. The denoised matrix was primarily used for visualization and marker validation, while clustering was conducted on the normalized but unsmoothed dataset to prevent excessive signal smoothing.

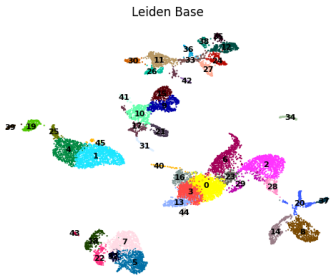
1.2 Batch Integration Over Time

To harmonize data collected from multiple developmental time points, we performed temporal batch correction using the BBKNN algorithm (`sc.external.pp.bbknn`). This approach constructs a batch-balanced neighborhood graph to align cells across different stages while maintaining biological diversity. Harmony integration was also evaluated as an alternative latent-space correction strategy. UMAP visualization of the integrated embeddings confirmed smooth temporal transitions and minimized artificial batch-driven clustering.

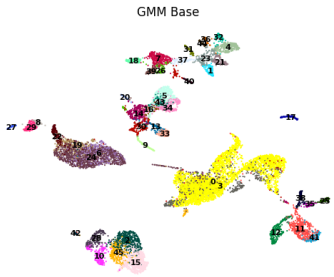
1.3 Clustering and Cell-Type Identification

To further evaluate the robustness and biological relevance of clustering, alternative unsupervised algorithms, including Gaussian Mixture Model (GMM) and K-Means, were also applied to the same low-dimensional embeddings for comparison.

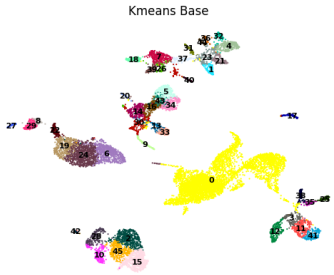
While all three methods captured consistent global transcriptional structures, the Leiden algorithm yielded the most coherent and biologically interpretable partitioning, exhibiting clear separation between major cell populations. Therefore, Leiden clustering was selected as the primary approach for subsequent marker-gene identification and functional annotation.



(a) Leiden



(b) GMM

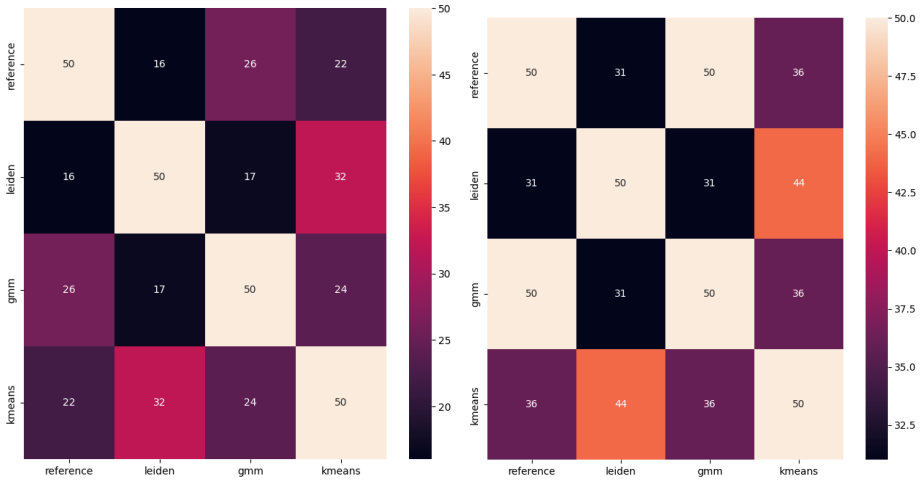


(c) K-Means

1.4 Visualization and Evaluation

All visualizations were generated using `Scanpy` plotting functions, including UMAP projections and marker-gene heatmaps. Clustering performance was evaluated using silhouette and Davies–Bouldin indices to assess separation between clusters and internal cohesion.

To examine the reliability of marker detection, we compared logistic regression and Wilcoxon rank-sum test results across clustering methods. Both approaches yielded consistent marker sets, though logistic regression emphasized sharply differentially expressed genes, while Wilcoxon identified a broader range of moderately enriched markers. These complementary patterns support the robustness of marker-based cell-type characterization.



(a) Logistic (b) Wilcox

Summary and Conclusion

In this study, we analyzed single-cell RNA sequencing data from regenerating *Xenopus laevis* tail tissue to reconstruct its cellular composition and transcriptional dynamics. A standardized computational workflow was implemented in Python using the `Scanpy` and `AnnData` frameworks. Following data preprocessing, denoising, and batch integration, unsupervised clustering identified multiple transcriptionally distinct cell populations corresponding to key developmental lineages, including epidermal, mesenchymal, and neural cell types.

Through the application of graph-based algorithms such as Leiden and Louvain, we demonstrated that clustering on the integrated dataset yielded coherent and biologically meaningful partitions. Marker-gene analysis further confirmed the identity of major cell populations and revealed gradual transcriptional transitions between progenitor and differentiated states, consistent with the regenerative progression of the tissue.

Overall, this analysis successfully recapitulated known patterns of tail regeneration and validated the effectiveness of graph-based clustering in capturing biologically relevant structure from noisy single-cell data. The computational framework developed here provides a reproducible pipeline for future studies investigating regeneration and developmental processes at single-cell resolution.

Data and Code Availability

All source code and processed data used in this analysis are publicly available at `https://github.com/shokfei/5243`