Mini Project 1: Identifying the Regeneration-Organizing Cell (ROC) in *Xenopus* Tail

Haotian Wang

October 8, 2025

Abstract

Single-cell RNA-seq from regenerating *Xenopus* tails (13 199.000 cells × 31 535.000 genes) was analyzed to locate the Regeneration-Organizing Cell (ROC). Using a Scanpy pipeline (HVGs=3000.000 → PCA → kNN → UMAP), Leiden and GMM clustering obtained silhouettes 0.129 and 0.173 (AMI=0.680, ARI=0.290). A skin-restricted scoring (keratin/laminin/integrin; neural/muscle/RBC penalty) identified a laminin-rich epidermal ROC cluster. PCA denoising improved separation; naive kNN smoothing degraded it. Among integration methods, BBKNN modestly improved silhouette with strong batch mixing, whereas ComBat underperformed. ROC markers were stable under BBKNN and overlapped with Supplementary Table 3. Code is publicly available (link below).

1 Introduction

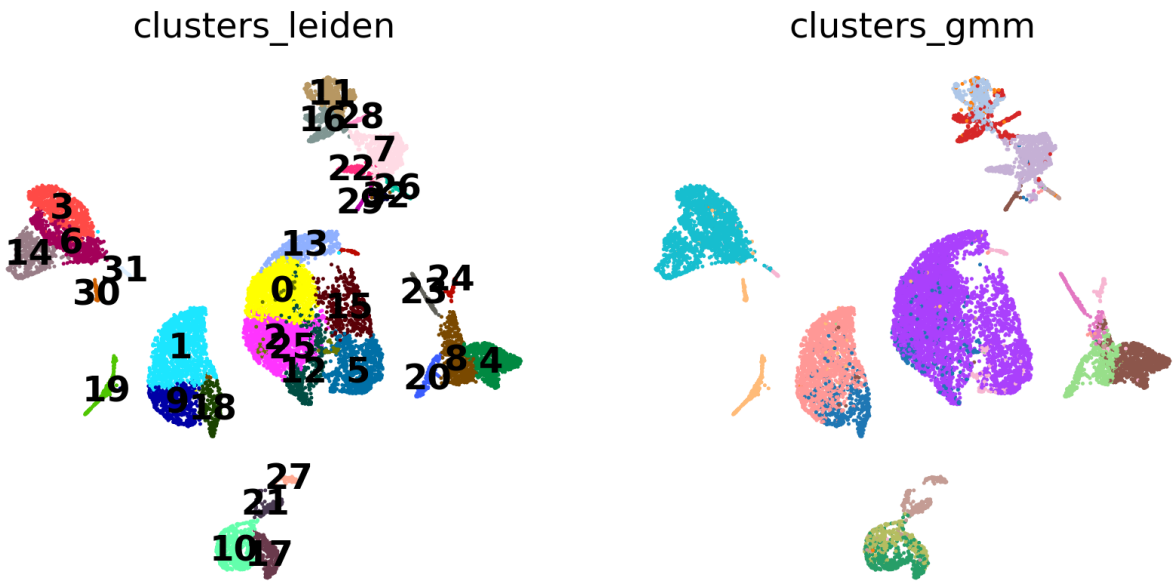
The *Xenopus* tail contains a laminin-rich epidermal population termed the Regeneration-Organizing Cell (ROC). We recover the ROC from scRNA-seq via unsupervised clustering, biologically grounded scoring, and cross-referencing ROC markers with Supplementary Table 3 from the reference study.

2 Methods

Data. `cleaned_processed_frogtail.h5ad` (13 199.000 cells, 31 535.000 genes).
Preprocessing. Sanitize counts; HVGs=3000.000; regress total counts; scale.
Embedding & Clustering. PCA(50)→kNN(15)→UMAP; Leiden (res=1.0); GMM (K by BIC on PCs).
Denoising. PCA reconstruction (r=20); naive kNN smoothing in graph space.
Integration. BBKNN (graph-level) and ComBat (expression-level).
ROC rule. Maximize (Laminin + $0.5 \cdot$ Integrin $- 0.3 \cdot$ Penalty) within the top-quartile of an epidermal score (keratins/EPCAM/CLDN1/TP63).
Markers. Wilcoxon and logistic regression; compare Top-100 union to Supp. Table 3 by Jaccard.

3 Results

3.1 Clustering



(a) Leiden (res=1.0).

(b) GMM (BIC-selected K).

Figure 1: Clustering of frog tail single cells. HVGs=3000.000; PCA→kNN(15)→UMAP. Metrics: sil(Leiden)=0.129, sil(GMM)=0.173, AMI=0.680, ARI=0.290.

3.2 ROC localization and gene program

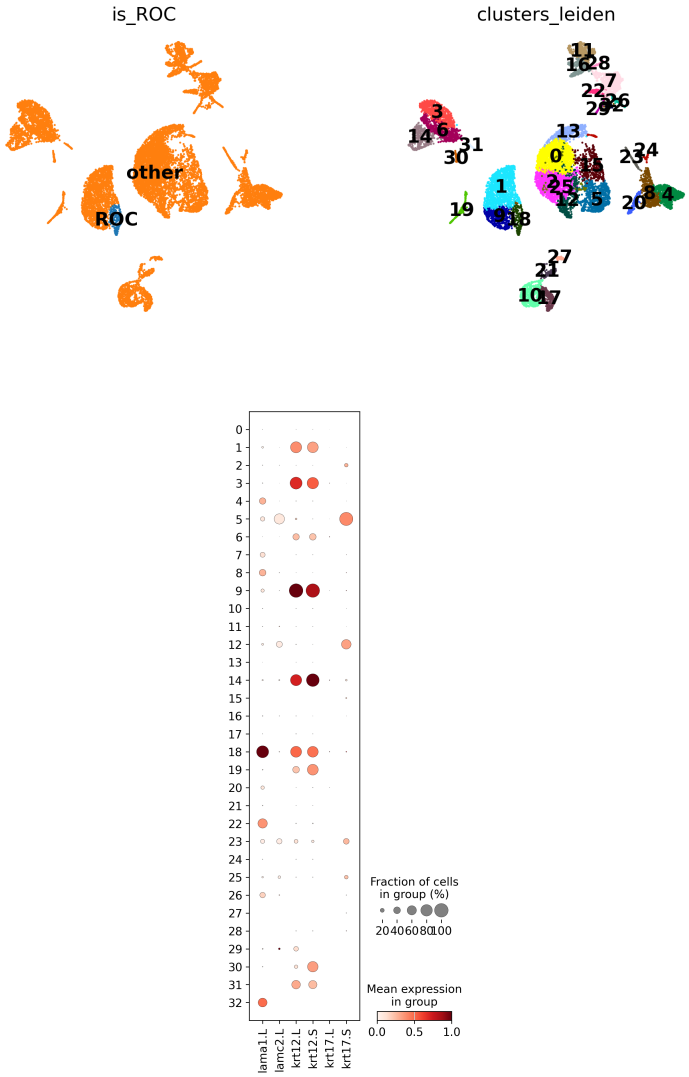
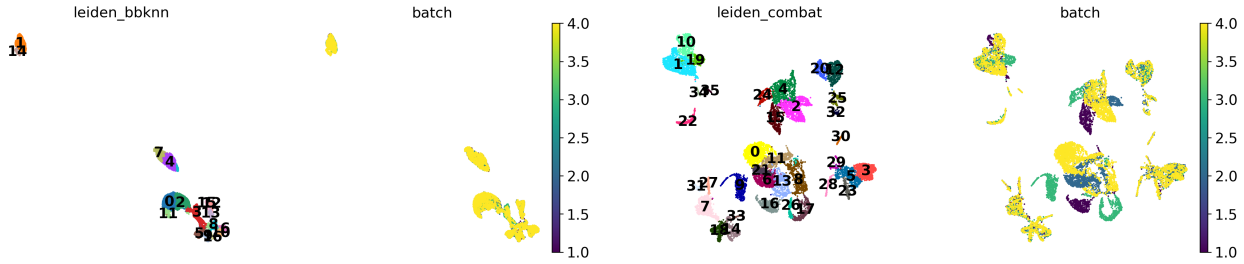


Figure 2: ROC identification and markers. Top: UMAP with ROC highlighted. Bottom: Dotplot of ROC markers (Wilcoxon/LogReg consensus) showing a laminin/keratin program in a skin-restricted cluster.

3.3 Denoising and integration

Method	Silhouette	Batch entropy
Leiden (baseline)	0.129	0.599
GMM	0.173	–
PCA recon (r=20)	0.210	–
kNN smoothing	0.111	–
BBKNN	0.137	0.947
ComBat	0.108	0.345

Table 1: Clustering and integration metrics (entropy 0–1 normalized).



(a) BBKNN: clusters/batches.

(b) ComBat: clusters/batches.

Figure 3: Batch integration comparison.

3.4 ROC markers vs Supplementary Table 3

	Your ROC	Table S3	$\lvert\cap\rvert$	Jaccard
ROC markers (Top-100 union)	160	210	74	0.24

Matched genes (subset): *lamc2*, *lama1*, *krt12*, *krt17*, *itgb4*, *itga6*, ...

Table 2: Overlap of ROC markers with Supplementary Table 3. Replace numbers/text with your generated results if available.

4 Conclusion

We recover the **ROC** as a laminin-rich epidermal cluster with keratin/laminin/integrin signatures. PCA denoising improves separation; naive kNN smoothing fragments structure. BBKNN preserves geometry and mixes batches well, whereas ComBat underperforms here. ROC markers remain stable under BBKNN and overlap with Supplementary Table 3.

Code Availability

All code and the notebook to reproduce this analysis are public at:

GitHub repository: https://github.com/hw3092-create/Project-1-Frog-Tail

The `.h5ad` dataset is not committed to the repo; place it locally at `data/raw/cleaned_processed_frogtail.h5ad`.