Single-Cell Analysis of Regeneration-Organizing Cells in Xenopus Tail

Jiayi Dan UNI: jd4243

Abstract

The regeneration-organizing cell (ROC) displays a unique gene expression profile and plays a pivotal role in the regeneration of Xenopus tail. In this study, we applied single-cell analyses to systematically investigate the unique gene expression profiles of this newly identified regeneration-organizing cell. Clustering and visualization revealed distinct cellular populations, with the ROC cluster clearly separated, confirming its transcriptional uniqueness. Clustering quality was further validated through multiple evaluation metrics. Using logistic regression and the Wilcoxon rank-sum test, we identified marker genes in ROCs. While genes such as apoc1.like.L are highly expressed across multiple cell types, pltp.S and others serve as specific markers of ROCs. We applied multiple denoising and batch integration techniques to assess the robustness of our findings. Notably, Denoising with MAGIC and Harmony correction improved clustering metrics, while the top ROC marker genes remained consistent with previous results,which reinforce the reliability of our pipeline. Code is available at https://github.com/DanJiayi/STAT5243_Project1.

Keywords: Regeneration-organizing cell, Single-cell analyses, Xenopus tail regeneration

1 Introduction

Xenopus laevis tadpoles exhibit remarkable regenerative capacity, particularly in their tails. Recent studies have identified a novel epidermal cell type, the regeneration-organizing cell (ROC), which displays a unique gene expression profile and plays a central role in orchestrating tissue regeneration. Previous work [1] using single-cell transcriptomics has identified the existence of ROCs and provided a mechanistic understanding of the initiation and organization of tail regeneration. However, systematic pipelines for single-cell analysis of ROCs have not yet been fully established. In this study, we aim to develop a ROC-focused single-cell analysis framework that integrates clustering with gene expression profiling, and further evaluate the reliability of our findings through extensive denoising and batch integration techniques.

2 Methods

Data preprocessing. The data with use is consistent with [1]. Raw count matrices were normalized to the same total counts per cell, and log-transformed to stabilize variance. Genes expressed in fewer than 3 cells and cells with fewer than 200 detected genes were removed. Highly variable genes were then identified (top 2,300 genes ranked by dispersion), and used for downstream tasks and analyses.

Clustering Analysis. To perform clustering, we first reduced the dimensionality of the data using principal component analysis (PCA). Based on the top 30 PCs, we constructed a k-nearest neighbor graph with k = 15. Using this graph, we performed community detection with both the Louvain and Leiden algorithms, generating alternative cluster partitions of the data. Parameters was adjusted to yield a number of clusters consistent with the 46 known clusters in the dataset, and the clustering results was visualized with UMAP. To quantitatively assess the clustering quality, we compared the inferred clusters against manually annotated labels using multiple metrics, including the adjusted Rand index (ARI), Rand index, silhouette score, and Calinski–Harabasz score.

Marker selection and Gene Ontology analysis. We applied differential expression analysis to identify marker genes for ROC. Two complementary statistical approaches were employed: logistic regression and the Wilcoxon rank-sum test. For each cluster, the top 50 genes ranked by each method were extracted, and we compared their expression between ROC and other clusters to select ROC-specific marker genes. Functional enrichment of the selected marker genes was performed using g:Profiler, restricted to Gene Ontology biological process, molecular function, and cellular component categories.

Data denoising. To mitigate technical noise and enhance the biological signal, we applied two complementary denoising approaches. First, we used scVI, a variational autoencoder–based framework, to learn a latent representation of the data and generate normalized denoised expression values (library size set to 1e4, trained for 100 epochs). Second, we applied MAGIC, a diffusion-based imputation method, to recover gene–gene relationships and smooth expression profiles. Both denoised datasets were saved for downstream analyses. After denoising, we re-performed clustering and marker selection, and compared the results with the original analysis to validate the improvement in clustering performance and the robustness of the identified marker genes.

Batch integration. For batch effect correction, we used two complementary approaches. Harmony integrates single-cell data by iteratively adjusting the principal component embeddings to remove batch-specific variation while preserving biological structure. Specifically, PCA was first performed to extract the top 50 components, and Harmony was applied using batch labels as covariates, yielding corrected embeddings. As an alternative, we applied BBKNN (Batch Balanced KNN), which modifies the neighborhood graph construction to ensure balanced representation of cells from each batch. Similarly, We re-performed clustering and tested their effectiveness by comparing with the previous results.

Code availability. Code supporting this study is available at https://github.com/DanJiayi/STAT5243_Project1.

3 Results

3.1 Clustering Analysis and Visualization

3.1.1 Visualization of clustering results

To evaluate the clustering performance, we applied both the Louvain and Leiden algorithms after PCA-based dimensionality reduction (all clustering in this study was tuned to match the number of clusters in the reference (46)). The UMAP visualization (Fig. 1) shows that both methods separated the cells into distinct clusters. The strong separation among clusters suggests that cell types are characterized by distinct gene expression signatures, with certain lineage-specific markers driving the divergence. For example, immune-related clusters were well distinguished from epithelial and mesenchymal populations, reflecting the underlying functional heterogeneity.

The concordance with the reference cluster annotations (Fig. 2) further supports that these clustering approaches effectively recover biologically relevant cellular identities. Notably, both clustering methods successfully distinguished ROCs (highlighted by the red box in the figure), underscoring the unique gene expression characteristics of these cells.

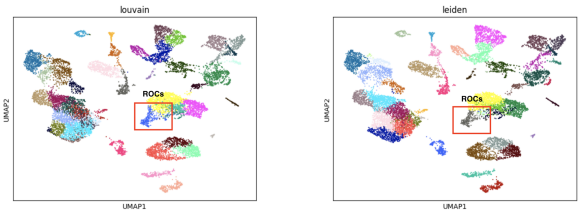

Fig. 1: Clustering results.

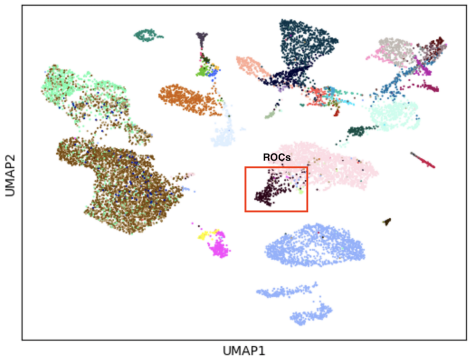

Fig. 2: Reference cluster annotations.

3.1.2 Clustering evaluation metrics

To quantitatively evaluate clustering performance, we compared the Louvain and Leiden algorithms using four common metrics: Adjusted Rand Index (ARI), Rand Index, Silhouette score, and Calinski–Harabasz score. As shown in Table 1, both methods achieved comparable results. Louvain and Leiden yielded similar ARI values and high Rand Index values (0.915), indicating a reasonable consistency with the reference labels. Leiden slightly outperformed Louvain in ARI and Silhouette score, suggesting that Leiden provides more coherent clusters. Interestingly, Louvain obtained a slightly higher CH score, reflecting that it may produce clusters with more distinct separation in terms of variance-based criteria. Overall, the results indicate comparable clustering performance between the two methods

Table 1: Clustering metrics.

Method	ARI	Rand Index	Silhouette	CH Score
Louvain	0.3056	0.9140	0.1977	1880
Leiden	0.3230	0.9151	0.2056	1867

3.2 Marker Selection and Gene Analysis

3.2.1 Marker Selection

We applied two complementary strategies: logistic regression and Wilcoxon rank-sum test, to identify marker genes of the ROC cluster. From the top 50 markers identified by each method, we observed a substantial overlap of 29 genes, highlighting a consistent marker signature across statistical approaches. When compared with the reference list provided in Supplementary Table 3 of [1], 8 of the logistic regression–derived markers and 5 of the Wilcoxon-derived markers were retained.

Using logistic-regression–based markers as an illustrative case, the five most highly expressed genes were apoc1.like.L, pltp.S, fn1.S, Xetrov90029035m.L, and id3.L. As shown in Fig, 3, apoc1.like.L, fn1.S, and id3.L also showed high expression in the Epidermis or other cell types, suggesting that they are not exclusive to ROCs. In contrast, pltp.S and Xetrov90029035m.L were specifically expressed in the ROC cluster. Importantly, both of these ROC-specific genes are consistent with the findings reported in Supplementary Table 3 of [1], with pltp.S also being independently identified by the Wilcoxon-based marker selection. These results highlight pltp.S and Xetrov90029035m.L as specific markers of ROCs.

3.2.2 Gene Ontology analysis

GO enrichment analysis revealed that the logistic-regression–based marker genes were strongly associated with extracellular matrix (ECM) organization. Among the most significant terms, "basement membrane" (GO:0005604, p = 4.30×10^{-10}) was enriched with 8 marker genes, highlighting a strong link between ROCs and basement membrane formation. Given the basement membrane's critical role in maintaining tissue architecture, providing structural support, and mediating cell adhesion, this finding suggests that ROCs may contribute specifically to ECM remodeling at epithelial or endothelial interfaces. Other enriched terms, such as "extracellular matrix," "extracellular region," and "collagen-containing extracellular matrix," further reinforce this ECM-associated signature.

3.3 Data denoising

We assessed the impact of scVI and MAGIC denoising on clustering performance. As shown in Table 3 and Table 4, both methods maintained ARI and Rand Index values close to original results, indicating that denoising preserved concordance with known labels. In contrast, MAGIC substantially increased the Silhouette and Calinski–Harabasz (CH) scores, suggesting that clusters became more compact and well-separated. These results demonstrate that denoising, particularly with MAGIC, can enhance the effectiveness of gene expression

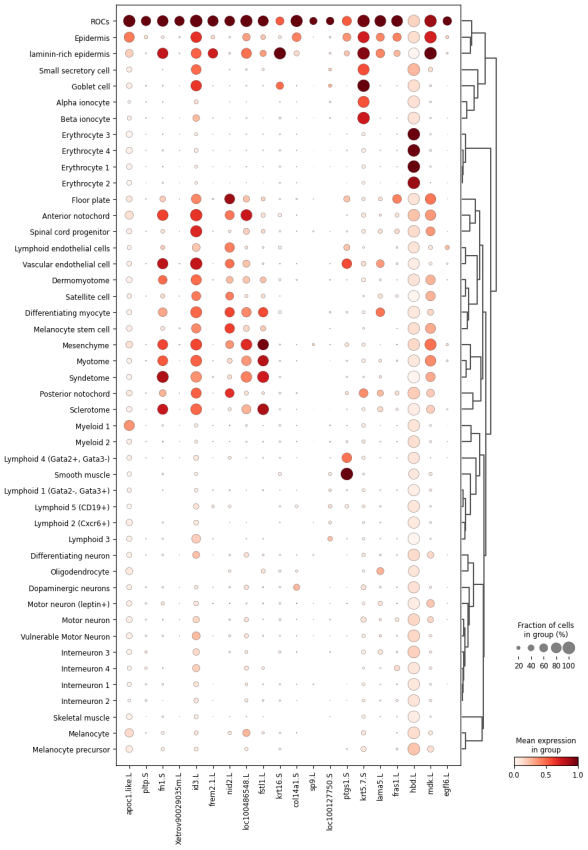

Fig. 3: Expression of top ROC marker genes.

Table 2: Top 10 significantly enriched GO terms from ROC marker genes.

Name	p-value	Intersection
Basement membrane	4.30×10^{-10}	8
Extracellular region	1.67×10^{-8}	25
Extracellular matrix	1.79×10^{-7}	11
External encapsulating structure	1.83×10^{-7}	11
Extracellular matrix structural constituent	1.86×10^{-7}	8
Extracellular space	5.58×10^{-7}	21
Collagen-containing extracellular matrix	2.36×10^{-6}	9
Structural molecule activity	1.06×10^{-5}	13
Anatomical structure morphogenesis	1.13×10^{-5}	19
Cell motility	1.01×10^{-4}	15

signals for downstream analyses while largely preserving label consistency.

Table 3: Clustering metrics - scVI

Method	ARI	Rand Index	Silhouette	CH Score
Louvain	0.3067	0.9152	0.2333	2959
Leiden	0.2594	0.9129	0.2479	2879

Table 4: Clustering metrics - MAGIC

Method	ARI	Rand Index	Silhouette	CH Score
Louvain	0.3077	0.9141	0.3801	5076
Leiden	0.2454	0.9110	0.3681	5147

After denoising the data using the two methods separately, we re-identified the marker genes for ROCs. Taking logistic regression–based markers as an example, although only 26 and 31 of the top 50 expressed genes overlapped with the previous results after each denoising method, the previously highest-expressed genes—apoc1.like.L, pltp.S, and fn1.S—remained highly expressed. Notably, pltp.S continued to be a ROC-specific marker. These findings confirm the robustness of our earlier conclusions and demonstrate that denoising has a substantial impact on marker identification, highlighting its necessity.

3.4 Batch integration

We further adopted two batch correction strategies, Harmony and BBKNN. As shown in Table 5 and Table 6, Harmony correction yielded ARI values slightly higher than the raw data (0.3195 vs. 0.3056 for Louvain; 0.3381 vs. 0.3230 for Leiden), suggesting modest improvement in alignment with the reference labels. In contrast, BBKNN markedly reduced the ARI (0.2350 for Louvain; 0.2688 for Leiden) and even produced negative Silhouette scores, indicating that the corrected embedding led to poorly separated clusters. Overall, batch integration with Harmony improved clustering performance by effectively eliminating batch-specific variation, thereby enhancing the biological signal in the data representation.

Table 5: Clustering metrics - Harmony.

Method	ARI	Rand Index	Silhouette	CH Score
Louvain	0.3195	0.9156	0.2005	2030
Leiden	0.3381	0.9170	0.2013	1964

Since batch integration methods such as Harmony and BBKNN operate on the latent embedding or neighborhood graph rather than altering the original expression matrix, the identification of marker genes remains unchanged. Therefore, we did not repeat the marker analysis after batch integration in this study.

Table 6: Clustering metrics - BBKNN.

Method	ARI	Rand Index	Silhouette	CH Score
Louvain	0.2350	0.9050	-0.1907	349
Leiden	0.2688	0.9055	-0.2298	285

4 Conlusion

In this work, we applied single-cell analyses to systematically investigate the unique gene expression profiles of regeneration-organizing cells (ROCs). We visually demonstrated the transcriptional uniqueness of ROCs using multiple clustering methods and visualization techniques, and validated the robustness of our results through various evaluation metrics. Using logistic regression and related approaches, we identified the top 50 ROC-expressed genes and analyzed their differential expression. The results of the two methods were highly consistent, revealing that genes such as pltp.S serve as specific markers of ROCs. GO enrichment analysis further indicated that these highly expressed genes are strongly associated with extracellular matrix organization. Finally, the application of denoising and batch integration techniques significantly improved clustering metrics while preserving the consistency of key marker genes, confirming the reliability and reproducibility of our pipeline and findings.

References

[1] Can Aztekin, TW Hiscock, JC Marioni, JB Gurdon, BD Simons, and Jerome Jullien. Identification of a regeneration-organizing cell in the xenopus tail. *Science*, 364(6441):653–658, 2019.

[2] Malte D Luecken, Scott Gigante, Daniel B Burkhardt, Robrecht Cannoodt, Daniel C Strobl, Nikolay S Markov, Luke Zappia, Giovanni Palla, Wesley Lewis, Daniel Dimitrov, et al. Defining and benchmarking open problems in single-cell analysis. *Nature Biotechnology*, pages 1–6, 2025.