Project 1: Find the Frog's Source of Regenerative Power

Xiaocong Xuan (xx2438)

October 5, 2024

1 Abstract

This study identified Regenerative Organizing Cells (ROCs) in the frog tail and characterized their unique genetic markers. Using single-cell RNA sequencing, we applied PCA clustering algorithms to distinguish cell populations, followed by marker selection with Wilcoxon rank-sum tests and logistic regression. Significant overlap between these analyses and comparison with Supplementary Table 3 highlighted key genes involved in regeneration, while Gene Ontology enrichment confirmed the crucial role of ROCs in tissue repair processes.

2 Introduction

The study of amphibian regeneration, particularly in Xenopus laevis tadpoles, offers valuable insights into tissue repair at the cellular level. These tadpoles regenerate their tails through complex signaling pathways, with regeneration-organizing cells (ROCs) playing a key role. This project aims to identify ROCs in the frog tail's skin and determine the genetic features that distinguish them from other cells, contributing to our understanding of regenerative biology.

3 Methodology

In this project, we set out to identify the regeneration-organizing cells (ROCs) in the frog tail using advanced single-cell RNA sequencing, utilizing Scanpy for processing and analysis. We began by loading the single-cell data from a '.h5ad' file, focusing specifically on the regenerative phase by filtering the dataset to include only the cells labeled "DaysPostAmputation" equal to 0, ensuring our analysis targeted the cells directly involved in regeneration. The initial preprocessing involved log-normalization to stabilize gene expression variance across cells, followed by selecting highly variable genes (HVGs) to reduce dimensionality and retain only the most informative features. We then applied standard scaling to normalize the expression values across all cells, centering the data and ensuring unit variance, essential for accurate clustering.

Next, we performed Principal Component Analysis (PCA) to identify key components, focusing on capturing the major sources of variance within the data. This step allowed us to distill the most significant variations, laying a solid foundation for clustering. Using the PCA embeddings, we computed the nearest neighbors with parameters set to 'n_neighbors=15' and 'n_pcs=20'. We then applied both the Louvain and Leiden clustering methods to gain a comprehensive understanding of the data's structure. To visualize the clusters, we used Uniform Manifold Approximation and Projection (UMAP), which provided an effective way to explore the data in a lower-dimensional space. To evaluate the quality of the clusters, we calculated silhouette scores and the Davies-Bouldin Index—quantitative measures that helped us assess how well the cells were grouped and separated, ensuring that we could confidently identify distinct cell populations.

To identify marker genes for the Regenerative-Organizing Cell (ROC), we employed both the Wilcoxon rank-sum test and logistic regression analysis. The Wilcoxon rank-sum test highlighted genes significantly upregulated in ROC compared to other cells, while logistic regression identified genes most strongly associated with ROC by analyzing model coefficients and selecting the top markers. Comparing the results from both methods allowed us to identify common markers, with those identified by both considered the most robust. Dotplot visualizations were used to illustrate the expression patterns of these marker genes across clusters, providing a clear view of their distribution. Finally, we compared the combined set of top marker genes to those listed in Supplementary Table 3, identifying overlapping and unique genes linked to regeneration, thereby providing deeper insights into the genes consistently involved in the regenerative process.

To better understand the biological relevance of the identified genes, we performed Gene Ontology (GO) enrichment analysis using Enrichr from the ‘gseapy’ library. This allowed us to identify the biological processes associated with these genes, providing a clearer picture of the pathways likely contributing to the regenerative capabilities observed in the Xenopus tadpole.

The complete code used for data processing and analysis is available in the following GitHub repository:[https://github.com/xx23438/STATGR5243-Proj1].

4 Results

The clustering analysis of the frog tail regeneration data revealed distinct groups of cells, each characterized by unique marker genes that play specific roles in the regenerative process. We used both Louvain and Leiden clustering algorithms on PCA-transformed data and visualized the clusters using Uniform Manifold Approximation and Projection (UMAP), which effectively represented the cell groupings in two dimensions. Figure 1 shows the UMAP of Louvain clustering, while Figure 2 displays the UMAP of Leiden clustering. To evaluate the quality of these clusters, we calculated silhouette scores and the Davies-Bouldin Index, obtaining silhouette scores of 0.162 for Louvain and 0.160 for Leiden, and Davies-Bouldin Index values of 1.56 for Louvain and 1.65 for Leiden.

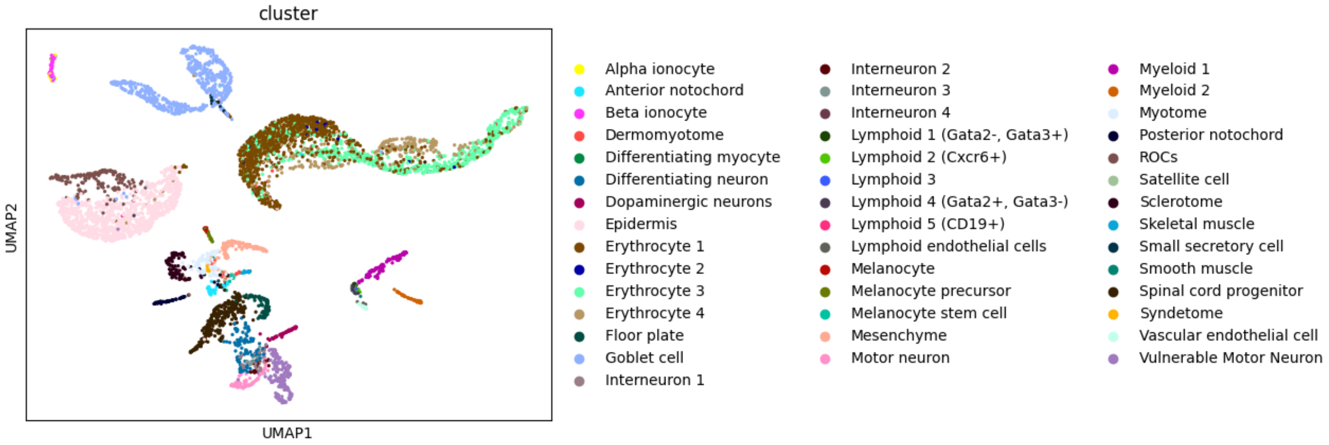

Figure 1: UMAP: Louvain Clustering

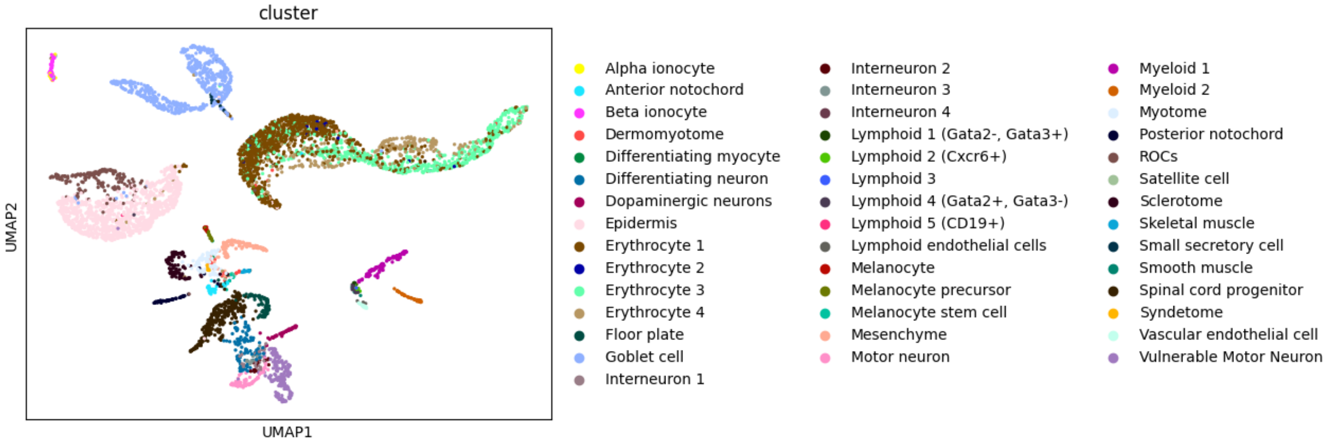

Figure 2: UMAP: Leiden Clustering

The silhouette scores, while modest, suggest a moderate level of cohesion within clusters and separation between clusters, indicating that the identified groups do capture some meaningful biological distinctions among the cells. The Davies-Bouldin Index values, which ideally should be low, also reflect moderate cluster quality, with Louvain performing slightly better than Leiden in terms of cluster compactness and separation. These metrics indicate that the clusters formed by both methods are reasonably well-defined, giving us confidence that we have identified biologically relevant and distinct cell populations involved in the regenerative process, albeit with some overlap or noise present. These distinct clusters set the stage for further exploration of their specific roles in tail regeneration.

The gene analysis revealed detailed expression patterns across the identified clusters, enabling us to locate and characterize the Regenerative Organizing Cells (ROCs) in the frog tail. Figures 3 and 4, which illustrate the top marker gene expression across clusters (Logistic Regression) - Parts 1 and 2, highlight the unique profiles defining each cluster. The dot plots present the expression levels of the top marker genes across different clusters, with individual marker genes along the x-axis and clusters labeled with inferred biological roles along the y-axis. Dot size indicates the fraction of cells expressing each gene, while color intensity reflects average expression levels. Larger, darker dots denote genes highly expressed by a significant proportion of cells in a cluster, emphasizing their importance in that cluster's function. For example, the cluster containing ROCs showed significant enrichment of genes such as 'loc100493805' and 'myct1.L', which are known for their roles in signaling and tissue reorganization, establishing these cells as central coordinators of regeneration.

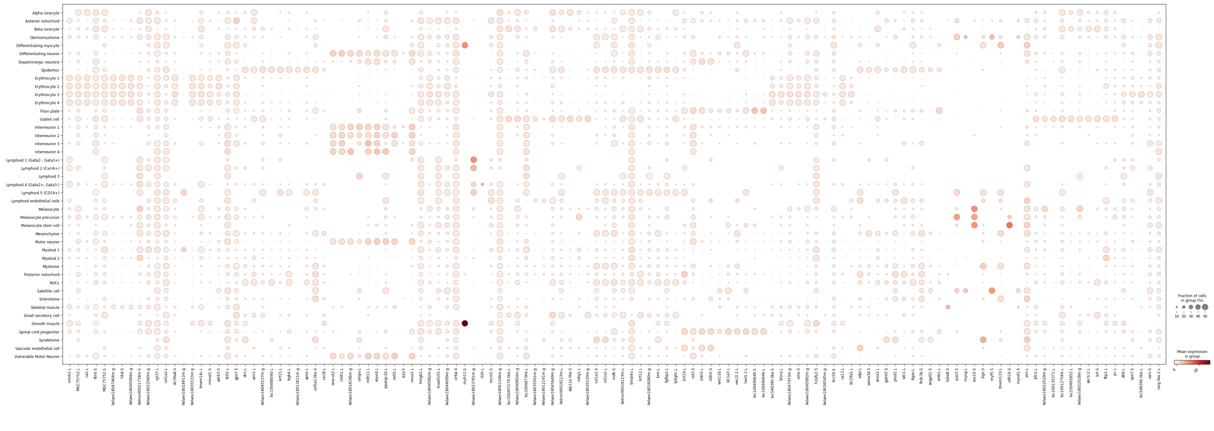

Figure 3: Top Marker Genes Expression Across Clusters (Logistic Regression) - Part 1

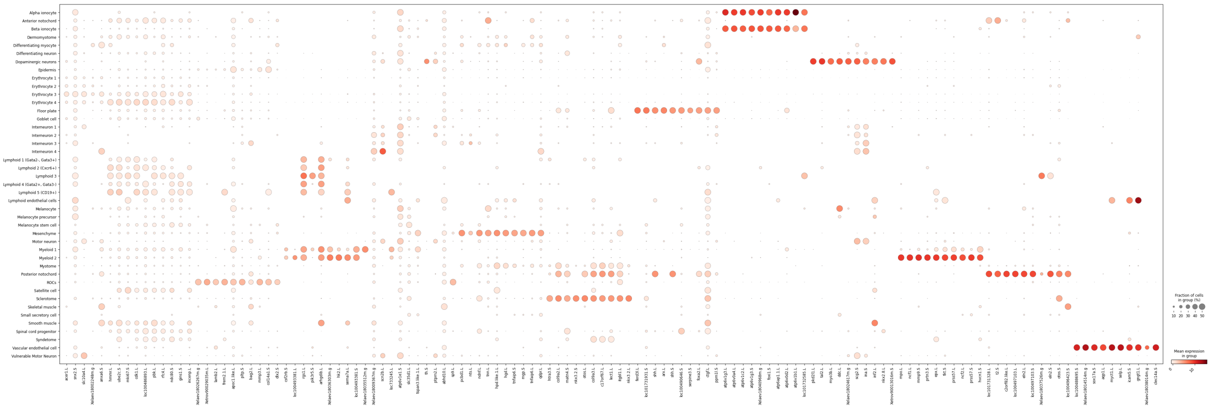

Figure 4: Top Marker Genes Expression Across Clusters (Logistic Regression) - Part 2

The overlap rate was substantial, with nearly half of the genes identified as common between the differential expression analysis and logistic regression for ROC clustering. Notably shared markers such as ‘bmp4.S’, ‘fgf9.L’, and ‘Xetrov90029035m.L’ were consistently associated with specific clusters, particularly enriched in ROC clusters, indicating their role in key functions during the regeneration process, such as tissue remodeling and cell proliferation. A comparison with Supplementary Table 3 revealed 43 overlapping genes, underscoring their importance in regeneration. The presence of regeneration-specific markers, such as ‘fgfr4.S’, ‘bmp4.S’, and ‘fgf9.S’, in both our analysis and Supplementary Table 3 suggests their vital role in cell proliferation and tissue regrowth. The enrichment of these regeneration-specific genes in ROCs highlights their pivotal role in coordinating the overall regenerative response.

5 Conclusion

This study focused on identifying Regenerative Organizing Cells (ROCs) in the frog tail and determining the genetic characteristics that distinguish them from other cell types. Utilizing single-cell RNA sequencing, we applied Principal Component Analysis (PCA) to reduce dimensionality and capture key sources of variance, followed by Louvain and Leiden clustering to define distinct cell populations involved in the regenerative process. To identify the top marker genes specific to the ROC clusters, we conducted a Wilcoxon rank-sum test for differential expression and used logistic regression to determine genes strongly associated with ROCs. Comparing our findings with Supplementary Table 3, we found 43 overlapping genes, highlighting their significant role in tissue repair and regeneration. Gene Ontology enrichment analysis further elucidated the biological pathways involved, confirming the pivotal role of ROCs in coordinating tissue remodeling and cell proliferation during regeneration.