Identification of Regenerative Organizing Cells in Tadpoles

Ivan Gorbunov

October 2024

Abstract

The identification of Regenerative Organizing Cells (ROCs) is incredibly important to understanding the functions of tissue regeneration in *Xenopus laevis* tadpoles. In this study, single-cell RNA sequencing data was used to analyze gene expression across multiple developmental stages, optimizing for ROC gene detection. Following logarithmic transformation and selection of 4500 highly variable genes with a minimum dispersion of 1.0, the data was scaled. Principal Component Analysis (PCA) was applied for dimensionality reduction. Clustering was performed using KNN+Leiden with $n_{neighbors} = 35$ and Kmeans with $K = 20$. Evaluation using UMAP visualizations, silhouette scores, adjusted Rand index (ARI), and rand index identified these settings as optimal for distinguishing clusters.

ROC clusters were identified based on mean gene expression, and gene marker selection was done using Wilcoxon and logistic regression techniques. For KNN clustering, logistic regression initially outperformed Wilcoxon in identifying ROC markers, with 25 of the top 70 genes matching the reference dataset. However, as the marker list was expanded to 300 genes, the performance of both methods converged, with KNN Wilcoxon identifying 40 of the 49 reference genes. In contrast, for Kmeans clustering, Wilcoxon consistently outperformed logistic regression in larger marker lists, identifying 42 of the 49 reference genes. The findings highlight the effectiveness of combining dimensionality reduction, clustering, and marker selection methods to accurately identify ROC-specific genes.

Introduction

Regenerative biology aims to understand the cellular and molecular mechanisms that enable certain organisms to regenerate damaged tissues. *Xenopus laevis* tadpoles are widely studied for their remarkable ability to regenerate lost appendages, specifically their tails, through a complex system of cell signaling and tissue reformation. This regenerative capability is largely controlled by a specialized population of cells known as Regenerative Organizing Cells (ROCs). These cells play a very important role in the early stages of tissue regrowth, particularly in forming the wound epidermis, which is fundamental for successful tissue regeneration.

Even with previous research identifying the role of the wound epidermis in regeneration, the specific cellular and molecular characteristics of ROCs are still poorly understood. Single-cell RNA sequencing provides a high-resolution approach to investigating the RNA molecule profiles of individual cells. It offers the opportunity to capture gene expression patterns related to ROCs. By leveraging RNA sequencing data from multiple developmental stages and carefully selecting highly variable genes, this project aims to identify

the key genetic markers that define ROCs.

In this project, RNA sequence data from various developmental stages of *Xenopus laevis* were used to optimize the identification of ROC-related genes. Dimensionality reduction was done through PCA, followed by clustering using KNN+Leiden and Kmeans algorithms. Marker gene selection was performed using both Wilcoxon and logistic regression methods to determine the most robust markers of ROCs. The analysis aims to evaluate the relative performance of different clustering and marker selection methods.

Methods

Data Preprocessing

The single-cell RNA sequencing dataset used in this study was obtained from *Xenopus laevis* tadpoles, covering multiple developmental stages. The raw data were logarithmically transformed to normalize expression values across all cells. Highly variable genes were selected, with the threshold set at 4500 genes and a minimum dispersion of 1.0, to capture the largest set of ROC-related genes without over representing highly variable genes. Attempts to subset the data into smaller groups of stages resulted in the need for higher variable gene thresholds, which led to less optimal gene representation. The final subsetted data set included 49 out of 50 ROC marker genes give by Supplementary Table 3 ROC sheet.

Data Scaling and Dimensionality Reduction

The dataset was scaled to ensure equal weighting of gene expression across the dataset. Principal Component Analysis (PCA) was used for dimensionality reduction. The clustering analysis was based on 7 principal components, as these captured sufficient variance while keeping the data manageable.

Clustering

Two clustering algorithms were applied to the processed data: KNN+Leiden and Kmeans. The KNN+Leiden clustering was performed using a neighbor parameter of $n_{neighbors} = 35$, which provided the best results in terms of visual separation on UMAP plots and clustering evaluation metrics such as silhouette scores, adjusted Rand index (ARI), and rand index. Kmeans clustering was also applied, with $K = 20$, which produced the most distinct clusters.

Both clustering approaches were evaluated based on UMAP visualizations, silhouette scores, adjusted Rand index (ARI), and rand index to determine their effectiveness in separating distinct cell populations. More UMAP visualizations and all the clustering metrics are available in the code file linked in the *Code Availability* section.

Identification of ROC Clusters

To identify the ROC clusters, mean gene expression profiles were analyzed across all clusters. The cluster corresponding to the ROC population was determined based on the expression of known marker genes such as *Lef1* and *Tp63* and others provided in the reference file, with a total of 49 of the 50 present in the data. These genes were cross-referenced with marker genes from the reference dataset to ensure accurate identification of the ROC cluster.

Gene Marker Selection

Marker gene selection was performed using two methods: the Wilcoxon rank-sum test and logistic regression. The top 70 marker genes were selected to account for variability between runs, as some marker

genes might fall just outside the top 50. For KNN+Leiden clustering, logistic regression initially performed better than Wilcoxon in identifying key ROC markers, with 25 of the top 70 genes matching the reference set, while Wilcoxon identified only 7. However, as the marker list expanded to 300 genes, Wilcoxon and logistic regression converged in performance, identifying 40 and 38 reference genes, respectively.

In contrast, for Kmeans clustering, Wilcoxon consistently outperformed logistic regression with larger marker lists. For the top 300 marker genes, Wilcoxon identified 42 of the 49 reference genes, while logistic regression identified 31.

Evaluation and Comparison

The clustering and marker selection results were evaluated by comparing the identified marker genes against the reference gene list from Supplementary Table 3 ROC sheet. Performance was assessed based on the number of correctly identified ROC-specific genes, with particular attention to the top 70 and top 300 marker genes. The clustering and marker selection methods were also evaluated using standard clustering metrics such as silhouette scores, adjusted Rand index (ARI), and rand index.

Code Availability

The code used for this analysis, which includes additional plots and metrics for clustering evaluation, is publicly available on GitHub and can be accessed here.

Results

Clustering Analysis

KNN+Leiden clustering with $n_{neighbors} = 35$ was the best clustering algorithm in identifying the cluster with the top marker genes. The RNA sequence data was clustered using KNN+Leiden with $n_{neighbors} = 35$, resulting in 58 distinct clusters. These clusters were visualized using UMAP, as shown in Figure 1. Each cluster is represented by a different color, and distinct cell populations can be identified.

All metrics used to evaluate clustering performance, such as silhouette scores, adjusted Rand index (ARI), and others, are available in the code file linked in the *Code Availability* section.

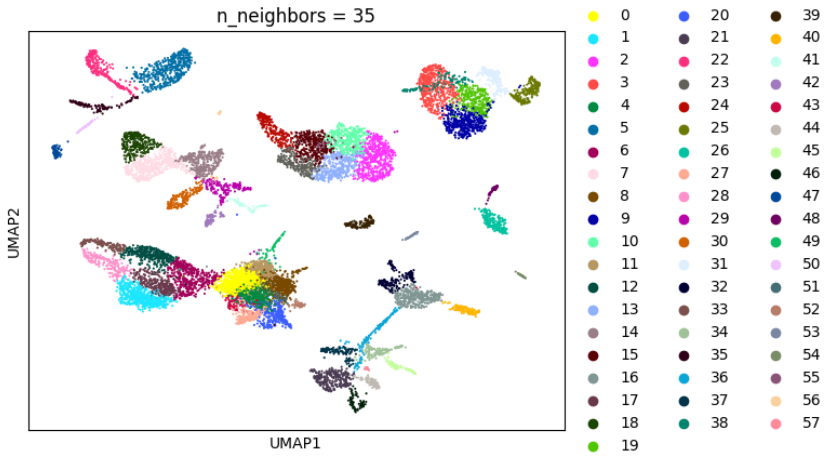

Figure 1: UMAP visualization using $n_{neighbors} = 35$.

Marker Gene Expression

To identify genes that distinguish the ROC population, marker gene selection was performed using both the Wilcoxon rank-sum test and logistic regression. As shown in Figure 2, the mean expression levels of selected marker genes are displayed, with *Wnt5a*, *Fgf10*, and *Lef1* among the top genes highly expressed in the ROC cluster. The marker gene selection results were compared to the reference dataset from Supplementary Table 3 ROC sheet.

KNN+Leiden clustering with logistic regression marker identification was the better method especially in its ability to identify the very top marker genes. For KNN+Leiden clustering, logistic regression initially identified 25 of the top 70 marker genes, while Wilcoxon identified only 7. As the marker list expanded to 300 genes, Wilcoxon and logistic regression converged, identifying 40 and 38 reference genes, respectively. For Kmeans clustering similarly for small gene marker lists logistic regression outperformed Wilcoxon. However, for larger marker lists, 300, Wilcoxon consistently outperformed logistic regression, identifying 42 of the 49 reference genes, while logistic regression identified 31.

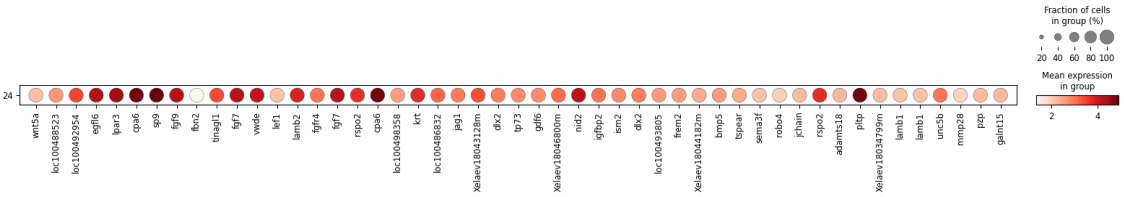

Figure 2: Dot plot showing mean expression of selected genes across clusters.

Conclusion

This study aimed to identify Regenerative Organizing Cells (ROCs) in *Xenopus laevis* using RNA sequence data across various developmental stages. Through careful selection of highly variable genes and the use of dimensionality reduction and clustering algorithms, specific cell populations were identified, focusing on the ROC cluster. KNN+Leiden and Kmeans clustering methods were evaluated, with $n_{neighbors} = 35$ for KNN and $n_{clusters} = 20$ for Kmeans providing the best results.

Interestingly, KNN combined with logistic regression, despite producing worse UMAP plots and clustering metrics, outperformed other methods in identifying the top most variable genes, particularly in shorter marker lists. These short lists seem to carry more importance when pinpointing key ROC-specific genes. In contrast, as the marker lists expanded, Wilcoxon began to converge with logistic regression in KNN clustering performance. For Kmeans clustering, Wilcoxon consistently outperformed logistic regression, but only for larger marker lists.

Overall, the combination of clustering and marker gene selection techniques successfully identified ROC specific markers, providing insights into the molecular drivers of regeneration in *Xenopus laevis*. These findings setup a way for further investigation into the role of these cells in regeneration, with the potential to inform future studies in regenerative biology and therapeutic approaches.