Replication of ROCs Identification Techniques

Audrey Christensen

October 2024

1 Abstract

In this report, we replicate the findings of the paper "Identification of a regeneration-organizing cell in the Xenopus tail" by applying various clustering methods to the original single-cell RNA-seq data. Using techniques such as Leiden, k-means, and hierarchical clustering, we compare our clusters against those presented by the original authors, treating their results as ground truth. We evaluate the performance of our methods using metrics like Adjusted Rand Index (ARI), Rand Index, and silhouette score, and perform gene marker analysis to identify genes expressed by ROC cells. We found that our clusters were able to achieve a maximum ARI of .69, indicating a moderate recovery of the original clustering results. Through gene marker analysis, we can confirm that our Leiden clustering method successfully identified the ROCs. Our gene marker techniques do not seem to be fully aligned with the techniques used in the original paper, as there is minimal overlap between the genes we detect in the ROCs and the genes the paper provides in the ROCs.

2 Introduction

In the original paper, the authors identified distinct cell populations within regenerating Xenopus tails using single-cell RNA-seq data, specifically focusing on the regeneration-organizing cells (ROCs). They aim to identify the specific cell that allows regeneration-capable frogs to grow their tails back by analyzing samples collected from both regeneration-capable and regeneration-incapable frogs. The frogs' tails were amputated and cell samples were collected before, on the day of, and on each day after amputation.

Our objective in this report is first, to replicate the clustering results from the original study using alternative clustering algorithms, and second, to explore the gene expression patterns of the ROC cells to validate the biological findings. We treat the clusters identified by the authors as the objective truth and compute clustering performance metrics with respect to these reference clusters. Using the raw gene expression counts provided by the original authors, we applied several clustering techniques, including Leiden, k-means, and hierarchical

clustering, followed by gene marker analysis to identify genes highly expressed in ROC cells.

3 Methods

We processed the data using Scanpy and AnnData, starting with normalization, where gene expression counts were scaled to a target value of 10,000. To focus on the most informative genes, we selected highly variable genes based on a minimum mean expression of 0.05, a maximum mean of 0.8, and a minimum dispersion of 0.65. After subsetting the data to include only these highly variable genes, we scaled the dataset for further analysis.

For dimensionality reduction, we applied Principal Component Analysis (PCA). Initially, we used the elbow method to select the number of PCs, but after observing that retaining more PCs improved clustering performance, we opted to use the maximum number of PCs available, which resulted in better metrics.

We then performed three different clustering techniques: Leiden clustering, k-means, and hierarchical clustering. For Leiden clustering, we began with 20 neighbors, but through experimentation, we found that increasing the number of neighbors as high as 50 led to a higher ARI and a slightly lower silhouette score. Ultimately, we chose to use 30 neighbors as this configuration maximized ARI while maintaining acceptable silhouette scores. For k-means, we manually set the number of clusters to 36, matching the number of clusters in the original study. Similarly, hierarchical clustering was performed with 36 clusters, and arguments were chosen based on a combination of default settings and insights from the supplemental materials accompanying the original paper. We computed ARI, RI, and silhouette scores for each clustering method, with Leiden having the most favorable scores for each category.

To visualize the clusters, we used UMAP, projecting the cells in two ways: first by coloring cells according to the original clusters from the paper, and second by coloring based on the clusters we identified using our methods. This allowed us to compare the similarity between our clustering and the original study's results. Because the Leiden clustering had the highest ARI of each clustering method, we used those cluster results for any subsequent analysis. We identified the particular cluster that was present post amputation in regeneration-capable frogs but was absent in regeneration-incapable frogs and used that as a proxy for the ROCs cluster provided by the original paper.

For gene marker analysis, we utilized the built in gene marker analysis function in ScanPy, choosing the 50 most expressed genes for both the ROCs and our Leiden cluster proxy, and using the t-test and wilcoxon methods. We compared the genes identified for the ROCs and the Leiden cluster proxy, and all of the genes matched, indicating that our Leiden cluster is in fact the ROC. Upon comparing this list of genes to the paper's provided list of genes, there was no overlap, which we later realized was due to suffixes such as .L and .S on the genes, indicating that they were from large or short chromosomes. For sim-

plicity, we stripped the suffixes for this comparison because the genes provided for the ROC do not have such suffixes, but we understand that there may be further complexities that need to be accounted for.

The full code for this analysis is available at this GitHub link.

4 Results

Our clustering analysis yielded mixed results across methods. In the case of Leiden clustering, using 30 neighbors resulted in an ARI of .695 and a silhouette score of .315, indicating a strong alignment with the original clusters but less distinct separation between them. Increasing the number of neighbors improved ARI but came at the cost of cluster distinctiveness, as evidenced by the declining silhouette scores. K-means clustering produced an ARI of .488, which was lower than Leiden but still showed moderate consistency with the ground truth clusters. Hierarchical clustering and k-means performed roughly the same, with hierarchical clustering producing an ARI of .494. The clusters are visualized in Figure 1, which displays the clusters colored by the clustering results of the original paper as well as Figures 2, 3, and 4, which show the clusters colored by Leiden, k-means, and hierarchical clustering. In our analysis, we found that Leiden cluster 9 is most likely to be the ROC.

In terms of gene marker analysis, our results indicated that our Leiden clustering and analysis correctly identified the ROCs. When comparing the gene markers that we identified for the ROCs and our Leiden cluster proxy, we saw a complete overlap of all gene markers. The ROCs cluster as well as the top 5 of the identified genes are highlighted in Figure 5. There was no difference in genes identified using the wilcoxon versus t-test method. In comparing our gene markers to that provided by the paper's authors, only 9 genes out of the 50 matched. This indicates that our gene marker methods differed from theirs, but the match between our Leiden clustering and their ROC cluster does indicate that we are analyzing the same group of cells.

5 Conclusion

In conclusion, we had moderate success in replicating the clusters described in the original study by applying alternative clustering techniques. Leiden clustering, with 30 neighbors and full PCA dimensionality, yielded the highest ARI scores, indicating that these clusters most closely matched the original findings. However, this came at the expense of silhouette scores, suggesting that the clusters were not as well-separated. K-means and hierarchical clustering performed less effectively, but still captured some of the key cell populations. However, for our cluster of interest, our gene marker analysis indicated an exact match between our clustering results and that of the paper. We also succeeded in matching 9 out of 50 gene markers of the original ROC cells, and hope to explore gene marker analysis methods further.

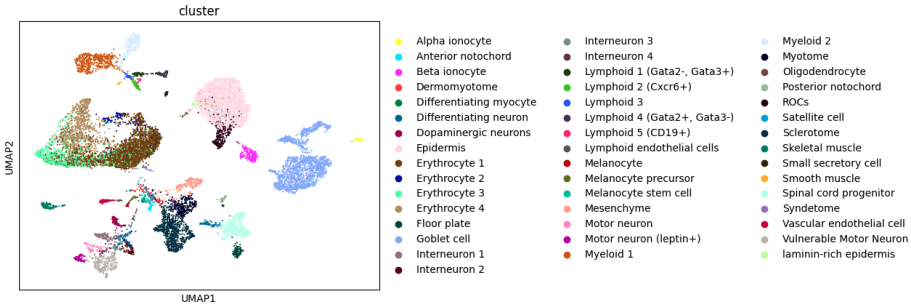

Figure 1: Original Clusters

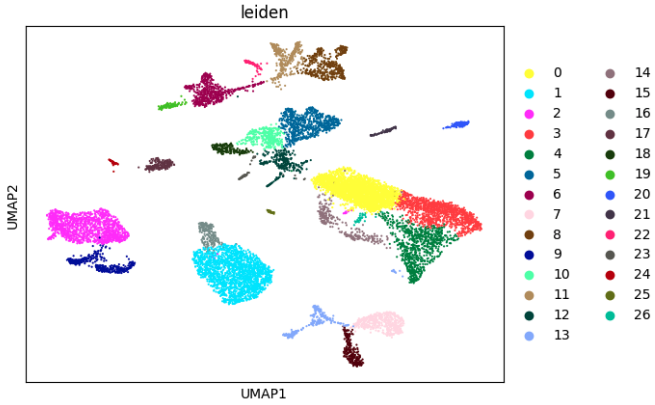

Figure 2: Leiden Clusters

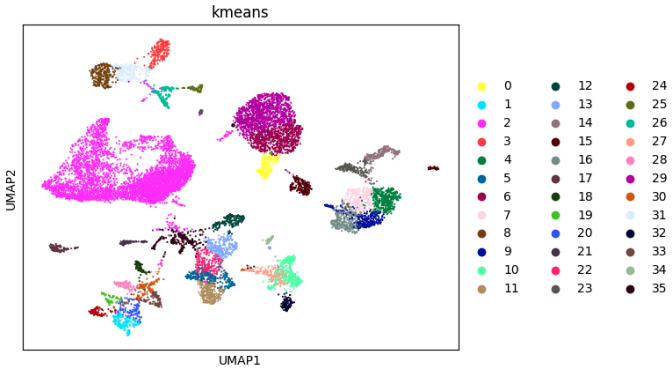

Figure 3: K-means Clusters

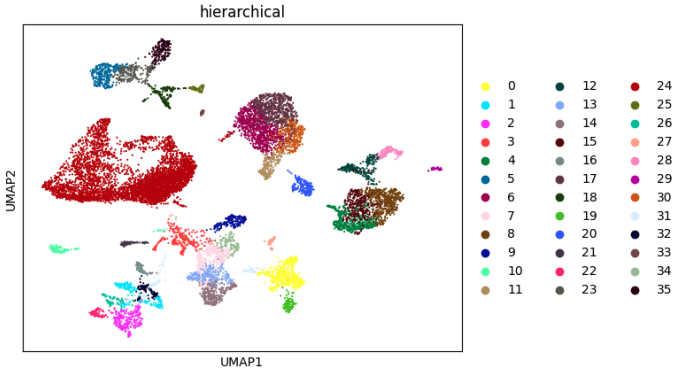

Figure 4: Hierarchical Clusters

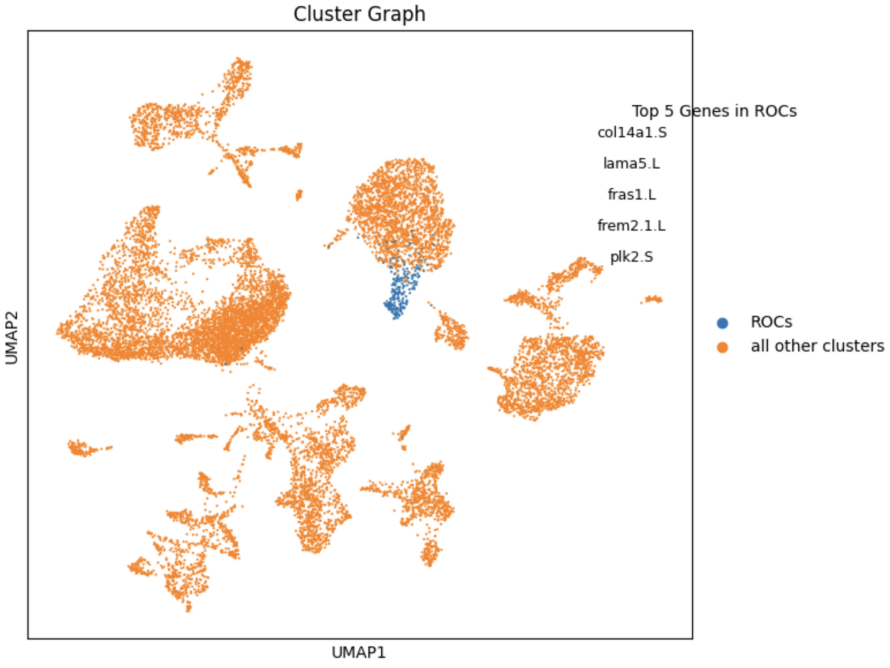

Figure 5: ROCs with Gene Markers