Replicating analysis in “Identification of a regeneration-organizing cell in the *Xenopus* tail”

Project 1 - Applied Data Science GR5423
Val Rogel

October 8, 2024

1 Abstract

Several studies have examined cellular characteristics of regenerative species, but the mechanisms underlying regeneration continue to be examined. One such species is the *Xenopus laevis* tadpole, which can regenerate its tail after amputation. By examining the cellular makeup of tadpoles that have undergone regeneration after amputation, we can identify the involved cells. Using single-cell RNA sequencing methods I aim to identify the regeneration-organizing cell, unknown previous to 2019. Marker identification methods, including the Wilcoxon tie-corrected process, reveal specific gene markers that are more highly expressed in these regeneration-organizing cells.

2 Introduction

Regeneration research has promise in biology, specifically with regards to the development of cell replacement therpapies. Regeneration has three key stages: formation of specializied wound epidermis, blastema or regenerative bud formation, and outgrowth. In contrast with mammals, *Xenopus laevis* tadpoles have high regenerative potential, specifically in their tails (Aztekin, et. al, 2019). Cells attributed to such regeneration are called regeneration-organizing cells (ROCs), a term coined by Aztekin and fellow researchers to describe the group of cells that migrate from the epidermis to the amputation plane within 24 hours after tail amputation.

It is important to note that *Xenopus* tadpoles do not always regenerate; they can be found in two developmental stages: regeneration-competent and regeneration- incompetent. The lack of ROC migration to the amputation plane in regeneration-incompetent *Xenopus* lends some clues as to which biological mechanisms are responsible for facilitating the regeneration process.

Identifying cells involved in regeneration requires the use of single-cell RNA sequencing practices including principal component analysis (PCA), clustering, and gene marker identification. Scanpy, a Python package, offers single-cell RNA sequencing methods that can be implemented in Python to complete preprocessing, PCA, clustering, data visualization, and finally, marker selection.

Marker genes have expression profiles that allow a desired cell type to be distinguished from other cells in a data set, even those found within a similar class of cells. Several processes exist to identify gene markers in both Python and R, and Seurat and Scanpy use a ’one versus rest’ strategy to compare different clusters (Pullin and McCarthy, 2024).

3 Methods

First, I imported the cleaned and processed data from Aztekin et al. and initalized an annotated data object. I used this data as the basis for my methodology.

3.1 Data Preprocessing

This decision was informed by the fact that in ”successful regeneration, ROCs appear in the amputation plane within 24 hours” (Aztekin et. al, 2019). Therefore, I thought it best to select a point at 1 day (24 hours) or later. After plotting a histogram of the data points relative to their respective time frames, I saw that 2 days post-amputation had the lowest count. I elected to choose the most conservative count because I hypothesized it might lead to the fewest low-quality cells. Therefore, I limited my data to the time period of 2 days post-amputation.

When all cells are included in the dataset, the analysis can suffer due to a higher level of noise in our dataset. Therefore, by having the minimum number of cells expressing a gene as 2, dataset’s noise is reduced without filtering out all rare cells; rarer cells can still be important to representing the cells that are collected. When filtering the cells, I chose the minimum number of counts to equal 200. As someone without a background in biology, I felt it would be important to reduce noise that I might not be able to effectively interpret. However, at this stage, it is unclear the ROC cell’s prevalence in the data set. Therefore, a higher threshold (say 1000) might accidentally filter out important cells in this analysis.

I made an attempt to remove the mitochondrial genes that are in the dataset because a higher proportion of mitochondrial genes can suggest lower-quality cells that are dying or damaged. However, the traditional string starter ”MT-” was not in the dataset. This could be suggestive of the fact that mitochondrial genes were labeled differently, but more likely, as the violin plots demonstrate, there may not have been any counts of these mitochondrial genes. I was left with 2310 cells after attempting this filtering.

Cell counts were normalized to a target sum of 10,000, which is often used in other scanpy-based single cell RNA-sequencing research (Li, et. al, 2020). The data was then normalized with a base-2 logarithm, in accordance with methods used with this dataset in the literature (Aztekin, et. al, 2019). Finally, the 15% most highly variable genes were selected.

3.2 PCA

Dimensionality reduction was completed by way of principal component analysis (PCA). The data was scaled prior to PCA. An ”elbow” in the variance ratio plot was determined, which indicate the principal component (PC) point after which additional PCs have less significant contributions to the explanation of the variance.

We can calculate where the principal components start to elbow by either of two metrics (traditionally that which is larger): the point where the principal components only contribute 5% of standard deviation and the principal components cumulatively contribute 90% of the standard deviation *or* the point where the percent change in variation between the consecutive PCs is less than 0.1% (Khetani).

3.3 Clustering

K-means clustering was carried out with 15 PCs and UMAP and tSNE visualizations were produced. For the latter, perplexity was selected to be 30 with a learning rate of 200 due to the fact that they are both the default value.

Leiden clustering was carried out under three resolutions: .05, .5, and 1, and visualized with UMAPs.

3.4 Clustering metrics

I computed the Rand index, adjusted Rand index, silhouette score, and normalized mutual information (NMI) for the k-nearest neighbors and Leiden clusters.

The adjusted Rand index proposed by Hubert and Arabie (1985) confronts issues with the Rand index, namely that ”the Rand statistic approaches its upper limit of unity as the number of clusters increases” (Santos and Embrechts, 2009). For this reason, the adjusted Rand index tends to be a more popular choice than its unadjusted counterpart for this type of research.

A paper designed to compare and measure the effectiveness of 14 clustering methods used four ”evaluation measures... to quantify the concordance of clustering results on each scRNA- seq dataset with respect to their predefined cell-type annotations” and these metrics included the Rand index (completed above), normalized mutual information, Fowlkes–Mallows index, and Jaccard index (Yu et. al, 2022). For this reason, I chose to also calculate the normalized mutual information metric. According to the scikit learn documentation, while the adjusted mutual information metric also exists, the former is more commonly used in existing research.

3.5 Marker identification

The two marker selection methods used in this analysis are Scanpy’s Wilcoxon tie-corrected (TC) test and the Student t-test. In a comparison of marker gene selection methods, the highest performance came from ”edgeR, scran’s Binomial-any method, Wilcoxon rank-sum based methods (Seurat and Scanpy), Student’s t-test, and Rank-corr” (Pullin and McCarthy, 2024). I selected Wilcoxon TC because it had one of the highest F1 scores in the analysis. In contrast, the t-test had an F1 score in the bottom 50% in the methods for which such scores were calculated. Scanpy’s methods for selecting marker genes only focuses on up-regulated marker genes by default (rather than both up- and down-regulated genes) and selects those with the largest overall effect sizes (Pullin and McCarthy, 2024).

Then I performed differential gene expression analysis grouped by cluster with the parameter groups set to ”ROCs” with the Wilcoxon TC and Student’s t-test.

3.6 Identifying cells

Finally, I identified the cells expressing the marker genes at relatively high levels (those which could potentially be identified as ROCs). A reference UMAP was created that showed the ROCs in a UMAP in the context of the other skin cells, such as alpha and beta ionocytes, melanocytes, and epidermal cells. This reference UMAP classified ROCs based on their cluster labels provided in the dataset.

Following the reference UMAP, UMAPs based on the cell classification via the marker genes were generated at expression threshold .4. The top 20 markers from the Wilcoxon TC and knn methods were used classify the cells.

3.7 Code availability

Code link: https://github.com/valcsuremm/Project-1

4 Results

4.1 Clustering

With regards to evaluating the ”elbow” of principal component analysis, I ultimately selected the point where the percent change in variation between the consecutive PCs is less than 0.1%. I selected against the metric of the point where the principal components only contribute 5% of standard deviation and the principal components cumulatively contribute 90% of the standard deviation. This is due to the fact that greater than 50 components cumulatively contributed to 90% of the standard deviation. To constrain the number of relevant principal components, PCs 1 through 8 were selected; however, it is important to note that they cumulatively only contributed to 27.9% of the variance.

Clustering metrics calculated for the k-nearest neighbors clustering yielded a Rand index of .3199 and an adjusted Rand index of .7615. The adjusted Rand index’s value suggests that there is some overlap with the cluster labels and the way that the k-means clustering method assigns clustering labels.

The knn clustering silhouette score of .0322 suggests that the clustering method did not perform very well. However, The range of NMI is [0,1], so the value of .62 suggests that there is a medium to medium-high degree of overlap between the k means clustering and the original clustering labels.

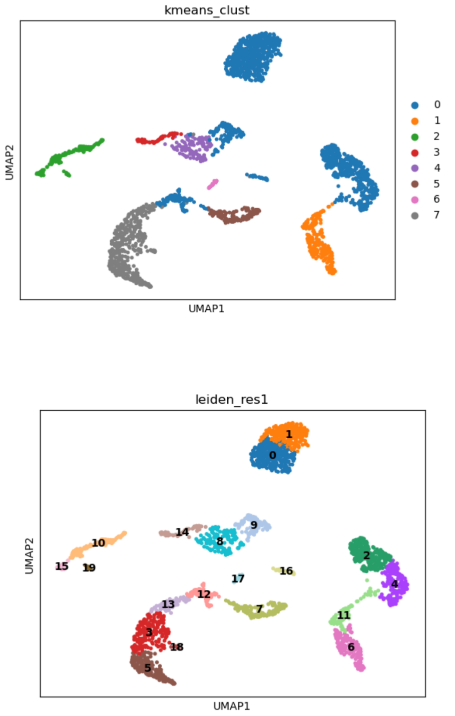

Figure 1: *UMAP representing k-nearest neighbor clusters with default settings. There are 7 clusters (above). UMAP representing Leiden clusters with resolution 1. There are 19 clusters (below).*

The Leiden clustering's Rand index and adjusted Rand index are .9202 and .5603 respectively, suggesting higher performance for Leiden clustering than knn. Similarly, to the k-nearest neighbor clustering, the Silhouette's criteria for well-clustered data was not met, and is equal to -.0336. It appears that the data points inside the Leiden clusters are not very similar to each other (based on the silhouette score), but the pairs that are assigned to various clusters tend to be similar between the ground truth and the Leiden clusters. Finally, the NMI score was 0.7660.

4.2 Marker selection and gene analysis

Through the Wilcoxon tie-corrected and Student's t-test marker selection processes, I was able to produce a list of the marker genes associated with the ROCs. The top ten genes from each clustering method are listed below with the homolog remains appended to each name.

Top ten marker genes for marker ID methods	
Wilcoxon TC	t-test
sp9.L	apoc1.like.L
fgf9.L	azin2.S
loc100492954.S	mdk.S
lpar3.L	mdk.L
Xelaev18026267m.g	epcam.L
cpa6.L	col14a1.S
Xetrov90029035m.L	id3.S
fgf7.S	epcam.S
pltp.S	krt5.7.S
mmp3.L	s100a11.L

All have a p-value < .05 and the top ten marker genes for the methods have no overlap.

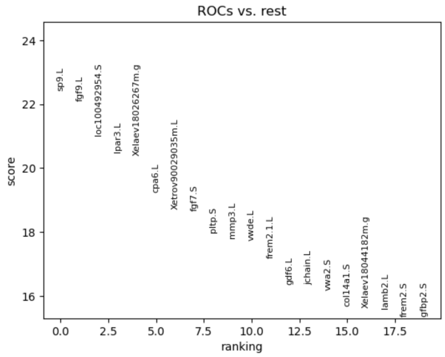

Figure 2: *Highly differentially expressed marker genes for each group as determined by the Wilcoxon TC method*

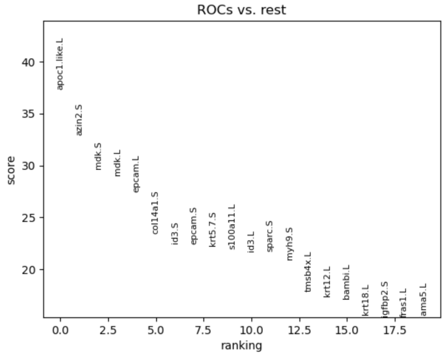

Figure 3: *Highly differentially expressed marker genes for each group as determined by the Student's t-test*

Amongst the top fifty marker genes, the two methods share ten: Xelaev18026267m.g, frem2.1.L, col14a1.S, igfbp2.S, cldn6.2.S, krt18.L, bambi.S, fras1.L, lama5.L, and lama1.L.

When we ignore the homolog, we can compare the marker genes to that of Supplementary Table 3 (S3) (Aztekin et al., 2019). 3 of the marker genes found through the t-test match those in S3: igfbp2, egfl6, and frem2. In contrast, 22 of the marker genes found by the Wilcoxon TC method match those in S3. Finally, the t-test marker genes and the Wilcoxon TC marker genes share 9 between them when homologs at the tail of the gene name are disregarded.

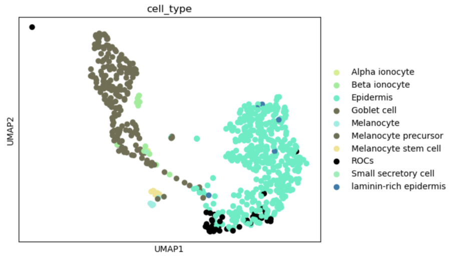

Figure 4: *UMAP based upon the provided cluster labels of the various skin cells. Points representing ROCs are black. (Aztekin, et. al, 2019)*

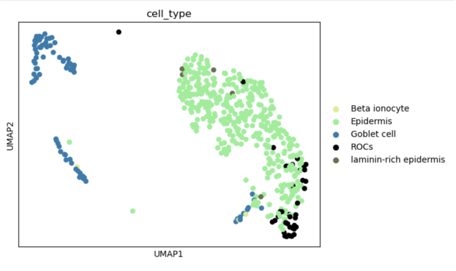

Figure 5: *ROCs (black) identified among skin cells using the Wilcoxon TC method.*

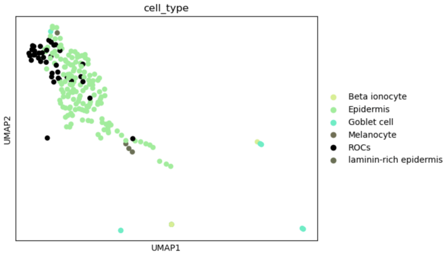

Figure 6: *ROCs (black) identified among skin cells using the Student's t-score method.*

The UMAPs that contain the ROCs identified amongst the skin cells indicate that not all cells may have passed the expression threshold to be included in the visualization. Each gene marker identification process (Wilcoxin TC and Student's t-test) was based upon the top 20 markers determined previously.

5 Conclusion

Investigating the gene markers used to identify regeneration-organizing cells presented some challenges. The lack of cells having a high percentage of mitochondrial genes may suggest a dataset that had been cleaned of those genes in advance, human error in the data preprocessing, or different naming conventions in the original dataset. In addition, the constraint of principal components to merely 8, while in line with finding the "elbow" only led to a cumulative contribution of approximately 28% of the variance in the dataset. This may pose a problem because important information that appears in rare subtypes of cells may be lost.

Based on the clustering metrics used to evaluate the k-nearest neighbors and Leiden clusters, it appears that Leiden clustering was more effective at representing the relationships between pairs of data points. This was demonstrated through relatively high Rand, adjusted Rand, and NMI scores, while the silhouette scores in both clustering processes were decidedly low.

Relative to Figure 3, the UMAP based upon provided cluster labels, it is evident that the identification of ROCs based on the genetic markers uncovered by the Wilcoxin tie-corrected (Figure 4) and Student's t-test (Figure 5) methods were somewhat, but not entirely effective at identifying the ROCs. Throughout the analysis, it was evident that the UMAP is extremely sensitive to its resolution, so the visualization is beholden to this feature. Finally, the different outcomes in the gene marker selection processes revealed a strong inconsistency between the Wilcoxon tie-corrected and the Student's t-test method, despite the fact that they were both implemented with the Scanpy package. This reveals that, without robust knowledge of the proper marker selection methods, researchers might be led astray by inaccurate markers.

6 References

C. Aztekin et al. ,Identification of a regeneration-organizing cell in the Xenopus tail.Science364,653-658(2019).

Khetani, R. (n.d.). Elbow plot: Quantitative approach. Introduction to Single-cell RNA-seq - ARCHIVED. Core.

Li, J., Yu, C., Ma, L. et al. Comparison of Scanpy-based algorithms to remove the batch effect from single-cell RNA-seq data. Cell Regen 9, 10 (2020).

Pullin, J.M., McCarthy, D.J. A comparison of marker gene selection methods for single-cell RNA sequencing data. Genome Biol 25, 56 (2024).

Santos, J. M., & Embrechts, M. (2009, September). On the use of the adjusted rand index as a metric for evaluating supervised classification. In International conference on artificial neural networks (pp. 175-184). Berlin, Heidelberg: Springer Berlin Heidelberg.

Yu, L., Cao, Y., Yang, J.Y.H. et al. Benchmarking clustering algorithms on estimating the number of cell types from single-cell RNA-sequencing data. Genome Biol 23, 49 (2022).