Identifying the Regenerative Organizing Cell in the Frog Tail

Abstract

In this project, we analyzed single-cell RNA sequencing data to explore the mechanisms behind tissue regeneration. Dimensionality reduction techniques, including Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP), were applied to visualize the cellular diversity. Clustering was performed using the Louvain method, which helped identify distinct cell populations. Marker genes were selected for each cluster and compared against established gene sets, such as those found in Supplementary Table 3, to pinpoint regenerative organizing cells (ROCs). Various clustering metrics, including silhouette score, Adjusted Rand Index (ARI), and Rand Index (RI), were used to assess clustering performance.

Introduction

Regeneration, the process by which organisms restore damaged tissues, has long fascinated scientists due to its potential applications in medicine, particularly for humans who have limited regenerative abilities. Some animals, such as axolotl, possess remarkable regenerative capabilities, allowing them to restore the entire body parts. Understanding the complex regenerative mechanisms behind these abilities could pave the way for advances in wound healing, organ transplantation, and other related areas.

Recently, advances in computation enables us to view cells with methods like single-cell RNA sequencing (scRNA-seq) for cells and genes of regenerative tissues. By examining the gene expression profiles of individual cells, researchers can identify key genes involved in regeneration.

In this project, we focused on analyzing single-cell RNA sequencing data from regenerating tissues collected on day 0, with the aim of identifying the key cellular populations and genes responsible for the regenerative process. Dimensionality reduction and clustering techniques were employed to uncover distinct cell populations, including the regeneration-organizing cells (ROC) that play an important role in the regeneration. This work builds on the study "Identification of a regeneration-organizing cell in the Xenopus tail", which highlights the critical role of ROCs in regeneration. The goal of this project is to replicate and compare the findings from that study, while exploring similarities and differences in the underlying biological mechanisms.

Methods

Data Loading: The dataset analyzed in this study was loaded using the Scanpy library, focusing on a time point Day 0 of amputation. Data was filtered to include only cells present at this time point. The single-cell RNA sequencing data consisted of 5,302 cells and 23,111 genes after filtering for genes expressed in at least 5 cells. A layer of raw counts was preserved for later use.

Dimension Reduction: Principal component analysis (PCA) is used to reduce the dimensionality of the dataset. Before PCA, we normalized the total gene expression per cell, and then took the log of the counts. Highly variable genes (HVGs) were selected using the Seurat method, retaining the top 2000 genes.

Clustering: Louvain clustering was applied to the PCA-reduced dataset to identify distinct cellular populations. The neighborhood graph was computed using 40 principal components and 20 nearest neighbors. Louvain clustering was performed at a resolution of 1.0. We visualized the clusters using UMAP (Uniform Manifold Approximation and Projection).

Gene Marker Identification: For each cluster identified through Louvain clustering, marker genes were determined using the t-test method implemented in Scanpy. This analysis identified genes that were differentially expressed in each cluster, revealing key marker genes that could represent different cell types or states during regeneration. The marker genes were compared against a reference gene set from Supplementary Table 3. Lastly, Gene Ontology (GO) is used to compare with humans.

Metrics: The performance of the clustering was evaluated using some metrics like silhouette score, Adjusted Rand Index (ARI), and RAND-index. The silhouette score measured the cohesion and separation of clusters based on Louvain, while ARI and RAND-index compared the clustering results with ground truth labels. These metrics provided a quantitative evaluation of the quality of clustering and helped in assessing the biological relevance of the identified clusters. V-Measure is a metric measuring homogeneity and completeness.

Results:

1. Clustering and Dimensionality Reduction:

A group of colorful dots

Description automatically generated with medium confidence

PCA is used to reduce the dimension of the high-dimensional single-cell RNA sequencing data. In the plot, the cells are projected onto the first three principal components (PC1, PC2 and PC3), which capture the most variance in the data. Each point represents a single cell, and the different colors correspond to different clusters. Some clusters are well-separated from each other, suggesting that these cell populations are quite distinct in gene expression profiles; some clusters overlaps because they are similar or not fully separated by PCA.

A screenshot of a computer generated image

Description automatically generatedA colorful dots on a white background

Description automatically generated

However, UMAP tends to better capture complex relationships in the data because it shows a clearer separation between most clusters compared to PCA. For example, Cluster 0 (yellow), Cluster 1 (blue), and Cluster 27 (green) are distinctly separated from the other clusters, meaning that these groups of cells have distinct gene expression profiles. Also, highly variable genes (HVG) of UMAP can give us a different view. Although it looks weird, UMAP on HVG look closer at genes’ differences and clusters.

A graph with numbers and letters

Description automatically generatedMarker Gene Analysis:

This shows the top-ranked marker genes for each cluster, highlighting the genes that are most differentially expressed between a given cluster and the rest of the data. For example, in Cluster 12, the top marker gene is Xelaev18038242m.g. It has the highest score, indicating strong differential expression.

In our analysis, we focused on Cluster 0. Using Gene Ontology (GO) for the biological roles, we compared the identified marker genes with the gene sets listed in Supplementary Table 3, G2M, S, and G1. By excluding the common markers found in these phases, we identified a subset of ROC-specific markers for Cluster 0, which may be an ROC-specific marker.

t-SNE:A close-up of a blue and purple dot

Description automatically generated

This visualizes the expression of the ptrf.S gene across all cells. Each point represents a cell, and the color gradient indicates the expression level of ptrf.S: purple indicates low expression and yellow-green represents higher expression levels.

Metrics:

A screenshot of a computer code

Description automatically generatedRAND-Index measures the similarity between the predicted clusters and the ground truth, without adjustment. Silhouette Score evaluates the separation of clusters: the higher scores, the better without looking at it. It suggests that the clustering is not performing well. Maybe the reason is the clusters are overlapping or poorly separated. ARI measures clustering agreement with the ground truth, accounting for chance. 0.276 ARI shows at least some agreement between the clustering and the ground truth. V-Measure is a metric measuring homogeneity and completeness. A high V-Measure score means clusters are both homogeneous and complete. Data points in a single cluster belong to the same class, and all members of a class are assigned to the same cluster. All scores are not high as expected.

Conclusion:

This study identified distinct cellular populations involved in tissue regeneration, with a particular focus on ROCs. With the use of techniques, such as the Louvain method, PCA, UMAP, and GO, we uncovered key gene markers associated with different cell populations. The clustering performance, evaluated by metrics like silhouette score and ARI, indicated moderate agreement between the clustering results and the biological labels. Additionally, the marker gene analysis provided insights into the potential biological roles of these genes in the regenerative process. Future studies could focus on experimentally validating the identified markers and exploring their potential for tissue repair in regeneration.

Code Availability:

The code for this analysis is available on [GitHub](https://github.com/Haobo-Yang/Project1/blob/main/Project1finalversion.ipynb).