Find the frog's source of regenerative power

Xinrong Dong xd2323

Abstract

We reproduce and extend ROC analysis in Xenopus tail scRNA-seq. At the global level, Leiden, Louvain, and K-means reveal a consistent topology. In the skin subset, three complementary differential tests converge on a robust 84 gene majority consensus. The ROC gene-set score forms a single high signal hotspot on UMAP that aligns with an unsupervised cluster and matches the biological ROC annotation.

Introduction

Amphibian tadpoles such as Xenopus laevis can regrow a lost tail, which provides a possible model to uncover the cellular programs that enable complex tissue regeneration. In the paper *Identification of a regeneration-organizing cell in the Xenopus tail*, Dr. Aztekin and other scientists used single-cell RNA-seq across intact tails and multiple post-amputation time points to generate a pooled atlas of cell states. The key discovery from this atlas was an epidermal cell type, the regeneration organizing cell (ROC). ROCs can distinguish incompetent tails and expresses pro-regenerative ligands.

This project reproduces and extends the ROC analysis pipeline on the provided scRNA-seq dataset. We will cluster cells, identify and compare ROC marker genes, and evaluate how data denoising and batch integration over time affect clustering and marker selection.

Method

1. Code Availability

https://github.com/xr0706/5243-project-1

2. Data

We used a single-cell RNA-seq dataset stored as an AnnData object, in which the matrix has 13199 cells and 31535 genes. There are also metadata information with 13 columns, such as cell, Days Post Amputation, cluster, and so on. Basic diagnostics, including non-zero density and value ranges, confirm this data matrix is a typical high dimensional and sparse scRNA-seq input. Since the dataset has numeric matrix plus structured categorical covariates per cell, it is a structured data. There is no train or test split because the tasks we planned to perform are unsupervised, instead we evaluated clustering quality with internal indices.

Before the main analysis, a preprocessing pipeline for scRNA-seq is applied to the AnnData object. First, the raw counts matrix should be backed up, so that downstream DE analyses can revert to the unnormalized counts. Then the library-size normalization is used to scale each cell and followed by a log transform to stabilize the mean-variance relationship, which helps to reduce the dominance of extremely highly expressed genes. For quality control, we remove genes detected in fewer than 3 cells and cells with fewer than 200 detected genes. Finally, after correcting for the mean dependence of the original dispersion, 2300 highly variable genes (HVG) are selected. The goal in this part is to obtain an information dense gene subset with lower noise, improving clustering stability and biological interpretability.

3. Visualization and Clustering

Based on HVGs, we standardized the matrix, computed PCA with50 PCs, built a kNN graph, and run UMAP for

low-dimensional visualization. To ease interpretation across cell types, raw cluster names were regex mapped in some lineages, like Skin, Neural, and Somite. Then a global UMAP is plotted, and lineage specific UMAPs were recomputed with auto placed labels to inspect local structure and potential batch effects. On the same PCA graph, we fit Leiden, Louvain, and K-Means to obtain three labels. We used Silhouette as the primary algorithmic metric, because of its label-free and balancing compactness and separation. Besides, we report Calinski–Harabasz, Davies–Bouldin, ARI, Rand, and NMI to quantify agreement and stability. The goal of this part is to visually assess structure effects and quantitatively choose a robust clustering to support the following marker analysis.

4. Marker Selection and Gene Analysis

We performed marker discovery within the Skin subset. To avoid double-log transformation, raw counts were backed up and followed by library-size normalization and log1p. The annotated ROC cluster was considered as the positive class, and three complementary differential tests were used to find specific genes: Wilcoxon rank sum is single-cell friendly, variance overestimated t-test is more stable for small groups, and multinomial logistic regression can model multi-class structure. Based on these tests, we defined strict, majority, and union consensus sets.

To validate the program at the cluster level, we computed ROC score for each cell**,** and flagged clusters when their mean score exceeded the threshold we set. Our primary algorithmic metric is the agreement with *Table S3* ROC genes.

5. Data Denoising

In this part, we apply two complementary strategies on the normalized Skin subset. For PCA low rank reconstruction, we approximated X with the top 30 PCs to remove high dimensional random noise while keeping major variance. For kNN graph smoothing, we performed weighted neighbor average on the cell graph to denoise along the manifold and strengthen local consistency. After each denoising pass, we would run the same pipeline, from scale, PCA, neighbors, to Leiden**.** On the new cluster labels, we would perform Wilcoxon DE to obtain top markers from the same ROC candidate panel. Three primary algorithm performance metrics, Silhouette, Calinski–Harabasz, and Davies–Bouldin, are reported to quantify compactness and separation. Then we used marker overlap counts and Jaccard as the proxy to understand the stability and interpretability of markers.

6. Batch Integration over time

We treat days post-amputation as an explicit temporal batch factor and apply two complementary families of integration under the same HVG and PCA pipeline. The first method is Harmony, which aligns the potential representations at different time points and corrects global shifts in the principal component space. The second method is BBKNN, which operates at the graph level and constructs a time-balanced connectivity structure prior to embedding and clustering. Using both approaches in parallel reduced time-driven batch effects while avoiding excessive warping of biological structure. To assess model performance against the raw baseline, we reported standard internal clustering metrics, including Silhouette, Calinski–Harabasz, and Davies–Bouldin. Besides, we run differential expression and marker selection again on the updated cluster labels and quantify agreement with baseline marker sets through Jaccard.

Result

Each panel colors the same UMAP, embedding with a different label set including Leiden, Louvain, and K-means with k=25. The *major_lineage* is the aggregation. The global topology is stable across all methods, since the same islands appear in the same places. The *major_lineage* cleanly separates Immune, Neural, Red blood cells, Skin,

and Somite & others, which indicats good separation and limits batch dominance. For granularity, Leiden and K-means yield finer partitions with 25 clusters, whereas Louvain is coarser with 16 clusters, which is consistent with their typical behavior.

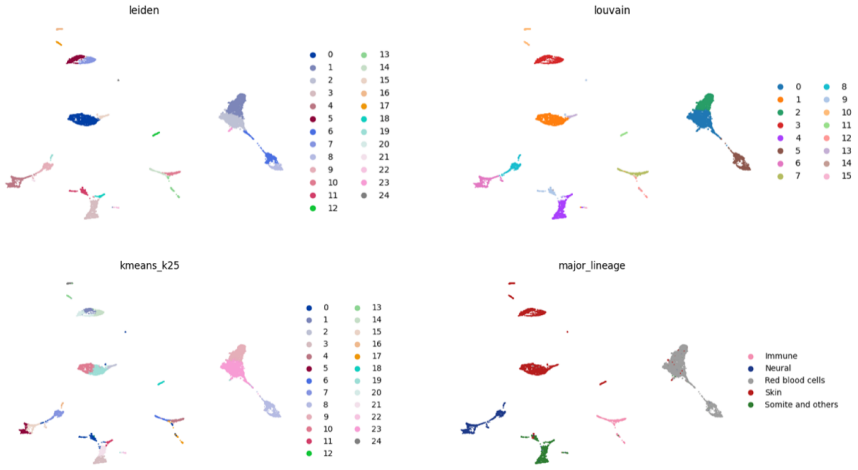

Figure 1 - UMAP

In fact, three clustering algorithms focus on different aspects. Louvain achieves the highest silhouette, 0.3240, and slightly better agreement with the reference labels, ARI = 0.5626, Rand = 0.9214, and NMI = 0.7292. Leiden delivers the sharpest boundaries, with the lowest Davies–Bouldin = 1.1467 and the strongest information agreement with the reference, NMI = 0.7363, Rand = 0.9243, and ARI = 0.5371 at a finer granularity. K-means yields the highest Calinski–Harabasz = 2916.23, indicating a strong within variance ratio, but shows a lower silhouette, 0.2966, and weaker alignment with the reference ARI = 0.4583 and NMI = 0.7199. Besides, pairwise concordance among the three is uniformly high. In general, these numbers suggest the global topology is consistent across methods, but with differences mainly in cluster count and boundary precision. Louvain is a good choice for a robust and interpretable coarse partition. Leiden is preferable for finer subtypes with cleaner boundaries. And K-means serves as a variance-driven baseline for comparison.

Each differential expression pipelines would produce its own ranked output list of genes. Based on this, we computed set overlaps to quantify cross-method agreement from Wilcoxon rank sum, t-test, and logistic regression. There are 54 genes selected by all three methods and 84 genes selected by at least two methods. Practically, the majority set of 84 genes is serves as a consensus marker panel, which was used to compute an ROC score. The genes obtained through multiple methods cover extracellular matrix and epithelial signaling pathways, including *FN1, LAMA5, COL14A1, COL27A1, NID2, FRAS1, FREM2, KRT8/KRT18, CLDN6, BAMBI, BMP2/4, ACVR1, JAG1, FGF7/FGF9, APOC1.LIKE, PLTP, GPX3, IFITM1*.

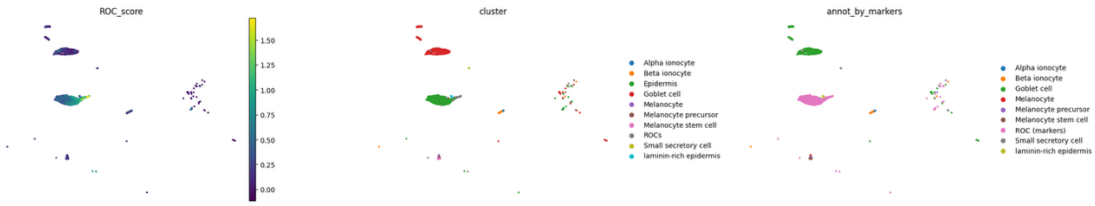

Figure 2 – ROC Clustering

The figure 2 shows the same UMAP embedding from three perspectives. The left panel colors each cell by its ROC gene-set score. The warmer colors indicate stronger enrichment and reveal a clear high score hot spot, while

most other regions remain low. The middle panel provides unsupervised cluster labels on the same coordinates, which means the high score hot spot maps almost entirely to a single cluster, indicating that the ROC signal is concentrated at the group level rather than scattered across clusters. The right panel shows biological annotations, and the cluster aligned with the hot spot is annotated as *ROC(markers)*, whereas distant lineages show little to no signal in the left panel. This figure demonstrates that the ROC gene set forms a coherent island in transcriptomic space, aligns with one unsupervised cluster, matches its biological annotation, and shows minimal spillover to other lineages, which is evidence of high specificity and discriminative power for the target cell population.

Besides, after comparing the ROC top markers obtained through three different methods with the marker genes listed in the Table S3, we got 19 genes from t-test in S3, 17 genes from Wilcoxon in S3, and 13 genes from logistic regression. The genes that selected by all three methods and in S3 are *CPA6, EGFL6, FGF7, FGF9, LPAR3, NID2, PLTP, SP9, TINAGL1, UNC5B, and VWDE.*

For the Denoising part, when compared with the baseline without noise reduction, we evaluated two types of noise reduction, PCA low-rank reconstruction and kNN graph smoothing. PCA low rank reconstruction weakened separability with the silhouette fell from 0.196 to 0.170, Calinski–Harabasz from 611 to 309, while Davies–Bouldin slightly improved to 1.578, which is a typical sign of over smoothing. kNN graph smoothing was gentler, with silhouette 0.171, CH 484, and DB 1.409. Importantly, its top 50 ROC markers overlapped the baseline more, with Jaccard = 0.351, than PCA low-rank 0.205.

To realize batch integration, with days post-amputation is the batch key, we presented that Harmony improved compactness and alignment, showing silhouette 0.215, CH 739, and DB 1.373. BBKNN gave slightly lower silhouette in 0.185, strong global structure with CH 758 and DB 1.473, and the highest marker reproducibility versus baseline, Jaccard = 0.611 vs Harmony 0.404.

Conclusion

This project reproduces and extends the ROC analysis from Xenopus tail single-cell RNA-seq. At the global clustering level, three clustering algorithms agree on the global topology. Louvain attains the highest silhouette, Leiden yields the sharpest boundaries, and K-means maximizes cluster separation in the variance sense, jointly delineating clean lineages. Within the skin subset, three complementary DE tests converge strongly, producing a majority consensus panel of 84 genes. The ROC gene set score forms a single high signal hot spot on the UMAP that maps similarly to an unsupervised cluster and matches the biological ROC annotation. For data denoising, PCA low-rank reconstruction tends to over-smooth, whereas kNN graph smoothing is gentler and preserves marker stability better. For batch integration over time, Harmony improves compactness and separation, while BBKNN offers the strongest marker reproducibility versus the baseline without over warping biology.

Overall, ROCs are distinct and reproducible in transcriptomic space. The mild denoising plus temporal integration jointly enhance clustering interpretability and marker robustness, which provides the pipeline for possible pathway analysis.