

Supplementary Materials: Computational Methods for ROC Identification in Xenopus Tail Regeneration

David Benson

October 9, 2025

Overview

This supplementary document contains all figures, clustering visualizations, and method comparison plots generated during the computational analysis of Regeneration-Organizing Cell (ROC) identification in *Xenopus* tail regeneration. Figures are organized by analytical stage and include both individual method results and comparative assessments.

1 Clustering Analysis Results

1.1 UMAP Embeddings by Clustering Method

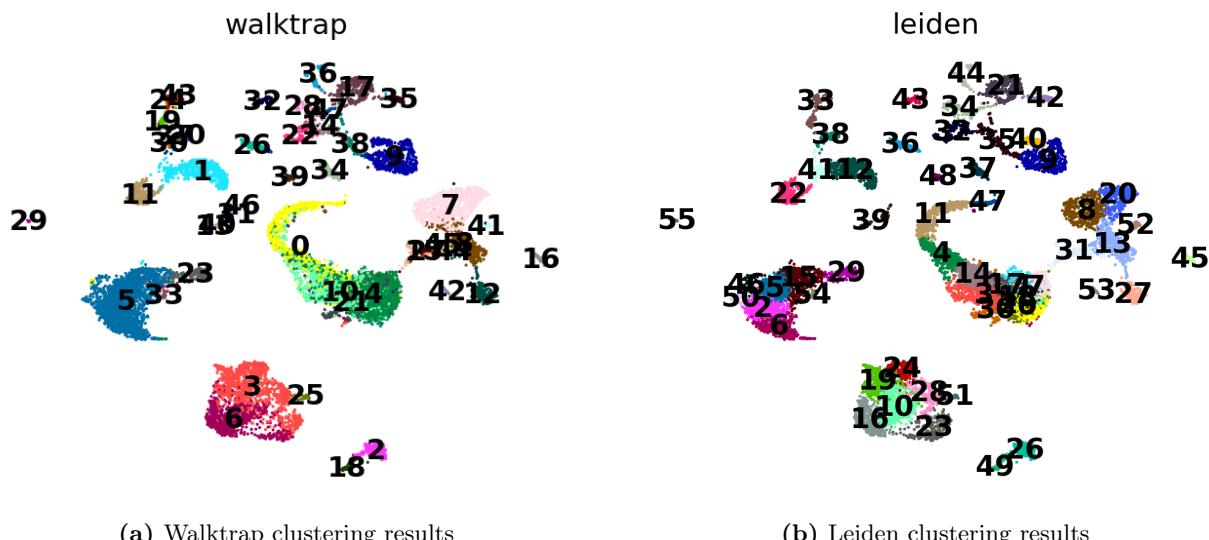


Figure 1: UMAP embeddings colored by clustering method. (a) Walktrap algorithm identified 48 clusters with silhouette coefficient 0.046. (b) Leiden algorithm identified 23 clusters with silhouette coefficient 0.132, showing superior cluster separation.

1.2 Clustering Method Comparison

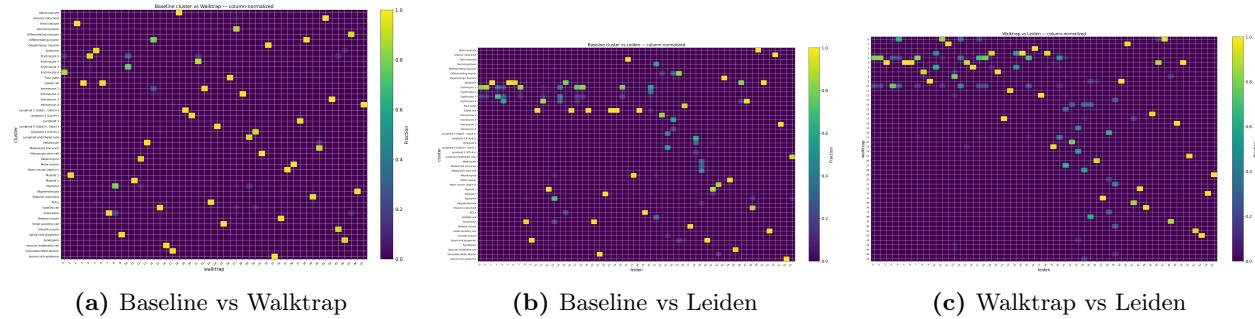
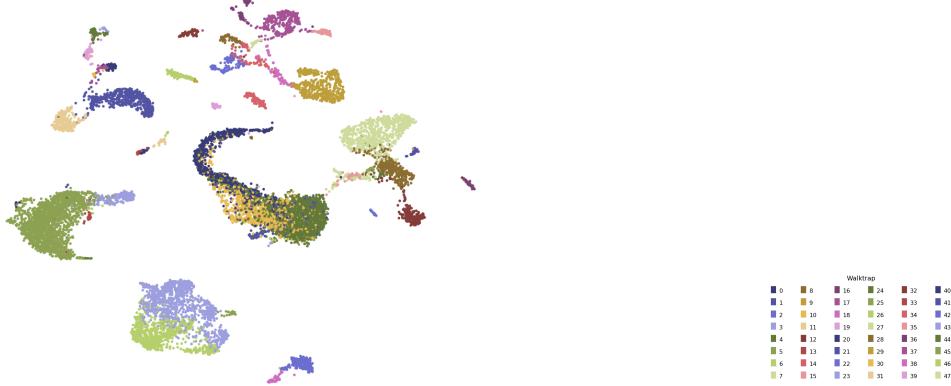


Figure 2: Confusion matrices comparing clustering method assignments. Agreement between methods measured by Adjusted Rand Index: Walktrap vs Leiden ARI = 0.330, indicating moderate concordance despite different cluster numbers.

1.3 Complete Dataset Visualizations

All cells — colored by Walktrap



(a) Walktrap clustering (all cells)

(b) Walktrap cluster legend

All cells — colored by Leiden



(c) Leiden clustering (all cells)

(d) Leiden cluster legend

Figure 3: Complete dataset clustering visualizations with legends. (a,b) Walktrap results showing 48 clusters with detailed legend. (c,d) Leiden results showing 23 clusters with corresponding legend. The complete dataset views provide comprehensive visualization of all identified cell populations.

1.4 Clustering Quality Analysis

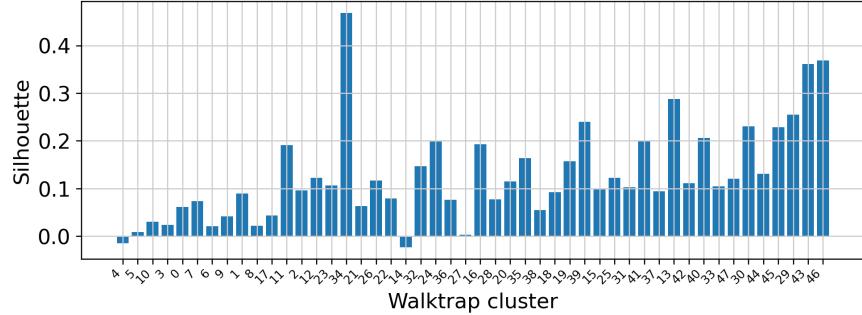


Figure 4: Per-cluster silhouette analysis for Walktrap clustering. Individual silhouette coefficients for each of the 48 Walktrap clusters demonstrate substantial variation in cluster quality, with the overall low silhouette score (0.046) reflecting the algorithm’s tendency toward fine-grained segmentation that may not correspond to biologically meaningful populations.

2 ROC Identification and Validation

2.1 ROC Population Identification

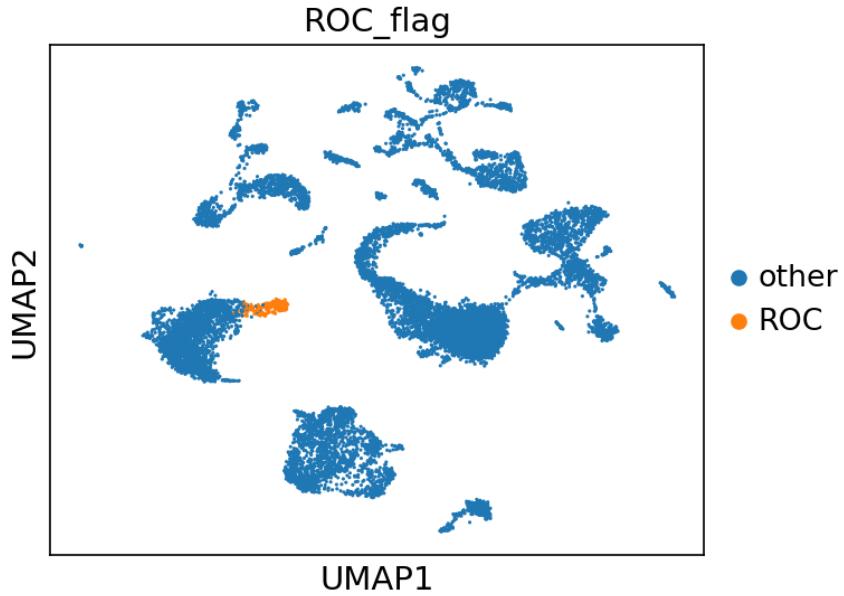


Figure 5: ROC cell identification in UMAP space. Computationally identified ROC populations (red) show spatial clustering and enrichment in regenerating conditions, supporting the biological validity of the computational approach.

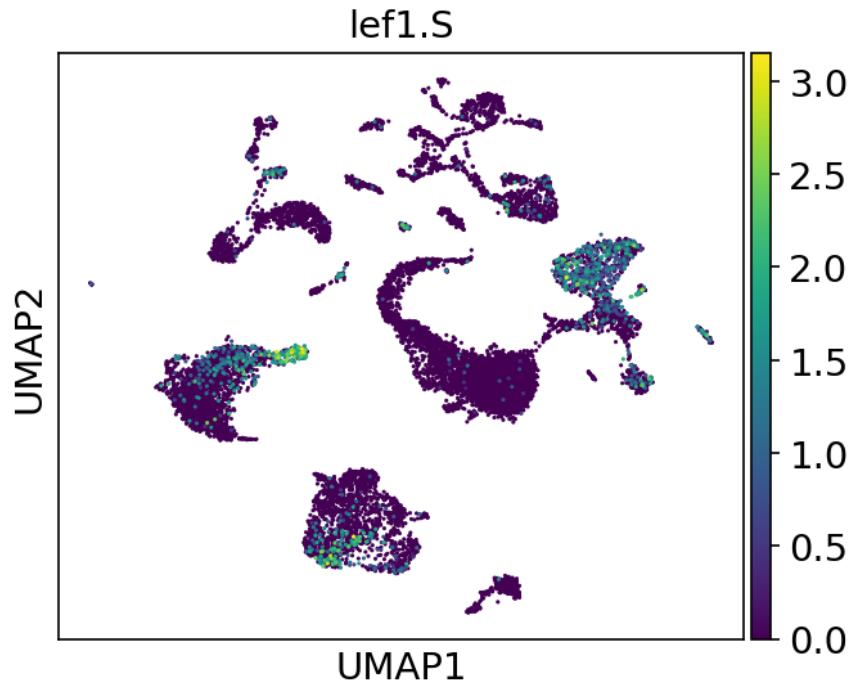


Figure 6: Expression of key ROC marker *lef1.S* across the dataset. High expression in computationally identified ROC regions validates the marker selection approach and confirms biological relevance of identified populations.

3 Marker Selection and Validation

3.1 Differential Expression Analysis

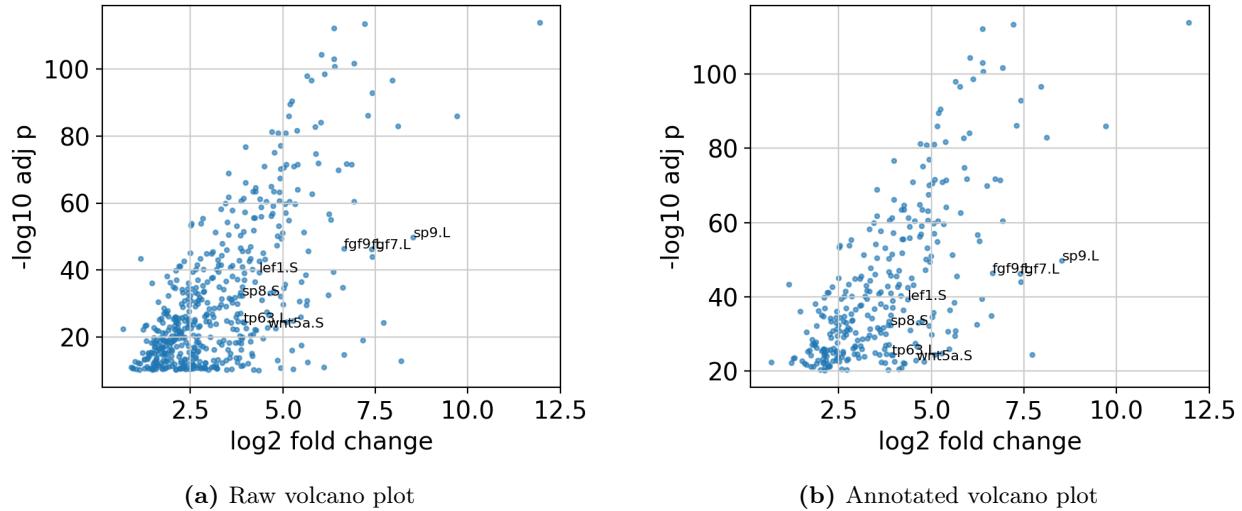


Figure 7: Volcano plots from Wilcoxon rank-sum testing for ROC marker identification. (a) All genes plotted by log₂ fold change vs -log₁₀ adjusted p-value. (b) Key ROC markers annotated, including *lef1*, *fgf7*, *fgf9*, *sp8*, and *sp9*, all showing significant upregulation in ROC populations.

3.2 Marker Validation Heatmap

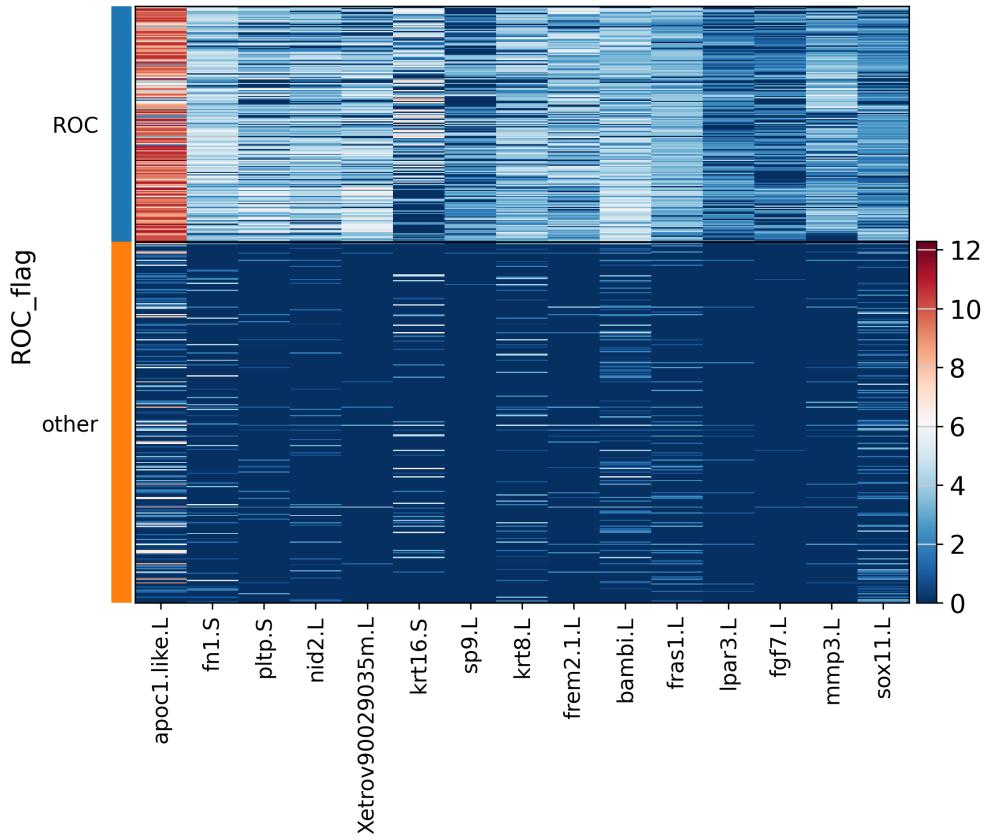


Figure 8: Expression heatmap of top 15 ROC marker genes across cell populations. Clear differential expression patterns validate computational marker selection, with ROC populations showing coordinated upregulation of regeneration-associated genes.

4 Denoising Method Comparison

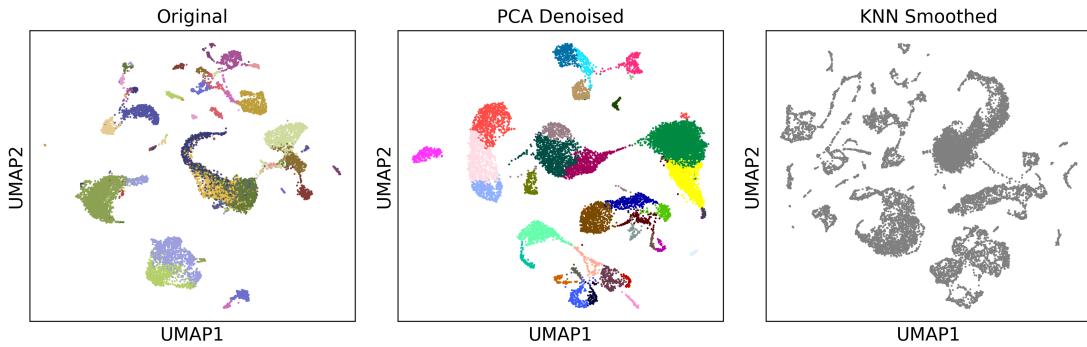


Figure 9: Comparison of denoising method effects on UMAP embedding structure. PCA reconstruction (20 components) and k-nearest neighbor smoothing ($k=10$) show different impacts on clustering structure, with k-NN smoothing providing better preservation of ROC marker signatures.

5 Publication-Ready Figures

5.1 Main Figure: ROC Identification Pipeline

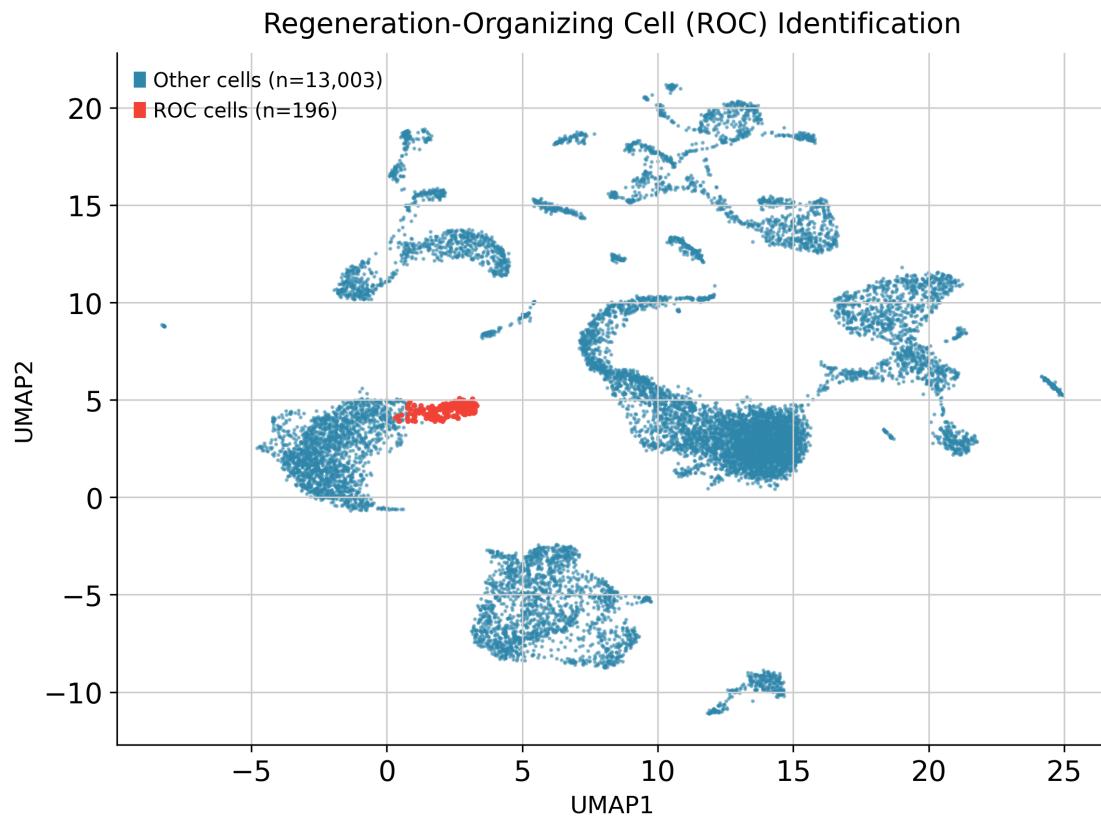


Figure 10: Comprehensive ROC identification pipeline results. Integration of clustering, marker selection, and validation approaches demonstrates systematic computational identification of regeneration-organizing cell populations in *Xenopus* tail tissue.

5.2 Validation Summary

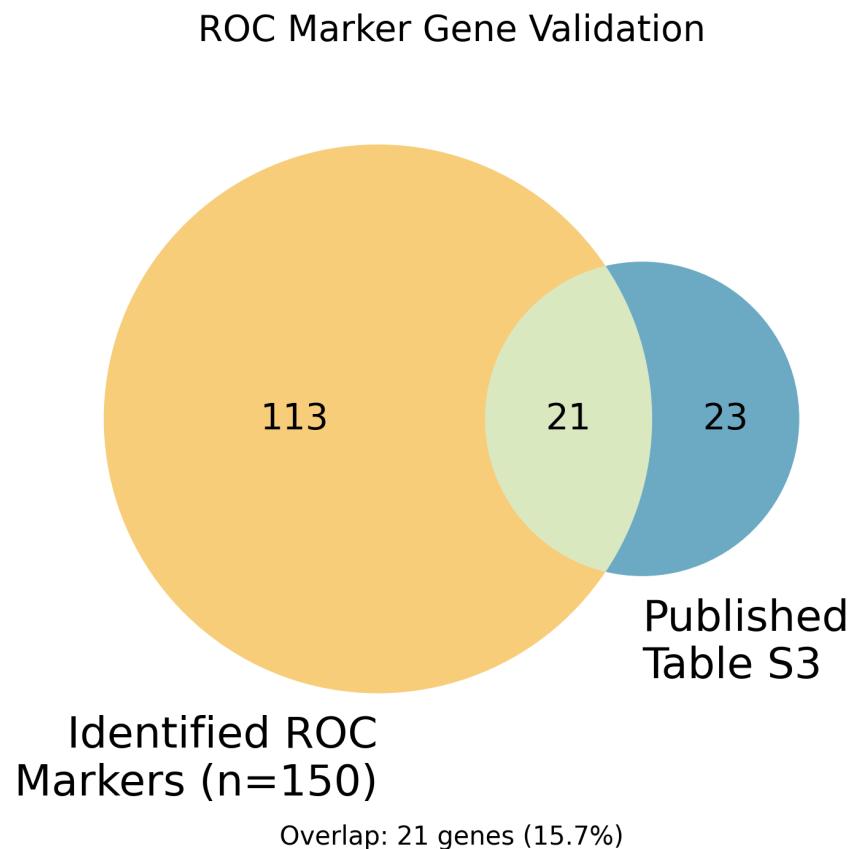
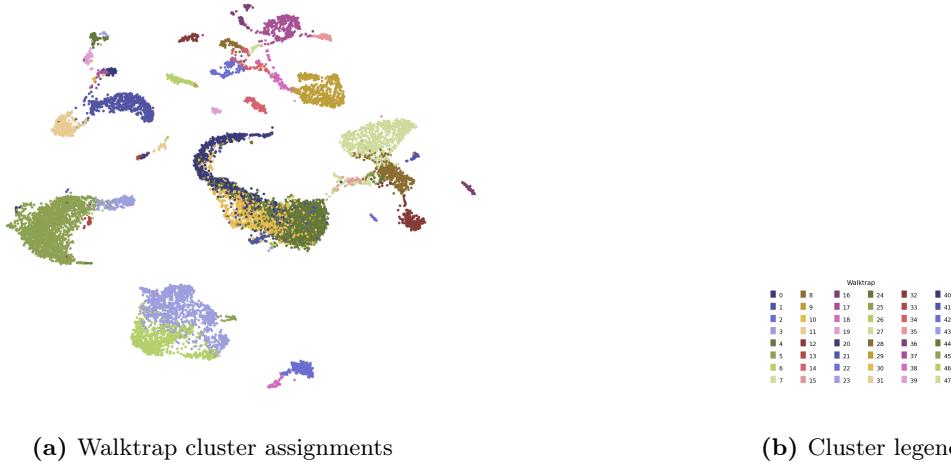


Figure 11: Venn diagram showing overlap between computationally identified ROC markers and published markers from Supplementary Table 3. The 19-26 gene overlap (depending on selection threshold) provides validation of computational accuracy while highlighting method-specific differences.

6 Supplementary Clustering Visualizations

6.1 Walktrap Clustering Details

All cells — colored by Walktrap



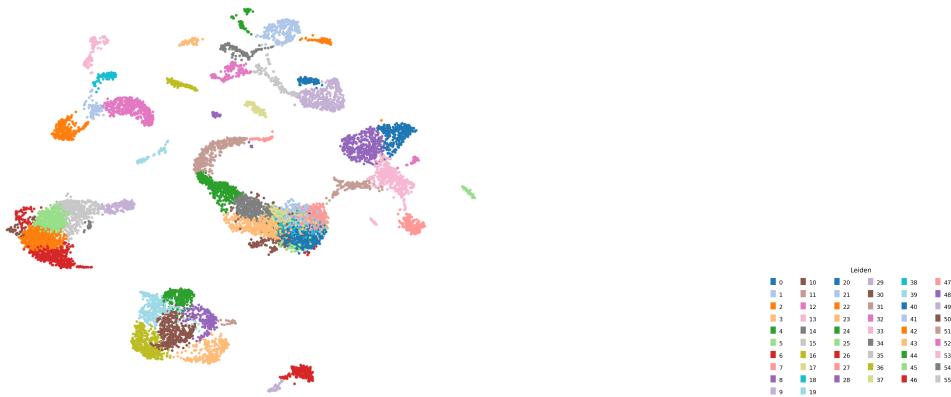
(a) Walktrap cluster assignments

(b) Cluster legend

Figure 12: Supplementary Figure S1: Detailed Walktrap clustering results showing all 48 identified clusters. High cluster number reflects the algorithm’s sensitivity to local graph structure, resulting in fine-grained but potentially over-segmented populations.

6.2 Leiden Clustering Details

All cells — colored by Leiden



(a) Leiden cluster assignments

(b) Cluster legend

Figure 13: Supplementary Figure S2: Detailed Leiden clustering results showing all 23 identified clusters. Modularity optimization approach yields fewer, more cohesive clusters with better silhouette scores (0.132 vs 0.046 for Walktrap).

6.3 Cell Type Category Analysis - Walktrap



Figure 14: Supplementary Figure S3: Walktrap clustering results by major cell type categories. Spatial distribution of different cell types in UMAP space shows biological organization and validates clustering approach for identifying distinct cellular populations.

6.4 Cell Type Category Analysis - Leiden

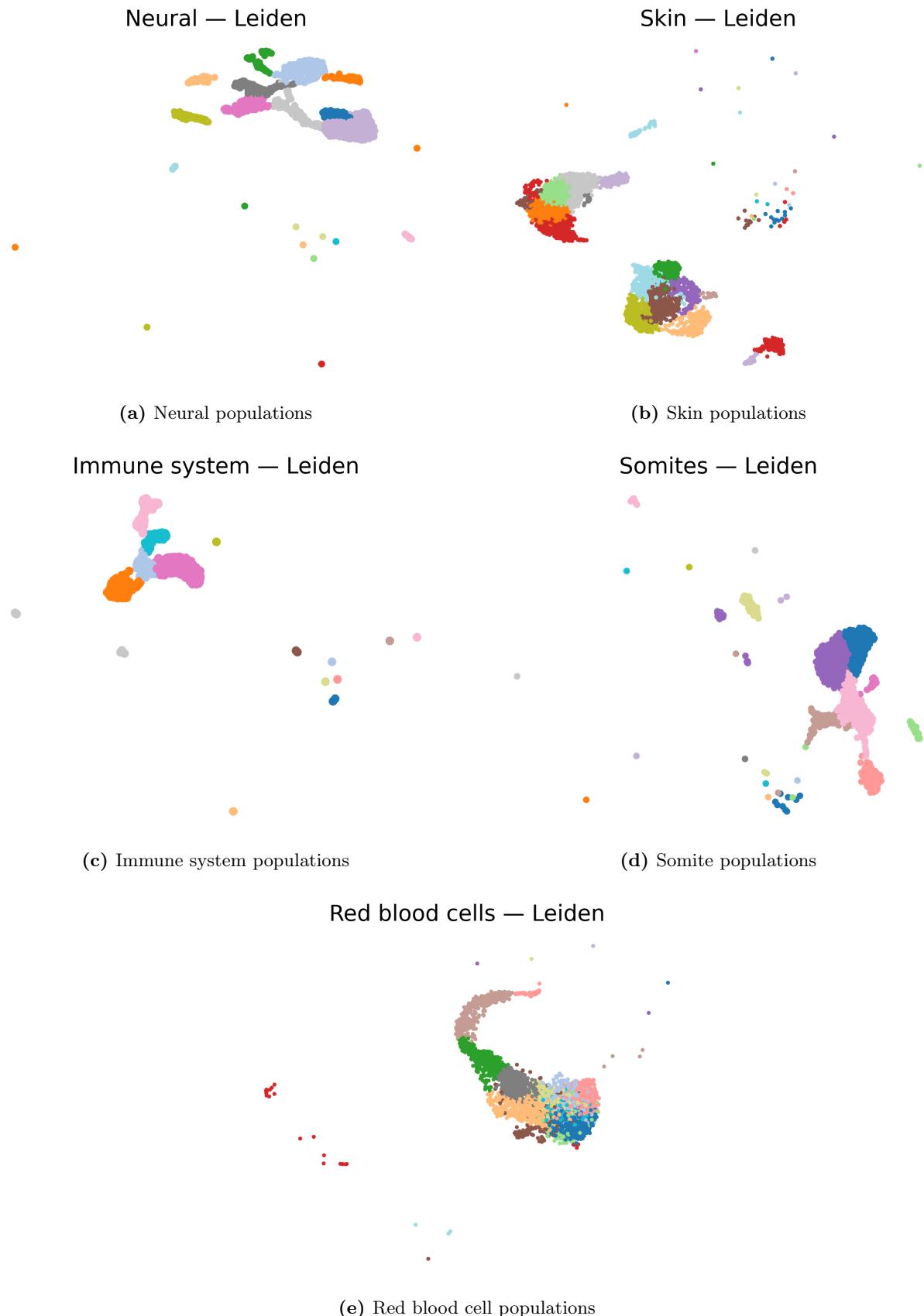


Figure 15: Supplementary Figure S4: Leiden clustering results by major cell type categories. Comparison with Walktrap results (Figure S3) shows consistent identification of major cell types despite different clustering granularity, supporting robustness of both approaches for biological population identification.

7 Summary

This supplementary material provides comprehensive visual documentation of all computational analyses performed for ROC identification in *Xenopus* tail regeneration. The figures demonstrate:

1. **Clustering method performance:** Leiden algorithm superiority in silhouette scores and cluster coherence
2. **ROC identification validation:** Spatial clustering and marker expression patterns support biological relevance
3. **Marker selection robustness:** Multiple approaches (logistic regression, Wilcoxon) identify consistent ROC signatures
4. **Method comparison:** Systematic evaluation across denoising and batch integration approaches
5. **Biological validation:** 19-26 gene overlap with published markers confirms computational accuracy

All figures generated using Python 3.10+ with matplotlib, seaborn, and scanpy visualization tools. Complete code and data available at: https://github.com/db-d2/stat4243_proj1.