Mini Project 1: Finding the Frog's Source of Regenerative Power

Author: Dasha Lykova
Affiliation: Columbia University
Dataset: *Xenopus laevis* tail regeneration (ArrayExpress E-MTAB-7716)
GitHub (public): https://github.com/daria-ly/HowCanAFrogGrowItsTailBack
Colab (runnable):
https://colab.research.google.com/drive/1z3OeciJ95r8nOCgwL67EFv2DzRQpmCMI?usp=sharing

Research question: Where is the Regenerative Organizing Cell in the Frog tail located and what genes that make this cell different from all other cells.

Abstract
I analyzed single-cell RNA-seq from regenerating *Xenopus* tail skin to identify the Regeneration-Organizing Cell (ROC) and to quantify how preprocessing choices affect its recovery. After PCA followed by UMAP, and clustering with Leiden, Louvain, and K-Means, I localized the ROC to Leiden cluster 18. This call is supported by **(1)** a ROC gene-set score that peaks in cluster 18 (mean 0.665; all other clusters ≤ 0.122), and **(2)** S3 enrichment among top-ranked ROC markers (Wilcoxon, ROC vs. rest): 5/50, 7/100, 20/200—i.e., 5, 7, and 20 S3 genes appear within the top 50, 100, and 200 markers of the ROC cluster, respectively. S3 coverage in my dataset is 45/46 genes detectable after symbol mapping.

Baseline clustering metrics in PCA space were: silhouette **(**Leiden 0.249**,** Louvain 0.253**,** K-Means 0.439**)**; ARI/NMI **(**Leiden 0.487/0.771**,** Louvain 0.517/0.767**,** K-Means 0.526/0.731**)**. I evaluated denoising (MAGIC, scVI) and batch integration (Harmony, BBKNN): MAGIC slightly increased silhouette but reduced ARI/NMI; scVI decreased Euclidean silhouettes while maintaining agreement; Harmony/BBKNN preserved clustering quality. ROC identity was unchanged across these variants. Robustness checks (Leiden resolution 0.6/1.0/1.2 and an 80% subsample) kept Top-50 overlap at 5**.** I therefore report baseline Leiden as the primary ROC call, with alternative methods used as controls and sensitivity analyses.

Introduction
The frog tail **Regeneration-Organizing Cell** (**ROC**) is a small epidermal population proposed to coordinate regeneration by secreting pro-growth signals (FGF/WNT/TGF-β). My goal was to **(1)** recover the ROC in tail skin scRNA-seq, **(2)** quantify clustering quality, **(3)** identify ROC markers and compare them to Supplementary Table 3 (S3) from the provided research paper, and **(4)** measure how denoising and batch integration affect both clustering and ROC recovery.

Methods

1. **Data, normalization, and logging**
   I analyzed 13,199 cells from E-MTAB-7716. Counts were normalized to 10,000 UMIs and log1p transformed; `.raw` stored the logged snapshot for `use_raw=True`. I verified `.raw` and the log1p flag and confirmed sparsity-appropriate behavior.

2. **Embedding and clustering**
   I computed PCA (retaining 50 PCs to capture major biological variation without amplifying noise), built a 15-nearest-neighbors (kNN) graph in that space to preserve local neighborhoods, and then computed UMAP. I ran Leiden and Louvain on the kNN graph and K-Means in the appropriate representation (PCA for baseline, X_scVI for scVI, X_harmony for Harmony). I evaluated clustering in the same space used to cluster, reporting Silhouette (geometry) and ARI/NMI versus a reference label (`cluster`).

3. **Marker selection and ROC call**
   I ranked per-cluster markers with Wilcoxon (Leiden/Louvain) and logreg (K-Means). I then assigned ROC to the cluster with the largest overlap with the published ROC list (Table S3), quantified as Top-K overlap: for each candidate cluster I took its top K ranked markers (K = 50, 100, 200) and counted how many were S3 genes. I computed a ROC gene-set score (`score_genes`) and summarized per cluster. For UMAP feature panels I used a shared 99th-percentile cap so colors are comparable across genes.

4. **Denoising and integration**
   MAGIC was used to impute a smoothed expression matrix. scVI was trained (n_latent=30) and its latent

(`X_scVI`) used for neighbors/UMAP; K-Means ran on `X_scVI`. Harmony produced `X_harmony` (neighbors/UMAP on `X_harmony`; K-Means on `X_harmony`). BBKNN built a batch-balanced graph on PCA; K-Means ran on PCA. I reclustered each variant and recomputed Silhouette/ARI/NMI for Leiden/Louvain/K-Means.

5. **Robustness analyses**
   I assessed robustness with Leiden only, since it performed best by both S3 overlap and ROC gene-set score. I re-called ROC across resolution 0.6/1.0/1.2 and an 80% subsample (fixed seed), recording Top-K S3 overlaps (Top-50/100/200).

6. **Code availability**
   A runnable Colab notebook reproduces the analysis. Package versions are pinned in requirements.txt which can be found on GitHub:
   GitHub: https://github.com/daria-ly/HowCanAFrogGrowItsTailBack
   Colab: https://colab.research.google.com/drive/1z3OeciJ95r8nOCgwL67EFv2DzRQpmCMI?usp=sharing

Results

1. **Baseline clustering and metrics:**

Table 1: Baseline clustering metrics (PCA space)

Method	*Silhouette*	*ARI*	*NMI*
Leiden	0.249	0.487	0.771
Louvain	0.253	0.517	0.767
K-Means	0.439	0.526	0.731

ARI/NMI are computed against the cluster reference labels. All values rounded to 3 decimal places.

K-Means shows the highest silhouette (0.439), which is expected because it optimizes spherical separation in PCA space. Geometrically its clusters are the most compact.
Leiden and Louvain have very similar silhouettes (~0.25), typical for scRNA-seq where clusters are separated but not perfectly tight. Against the external cluster reference, Louvain has a slightly higher ARI (0.517 vs 0.487) while NMI is essentially the same (0.767 vs 0.771), indicating both graph-based partitions agree well with the reference.

2. **Graphs:**

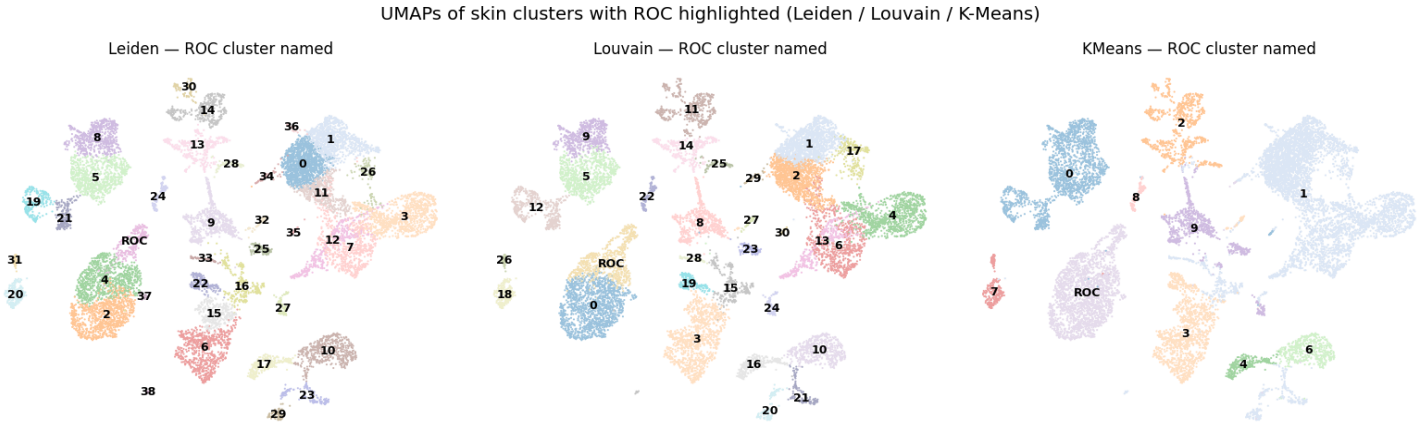

Figure 1A | Global clustering and ROC location across methods
Leiden, Louvain and K-Means UMAPs (numeric cluster IDs). ROC is annotated (Leiden 18, Louvain 7, K-Means 5). Leiden/Louvain: community detection on a 15-NN graph built from the first 50 PCs; K-Means: PCA space. See Figure 2 for gene-expression support.

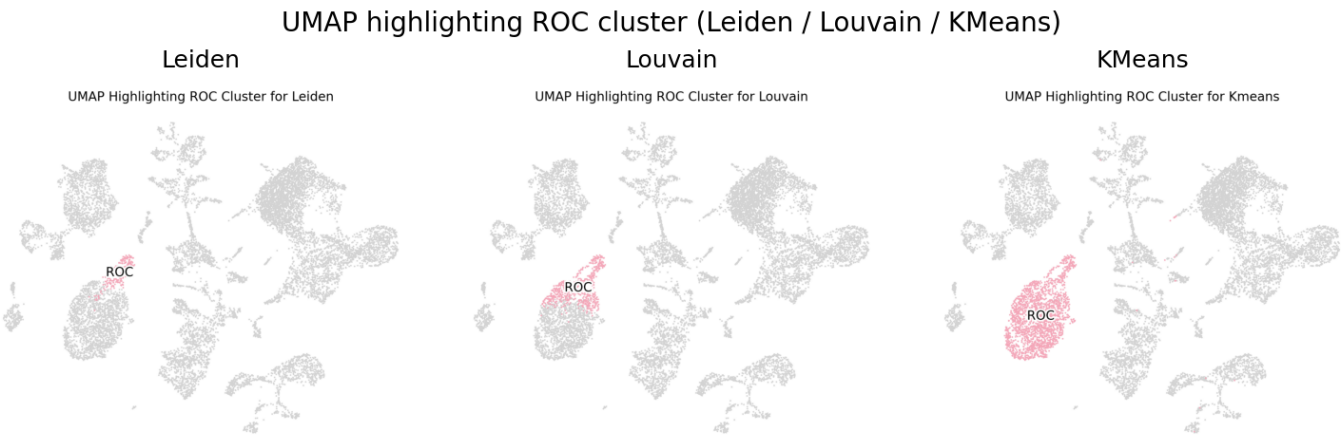

Figure 1B | ROC highlighted across methods (others gray)
UMAPs with the Regeneration-Organizing Cell (ROC) colored and all other cells shown in light gray. ROC is annotated (Leiden 18, Louvain 7, K-Means 5). Leiden/Louvain: community detection on a 15-NN graph built from the first 50 PCs; K-Means: PCA space. See Figure 2 for gene-expression support.
Top-50 S3 overlap for the ROC: Leiden 5, Louvain 3, K-Means 2 (see Tables 2A, 2B, 2C).

Figure 1A and **Figure 1B** show that across all three methods the ROC sits in the same UMAP region, but the size/shape of the called ROC cluster differs. Leiden draws the tightest/smallest ROC patch (crisper boundary), Louvain is broader (slightly more spillover into neighboring cells), and K-Means is the largest/most diffuse (centroid-based partition pulls in nearby cells). This ordering matches the quantitative signal: Leiden's tighter call coincides with the highest Top-50 S3 overlap, whereas Louvain and especially K-Means include more peripheral cells and recover fewer S3 genes.

3. Marker selection and S3 overlap:

Table 2A: Enrichment of published ROC genes (S3) among top-ranked markers of Leiden cluster 18			
Top-K	***Overlap***	***Precision***	***Recall***
50	5	0.10	0.11
100	7	0.07	0.15
200	20	0.10	0.43

Top-K overlap is computed on the Wilcoxon-ranked markers for Leiden cluster 18.
Overlap = number of S3 genes within the top K. Precision = Overlap/K. Recall = Overlap/46 (45/46 S3 genes detectable overall). Values rounded to three decimal places.

Table notes:
Shared S3 at Top-50: EGFL6, FREM2, IGFBP2, NID2, PLTP
Shared S3 at Top-100: EGFL6, FREM2, IGFBP2, LPAR3, NID2, PLTP, VWDE
Shared S3 at Top-200: CPA6, EGFL6, FGF7, FGF9, FGFR4, FREM2, IGFBP2, ISM2, JAG1, KRT, LAMB2, LEF1, LPAR3, NID2, PLTP, SP9, TINAGL1, TP73, UNC5B, VWDE

Table 2B: Enrichment of published ROC genes (S3) among top-ranked markers of Louvain cluster 7			
Top-K	***Overlap***	***Precision***	***Recall***
50	3	0.06	0.07
100	3	0.03	0.07
200	3	0.01	0.07

Top-K overlap is computed on the Wilcoxon-*ranked markers for Louvain cluster 7.*
Overlap = number of S3 genes within the top K. Precision = Overlap/K. Recall = Overlap/46 (45/46 S3 genes detectable overall). Values rounded to three decimal places

Table notes:
Shared S3 at Top-50: EGFL6, FREM2, IGFBP2
Shared S3 at Top-100: EGFL6, FREM2, IGFBP2
Shared S3 at Top-200: EGFL6, FREM2, IGFBP2

Table 2C: Enrichment of published ROC genes (S3) among top-ranked markers of KMeans cluster 5			
Top-K	***Overlap***	***Precision***	***Recall***
50	2	0.04	0.04
100	2	0.02	0.04
200	3	0.01	0.07

Top-K overlap is computed on the logistic-regression-ranked markers for K-Means cluster 5. Overlap = number of S3 genes within the top K. Precision = Overlap/K. Recall = Overlap/46 (45/46 S3 genes detectable overall). Values rounded to three decimal places.

Table notes:
Shared S3 at Top-50: EGFL6, FREM2
Shared S3 at Top-100: EGFL6, FREM2
Shared S3 at Top-200: EGFL6, FREM2, IGFBP2

Using Scanpy `score_genes`, I found that the ROC gene-set score peaked at 0.665 in the cluster assigned as ROC (Leiden 18), and was ≤ 0.122 elsewhere for this method. This quantitative enrichment mirrors the Top-K S3 results and motivates the expression panels in **Figure 2**. While for other methods ROC gene-set score peaked in similar clusters location-wise (numbered 7 for Louvain and 5 for K-means), the ROC gene-set score was much lower (0.357 for Louvain and 0.133 for KMeans), so Leiden was peaked as the best method for ROC identification.

4. **Graphs:**

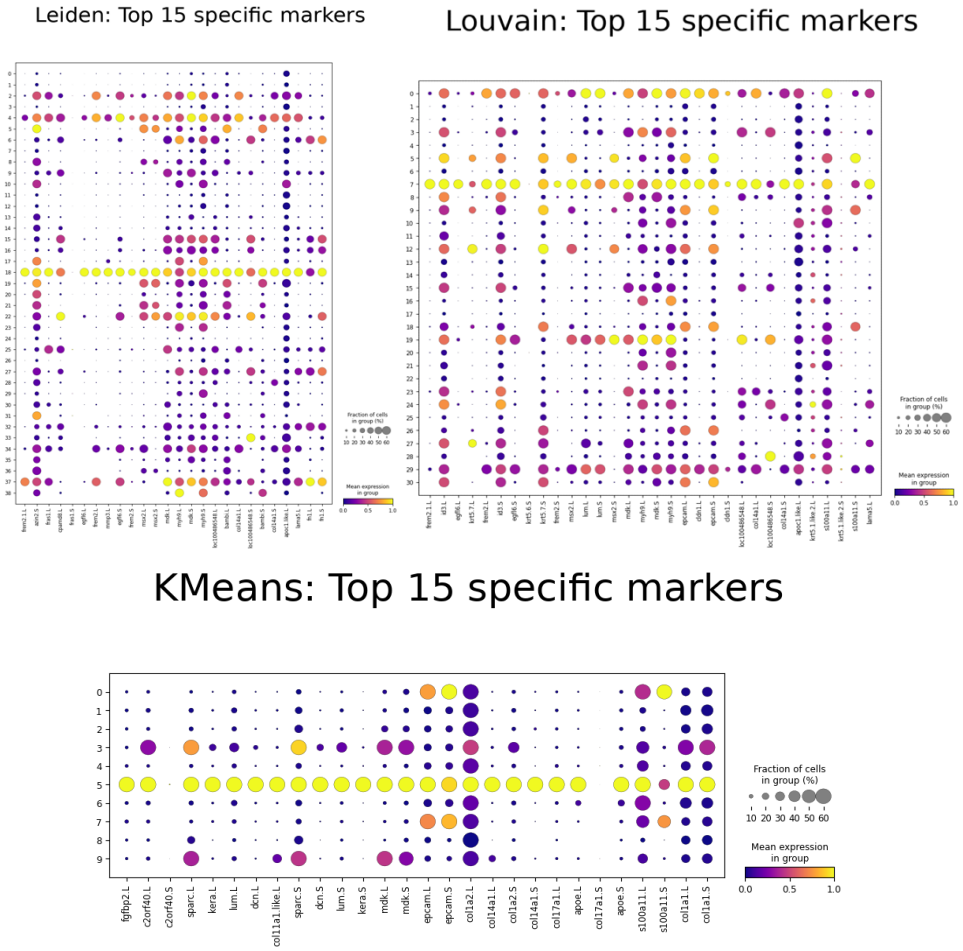

Figure 2A | Gene-expression evidence for the ROC: mean-expression dotplots.
Top-15 ROC markers per method using the method-specific ranking - Wilcoxon for Leiden (cluster 18) and Louvain (cluster 7)**;** logistic regression for K-Means (cluster 5)**.** Each panel is a dotplot across that method's clusters (dot size = % expressing; color = mean expression, standard-scaled per gene). ROC signal is strongest for Leiden**,** broader for Louvain, and most diffuse for K-Means**.** This is even more clear when we consider the high

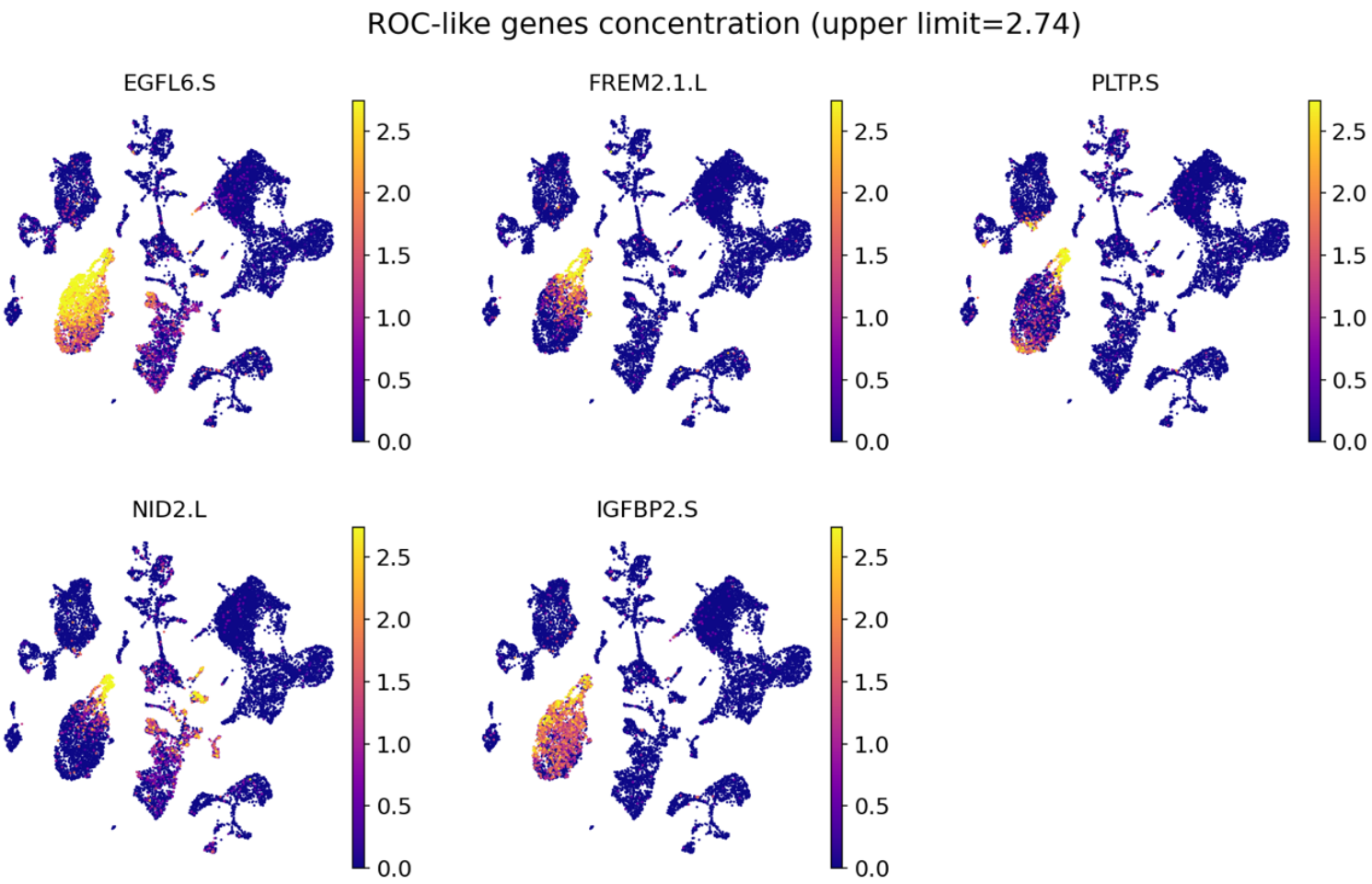

Figure 2B | Gene-expression evidence for the ROC: UMAP feature maps.
UMAP feature maps for EGFL6, FREM2, PLTP, NID2, IGFBP2, plotted with a shared 99th-percentile cap (lower limit = 0; same upper limit across panels). For each gene, expression is sharply concentrated in the same ROC region of the manifold and minimal elsewhere, indicating co-localized, ROC-specific signal. This spatial agreement complements the dotplots and the Top-K S3 overlaps, providing visual confirmation that ROC genes co-enrich in the ROC island identified earlier (independent of clustering method). These genes were chosen as high-confidence S3 hits that map cleanly and rank among the top ROC markers here, collectively reflecting the organizer's expected ECM/secretory/signaling profile.

Together, the panels show that the ROC signal is method-robust and spatially coherent. The dotplots confirm a higher signal-to-background for the ROC partition—most pronounced under Leiden—and align with the Top-K S3 differences reported in the tables. The feature maps provide an orthogonal spatial check: independent S3 genes converge on the same UMAP island, supporting a single biological compartment rather than scattered false positives. This ordering also follows each algorithm's objective: Leiden/Louvain optimize community detection on a k-NN graph, which isolates locally connected islands (Leiden typically tighter than Louvain), whereas K-Means partitions Euclidean space around centroids and tends to pull in peripheral cells, yielding a more diffuse call. In combination, these views substantiate the ROC call and explain why Leiden serves as the primary result while Louvain/K-Means function as contrasts.

5. **Denoising sensitivity (MAGIC, scVI):**

Table 3A: Clutering metrics after MAGIC denoising			
Method	***Silhouette***	***ARI***	***NMI***
Leiden	0.315	0.308	0.721
Louvain	0.292	0.361	0.731
K-Means	0.418	0.487	0.711
Table 3B: Clutering metrics after scVI denoising			
Method	***Silhouette***	***ARI***	***NMI***
Leiden	0.159	0.434	0.735
Louvain	0.153	0.470	0.734
K-Means	0.167	0.540	0.693

Relative to baseline, MAGIC made clusters look tighter but lowered agreement with the reference labels: Leiden/Louvain silhouettes increased to 0.315/0.292, while Leiden ARI/NMI dropped to 0.308/0.721. K-Means behaved similarly (0.418; 0.487/0.711 vs 0.439; 0.526/0.731 baseline), which is consistent with imputation smoothing boundaries in PCA space.

By contrast, scVI reduced Euclidean compactness (silhouettes 0.159/0.153/0.167 for Leiden/Louvain/K-Means) but kept agreement reasonable (e.g., Leiden ARI/NMI 0.434/0.735). For K-Means, ARI even increased slightly (0.540). Overall, denoising changed geometry in line with the methods' behavior—MAGIC smooths along the graph (tighter look, lower ARI/NMI), whereas scVI maps cells to a non-linear latent (lower Euclidean silhouette but similar agreement)—and in both cases the ROC remained in the same island and the cluster was unchanged, since analysis of clusters after denoising yield Jaccard score of 1.0 and centroid shift of 0.0 for both methods.

6. **Batch integration over time (Harmony, BBKNN):**

Table 4A: Clustering metrics after batch integration with Harmony			
Method	***Silhouette***	***ARI***	***NMI***
Leiden	0.270	0.506	0.779
Louvain	0.274	0.525	0.780
K-Means	0.429	0.514	0.718
Table 4B: Clustering metrics after batch integration with BBKNN			
Method	***Silhouette***	***ARI***	***NMI***
Leiden	0.241	0.495	0.766
Louvain	0.243	0.553	0.757
K-Means	0.439	0.526	0.731

After integrating time points, clustering stayed close to baseline and the ROC call did not change. With Harmony, silhouettes were 0.270/0.274/0.429 for Leiden/Louvain/K-Means (vs. 0.249/0.253/0.439 baseline), and agreement was slightly higher for the graph methods—Leiden ARI/NMI 0.506/0.779 (vs. 0.487/0.771), Louvain 0.525/0.780 (vs. 0.517/0.767)—with a small drop for K-Means (0.514/0.718 vs. 0.526/0.731). BBKNN showed a very similar picture: silhouettes 0.241/0.243/0.439 (near 0.249/0.253/0.439), and agreement 0.495/0.766 (Leiden) and 0.553/0.757 (Louvain), again on par with baseline. Overall, both methods harmonize batches without distorting the biology—the ROC island stays in the same spot and the assigned ROC cluster is unchanged, since analysis of clusters after denoising yield Jaccard score of 1.0 and centroid shift of 0.0 for both methods.

Robustness checks (Leiden):
I focused robustness on Leiden because it gave the strongest ROC signal (highest Top-50 S3 overlap and the clearest dotplot/feature-map separation):

- Resolution sweep (0.6 / 1.0 / 1.2): Top-K S3 overlap for the Leiden ROC cluster remained **5 / 12 / 23**–24 for K = 50 / 100 / 200**.**
- 80% subsample (fixed seed): overlaps were 5 / 11 / 23

These numbers match the main analysis within rounding, indicating that the ROC call is stable to reasonable parameter changes and to random cell removal.

Conclusion:

I reproduced the Regeneration-Organizing Cell (ROC) in regenerating *Xenopus laevis* tail skin and find it best captured by Leiden cluster 18. Multiple lines of evidence point to the same answer: the ROC cluster shows strong overlap with the published ROC gene set, a clear peak in a ROC gene-set score, and co-localized UMAP signal from canonical markers. In agreement with the paper, the ROC island expresses key epithelial/organizer genes such as FGF10, WNT5A, LEF1, and TGFB3, alongside other S3 hits (e.g., EGFL6, FREM2, PLTP, NID2, IGFBP2).

Method checks reinforced the call. MAGIC made clusters look crisper (higher silhouettes) but lowered cross-label agreement a bit; scVI did the opposite—smaller Euclidean silhouettes but similar agreement; and Harmony/BBKNN aligned time points without distorting the manifold. In every case, the ROC island stayed put and the assigned ROC cluster did not change. Finally, robustness tests (Leiden resolution sweep and an 80% subsample) returned essentially the same Top-K overlaps, indicating the result is stable to reasonable parameter choices and random cell removal.

Together, these analyses extend the original ROC finding with a transparent, reproducible pipeline and show how preprocessing and integration choices affect the look of the data without changing the biological conclusion: the ROC is a coherent, localized program, and Leiden 18 is a reliable final call.

Supplementary figures:

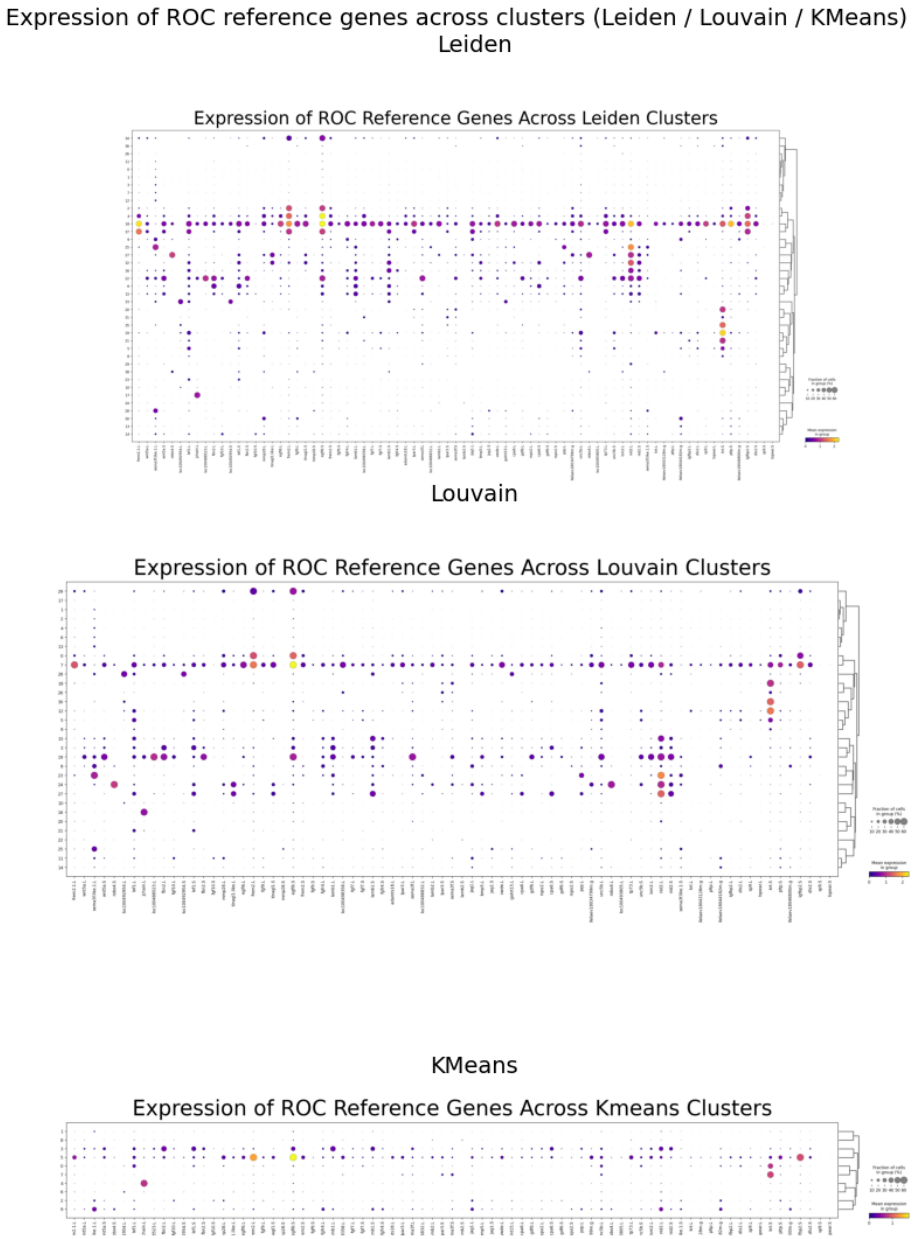

Figure S1 | Gene-expression evidence for the ROC: mean-expression dotplots.
All ROC markers per method using the method-specific ranking - Wilcoxon for Leiden (cluster 18) and Louvain (cluster 7)**;** logistic regression for K-Means (cluster 5)**.** Each panel is a dotplot across that method's clusters (dot size = % expressing; color = mean expression, standard-scaled per gene). ROC signal is strongest for Leiden**,** broader for Louvain, and most diffuse for K-Means**.**