Evaluating Clustering, Denoising, and Marker Selection Pipelines for Regeneration-Organizing Cells in *Xenopus* Tail scRNA-seq

Nicolas Garcia

October 15, 2025

Abstract

Single-cell RNA-seq of regenerating *Xenopus laevis* tails enables identification of Regeneration-Organizing Cells (ROCs), a specialized population proposed to coordinate regeneration. Here we systematically evaluate how preprocessing, denoising, batch integration, and clustering choices affect agreement with curated cell-type annotations and robustness of ROC marker genes. Using the E-MTAB-7716 dataset with provided `cluster` labels (including "ROCs"), we compare five pipelines (Base, MAGIC, scVI, Harmony, BBKNN), each combined with Leiden and K-means clustering. We quantify performance using Adjusted Rand Index (ARI) and internal metrics (Silhouette, Calinski–Harabasz, Davies–Bouldin), and identify ROC markers with Wilcoxon rank-sum tests and multinomial logistic regression. Marker sets are compared against the 50 ROC-associated genes from Supplementary Table 3 and across pipelines via top-K recovery curves and binary overlap maps. scVI+Leiden achieves the highest ARI (~ 0.59), MAGIC+K-means shows the strongest internal structure, and denoised pipelines (MAGIC, scVI) substantially improve enrichment of known ROC markers while preserving a stable core ROC signature. These results illustrate how methodological choices reshape clustering and marker discovery while confirming that key ROC features are robust.

1 Introduction

Tail regeneration in *Xenopus* involves coordinated activity of multiple cell types. Previous scRNA-seq work defined Regeneration-Organizing Cells (ROCs) as a distinct epidermal-like population enriched for signaling and extracellular matrix genes, supported by a curated ROC marker list (Supplementary Table 3). Because scRNA-seq analyses are sensitive to normalization, denoising, batch integration, and clustering, it is important to assess how these choices influence identification of the ROC population and its markers.

In this project we reanalyse the published dataset to (i) compare clustering pipelines against curated cell-type labels, (ii) assess how denoising and integration affect ROC marker selection using two complementary methods, and (iii) quantify concordance with the published ROC markers across pipelines.

2 Methods

2.1 Data, preprocessing and visualization

We used the *Xenopus* tail regeneration dataset (E-MTAB-7716), provided as a sparse count matrix with annotations, to construct an `AnnData` object containing raw counts, gene names, barcodes, and metadata (developmental stage, days post-amputation, batch, and published `cluster` labels). The label set includes a dedicated "ROCs" cluster (254 cells), used as the reference ROC population. Counts were normalized to 10^4 per cell, log-transformed as $\log_2(x+1)$, and filtered to remove low-quality cells and genes. Highly variable genes were selected using a Fano factor–based method with mean-expression bounds, and PCA (50 components) was performed on these HVGs, while raw counts were retained in `layers["counts"]`. UMAPs colored by tissue categories and ROC identity reproduced the major compartments and ROC positioning reported in Figure 1B of the original study.

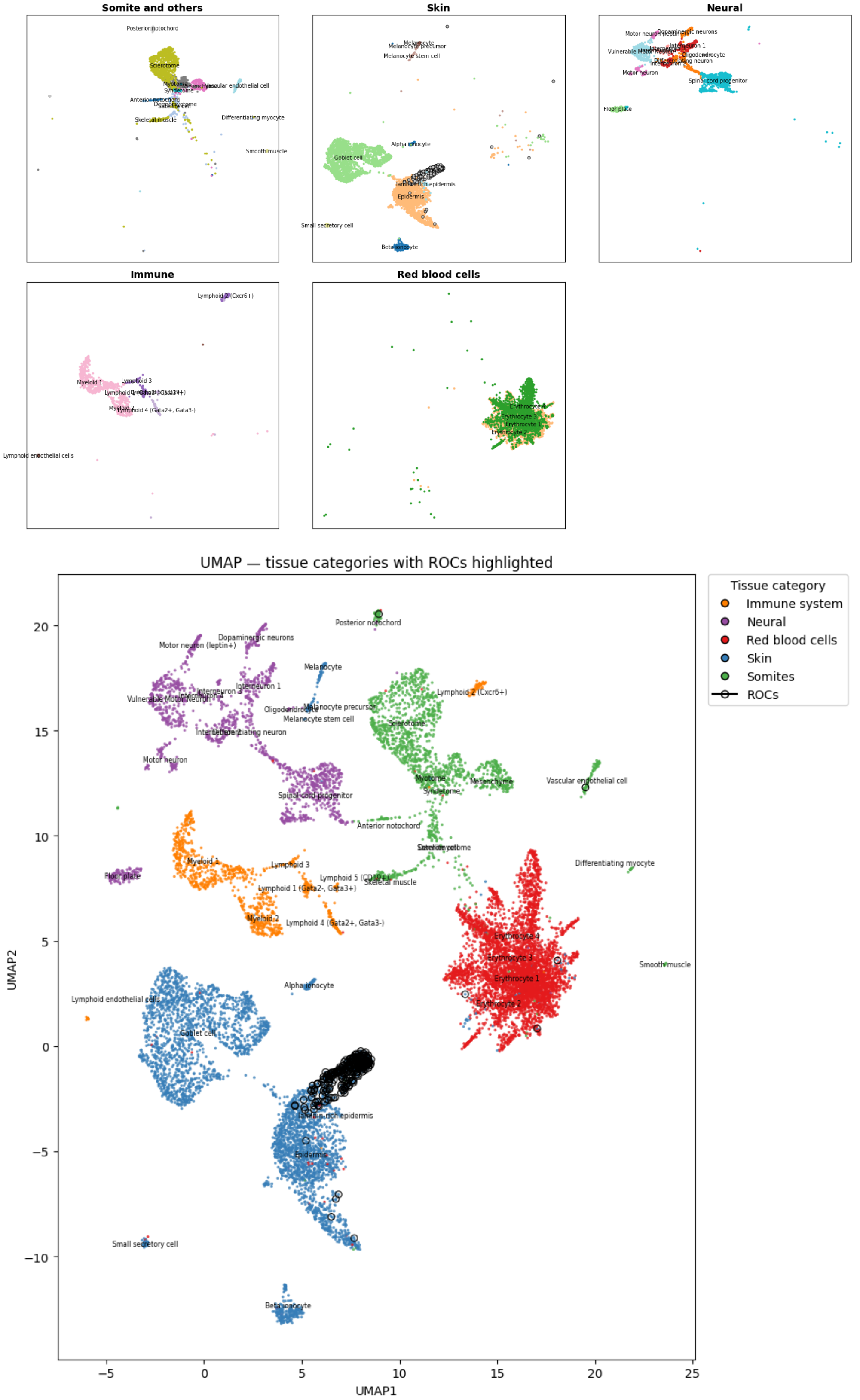

Figure 1: Reconstruction of the published UMAP visualization. Top: faceted UMAP by broad tissue category, showing separation of somite/mesenchyme, skin, neural, immune, and erythroid compartments. Bottom: UMAP colored by tissue category with ROCs outlined, confirming that our preprocessing and embedding reproduce the original Figure 1B structure and ROC localization.

2.2 Pipelines, embedding, and clustering

From this baseline we constructed five pipelines: (1) **Base**, using the log-normalized HVGs and PCA; (2) **MAGIC**, with MAGIC denoising followed by PCA and graph construction on imputed values; (3) **scVI**, with a 20-dimensional variational latent space fit on counts with batch as a covariate; (4) **Harmony**, applying Harmony integration on PCs; and (5) **BBKNN**, using a batch-balanced k-NN graph on PCs. MAGIC and scVI primarily modify expression (denoising), whereas Harmony and BBKNN primarily modify neighborhood structure (integration).

For each pipeline we computed a k-NN graph and UMAP embedding (cosine distance, $k = 20$, `min\_dist=0.5`) for visualization, and ran both Leiden and K-means clustering using the relevant space (PCA, scVI latent, Harmony PCs, or BBKNN graph). Curated `cluster` labels were not used during clustering.

2.3 Clustering metrics

Clustering performance was assessed using the Adjusted Rand Index (ARI) between each clustering and the curated `cluster` labels, plus Silhouette, Calinski–Harabasz (CH), and Davies–Bouldin (DB) indices as internal validation metrics. Silhouette and CH are higher-is-better; DB is lower-is-better. For figure visualization, internal metrics were linearly scaled to $[0, 1]$ per metric and DB inverted so that higher scores always indicate better structure.

The main configurations and metrics used in the summary figure are listed in Table 1.

Pipeline	Method	k	ARI	Silhouette	DB
scVI	scvi_leiden	29	0.593	0.141	1.790
Harmony	harmony_kmeans	46	0.536	0.325	1.125
BBKNN	bbknn_kmeans	46	0.526	0.349	1.152
Base	base_kmeans	46	0.526	0.349	1.152
BBKNN	bbknn_leiden	34	0.511	0.246	1.227
Base	base_leiden	36	0.484	0.242	1.318
Harmony	harmony_leiden	38	0.484	0.255	1.153
MAGIC	magic_kmeans	46	0.479	0.572	0.750
scVI	scvi_kmeans	46	0.333	0.133	1.875
MAGIC	magic_leiden	58	0.285	0.297	1.156

Table 1: Clustering metrics for selected pipeline+method combinations.

2.4 ROC marker selection and comparison

ROCs were defined purely by the curated label: cells with `cluster == "ROCs"`. For each pipeline we ran Scanpy's `rank_genes_groups` twice:

1. Wilcoxon rank-sum: ROCs vs. all other cells.
2. Multinomial logistic regression: all curated clusters jointly; ROC markers from ROC coefficients.

A shared helper (`run_roc_markers`) extracted the top 200 ROC markers from each method, computed their overlap, canonicalized gene names, and checked membership in the 50-gene ROC list from Supplementary Table 3. For each pipeline and method we counted how many Table 3 genes appeared in the top-K ROC markers for $K = 50, 100, 150, 200$, generated recovery-vs-K curves, and visualized binary presence/absence matrices for Table 3 genes across pipelines.

2.5 Code availability

All analysis code (preprocessing, clustering, metrics, marker selection, and figure generation) is available at: https://github.com/nicolas-garc/STATGR5243_Project1_FrogTail/tree/main

3 Results

3.1 Clustering performance

Figure 2 (from `Clustering_Metrics.png`) summarizes ARI and normalized internal metrics. scVI+Leiden shows the highest ARI (0.5929) relative to curated labels, indicating that the scVI latent space best aligns unsupervised clusters with the provided biology. Base, Harmony, and BBKNN with K-means all reach ARI ≈ 0.53 with similar Silhouette and CH values, suggesting that mild integration leaves cluster structure largely consistent with the baseline. MAGIC+K-means displays the strongest internal structure (Silhouette 0.57, highest CH, lowest DB 0.75) but reduced ARI (0.48), implying that extremely compact MAGIC clusters do not fully match the original partitioning. scVI+K-means and MAGIC+Leiden underperform in ARI and internal metrics relative to the best configurations.

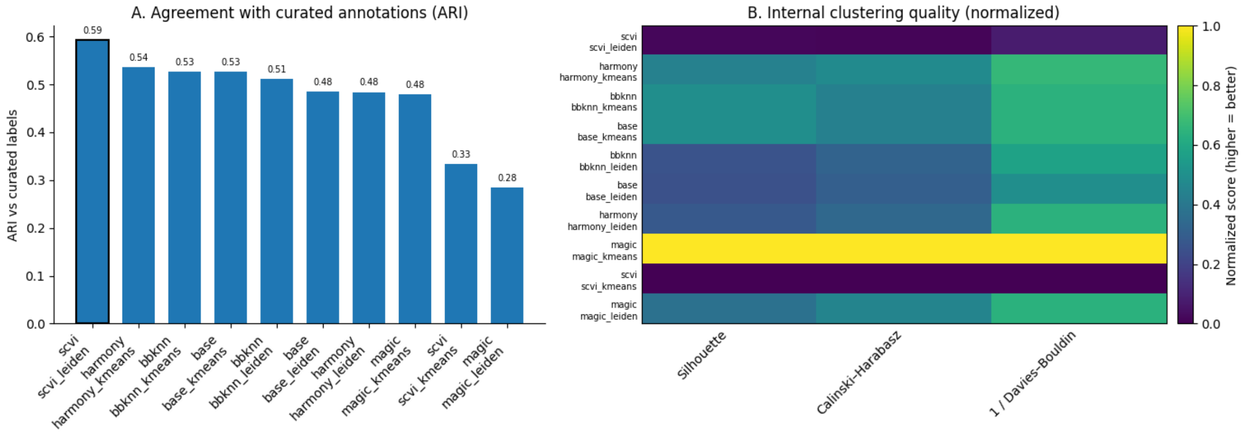

Figure 2: Summary of clustering quality. Left: ARI vs. curated labels for each pipeline+method. Right: normalized internal metrics (Silhouette, Calinski–Harabasz, inverse Davies–Bouldin), scaled so higher is better.

Overall, scVI+Leiden provides the best match to curated biology, while MAGIC+K-means optimizes geometric separation. Harmony and BBKNN behave similarly to the baseline, indicating that moderate batch integration preserves the major cell-type structure.

3.2 ROC marker recovery

We next examined how pipelines affect identification of ROC markers using the fixed ROC label set. For Wilcoxon-based top-50 markers (Figure 3), the Base, Harmony, and BBKNN pipelines each recover only 5–6 of 50 Table 3 genes, whereas MAGIC and scVI recover 18–19, showing that denoising substantially enriches known ROC markers among the highest-ranked genes.

Extending to $K = 100$ and $K = 200$ (Figures 4 and 5), MAGIC and scVI continue to outperform: by 200 markers they recover $\sim$37–39 of 50 ROC genes, whereas other pipelines plateau lower. Binary heatmaps at each K demonstrate that a core subset of canonical ROC genes (e.g. *EGFL6*, *LPAR3*, *NID2*, *PLTP*, *SP9*, *VWDE*) is shared across pipelines, while additional ROC genes are preferentially captured and ranked higher by the denoised pipelines.

Recovery-vs-K curves (Figure 6) summarize this behavior across thresholds. For Wilcoxon, MAGIC and scVI dominate over the full range of K, while Base, Harmony, and BBKNN grow slowly. Logistic regression curves show similar ordering though with smaller counts at low K, reflecting the more conservative nature of the multiclass model. Thus, denoising pipelines improve ROC marker recall in a way that is robust to the exact cutoff.

Across all analyses, Wilcoxon- and logreg-derived marker sets agree on the major ROC genes, and cross-pipeline comparisons reveal a consensus ROC program plus a small number of method-specific candidates.

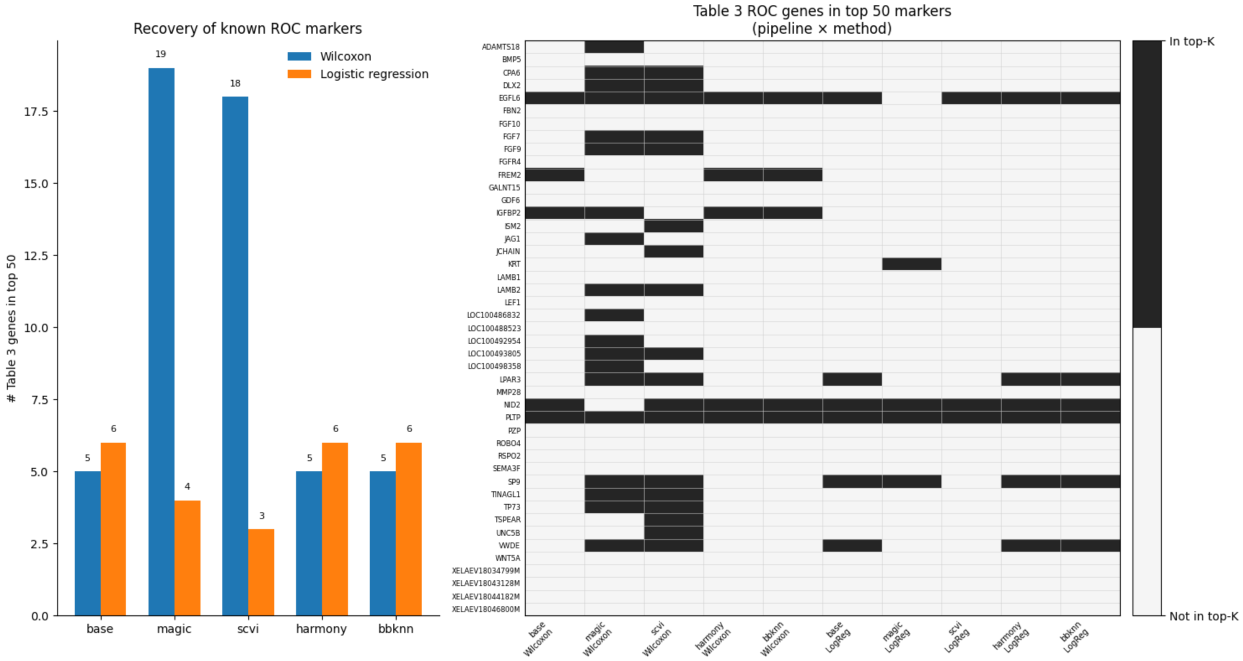

Figure 3: Top-50 ROC marker analysis. Left: number of Table 3 genes in top-50 ROC markers for each pipeline and method. Right: binary heatmap showing which Table 3 genes appear in each top-50 list.

Differences between pipelines mainly influence ranking strength and sensitivity rather than completely changing ROC identity.

4 Conclusion

On the *Xenopus* tail scRNA-seq dataset, scVI+Leiden provides the highest concordance with curated cell-type labels, while MAGIC+K-means offers the strongest internal clustering structure. Harmony and BBKNN maintain performance similar to the baseline and demonstrate that batch integration does not disrupt the major biological signal. Importantly, denoising with MAGIC or scVI markedly improves enrichment of known ROC markers across a wide range of top-K thresholds, and all pipelines recover a stable core ROC gene signature. These results show that biologically meaningful ROC features are robust but that preprocessing and clustering choices modulate clarity and ranking of markers; combining latent models, graph-based clustering, and complementary marker selection methods yields a reliable strategy for characterizing regeneration-organizing cells.

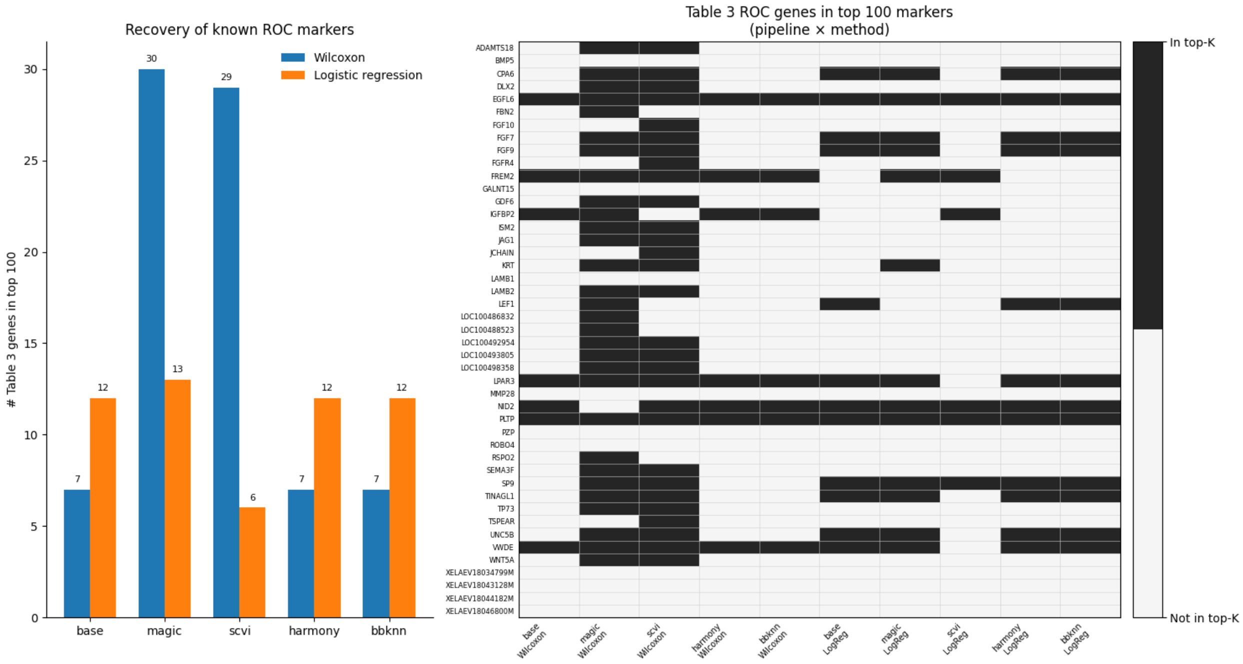

Figure 4: Top-100 ROC markers: Table 3 recovery counts and binary membership across pipelines.

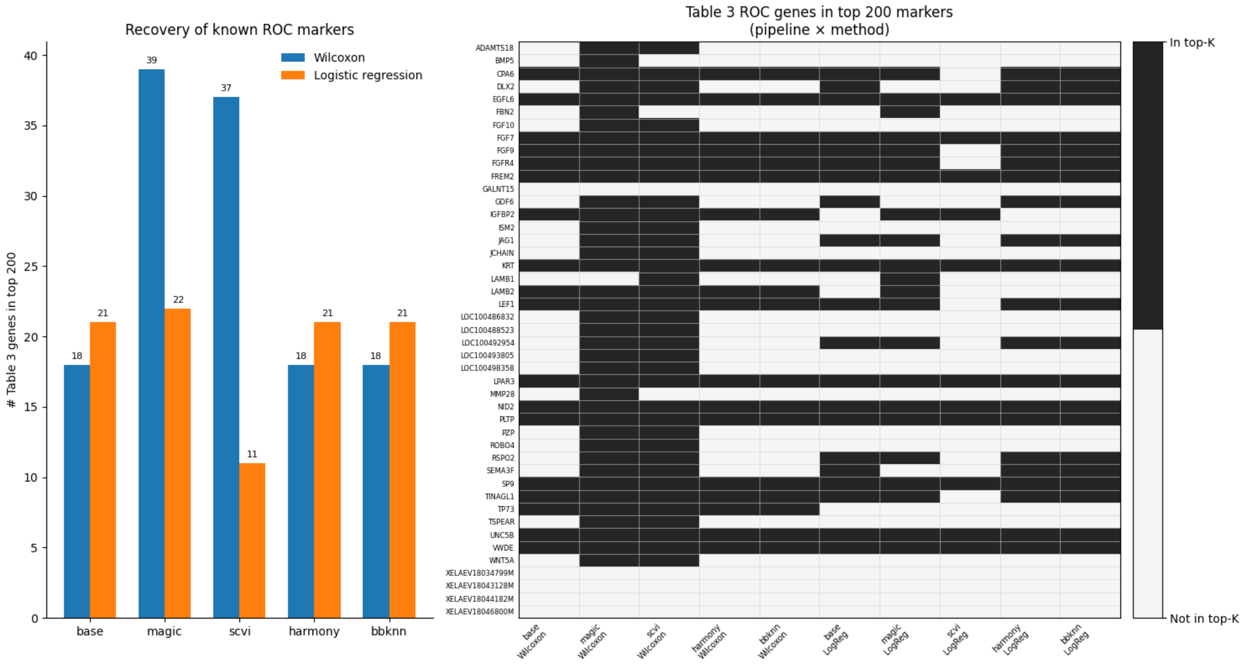

Figure 5: Top-200 ROC markers: Table 3 recovery counts and binary membership across pipelines.

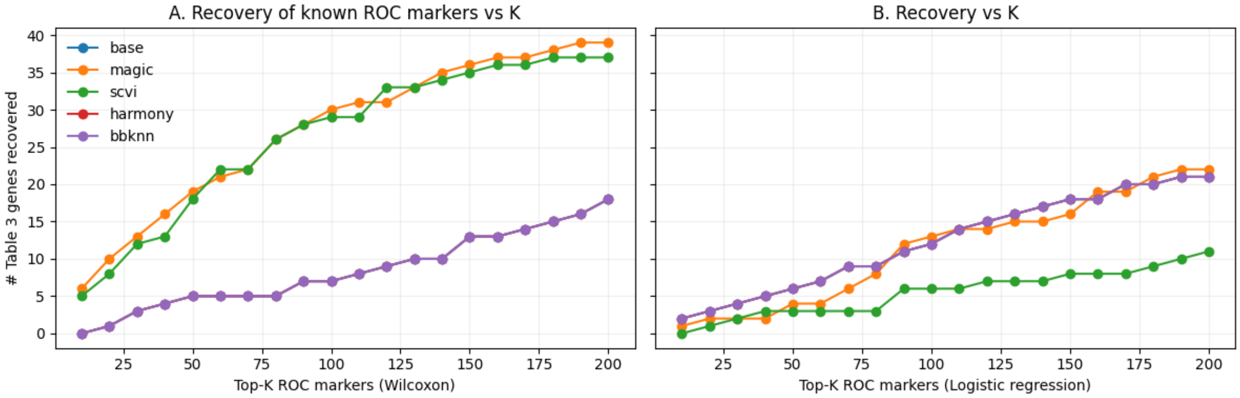

Figure 6: Recovery of Table 3 ROC genes as a function of top-K ROC markers for each pipeline. Left: Wilcoxon; right: logistic regression.