Clustering and Marker Gene Analysis of the FrogTail Dataset

Yuzhi Yao

10/10/2025

Abstract

We analyzed the FrogTail single-cell dataset using three clustering approaches—PCA combined with Louvain, Leiden, and k-means—to assess how algorithmic choices influence data partitioning and biological interpretation. Dimensionality reduction was performed through PCA followed by UMAP visualization. Clustering performance was evaluated using internal validation indices and Adjusted Rand Index (ARI). Marker genes were identified through Wilcoxon rank-sum test for biological annotation. Leiden and Louvain produced comparable cluster structures, while k-means achieved higher compactness but lower biological consistency.

Introduction

Clustering analysis is essential for revealing cellular heterogeneity in single-cell transcriptomics. The outcome, however, depends strongly on both dimensionality reduction and clustering algorithms. Graph-based approaches such as Louvain and Leiden optimize community modularity, while centroid-based k-means relies on Euclidean partitioning. This project compares these methods quantitatively and evaluates their biological validity through marker gene identification.

Methods

Data processing and visualization

Raw gene expression data were normalized and log-transformed before Principal Component Analysis (PCA). The top 20 components were retained for downstream analyses.Neighbor graph was computed using 15 neighbors and 20 PCs UMAP was applied to visualize the structure of the reduced data, with parameters adjusted to reproduce the topology shown in the reference paper. All plots were generated using Scanpy and Matplotlib.

Clustering and performance evaluation

Three clustering methods were applied to the PCA output:

- **PCA + Louvain:** modularity-based graph clustering.
- **PCA + Leiden:** refined graph-based clustering with higher stability.
- **PCA + k-means:** centroid-based partitioning in feature space.

Internal clustering quality was measured using Silhouette score, Davies–Bouldin index, and Calinski–Harabasz index. The Adjusted Rand Index (ARI) quantified consistency between cluster assignments from different methods.

Marker gene identification

Marker genes were detected using Wilcoxon rank-sum test with Scanpy's `rank_genes_groups()` function. For each method, genes with the highest log fold-change and score were selected for visualization. Cluster heatmaps were generated to illustrate marker specificity and guide biological interpretation.

Code availability

All analysis scripts and figures are available at: `https://github.com/linfeihe/FrogTailProject`

Results

Clustering visualization

UMAP projections show distinct cluster separations under each method (Figure 1). Louvain and Leiden produce visually coherent modular structures, whereas k-means yields more spherical and numerous partitions (35 clusters). Leiden forms 37 clusters and maintains better global continuity.

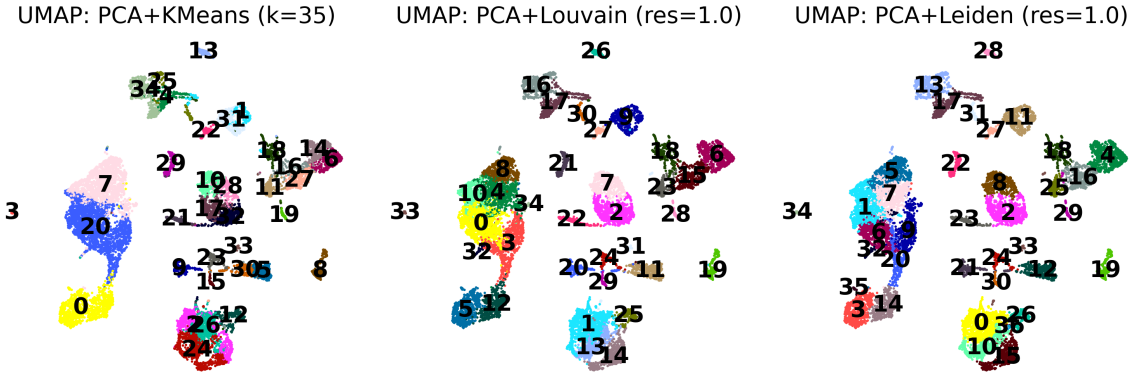

Figure 1: UMAP visualization of PCA-reduced data clustered using (a) k-means, (b) Louvain, and (c) Leiden. Leiden clusters are well-separated while preserving smooth transitions across cell states.

Quantitative comparison of clustering metrics

The quantitative comparison in Table 1 summarizes the internal and external validation scores derived from the dataset files. K-means achieved the highest Silhouette (0.29) and Calinski–Harabasz (3718) values, suggesting greater compactness and between-group separation. Leiden showed the lowest Davies–Bouldin index (1.25), indicating improved cluster distinction. Louvain's performance was close to Leiden in both metrics but slightly less compact. Cross-method ARI analysis (Table 2) confirmed that Leiden and Louvain cluster assignments were most consistent (ARI = 0.79), while both diverged substantially from k-means (ARI = 0.47–0.52).

Table 1: Internal clustering metrics derived from `clustering_internal_metrics.csv`.

Method	Silhouette	Davies–Bouldin	Calinski–Harabasz	n_clusters
PCA + k-means	**0.2923**	1.3071	**3717.96**	35
PCA + Leiden	0.2487	**1.2533**	3065.72	37
PCA + Louvain	0.2436	1.3024	3166.40	35

Table 2: Pairwise Adjusted Rand Index (ARI) values from `clustering_method_ARI.csv`.

Comparison	ARI
Louvain vs Leiden	**0.7855**
Louvain vs k-means	0.5185
Leiden vs k-means	0.4734

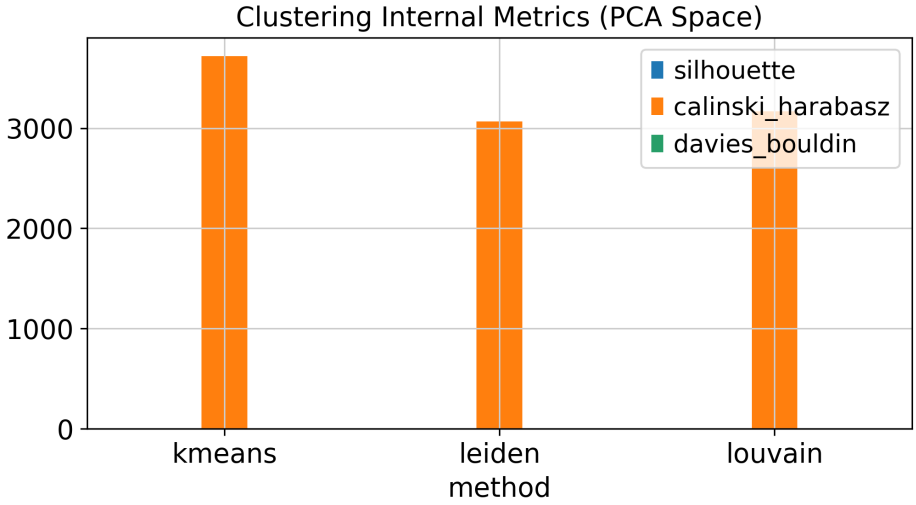

Figure 2: Comparison of internal clustering metrics across methods. Values correspond to those listed in Table 1.

Marker gene analysis

Marker gene results support the quantitative findings. Leiden clusters exhibited consistent differential expression with moderate log fold-changes (4.5–6.5), indicating stable biological signatures. Louvain identified similar but slightly noisier marker sets. K-means produced markers with higher fold-changes (up to 7.8) but broader expression across clusters, reducing interpretability.

For example, in Leiden clustering, *lum.L*, *loc101734526.L*, and *s100a11.L* were among the top markers (logFC 4.5–6.5). In contrast, k-means yielded genes such as *Xelaev18045082m.g* and *Xelaev18047452m.g* with logFC values exceeding 7.8, but without clear cluster specificity (Figure 3).

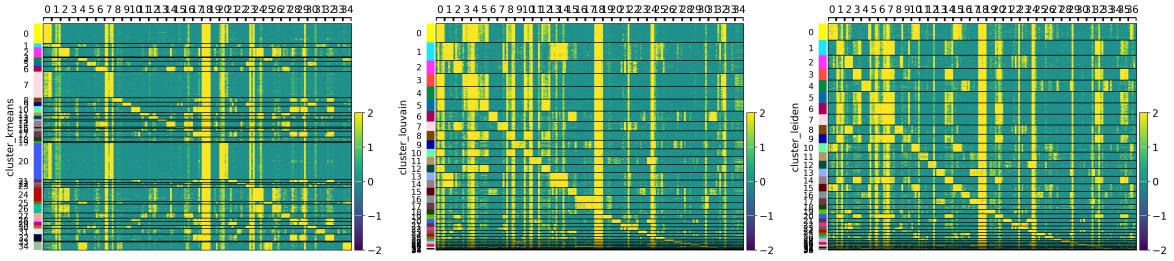

Figure 3: Heatmaps of top marker genes for (a) k-means, (b) Louvain, and (c) Leiden clustering. Leiden demonstrates higher marker specificity and less cross-cluster expression.

Conclusion

All three PCA-based clustering pipelines successfully identified structure within the FrogTail dataset, but their performance varied in compactness and biological interpretability. K-means achieved higher quan-

titative compactness but lacked consistent marker specificity. Leiden offered the best balance between numerical metrics and biologically coherent clustering, aligning closely with Louvain while maintaining better cluster separation. Overall, Leiden provides a stable and interpretable framework for single-cell clustering in this dataset.