Improving Clustering Coherence in scRNA-seq Data with Prior Knowledge and Pairwise Constraints

Davi Guimarães[1], Mateus Pereira[1,3], Kele Belloze[1], Marcel Pedroso[2], Eduardo Bezerra[1]

[1] *Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (Cefet/RJ)*
[2] *Fundação Oswaldo Cruz (Fiocruz)*
[3] *IBM Research*

The high intratumoral heterogeneity of breast cancer directly impacts patient prognosis and disease burden. Single-cell RNA sequencing (scRNA-seq) has emerged as a powerful tool to dissect this heterogeneity at the cellular level [Guo et al. 2020]. However, the large dimensionality and sparsity of scRNA-seq data demand advanced computational strategies.

In this study, we explore unsupervised and semi-supervised machine learning approaches to identify biologically relevant cellular subpopulations in breast tumors. We developed an analysis pipeline based on public scRNA-seq data (GSE75688), comprising 549 cells from 11 breast cancer patients [National Center for Biotechnology Information 2025]. The pipeline includes quality control, selection of highly variable genes (HVGs), dimensionality reduction via Principal Component Analysis (PCA), and clustering.

The experimental phase aimed to assess the impact of incorporating prior knowledge into the task of clustering single-cell gene expression data. After quality control and normalization, the top 500 HVGs were selected to capture the main signals of cellular heterogeneity. Dimensionality reduction was then performed using PCA, retaining the first 50 components to preserve the global structure of the data while mitigating the effects of high dimensionality.

Clustering was carried out by fixing the number of clusters at $k = 6$, reflecting a preliminary estimate of cellular diversity in the sample. Two algorithms were compared: standard K-Means and COP-KMeans, which incorporates supervised constraints. For COP-KMeans, 51 must-link constraints (enforcing that pairs of cells belong to the same cluster) and 75 cannot-link constraints (forcing separation of cells from different types) were used. These constraints were derived from known biological annotations, simulating a realistic semi-supervised scenario [Cai et al. 2023].

Both algorithms were executed 10 times with different random initializations to mitigate bias due to randomness and to enable a statistically robust comparison of results. Clustering quality was evaluated using the Normalized Mutual Information (NMI) met-

ric, which is widely used to quantify the similarity between algorithm-assigned labels and ground truth annotations. NMI measures the agreement between the predicted cluster assignments and the ground truth labels, accounting for chance overlap. It ranges from 0 (no mutual information) to 1 (perfect match), and is particularly suitable for evaluating unsupervised clustering methods where label alignment is not guaranteed. Higher NMI values indicate better correspondence with known cell types.

COP-KMeans achieved an average NMI of 0.4421, outperforming k-Means (average NMI = 0.4286) in alignment with known cell labels. To assess whether the COP-KMeans algorithm yields significantly higher NMI values compared to standard k-Means, we conducted a paired t-test. The results revealed a statistically significant difference between the methods ($t = 3.014$, $p = 0.0056$, one-tailed), with COP-KMeans showing superior performance. These findings support the hypothesis that incorporating constraints into the clustering process positively contributes to the quality of the resulting partitions.

We conclude that semi-supervised learning provides a promising framework for deciphering tumor heterogeneity in scRNA-seq data, offering insights into the cellular mechanisms that underlie disease progression and burden. Future analyses will investigate more powerful semi-supervised clustering algorithms, such as Deep embedded clustering [Ren et al. 2019].

Keywords: scRNA-seq, Breast cancer, Tumor heterogeneity, Semi-supervised learning.

References

[Cai et al. 2023] Cai, J., Hao, J., Yang, H., Zhao, X., and Yang, Y. (2023). A review on semi-supervised clustering. *Information Sciences*, 632:164–200.

[Guo et al. 2020] Guo, C. et al. (2020). Single-cell transcriptome analysis reveals tumor heterogeneity and tumor microenvironment in breast cancer. *arXiv preprint arXiv:2005.06692*.

[National Center for Biotechnology Information 2025] National Center for Biotechnology Information (2025). Geo accession viewer: Gse75688. `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE75688`. Acesso em: 31 de julho de 2025.

[Ren et al. 2019] Ren, Y., Hu, K., Dai, X., Pan, L., Hoi, S. C., and Xu, Z. (2019). Semi-supervised deep embedded clustering. *Neurocomputing*, 325:121–130.