Dimensionality Reduction Methods for Single Cell RNA Sequencing Data

A Comparative Review

CPSC 545
Riya Saju
Malki Wijesinghe
04/12/2023

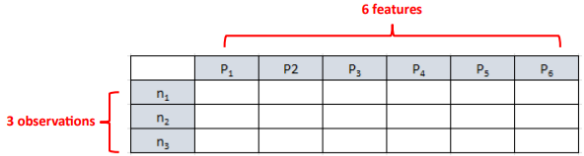
Outline

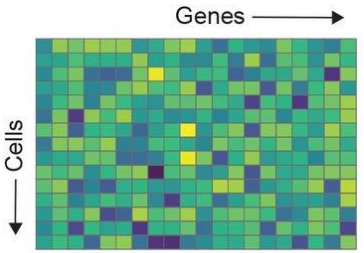
1. Introduction
2. Objective
3. Methods
4. Data Description
5. Results
6. Conclusions
7. References

Introduction - Curse of Dimensionality

The Curse of Dimensionality

- **Sparse Data:** As dimensions increase, data becomes sparse—fewer data points relative to the space they inhabit.
- **Increased Computational Demands:** More dimensions lead to higher computational costs for processing and analysis.
- **Overfitting Risk:** High-dimensional spaces raise the risk of overfitting, where models become too specific to training data and perform poorly on new data.





(Lu et al., 2021)

Key Takeaway: Dimension reduction methods alleviate the curse of dimensionality by condensing data, making it more manageable, computationally efficient, and less prone to overfitting.

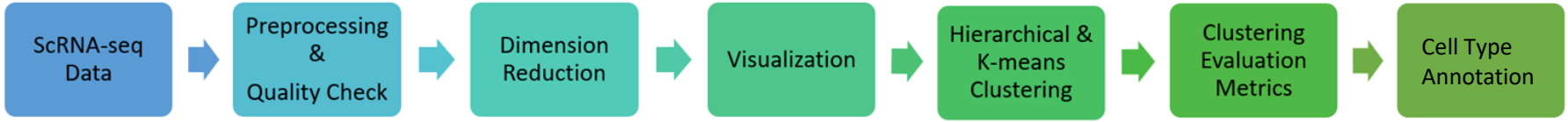
Objective

Compare and evaluate various dimensionality reduction methods used in single-cell RNA data analysis, aiming to validate the robustness of these methods and identify the most effective techniques for interpreting complex gene expression profiles.

Background

Term	Description
Single cell RNA Seq	Examines the nucleic acid sequence information from individual cells providing a higher resolution of cellular differences and a better understanding of the function of an individual cell in the context of its microenvironment.
Gene Expression matrix	Each entry in the matrix represents the number of reads (expression level) of a particular gene in a given sample (cell)
Dimension Reduction	It aims to reduce the number of separate dimensions in the data and this is possible because different genes are correlated if they are affected by the same biological process.
Clustering	Used to empirically define groups of cells with similar expression profiles and allows us to describe population heterogeneity in terms of discrete labels that are easily understood.
Normalized Mutual Information (NMI)	Quantifies the amount of information obtained about one clustering when the other clustering is known. 1 - Two clusterings are identical 0 - No mutual information or similarity between the clusterings
Adjusted Rand Index (ARI)	Measure of the similarity between two data clusterings 1 - Perfect agreement between the two clusterings, 0 - A random agreement

Methodology



- Remove non expressed Genes
- Normalization
- QC Plots

Linear

- **PCA** (Principal component analysis)
- **ICA** (Independent component analysis)
- **IPCA** (Independent principal component analysis)
- **NMF** (Non-negative matrix factorization)

Non-Linear

- **tSNE** (t-distributed stochastic neighbor embedding)
- **UMAP** (Uniform manifold approximation and projection)
- **Diffusion Map**
- **MDS** (Multidimensional scaling)

Deep Learning

- **Autoencoder**

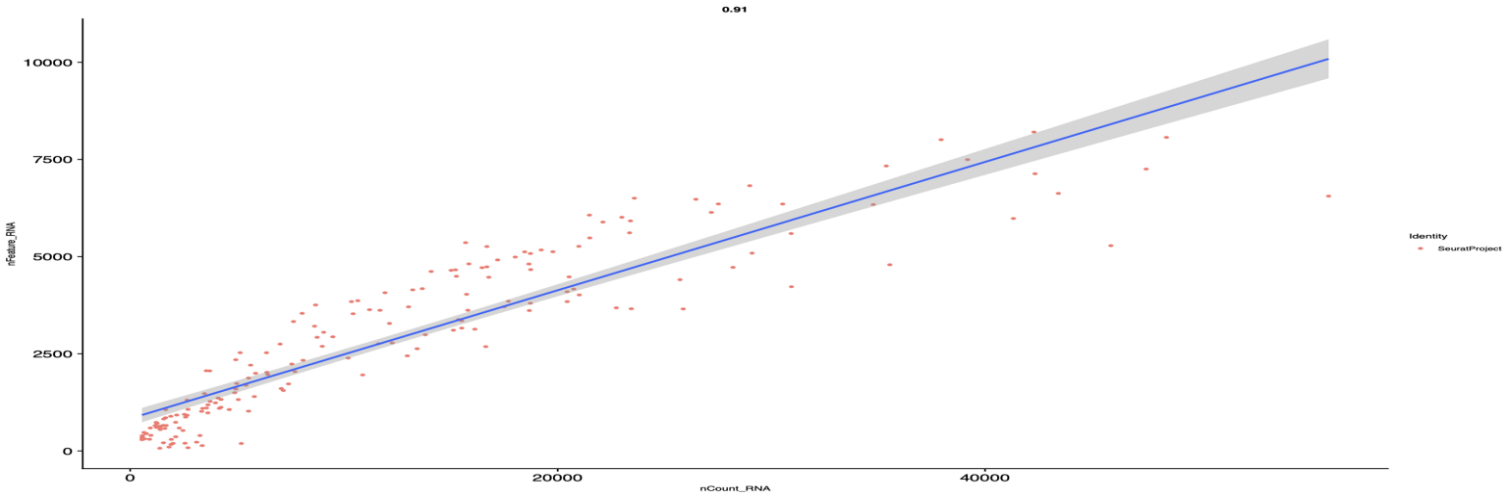
- ARI
- NMI

Data Description

Filtered dataset - Gene Exp. by Cell Ranger	Number of Cells	Number of Genes	
		Original	Preprocessed
Brain Tumor - male, 71	182	36,601	19,181
Breast Cancer - female, 65	687	36,601	21,667
Hodgkin's Lymphoma - male, 19	3049	1,253	1,165

Data Source : brain tumour, breast cancer, hodgkin's lymphoma

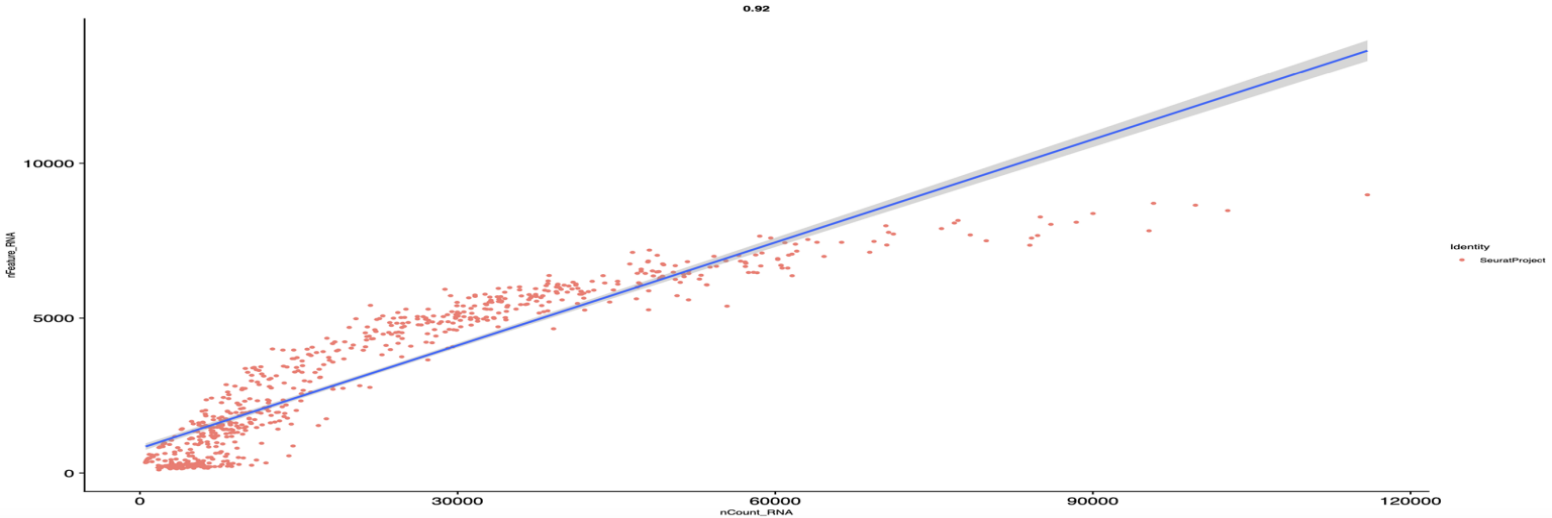
QC - Brain Tumour



Scatter plot of two QC metrics - nCount_RNA and nFeature_RNA

Fraction Reads in Cells	86.2%
Reads Mapped to Genome	96.2%

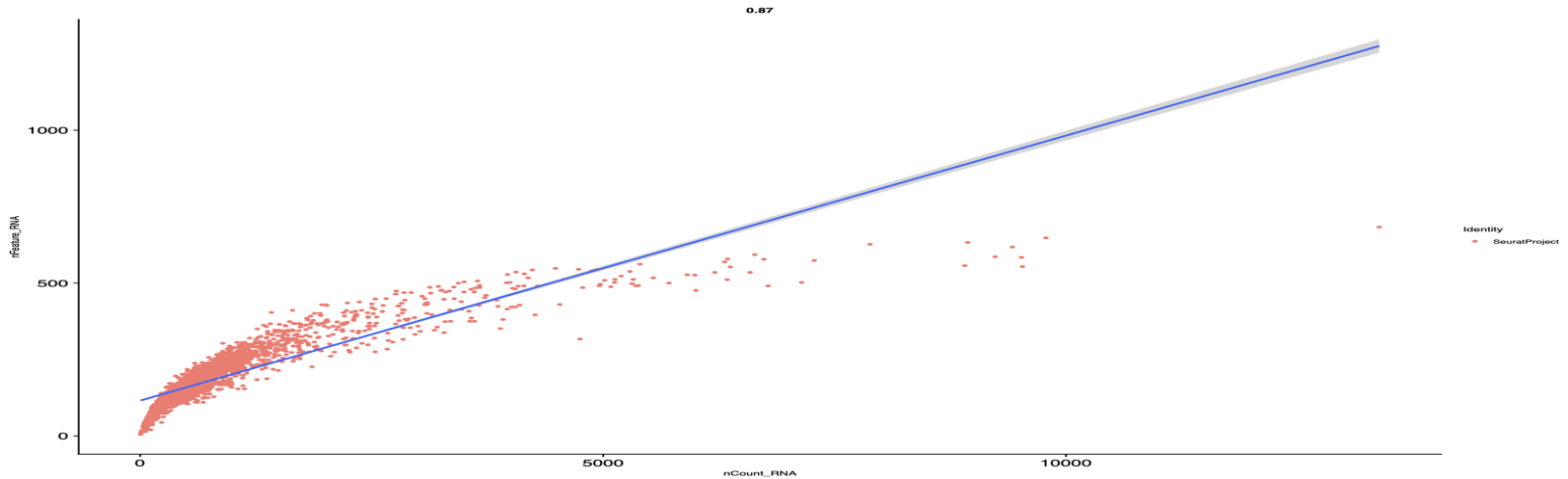
QC - Breast Cancer



Scatter plot of two QC metrics - nCount_RNA and nFeature_RNA

Fraction Reads in Cells	91.8%
Reads Mapped to Genome	96.6%

QC – Hodgkin's Lymphoma

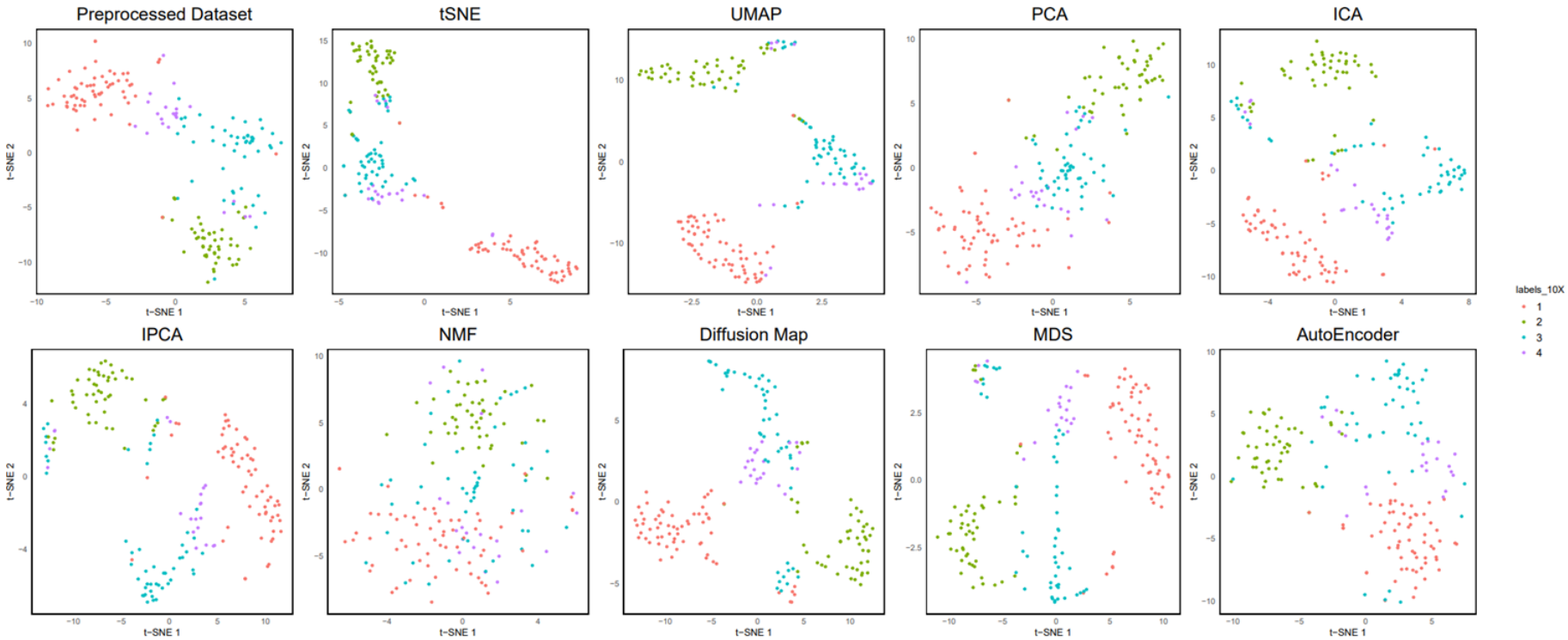


Scatter plot of two QC metrics - nCount_RNA and nFeature_RNA

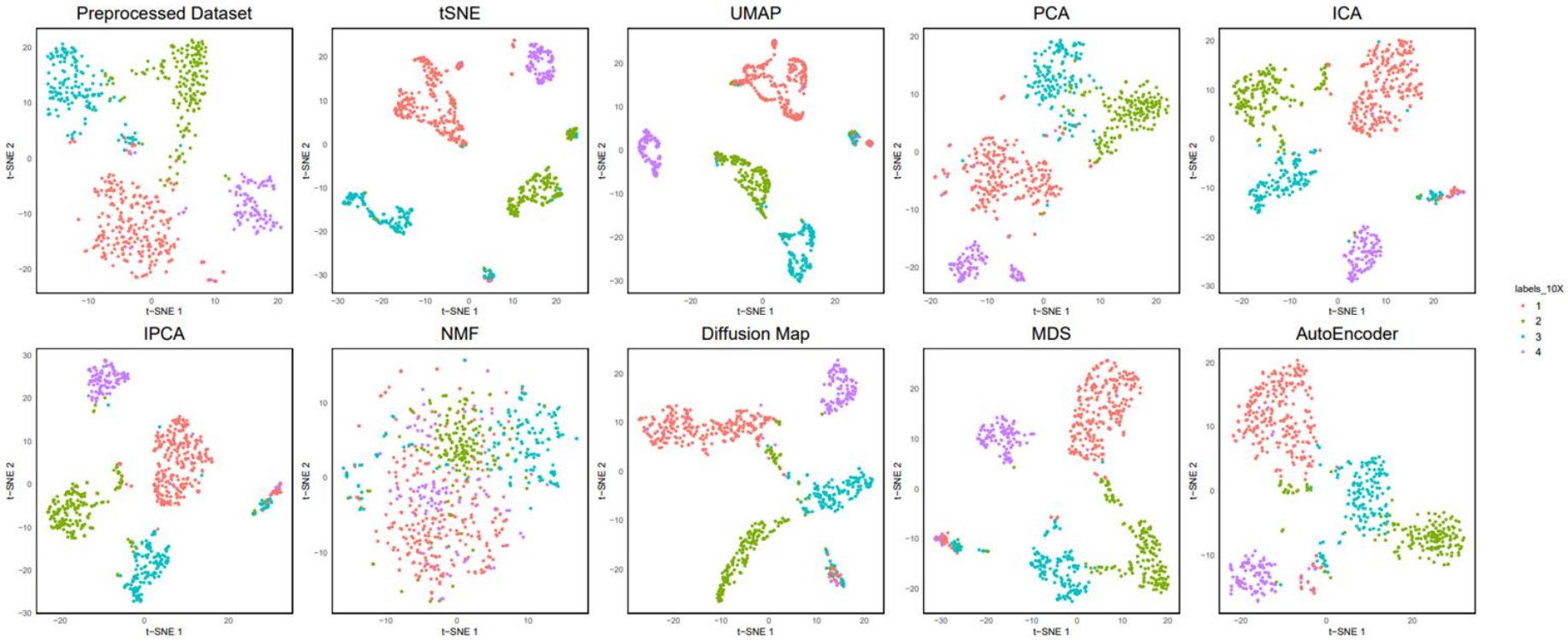
Fraction Reads in Cells	93.5%
Reads Mapped to Genome	98.6%

Results

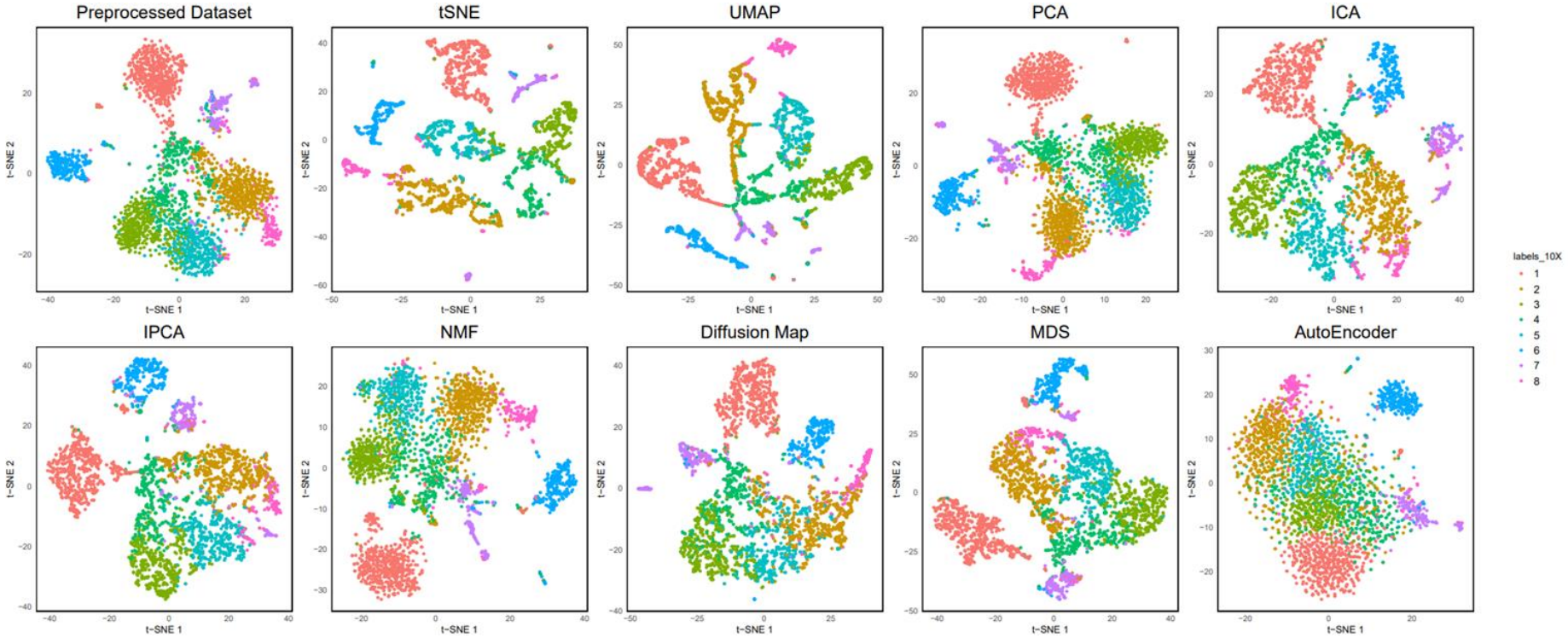
Brain Tumor - Dimension Reduction Before & After with 10XGenomics Labels



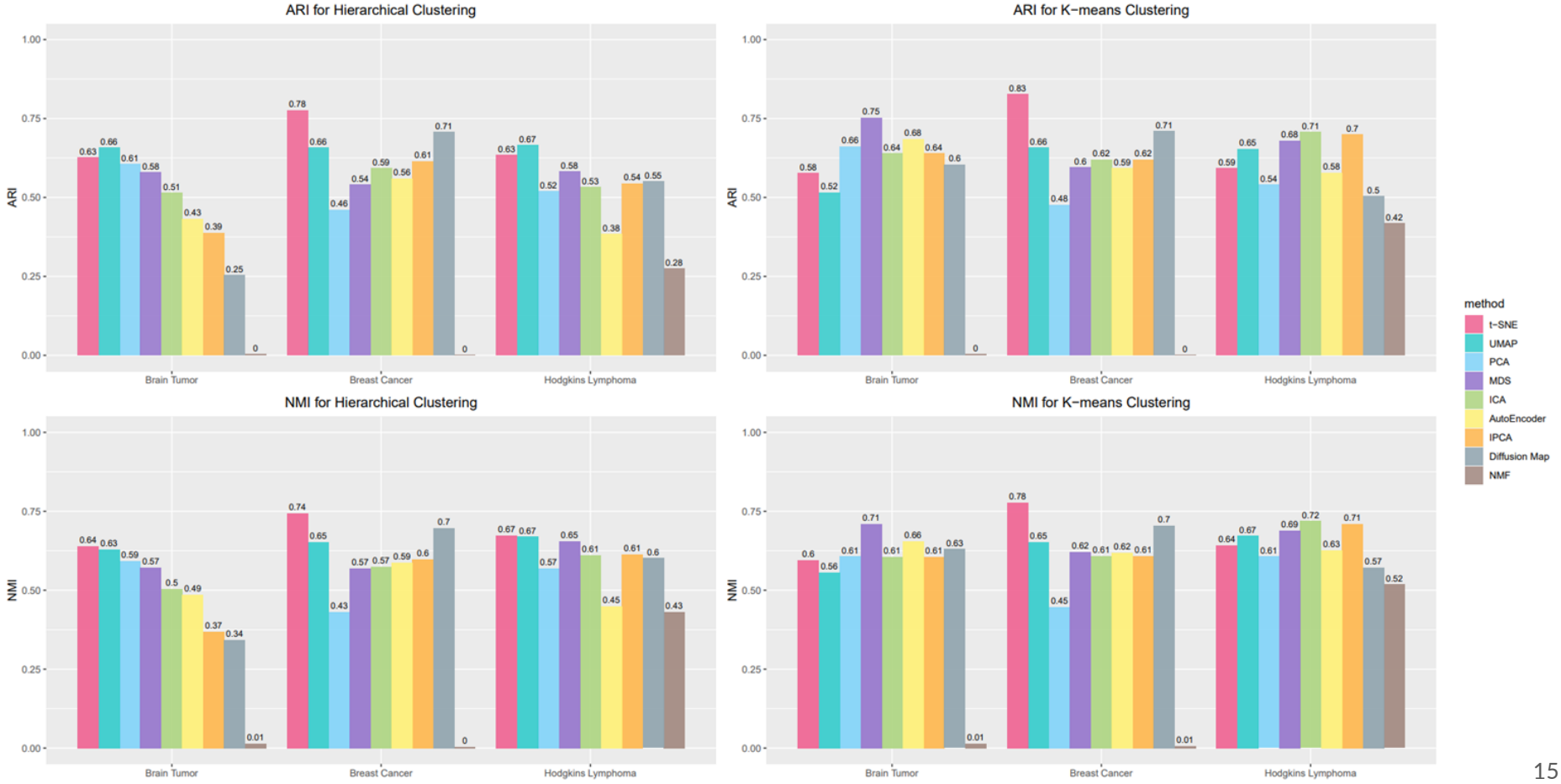
Breast Cancer - Dimension Reduction Before & After with 10XGenomics Labels



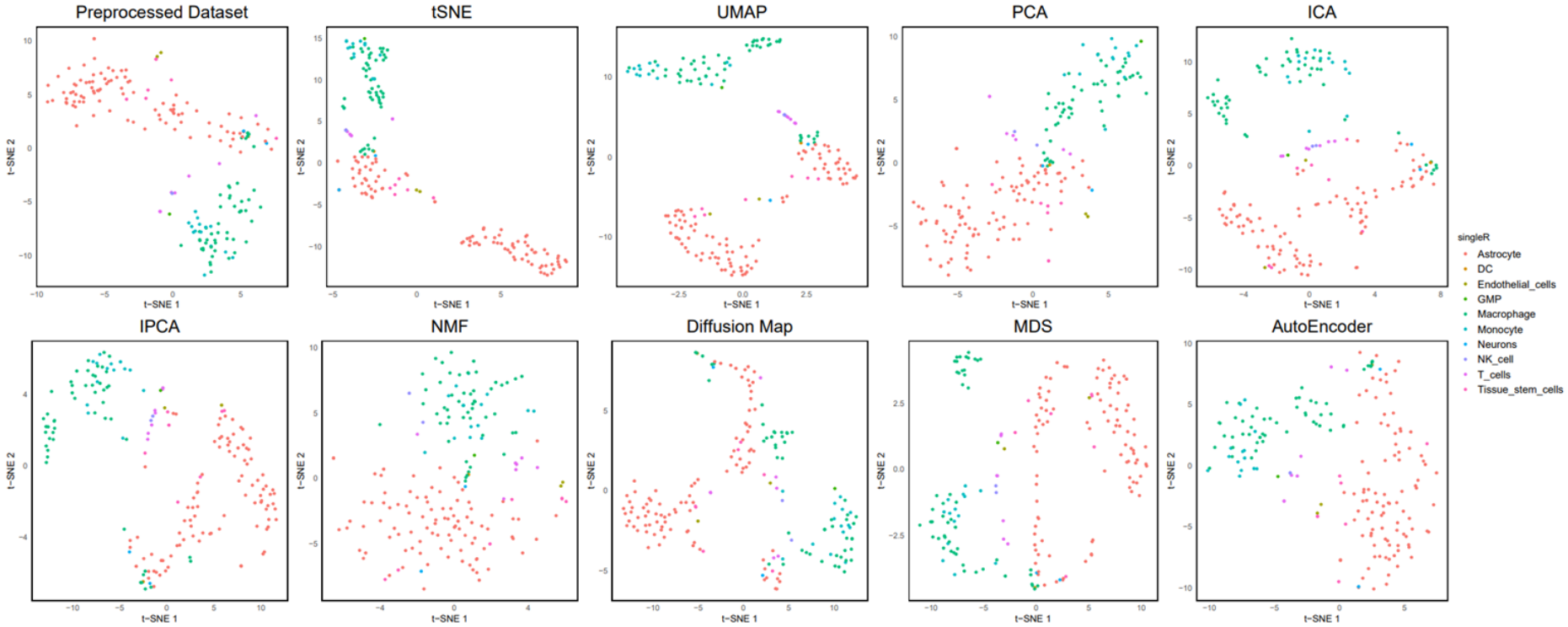
Hodgkins Lymphoma - Dimension Reduction Before & After with 10XGenomics Labels



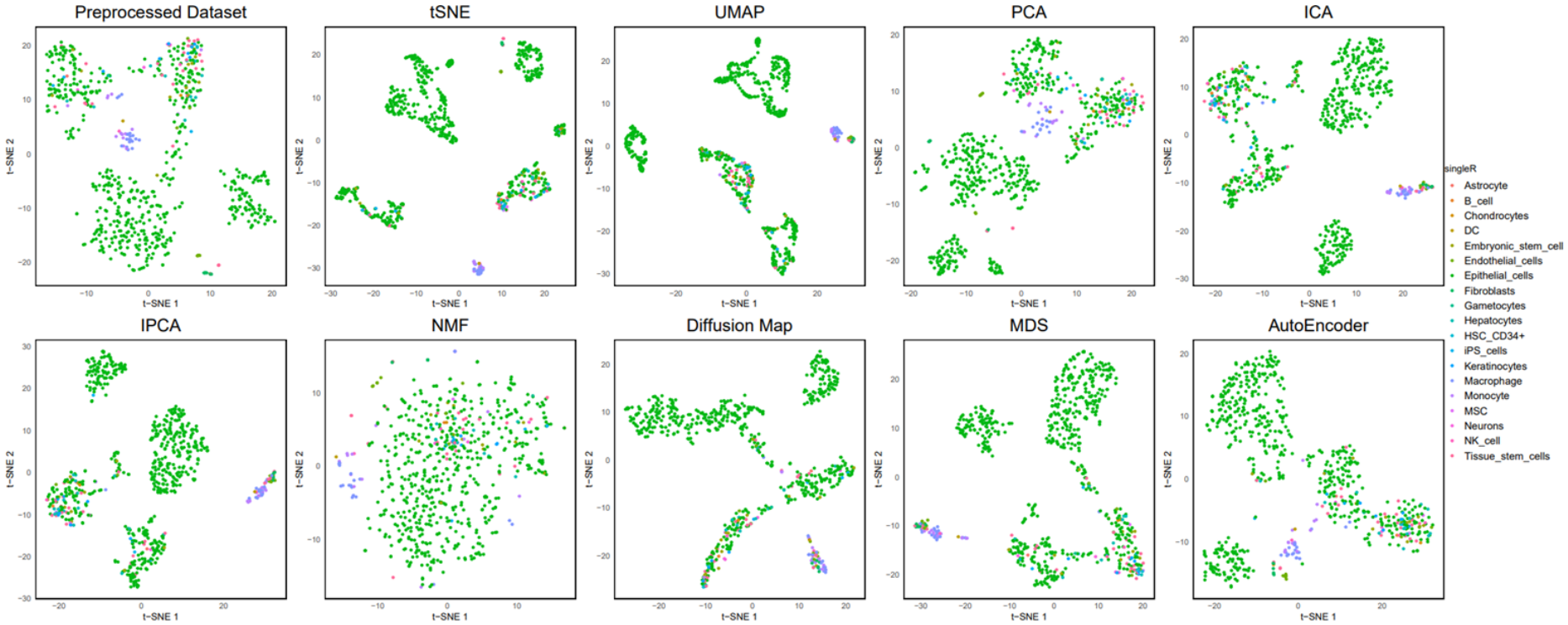
Clustering Summary with 10xGenomics Labels



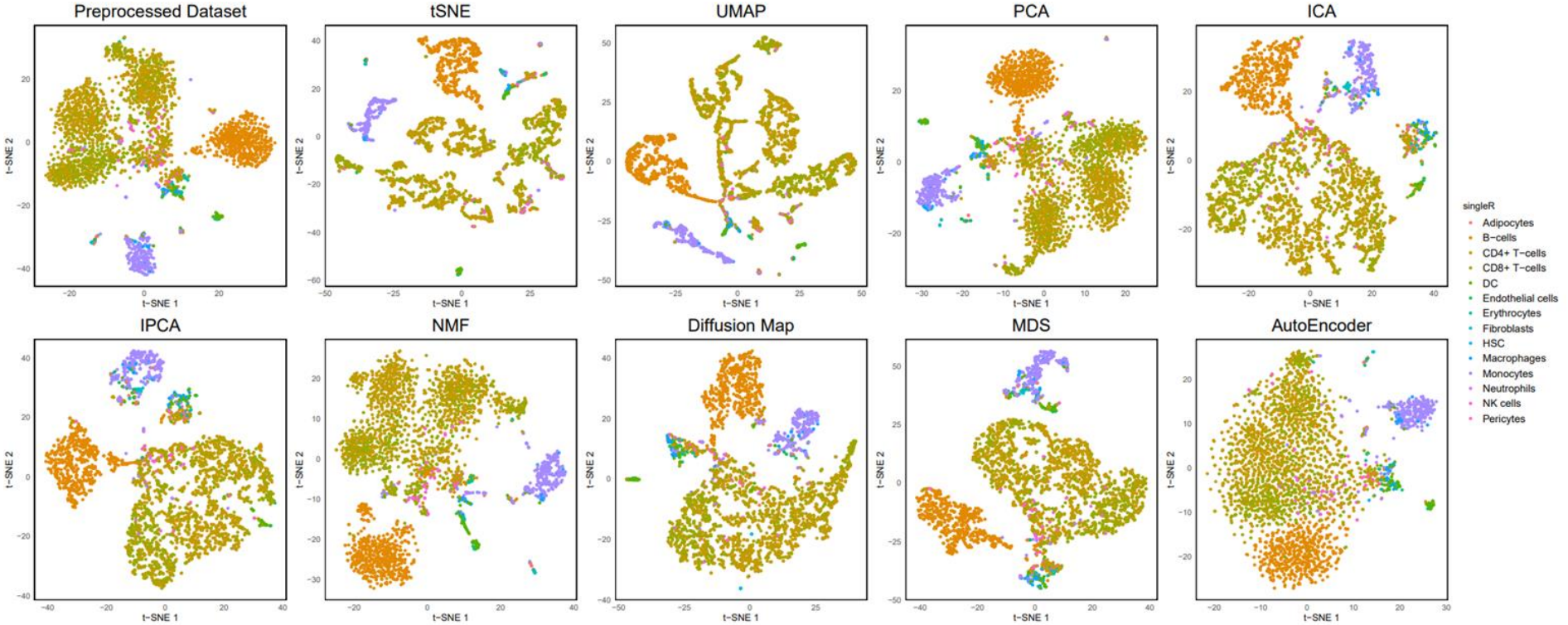
Brain Tumor - Dimension Reduction Before & After with SingleR Cell Annotations



Breast Cancer - Dimension Reduction Before & After with SingleR Cell Annotations



Hodgkins Lymphoma - Dimension Reduction Before & After with SingleR Cell Annotations



Conclusions

- NMF works better in low dimensional data compared to large scale high dimensional data
- UMAP & t-SNE performed well in all 3 datasets
- MDS & ICA gave slightly above average performance in all three datasets
- Diffusion map & IPCA performed well in 2/3 datasets
- Autoencoder & PCA displayed an average performance across the datasets
- K means clustering indicated a slightly better performance in the datasets used for this study
- Future work :
    - Using the singleR annotations as labels for clustering
    - Simulations on different distributions
    - Computational time comparison-larger datasets with high dimensions

References

1. Aqsazafar. (2022, December 18). *What is the curse of dimensionality? simplest explanation!*. MLTut. https://www.mltut.com/what-is-the-curse-of-dimensionality-simplest-explanation/
2. Yiu, T. (2021, September 29). *The curse of dimensionality*. Medium. https://towardsdatascience.com/the-curse-of-dimensionality-50dc6e49aa1e
3. Chapter 4 clustering algorithms and evaluations - uni-stuttgart.de. (n.d.). https://www.ims.uni-stuttgart.de/documents/team/schulte/theses/phd/algorithm.pdf
4. Ranjan, B., Schmidt, F., Sun, W., Park, J., Honardoost, M. A., Tan, J., Rayan, N. A., & Prabhakar, S. (2021, April 12). *Scconsensus: Combining supervised and unsupervised clustering for cell type identification in single-cell RNA sequencing data - BMC Bioinformatics*. BioMed Central. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-021-04028-4#Sec2
5. Sun, S., Zhu, J., Ma, Y., & Zhou, X. (2019, December 10). Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis - genome biology. BioMed Central. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1898-6
6. Xiang, R., Wang, W., Yang, L., Wang, S., Xu, C., & Chen, X. (2021, February 19). A comparison for dimensionality reduction methods of single-cell RNA-seq data. Frontiers. https://www.frontiersin.org/articles/10.3389/fgene.2021.646936/full
7. Oehm, D. (2018, July 28). PCA vs Autoencoders for dimensionality reduction: R-bloggers. R. https://www.r-bloggers.com/2018/07/pca-vs-autoencoders-for-dimensionality-reduction/
8. Yao, F., Coquery, J., & Cao, K.-A. L. (2012, February 3). Independent principal component analysis for biologically meaningful dimension reduction of large biological data sets - BMC Bioinformatics. BioMed Central. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-24
9. https://bioconductor.org/books/3.13/OSCA.basic/
10. https://en.wikipedia.org/wiki/Single-cell_sequencing#:~:text=These%20measurements%20may%20obscure%20critical,and%20phenotypes%20as%20of%202020.
11. https://ucdavis-bioinformatics-training.github.io/2019-single-cell-RNA-sequencing-Workshop-UCD_UCSF/data_reduction/Expression_Matrix.html#:~:text=Statistical%20analyses%20of%20scRNA%2Dseq,a%20given%20sample%20(cell).

THANK YOU!

QUESTIONS?